

# The **GOVERNMENT ANALYTICS** Handbook


**LEVERAGING DATA TO STRENGTHEN  
PUBLIC ADMINISTRATION**

*Edited by Daniel Rogger and Christian Schuster*



**WORLD BANK GROUP**




The top of the page features a decorative header. It includes a teal horizontal band with a pattern of binary code (0s and 1s) in a lighter shade. Above this band, there are several rows of small, light blue squares arranged in a grid-like pattern that tapers off towards the left.

# THE **GOVERNMENT ANALYTICS** HANDBOOK







# THE **GOVERNMENT ANALYTICS** HANDBOOK

**Leveraging Data to Strengthen  
Public Administration**

EDITED BY DANIEL ROGGER AND CHRISTIAN SCHUSTER

© 2023 International Bank for Reconstruction and Development / The World Bank  
1818 H Street NW, Washington, DC 20433  
Telephone: 202-473-1000; internet: [www.worldbank.org](http://www.worldbank.org)

Some rights reserved

1 2 3 4 26 25 24 23

This work is a product of the staff of The World Bank with external contributions. The findings, interpretations, and conclusions expressed in this work do not necessarily reflect the views of The World Bank, its Board of Executive Directors, or the governments they represent. The World Bank does not guarantee the accuracy, completeness, or currency of the data included in this work and does not assume responsibility for any errors, omissions, or discrepancies in the information, or liability with respect to the use of or failure to use the information, methods, processes, or conclusions set forth. The boundaries, colors, denominations, and other information shown on any map in this work do not imply any judgment on the part of The World Bank concerning the legal status of any territory or the endorsement or acceptance of such boundaries.

Nothing herein shall constitute or be construed or considered to be a limitation upon or waiver of the privileges and immunities of The World Bank, all of which are specifically reserved.

This book applies a streamlined editorial quality control check to disseminate the content with the least possible delay.

### Rights and Permissions



This work is available under the Creative Commons Attribution 3.0 IGO license (CC BY 3.0 IGO) <http://creativecommons.org/licenses/by/3.0/igo>. Under the Creative Commons Attribution license, you are free to copy, distribute, transmit, and adapt this work, including for commercial purposes, under the following conditions:

**Attribution**—Please cite the work as follows: Rogger, Daniel, and Christian Schuster, eds. 2023. *The Government Analytics Handbook: Leveraging Data to Strengthen Public Administration*. Washington, DC: World Bank. doi:10.1596/978-1-4648-1957-5. License: Creative Commons Attribution CC BY 3.0 IGO

**Translations**—If you create a translation of this work, please add the following disclaimer along with the attribution: *This translation was not created by The World Bank and should not be considered an official World Bank translation. The World Bank shall not be liable for any content or error in this translation.*

**Adaptations**—If you create an adaptation of this work, please add the following disclaimer along with the attribution: *This is an adaptation of an original work by The World Bank. Views and opinions expressed in the adaptation are the sole responsibility of the author or authors of the adaptation and are not endorsed by The World Bank.*

**Third-party content**—The World Bank does not necessarily own each component of the content contained within the work. The World Bank therefore does not warrant that the use of any third-party-owned individual component or part contained in the work will not infringe on the rights of those third parties. The risk of claims resulting from such infringement rests solely with you. If you wish to reuse a component of the work, it is your responsibility to determine whether permission is needed for that reuse and to obtain permission from the copyright owner. Examples of components can include, but are not limited to, tables, figures, or images.

All queries on rights and licenses should be addressed to World Bank Publications, The World Bank Group, 1818 H Street NW, Washington, DC 20433, USA; e-mail: [pubrights@worldbank.org](mailto:pubrights@worldbank.org).

ISBN (paper): 978-1-4648-1957-5

ISBN (electronic): 978-1-4648-1981-0

DOI: 10.1596/978-1-4648-1957-5

*Cover image:* © Pixel Matrix / Adobe Stock. Used with the permission of Pixel Matrix / Adobe Stock. Further permission required for reuse.

*Cover design:* Circle Graphics and Bill Pragluski, Critical Stages, LLC.

**Library of Congress Control Number: 2023941914**

# Contents

<i>Foreword</i> .....	xv
<i>Acknowledgments</i> .....	xvii
<i>About the Editors</i> .....	xix
<b>PART 1: OVERVIEW</b> .....	<b>1</b>
Chapter 1 Introduction to <i>The Government Analytics Handbook</i> .....	3
<i>Daniel Rogger and Christian Schuster</i>	
Chapter 2 How to Do Government Analytics: Lessons from the <i>Handbook</i> .....	13
<i>Daniel Rogger and Christian Schuster</i>	
Chapter 3 Government Analytics of the Future .....	43
<i>Daniel Rogger and Christian Schuster</i>	
<b>PART 2: FOUNDATIONAL THEMES IN GOVERNMENT ANALYTICS</b> .....	<b>57</b>
Chapter 4 Measuring What Matters: Principles for a Balanced Data Suite That Prioritizes Problem Solving and Learning .....	59
<i>Kate Bridges and Michael Woolcock</i>	
Chapter 5 Practical Tools for Effective Measurement and Analytics .....	83
<i>Maria Ruth Jones and Benjamin Daniels</i>	
Chapter 6 The Ethics of Measuring Public Administration .....	95
<i>Annabelle Wittels</i>	
Chapter 7 Measuring and Encouraging Performance Information Use in Government .....	121
<i>Donald Moynihan</i>	
Chapter 8 Understanding Corruption through Government Analytics .....	131
<i>James Anderson, David S. Bernstein, Galileu Kim, Francesca Recanatini, and Christian Schuster</i>	
<b>PART 3: GOVERNMENT ANALYTICS USING ADMINISTRATIVE DATA</b> .....	<b>149</b>
Chapter 9 Creating Data Infrastructures for Government Analytics .....	151
<i>Khuram Farooq and Galileu Kim</i>	
Chapter 10 Government Analytics Using Human Resources and Payroll Data .....	209
<i>Rafael Alves de Albuquerque Tavares, Daniel Ortega Nieto, and Eleanor Florence Woodhouse</i>	
Chapter 11 Government Analytics Using Expenditure Data .....	237
<i>Moritz Piatti-Fünfkirchen, James Brumby, and Ali Hashim</i>	

Chapter 12 Government Analytics Using Procurement Data . . . . .	259
<i>Serena Cocciolo, Sushmita Samaddar, and Mihaly Fazekas</i>	
Chapter 13 Government Analytics Using Data on the Quality of Administrative Processes . . . . .	285
<i>Jane Adjabeng, Eugenia Adomako-Gyasi, Moses Akrofi, Maxwell Ampofo, Margherita Fornasari, Ignatius Geegbae, Allan Kasapa, Jennifer Ljungqvist, Wilson Metronao Amevor, Felix Nyarko Ampong, Josiah Okyere Gyimah, Daniel Rogger, Nicholas Sampah, and Martin Williams</i>	
Chapter 14 Government Analytics Using Customs Data . . . . .	307
<i>Alice Duhaut</i>	
Chapter 15 Government Analytics Using Administrative Case Data . . . . .	327
<i>Michael Carlos Best, Alessandra Fenizia, and Adnan Qadir Khan</i>	
Chapter 16 Government Analytics Using Machine Learning . . . . .	345
<i>Sandeep Bhupatiraju, Daniel Chen, Slava Jankin, Galileu Kim, Maximilian Kupi, and Manuel Ramos Maqueda</i>	
Chapter 17 Government Analytics Using Data on Task and Project Completion . . . . .	365
<i>Imran Rasul, Daniel Rogger, Martin Williams, and Eleanor Florence Woodhouse</i>	
<b>PART 4: GOVERNMENT ANALYTICS USING PUBLIC SERVANT SURVEYS . . . . .</b>	<b>385</b>
Chapter 18 Surveys of Public Servants: The Global Landscape . . . . .	387
<i>Ayesha Khurshid and Christian Schuster</i>	
Chapter 19 Determining Survey Modes and Response Rates: Do Public Officials Respond Differently to Online and In-Person Surveys? . . . . .	399
<i>Xu Han, Camille Parker, Daniel Rogger, and Christian Schuster</i>	
Chapter 20 Determining Sample Sizes: How Many Public Officials Should Be Surveyed? . . . . .	423
<i>Robert Lipinski, Daniel Rogger, Christian Schuster, and Annabelle Wittels</i>	
Chapter 21 Designing Survey Questionnaires: Which Survey Measures Vary and for Whom? . . . . .	449
<i>Robert Lipinski, Daniel Rogger, Christian Schuster, and Annabelle Wittels</i>	
Chapter 22 Designing Survey Questionnaires: To What Types of Survey Questions Do Public Servants Not Respond? . . . . .	471
<i>Robert Lipinski, Daniel Rogger, and Christian Schuster</i>	
Chapter 23 Designing Survey Questionnaires: Should Surveys Ask about Public Servants' Perceptions of Their Organization or Their Individual Experience? . . . . .	497
<i>Kim Sass Mikkelsen and Camille Mercedes Parker</i>	
Chapter 24 Interpreting Survey Findings: Can Survey Results Be Compared across Organizations and Countries? . . . . .	525
<i>Robert Lipinski, Jan-Hinrik Meyer-Sahling, Kim Sass Mikkelsen, and Christian Schuster</i>	
Chapter 25 Making the Most of Public Servant Survey Results: Lessons from Six Governments . . . . .	549
<i>Christian Schuster, Annabelle Wittels, Nathan Borgelt, Horacio Coral, Matt Kerlogue, Conall Mac Michael, Alejandro Ramos, Nicole Steele, and David Widlake</i>	
Chapter 26 Using Survey Findings for Public Action: The Experience of the US Federal Government . . . . .	573
<i>Camille Hoover, Robin Klevins, Rosemary Miller, Maria Raviele, Daniel Rogger, Robert Seidner, and Kimberly Wells</i>	
<b>PART 5: GOVERNMENT ANALYTICS USING EXTERNAL ASSESSMENTS . . . . .</b>	<b>593</b>
Chapter 27 Government Analytics Using Household Surveys . . . . .	595
<i>Faisal Ali Baig, Zahid Hasnain, Turkan Mukhtarova, and Daniel Rogger</i>	

Chapter 28 Government Analytics Using Citizen Surveys: Lessons from the OECD Trust Survey .....	615
<i>Monica Brezzi and Santiago González</i>	
Chapter 29 Government Analytics Using Measures of Service Delivery .....	629
<i>Kathryn Andrews, Galileu Kim, Halsey Rogers, Jigyasa Sharma, and Sergio Venegas Marin</i>	
Chapter 30 Government Analytics Using Anthropological Methods .....	645
<i>Colin Hoag, Josiah Heyman, Kristin Asdal, Hilde Reinertsen, and Matthew Hull</i>	

## APPENDICES

Appendix A Checklist for Using Expansive and Qualified Measurement for Informed Problem Solving and Learning in Public Administration: Chapter 4 Appendix .....	659
Appendix B Framework for Evaluating the Ethics of Measuring and Tracking Public Sector Workers: Chapter 6 Appendix. ....	661
Appendix C Stages in the Creation of Expenditure Data: Chapter 11 Appendix .....	667
Appendix D Traditional Public Procurement Indicators: Chapter 12 Appendix .....	673
Appendix E Tools to Assess Quality of Government Processes: Chapter 13 Appendix .....	676
Appendix F Further Details of Analysis: Chapter 15 Appendix .....	686
Appendix G Further Details of Analysis: Chapter 19 Appendix .....	691
Appendix H Further Details of Analysis: Chapter 20 Appendix. ....	703
Appendix I Further Details of Survey Questions: Chapter 21 Appendix .....	713
Appendix J Framework for Coding the Complexity and Sensitivity of Questions in Public Administration Surveys: Chapter 22 Appendix. ....	731
Appendix K Referent Questions and Organizational and Commitment Measures: Chapter 23 Appendix .....	742
Appendix L Further Details of Surveys: Chapter 24 Appendix .....	746
Appendix M US Federal Government Best Practices Resources: Chapter 26 Appendix .....	753

## BOXES

7.1 Types of Performance Information Use .....	123
7.2 US Government Accountability Office Survey Measures of Performance Information Use .....	125
11.1 How Utilization of a Financial Management Information System in Malawi Supported Data Provenance and Helped Resolve a Major Corruption Episode .....	242
11.2 How the Government of Rwanda Uses Budget Tagging to Implement a High-Priority, Cross-Cutting Agenda .....	246
12.1 Types of Digitalization of Public Procurement Systems .....	262
12.2 What We Know about Corruption Risk Indicators .....	265
12.3 What We Know about Collusion and Cartel Screens .....	266
12.4 Examples of Common Issues with Data Quality and Completeness .....	274
12.5 What We Know about Green Public Procurement. ....	276
14.1 Interactions with Other Agencies: The Case of Malawi .....	310
14.2 ASYCUDA Data Structure .....	316
14.3 The Problem of Valuation. ....	321
14.4 Information and Customs Performance: The Case of Madagascar .....	322
16.1 The Precedent Case of the Netherlands' Automated Surveillance System .....	352
16.2 Leveraging Data and Technology to Improve Judicial Efficiency in Kenya .....	360
22.1 Applying the Coding Framework: Illustrative Examples from Romania .....	483
C.1 Cash Accounting versus Accrual Accounting. ....	668

## FIGURES

2.1 The Public Administration Production Function .....	15
2.2 Mapping Different Government Analytics Data in the Public Administration Production Function. ....	16
2.3 Wage Bill Projection and Policy Scenarios, Brazil, 2008–30 .....	20

2.4	Decision Tree for Surveys of Public Servants . . . . .	27
2.5	Average Difference between Survey Modes for Different Topics across Romanian Government Organizations . . . . .	28
2.6	Share of Public Servants Satisfied with Their Pay and/or Total Benefits, Various Countries. . . . .	32
4.1	Country Scores on Program for International Student Assessment Tests and Perceived Entrepreneurial Capability . . . . .	69
4.2	Four Risks with Corresponding Principles for Mitigating Them to Ensure a Balanced Data Suite. . . . .	70
5.1	DIME Analytics Uses an Iterative Process to Expand Technical Capacity throughout the Research Cycle . . . . .	85
5.2	Overview of the Tasks Involved in Development Research Data Work . . . . .	86
5.3	Workflow for High-Quality Survey Data Collection. . . . .	87
5.4	Summary Protocols for High-Quality Data Collection at Every Stage of the Survey Workflow. . . . .	88
6.1	Ethical Dimensions That Require Balancing in Data Collection Efforts on Public Sector Employees . . . . .	98
8.1	Examples of Corruption along the Public Administration Production Function. . . . .	133
8.2	Country Clusters Based on Relative Frequency of Bribes in Specific Areas . . . . .	140
8.3	Percentage of Firms Expected to Give Gifts in Meetings with Tax Officials, by Region. . . . .	144
9.1	Policy Objectives for Human Resources Management Information Systems and Respective Stakeholders . . . . .	155
9.2	Human Resources Management Information Systems Data Infrastructure Modules. . . . .	157
9.3	Human Resources Management Information Systems Reform Sequence . . . . .	160
9.4	Human Resources Management Information Systems Maturity . . . . .	161
9.5	Human Resources Management Information System, Luxembourg . . . . .	167
9.6	Luxembourg's Dashboard Modules . . . . .	170
9.7	Percentage of Positive Employee Engagement Scores from the Federal Employee Viewpoint Survey. . . . .	171
9.8	Sample Dashboard from Luxembourg's HR BICC . . . . .	176
9.9	Brazil's Federal Payroll, 2018 . . . . .	181
9.10	Brazil's Solution Workflow . . . . .	182
9.11	Comparison between Brazil's Legacy and New Payroll Workflows. . . . .	183
9.12	Instructions Tab in the EVS ART . . . . .	191
9.13	Landing Page of the EVS ART Dashboard. . . . .	192
9.14	Results Comparison in the EVS ART Dashboard . . . . .	193
9.15	Percentage of Positive Employee Engagement Scores from the Federal Employee Viewpoint Survey. . . . .	194
9.16	Percentage of Negative Employee Engagement Scores from the Federal Employee Viewpoint Survey. . . . .	195
9.17	Sample Action-Planning Tab in the EVS ART . . . . .	197
9.18	Identifying Challenges through the EVS ART . . . . .	200
9.19	Changes in Federal Employee Viewpoint Survey Index Measures, 2015–16 . . . . .	201
9.20	Changes in Federal Employee Viewpoint Survey Index Measures, 2015–19 . . . . .	202
9.21	Improving Measures of Accountability at the National Institute of Diabetes and Digestive and Kidney Diseases. . . . .	203
9.22	“Belief in Action” Scores from the Federal Employee Viewpoint Survey, 2014–20 . . . . .	205
9.23	Federal Employee Viewpoint Survey Participation Rates, 2014–20 . . . . .	206
9.24	The Ripple Effect . . . . .	206
10.1	Wage Bill and Pensions as a Percentage of Subnational States' Budget, Brazil, 1995–2019 . . . . .	211
10.2	Human Resources Microdata Quality Ladder . . . . .	218
10.3	Drivers of Wage Bill Variation, Brazilian Federal Government, 2008–18 . . . . .	220
10.4	Wage Bill Breakdown, by Sector, Brazilian State of Rio Grande do Norte, 2018. . . . .	220
10.5	Distribution of Civil Servants, by Career-Ladder Level, Brazilian Federal Government, 2018 . . . . .	221
10.6	Grade Progressions and Turnover, Uruguay Central Government, 2019 . . . . .	222
10.7	Distribution of Pay-for-Performance Allowances, Brazilian Federal Government, 2017 . . . . .	223
10.8	Measuring Pay Inequity in the Uruguayan and Brazilian Governments . . . . .	224

10.9	Inequity of Pay in Wage Components, Uruguay, 2019. . . . .	226
10.10	Gender Gap, by Ministry, Government of Uruguay, 2010–20 . . . . .	227
10.11	Career Types and Wages, by Career Group, Brazil . . . . .	228
10.12	Retirement Projections, Brazilian Federal Government, 2019–54. . . . .	229
10.13	Baseline Wage Bill Projection, Brazilian Federal Government, 2008–30 . . . . .	230
10.14	Decomposition of Wage Bill Projection between Current and New Employees, Brazilian Federal Government, 2018–30 . . . . .	230
10.15	Wage Bill Projection and Policy Scenarios, Brazil, 2008–30 . . . . .	231
10.16	Cumulative Fiscal Savings from Policy Scenarios, Brazil, 2019–30. . . . .	232
11.1	Use of Government Expenditure Data for Government Analytics . . . . .	239
11.2	Stages in the Execution Process That Create Government Expenditure Data . . . . .	241
11.3	Expenditure Transactions Profile, Bangladesh . . . . .	250
12.1	Data Map of Traditional Public Procurement Data . . . . .	264
12.2	Kraljic Matrix for Colombia . . . . .	267
12.3	Complexity Ladder for Analysis Tools in Procurement Monitoring and Evaluation. . . . .	268
12.4	National Agency for Public Procurement (ANAP) Dashboard, Romania. . . . .	269
12.5	Assessment of Bangladesh’s 10 Percent Rule . . . . .	270
12.6	Procurement Outcomes under Manual versus Electronic Government Procurement Systems, Bangladesh . . . . .	271
12.7	Obstacles and Challenges to Government Contracts, Croatia, Poland, and Romania, 2021 . . . . .	279
13.1	Diversity in Level of Procedural Adherence across Organizations and Divisions, Ghana. . . . .	296
13.2	Diversity in Content Scores across Organizations and Divisions, Ghana. . . . .	297
13.3	Relationship between Adherence to Procedure and Quality of Content, Ghana . . . . .	298
13.4	Average Number of SMART Objectives Identified in Appraisal Forms, Liberia. . . . .	301
13.5	Average Number of Relevant and Measurable Indicators Identified in Appraisal Forms, Liberia . . . . .	301
13.6	Diversity in Number of Days to Receive Requested Information from Organizations and Divisions, Ghana. . . . .	302
13.7	Diversity in Proportion of Required Files Submitted by Organizations and Divisions, Ghana . . . . .	303
14.1	Diagram of Customs Process. . . . .	312
B14.2.1	Example of an ASYCUDA Extract: Basic Variables. . . . .	316
B14.2.2	Example of an ASYCUDA Extract: Duty, Excise, and Value-Added Taxes Variables. . . . .	316
B14.2.3	Example of an ASYCUDA Extract: Time Stamps. . . . .	317
14.2	Value of Different Product Categories Imported to the United States for 50 Largest Ports of Entry, as Appraised by US Customs and Border Protection . . . . .	318
14.3	Border Compliance Times in Cross-Country and Regional View . . . . .	319
14.4	Example of Indicators to Measure Performance: Transit Time on the Northern Corridor . . . . .	320
B14.3.1	World Trade Organization Valuation Methods, Arranged Sequentially . . . . .	321
B14.4.1	Changes in Malagasy Customs Officials’ Performance. . . . .	323
15.1	Cross-Country Scale of the Three Sectors Discussed in the Chapter Relative to National Gross Domestic Product. . . . .	329
15.2	Expected Processing Time for Most Common Types of Pensions and Welfare Transfers, Italy . . . . .	334
15.3	Summary Statistics on the 25 Most Commonly Purchased Goods in the Punjab Online Procurement System, 2014–16 . . . . .	338
B16.2.1	Impact of One-Page Feedback Reports on Case Delays in Kenya . . . . .	361
17.1	Task Types across Organizations . . . . .	367
17.2	A Spectrum of Task Completion Measures, with Selected Examples. . . . .	379
18.1	Countries with Regular, Governmentwide Employee Surveys, Cumulative Count, 2002–21. . . . .	389
18.2	Management Practices Measured in Government Employee Surveys. . . . .	391
18.3	Employee Attitudes and Behaviors Measured in Government Employee Surveys . . . . .	391
18.4	Share of Public Servants Satisfied with Their Pay and/or Total Benefits, Various Countries. . . . .	396
19.1	Online and Face-to-Face Survey Response Rates, by Organization. . . . .	406
19.2	Online Survey Breakoff, by Page. . . . .	407



19.3	Average Item Difference, by Survey Topic .....	410
19.4	Average Modal Difference, by Organization .....	411
19.5	Organization Rankings, by Index and Mode .....	412
19.6	Distribution of Survey Mode Differences across Matched Individuals .....	413
19.7	Average Item Difference, by Managerial Status .....	414
19.8	Difference in Scores, by Response Rate .....	417
19.9	Average Item Difference, by Sample .....	418
20.1	Most Commonly Used Survey Measures of Attitudes and Behaviors across Civil Servant Surveys. ....	429
20.2	Most Commonly Used Survey Measures of Management Practices across Civil Servant Surveys .....	429
20.3	Full-Sample Distribution of Key Indicators: National and Institutional Values across Surveys .....	434
20.4	Simulated Distribution of Key Indicators: National-Level Values across Surveys and Sampling Proportions .....	439
20.5	Precision and Power of Simulated Distributions across Surveys. ....	443
21.1	Topics Measured in Government Employee Surveys .....	453
21.2	Distributions of Standardized Scores .....	461
21.3	Average Response Variance across Surveys and Question Topics .....	462
22.1	Share of Missing Responses, by Survey Module .....	473
22.2	Correlation between Subdimensions of Complexity and Sensitivity .....	484
22.3	Share of Missing Responses .....	485
22.4	Relationship between Complexity and Sensitivity Indexes and the Share of Missing Responses. ....	486
22.5	Relationship between Machine-Coded Complexity Scores: Correlograms and Scree Plots from Principal Component Analysis .....	491
22.6	Machine-Coded Complexity Indicators, Romania. ....	493
23.1	Intraclass Correlations for the Romanian Data .....	508
23.2	Intraclass Correlations for the Guatemalan Data .....	509
23.3	Organizational and Individual Referents in the Romanian Data .....	511
23.4	Organizational and Individual Referents in the Guatemalan Data .....	512
23.5	Estimates of Referent Effects on the Likelihood of Item Nonresponse. ....	513
23.6	Distributions of Referent Effects for Sensitive and Nonsensitive Questions .....	516
23.7	Response Score and Question-Referent Effects in the Romanian Data .....	517
24.1	Schematic Visual Representation of the Three Levels of Measurement Invariance: Configural, Metric, and Scalar .....	529
24.2	Measurement Invariance across Countries Classified by Region: Change in Model Fit Indexes across Configural-, Metric-, and Scalar-Invariance Models .....	536
24.3	Measurement Invariance across Countries Classified by Income Group: Change in Model Fit Indexes across Configural-, Metric-, and Scalar-Invariance Models .....	537
24.4	Measurement Invariance across Gender within Countries: Change in Model Fit Indexes across Configural-, Metric-, and Scalar-Invariance Models. ....	539
24.5	Measurement Invariance across Education Levels within Countries: Change in Model Fit Indexes across Configural-, Metric-, and Scalar-Invariance Models .....	540
24.6	Measurement Invariance across Public Administration Organizations within Countries: Change in Model Fit Indexes across Configural-, Metric-, and Scalar-Invariance Models .....	542
25.1	Share of Public Servants Satisfied with Their Pay and/or Total Benefits, Various Countries. ....	553
25.2	Results Report from the UK Civil Service People Survey .....	559
25.3	Results Report from Ireland, Top 5 Positive Results .....	559
25.4	International Benchmarking in Results Report from Ireland .....	560
25.5	State of the Service Report from Australia .....	561
25.6	Canada Public Service Employee Survey Dashboard .....	562
25.7	Canada Public Service Employee Survey Dashboard for Persons with a Disability .....	563
25.8	United Kingdom Civil Service People Survey Results Dashboard for Organizations and Teams. ....	563
25.9	Canada's Heat Maps for Survey Results of Units .....	564
25.10	Australian Public Service Commission Action Plan Template .....	566



25.11	Best Places to Work in the US Federal Government, 2022 Rankings .....	567
25.12	“At a Glance” Dashboard, Government of Ireland .....	568
26.1	Timeline of the Evolution of the Federal Employee Viewpoint Survey .....	579
26.2	Variation across and within Agencies in the Employee Engagement Index .....	580
26.3	The Architecture of Policy Action for the Federal Employee Viewpoint Survey .....	581
26.4	National Institute of Diabetes and Digestive and Kidney Diseases Staff Responses to Federal Employee Viewpoint Survey “Belief in Action” Question Compared to Organization-Level and Governmentwide Responses .....	586
26.5	National Institute of Diabetes and Digestive and Kidney Diseases Staff Responses to Federal Employee Viewpoint Survey “Accountability” Question Compared to Organization-Level and Governmentwide Responses .....	587
26.6	Trends in Negative and Positive Responses to Federal Employee Viewpoint Survey Questions, 2004–19 .....	590
27.1	Differences in Public Sector Employment, by Region, 2000–18 .....	603
27.2	Relative Size of the Public Sector Workforce, 2000–18 .....	604
27.3	Areas of Economic Activity as a Segment of Public Employment, by Region, 2000–18 .....	605
27.4	Public Sector Education and Health Care Employment, by Region, 2000–18 .....	605
27.5	Tertiary Education among Public and Private Sector Workers, by Region, 2000–18 .....	606
27.6	Share of Female Workers in the Public versus Private Sector, by Region, 2000–18 .....	607
27.7	Public Sector Wage Premium Compared to Country Income, by Region, 2000–18 .....	608
27.8	Public versus Private Sector Pay Compression Ratios, 2000–18 .....	609
27.9	Pay Inequity in the Brazilian Public Sector, 2020 .....	610
28.1	Determinants of Trust in the Civil Service in Norway .....	625
28.2	Trust Payoff Associated with Increase in Public Institutions’ Competence and Values in Finland, 2020 .....	625
29.1	MSD Health Indicators for a Selection of Countries in Africa .....	635
29.2	Updated Conceptual Framework and Questionnaire Modules for the Service Delivery Indicators Health Survey .....	636
29.3	Dimensions of the Global Education Policy Dashboard .....	637
29.4	Example of Global Education Policy Dashboard .....	641
C.1	Advances as a Share of Total Spending, Cambodia, 2016/17 .....	671
G.1	Distribution of Survey Mode Differences across Matched Individuals, Alternative .....	693
G.2	Survey Response and Breakoff Rates, by Mode .....	701
G.3	Mixed Model Results: Estimated Intercepts and Random Effects of Survey Mode Dummy across Organizations .....	701
H.1	Chile: Satisfaction Question .....	706
H.2	Chile: Motivation Question .....	706
H.3	Chile: Performance Review Question .....	707
H.4	Chile: Motivation Index .....	707
H.5	Chile: Leadership Index .....	708
H.6	Chile: Performance Index .....	708
H.7	Chile: Recruitment Question .....	709
H.8	Romania: Motivation Index .....	709
H.9	Romania: Leadership Index .....	710
H.10	Romania: Performance Index .....	710
H.11	Romania: Satisfaction Question .....	711
H.12	Romania: Motivation Question .....	711
H.13	Romania: Performance Review Question .....	712
H.14	Romania: Recruitment Question .....	712
I.1	Residuals before and after Box-Cox Transformation .....	729
L.1	Concept of Configural Invariance Visualized .....	747
L.2	Concept of Metric Invariance Visualized .....	748

L.3	Concept of Scalar Invariance Visualized . . . . .	749
L.4	Correlations between Questions from the Transformational Leadership and Salaries Sections of the Global Survey of Public Servants Administered across Seven Countries . . . . .	750
M.1	PMA Site on Career Development Programs . . . . .	753
M.2	PMA Site on Supervisory Career Development Programs . . . . .	754
M.3	PMA Site on Career Climbers Cohort. . . . .	755
M.4	OPM Link to PMA Resources . . . . .	756
M.5	EVS ART Online Dashboard. . . . .	757
M.6	EVS ART Results Dashboard Showing Institute Level . . . . .	758
M.7	EVS ART Results Dashboard Showing Office Level, with a Focus on Organizations. . . . .	759
M.8	EVS ART Results Dashboard Showing Office Level, with a Focus on Topics . . . . .	760

## MAPS

2.1	Variations in Productivity of Processing Social Security Cases, Subregions of Italy . . . . .	24
2.2	Subnational Patterns in Public Sector Employment, Indonesia, 2018. . . . .	34
12.1	Use of Electronic Government Procurements across the World, 2020 . . . . .	262
14.1	Customs and Other Import Duties as a Percentage of Tax Revenue . . . . .	309
B14.1.1	Location of Malawi. . . . .	310
B14.4.1	Location of Madagascar . . . . .	322
15.1	Variations in Productivity of Processing Social Security Cases, Subregions of Italy . . . . .	333
27.1	Subnational Patterns in Public Sector Employment, Indonesia, 2018. . . . .	611

## TABLES

2A.1	Mapping Government Analytics Data Sources to Chapters 10–30. . . . .	39
6.1	Types of Data Collected on Public Administration, with Examples . . . . .	97
8.1	Examples of Public Procurement Indicators . . . . .	137
9.1	Human Resources Modules and Associated Measures . . . . .	158
9.2	Project Timeline for Luxembourg's HR BICC. . . . .	177
10.1	Example of Payroll + Human Resources Microdata Set Showing Minimum Data Required . . . . .	217
11.1	Example of Expenditure Data, by Transactions . . . . .	240
11.2	Example of Classification of Functions of Government from the Health Sector . . . . .	244
11.3	Sample Output of FMIS Budget Coverage Estimation. . . . .	248
11.4	Template for a Government Expenditure Transactions Profile. . . . .	249
12.1	Examples of Public Procurement Indicators . . . . .	263
12.2	Regression Analysis of Framework Agreements versus Other Open Methods, Brazil . . . . .	270
13.1	Reasons for Incomplete Adoption or Nonadoption of the PMS Process, Liberia . . . . .	291
13.2	Completion of PMS Forms, Liberia, 2017–19. . . . .	295
13.3	Procedural Characteristics of Assessed Files, Ghana . . . . .	295
13.4	Content Characteristics of Assessed Files, Ghana . . . . .	296
13.5	Sufficiency of Information for Assessing Quality of PMS Forms, Liberia . . . . .	299
13.6	Challenges in Judging the Quality of PMS Forms, Liberia . . . . .	299
13.7	Formulating and Reporting on Objectives and Targets, Liberia. . . . .	300
13.8	Quality of Supervisors' Feedback, Liberia . . . . .	300
17.1	Selected Measures of Task Completion . . . . .	375
17.2	Selected Measures of Task Characteristics. . . . .	377
18.1	Countries with Regular, Governmentwide Employee Surveys, 2002–22 . . . . .	389
18.2	Methodological Choices in Public Servant Surveys. . . . .	393
19.1	Balance in Demographic Characteristics between Surveys. . . . .	406
19.2	Mean Modal Difference, by Level of Analysis . . . . .	409
19.3	Ordinary Least Squares Results: Individual Characteristics and Mean Survey Differences . . . . .	415
19.4	Mean Modal Differences at the National Level, by Weighting Approach. . . . .	418

20.1	Characteristics of Surveys Included for Simulations . . . . .	430
20.2	Overview of Survey Questions for All Items Included in the Simulations, by Survey. . . . .	431
21.1	Surveys Used in the Analysis. . . . .	457
21.2	Descriptive Statistics for Surveys of Public Servants. . . . .	458
21.3	Compare Models: ANOVAs, Nested. . . . .	463
22.1	Complexity Coding Framework. . . . .	481
22.2	Sensitivity Coding Framework. . . . .	482
22.3	The Impacts of Complexity and Sensitivity on Item Nonresponse . . . . .	487
22.4	Exploratory Factor Analysis . . . . .	487
22.5	Factor Analysis Regression . . . . .	489
22.6	Impact of Sensitivity, Complexity, and Unfamiliarity on Nonresponse Rate . . . . .	490
22.7	Impact of Machine-Coded Complexity on Nonresponse Rate. . . . .	492
22.8	Full Model: Impact of Sensitivity, Complexity, Unfamiliarity, and Machine-Coded Complexity . . . . .	492
23.1	Advantages and Disadvantages of Organizational and Individual Referents When Used for Calculating and Analyzing Organizational Aggregates. . . . .	501
23.2	Question Referents and Target Entities . . . . .	501
23.3	Estimated Differences in Relationships between Leadership Variables for Different Referents, Guatemala. . . . .	514
23.4	Standardized Question-Referent Effects, by Sensitivity . . . . .	515
23.5	Question-Referent Effects, by Years of Service, Romania. . . . .	518
24.1	Summary of the Seven Public Servant Surveys Used in the Chapter. . . . .	533
24.2	Basic Statistics on the Three Questions Aiming to Measure Transformational Leadership . . . . .	534
24.3	Distribution of Respondents, by Gender . . . . .	538
24.4	Distribution of Respondents, by Education Level. . . . .	540
24.5	Distribution of Respondents within Public Administration Organizations . . . . .	541
25.1	Comparing Country Approaches to Making the Most of Survey Results . . . . .	556
27.1	Variables Required for the Inclusion of a Survey in the WWBI . . . . .	601
28.1	Deconstructing Citizens' Trust in Public Institutions. . . . .	622
28.2	OECD Framework on Drivers of Trust in Public Institutions . . . . .	623
28.3	Examples of Questions on the Determinants of Public Trust . . . . .	624
29.1	Survey of Global Initiatives in Education and Health Care Delivery Indicators . . . . .	632
29.2	Indicators in the First Service Delivery Indicators Health Surveys . . . . .	636
29.3	Global Education Policy Dashboard Indicators. . . . .	638
A.1	Checklist for Using Expansive and Qualified Measurement for Informed Problem Solving and Learning in Public Administration . . . . .	660
C.1	Share of FMIS System Coverage, Pakistan, 2016/17 . . . . .	669
C.2	Budget Items Not Routed through the FMIS at the Federal Level, Pakistan, 2016/17. . . . .	670
C.3	Total Amount Processed through the FMIS and Values They Represent, Cambodia, 2016/17. . . . .	671
D.1	Public Procurement Indicators. . . . .	673
E.1	Tool to Assess Quality of Government Process, Ghana . . . . .	676
E.2	Tool to Assess Quality of Government Process, Liberia . . . . .	679
G.1	Mean Modal Differences, by Level of Analysis, Restricted Sample. . . . .	694
G.2	Cox-Weibull Hazard Model of Online Survey Breakoff . . . . .	694
G.3	OLS Results: Number of "Don't Know," Refusal, and Missing Responses . . . . .	695
G.4	OLS Results: Organizational Characteristics and Mean Survey Differences . . . . .	696
G.5	OLS Results: Propensity Score Matched Estimates . . . . .	697
G.6	OLS Results: Individual Characteristics and Mean Survey Differences . . . . .	698
G.7	OLS Results: Propensity Score Matched Estimates . . . . .	699
G.8	Survey Items, by Index . . . . .	699
H.1	Romania: Statistics for Chosen Indicators and Sampling Proportions . . . . .	703
H.2	Chile: Statistics for Chosen Indicators and Sampling Proportions. . . . .	704
H.3	Liberia: Statistics for Chosen Indicators and Sampling Proportions . . . . .	704

I.1	Survey Question Phrasing .....	713
I.2	Robustness Check: ANOVAs and Nested Model with Outliers .....	723
I.3	Compare Models, Full Data Set versus List-Wise Deletion: ANOVAs, $R^2$ .....	724
I.4	Regression Diagnostic Statistics .....	728
K.1	Organizational Commitment and Identification Using Other Measures .....	745
L.1	Distribution of Respondents: Survey Sample vs. Civil Service Population .....	746

# Foreword

The quality of governance has long been recognized as central to growth, progress toward equity, and social stability. The best-laid development plans founder on shortfalls in state capacity, corruption undermines confidence in the state management of resources, and poor service delivery generates frustration across all aspects of the daily life of a citizen. The world is now witnessing an unprecedented explosion of data that has revolutionized both analysis and outcomes across a range of fields. This volume, *The Government Analytics Handbook*, intends to advance that revolution with respect to how governments function.

Leveraging data more effectively leads to greater experimentation and new forms of analysis. It sharpens the questions asked about how the world functions. In the private sector, new measurement and analytics have helped firms confront the realities they face and experiment with novel responses. Academics, meanwhile, have mapped the determinants of private sector productivity ever more precisely. This has meant huge gains in both productivity and profitability for private sector firms that have best harnessed measurement techniques.

In the realm of international development, the World Bank's own drive to collect data on poverty helped to move to center stage efforts to amass and leverage data, spurring focused efforts to ameliorate poverty and catalyzing a whole academic subdiscipline. Similarly, our investments in a range of service-delivery indicators such as measures of the quality of schooling and health provision have been instrumental in our work in strengthening the quality of frontline government services.

Yet in the sphere of government, relatively limited work on the measurement of public administration has prevented a comparable evolution in this area. However, the digitization of administrative and other data now offers the opportunity to strengthen public administration in a way heretofore unimaginable, making this volume both timely and of great significance. By offering the first compendium of global progress in using new sources of information and new techniques to diagnose state administrations, it offers a vital resource to help governments around the world rapidly improve their approach to public administration.

The book also showcases policies to remedy weaknesses in a government's set of analytical tools and abilities. Improvements in these areas, in turn, can lead to improvements in the ways that governments function and the outputs and outcomes they produce.

Project teams across the World Bank—both in the Governance Global Practice and beyond—should capitalize on the ideas and toolkits that the *Handbook* contains as the basis for engagement with government clients to strengthen their government analytics abilities and, in turn, their management to improve their functions and delivery.

We also hope that this *Handbook* provides both inspiration and insights for governments interested in utilizing their administrative and survey data in new ways and repurposing them to deliver better public services. Through our work with governments around the world, we have seen firsthand the transformative power of government analytics. And we hope it acts as a basis for new communities of practice at the World Bank, in academia, and among governments.

Beyond the individual activities the *Handbook* can promote, it is also an opportunity to fundamentally change the approach to public administration that the world has taken for much of the history of government. Rather than acting as machinery for recordkeeping, government analytics offers the opportunity for governments to drive innovation and continuous experimentation across their societies and economies. Governments generate, receive, and connect data at an unprecedented scale. Analyzing these data in the way this *Handbook* recommends could fundamentally shift the role of the public sector in facilitating development.

But governments will need to use these data to change themselves as well. That calls for introspection by public administrations. Implementing government analytics requires more than just technical expertise. It also involves building a culture of data-driven decision-making, collaborating across government agencies, and addressing ethical and privacy concerns about data collection and use. The World Bank is committed to supporting governments in their efforts to harness the power of government analytics for the benefit of their citizens. By working together, we can unlock the potential of government analytics to drive inclusive and sustainable development. That effort is surely a pillar of a better tomorrow for development.

**Indermit Gill**

Chief Economist, World Bank Group  
Senior Vice President,  
Development Economics, The World Bank

**William F. Maloney**

Chief Economist, Latin America and the Caribbean Region  
Former Chief Economist, Equitable Growth, Finance,  
and Institutions Vice Presidency, The World Bank

# Acknowledgments

## In Memoriam

Steve Knack, who passed in 2019, was a cherished member of the World Bank's research community. He epitomized respect for, and partnership with, the World Bank's governance-focused operational colleagues. We are grateful for having had the opportunity to know Steve, be inspired by his approach to development, and learn from him. While we miss him and his warmth and collegiality, we take his spirit with us in trying to forge a stronger partnership between research and policy.

*The Government Analytics Handbook* is an output of the World Bank Bureaucracy Lab, a collaboration between the Development Impact Evaluation (DIME) Department of the Development Economics Vice Presidency and the World Bank Governance Global Practice. It has benefited greatly from a close collaboration among DIME, the Governance Global Practice, and the Office of the Chief Economist of the Equitable Growth, Finance, and Institutions (EFI) Vice Presidency. The editors are grateful to all of the members of the Lab past and present for their ongoing passion for, and stimulating approach to, strengthening government administration.

The book was edited by Daniel Rogger (senior economist, World Bank) and Christian Schuster (professor in public management, University College London), with support from a World Bank team consisting of Galileu Kim, Robert Lipinski, and Annabelle Wittels. It was prepared under the overall guidance of William F. Maloney during his tenure as EFI chief economist and subsequently Ayhan Kose, and Arianna Legovini, director of DIME. The editors are grateful to each of these individuals for their support of this project. The editors thank Tracey Lane, practice manager for the Public Administration and Institutional Reforms unit of the Governance Global Practice, under which the Bureaucracy Lab was located during the period in which the substantive development of the book occurred.

The book underwent rigorous internal and external review. The editors would like to express their gratitude to Geert Bouckaert (professor at the Public Governance Institute, KU Leuven) and Benjamin Lauderdale (professor of political science, University College London), who provided an academic peer review of most of the technical content of the book. Quality assurance in terms of technical competence and policy relevance was provided by internal reviewers at the World Bank. A thorough assessment of how to ensure the book project was relevant to policy makers was undertaken by Francisco Gaetani and Elizabeth Obeng-Yeboah. The book also benefited from inputs from Kerenssa Kay and other members of the Steering Committee, as well as numerous public officials. The editors would like to thank them all for their very helpful comments and suggestions. Finally, the editors are indebted to a vast number of public servants, researchers, and members of their departments and organizations, all of whom have shaped their work on government analytics. The lessons they provided are echoed throughout the *Handbook*.

**Postdoctoral researchers:** Galileu Kim and Annabelle Wittels

**Research assistant:** Robert Lipinski

**Academic peer reviewers:** Geert Bouckaert and Benjamin Lauderdale

**Policy peer reviewers:** Francisco Gaetani and Elizabeth Obeng-Yeboah

**World Bank peer reviewers:** Donna Andrews, Pierre Bachas, Jurgen Blum, Gero Carletto, Verena Fritz, Florence Kondylis, and Daniel Ortega Nieto

**Steering Committee:** Elsa Araya, Vincenzo Di Maro, Zahid Hasnain, Maria Ruth Jones, Kerenssa Kay, William F. Maloney, Gbemisola Oseni, and Michael Roscitt



# About the Editors

**Daniel Rogger** is a senior economist in the Development Impact Evaluation (DIME) Department at the World Bank. He manages the Governance and Institution Building unit of DIME and is colead of the World Bank's Bureaucracy Lab, a collaboration between DIME and the Governance Global Practice that aims to bridge research and policy to strengthen public administration. His research focuses on the organization of the delivery of public goods. He is a cofounder of the Worldwide Bureaucracy Indicators, Global Survey of Public Servants, and Microdata and Evidence for Government Action initiatives. He was a PhD scholar at the Institute for Fiscal Studies, where he is now an international research fellow. He holds a PhD in economics from University College London.

**Christian Schuster** is professor in public management at University College London. His core research interest lies in data and civil service management. His research is widely published, with more than 70 publications, and he has won a range of awards, including the 2018 Haldane Prize for the best article published in *Public Administration*. He cofounded the Global Survey of Public Servants, coleads the Centre for People Analytics in Government, and has collaborated in his research with more than 30 governments, as well as a range of international organizations. He was previously an economist with the World Bank and a visiting research scholar with the Inter-American Development Bank. He holds a PhD in government from the London School of Economics.



The background features a series of blue squares arranged in a grid-like pattern that curves upwards from the left. Overlaid on this is a stream of binary code (0s and 1s) that also curves upwards, creating a sense of digital flow and data movement.

# **PART 1**

## Overview



# CHAPTER 1

## Introduction to *The Government Analytics Handbook*

*Daniel Rogger and Christian Schuster*

### THE TRANSFORMATIVE POWER OF GOVERNMENT ANALYTICS

In 2016, Kim Wells, a senior member of the US federal government's Federal Employee Viewpoint Survey (FEVS) team—perhaps the world's most famous survey of public servants—had an appointment with an imposing ex-marine who had entered the public service as a manager. The marine still had a soldier's physicality about him as he entered the conference room with Kim, and you could see that he brought his military management style with him to his duties. For so many members of the US military, the idea of excellence is fundamental to how they see themselves and their work. His identity was rooted in the idea that he was an outstanding manager. And if you had asked him, he would have woven a narrative of success against all odds in the work that he and his team were doing.

Yet he was failing: failing to create a work environment for his staff in which they felt engaged, mentored, and safe. Kim had surveyed his entire team, giving each of them a chance to provide feedback and for that feedback to be compared with the experiences of other officials working under different managers. And the truth was that this burly ex-marine was failing his team, himself, and his country. He broke down in tears in front of Kim. His view of himself had been confronted by the survey data that gave his staff a voice they could not have had otherwise. He knew he needed to change and improve how he managed his team—and the survey data told him exactly how to go about doing it.

At roughly the same time, but more than four thousand miles to the south, Brazil's federal government was heading for financial catastrophe. Under the existing pay and pensions regime that compensated federal employees, wage costs were about to skyrocket. Seeing the impending danger through the dense wording of public contracting law was daunting. Mirian, a member of the Ministry of Economy, suspected something was wrong, but couldn't put a finger on what lay ahead. So, Mirian had a team calculate what the future of payroll and pensions looked like for every individual in the federal government under the existing regime. The danger suddenly seemed very real. As wages skyrocketed, funding for other inputs to government services would become unaffordable. Services would have to be stopped. Fortunately, Mirian had also asked the team to model other feasible scenarios. These cases gave the government the means to negotiate with politicians and other stakeholders and pass legislation to change compensation rules in time to avert catastrophe.

Four thousand miles east, the government of Nigeria had received debt relief from the Paris Club group of creditors in 2005 worth US\$18 billion. Many Nigerians wanted to know what would happen to those funds, including Amina Mohammed, a northern Nigerian who was almost invariably clothed in traditional dress and had a background in both engineering and civil society advocacy. The president asked Amina to join his team, and over the next few years she built one of the world's most innovative public sector tracking systems to follow the financial gains of debt relief through government.

From her office in the Presidency, Amina tracked every naira of those funds, combining budget, program, and audit data systems and sending teams to visit every project site. It was truly frontier analytics, showing where and how badly the government was failing. Some organizations fulfilled their commitments completely. Others did not. In 2006, for instance, the Ministry of Water Resources received US\$475 million to develop water infrastructure across the country. In return, it produced nothing. The ministry's officials seemed busy and budgetary releases were made. But when Amina's teams visited the sites that had been allocated funds all across the country, they could not find a single project that had been completed. Amina took this evidence to the president and won the political and bureaucratic space she needed to create new ways of spending government resources, such as a grants scheme to state governments that only paid out if water infrastructure was actually produced.

This book, *The Government Analytics Handbook*, is about enabling individuals like Kim, Mirian, and Amina to change their governments for the better. It draws on a moment in history when the world is capitalizing on innovations in measurement, data collection, and data analysis at an unprecedented scale. Never before has the world been able to build a richer picture of the realities of the public sector. The question for each and every public sector official, manager, and leader is what they are going to do with this revolution. How governments collect, analyze, and use microdata to improve the administration of government—or undertake what this *Handbook* calls government analytics—will determine how effective they are in this new world.

Government analytics can help solve big issues in public administration—governmentwide administrative challenges, as in the case of future fiscal liabilities from Brazil's payroll. But as important it can also help government organizations improve in small ways, addressing specific management challenges in specific teams in specific government organizations, as with the example of the former US marine. When small improvements happen across thousands of teams inside government—as enabled by regular governmentwide employee surveys, for instance—even small changes can transform government.

What do we mean by government analytics? It is the repurposing of administrative and survey data from within government to improve the way government functions. It uses microdata to diagnose the inputs, management practices, processes, outputs, or outcomes in public sector organizations, units inside such organizations, and/or public administration as a whole. These diagnoses can pinpoint how well government is functioning—or not. Microdata provide information about the characteristics of individual people or entities such as individual officials or departmental units in the case of government, or households, business enterprises, or farms in the case of the private sector. Such data can measure and study relationships among phenomena at a very granular scale, such as how the management practices of individual public service managers affect the productivity of their teams. Microdata can come from a range of sources: unit-level data obtained from sample surveys, wider censuses, and general administrative systems.

Government analytics is not restricted to governments with more advanced information technology (IT) platforms and employee records, like Brazil and the United States. Instead, it has been of use to governments around the world and at all stages of administrative development, as illustrated by Amina's efforts in Nigeria. We have had the good fortune of being collaborators and witnesses to many such improvements. Government analytics has, for instance, led to more rigorous merit recruitment procedures in Kosovo's government, better employee onboarding in Chile's government, staff mentoring in Ethiopia's government, higher-quality public service training in Ghana and Nepal, improved quality of management in Croatia's government, and better public procurement practices in Romania and Uruguay. The list goes on. Many more examples are contained across the 30 chapters of this *Handbook*.

Although many instances of government analytics are being carried out at an individual level, there is a lack of systematic practice in governments as a whole. This means that governments are missing out on the potential insights available to them for improving their public administrations at scale. This, in turn, means that money is being left on the table. Public revenues that could be spent more efficiently, with greater impact on the welfare of citizens, are simply not being spent as well as they could. It is time to pick up those funds and use them for a better society.

How can this *Handbook* help? By showcasing how effective and low cost analytics can be, the hope is that more governments will undertake analytics of their own administrations, and in a more systematic way. To make meaningful progress, we need to change the way we approach government, and this shift should reflect a broader change in what we expect to know about state institutions. Analytics have come to dominate discussions of many other spheres of life, and they should play a more significant role in efforts to strengthen the state.

Beyond any single government, there is a lack of systematic evidence on how to do analytics in a rigorous manner, and there are few carefully constructed global comparisons available. As a result, different governments tend to follow diverging practices even when undertaking similar analytics, limiting their ability to use objective benchmarks from other settings. For instance, as shown in chapter 18, different governments ask different questions in their employee surveys to measure the same concepts.

This *Handbook* aims to fill this gap. It presents frontier evidence and practitioner insights on how to leverage data to strengthen public administration. Across 30 chapters, it shows ways to transform the ability of governments to take a data-informed approach to diagnose and improve how public organizations work. An accompanying website contains tools for analytics, which enable readers to immediately apply insights from the *Handbook* in their own work ([www.worldbank.org/governmentanalytics](http://www.worldbank.org/governmentanalytics)). The *Handbook* covers many sources of microdata, ranging from administrative data, such as payroll, procurement, case, text, and human resources management information system (HRMIS) data; to public servant survey data; to data coming from external assessments, such as citizen and household surveys, or anthropological diagnostics of public administration. Methodologically, it covers both traditional qualitative and quantitative methods, as well as newer approaches, such as machine-learning diagnostics of unstructured text records from governments. To our knowledge, this is the first and most comprehensive volume of this kind.

## THE HIGH STAKES OF GOOD GOVERNMENT ANALYTICS

In their magisterial review of the formation of the state, Acemoglu and Robinson (2019, 341) note that “bureaucracy is vital to state capacity.” Whether a country’s laws and policies will indeed be implemented is determined by the quality of its public administration. Extensive research confirms that the quality of government administration affects institutional quality, safety guarantees, education opportunities, health care provision, and ultimately, the welfare of society and the economy (see, among others, Besley et al. 2022; Dahlström and Lapuente 2022; Finan, Olken, and Pande 2015; Pepinsky, Pierskalla, and Sacks 2017; Wilson 1989).

Significant opportunities exist for improving these outcomes through better administration, a variety of studies around the world show. For example, in the Russian Federation, researchers found that the quality of bureaucrats and their organizations accounted for two-thirds of the variation in the cost of public procurement contracts (Best, Hjort, and Szakonyi 2017). By reducing the prices paid by the worst-performing 25 percent of procurement agents to no more than that paid by the other 75 percent of agents, the Russian government could save approximately US\$10 billion each year—a sum equivalent to about 15 percent of Russia’s total public health care spending. Similarly, improving the quality of management in government organizations in Nigeria by a standardized unit could raise by 32 percent the likelihood

that a physical infrastructure project is completed (Rasul and Rogger 2018). In Italy, reassigning managers to place the best managers in the largest offices would boost productivity in social security claim processing by at least 7 percent (Fenizia 2022). In Pakistan, offering the best-performing tax collectors their top choice of job posting would increase tax revenue by 40 percent (Khan, Khwaja, and Olken 2019). Conversely, poor management of the administration leads to worse outcomes. In Brazil, politicized turnover of public personnel, including at schools, significantly lowers student learning (Akhtari, Moreira, and Trucco 2022). Corruption in education funds is associated with a 65 percent increase in dropout rates from schools (Ferraz, Finan, and Moreira 2012).

The essence of these studies is that good management in government can lead to significant and rapid improvements in government performance. Conversely, when public administration is weak or inefficient, programs and policies are far more likely to fail. The size of the findings of recent evaluations of public administration reforms indicates that there is perhaps no more effective means of improving public policy than by strengthening the quality of public administration.

The sheer scale of most countries' public administrations makes their quality important. The cost of wages for public sector employees is approximately 10 percent of gross domestic product (GDP) across the world, not even counting public sector pensions (World Bank 2020; World Bank Group 2019). That is a significant portion of the economy to ensure is managed effectively. Similarly, the assets the public administration manages directly are large. Across the world, public procurement of goods and services accounts for roughly 12 percent of GDP (Bosio and Djankov 2020). As the study of Russian procurement suggests, ensuring that governments are paying appropriate prices for these purchases would yield much more money to spend on health care or other welfare improvements.

A direct consequence of the size of government is its influence over the rest of the economy. Globally, the public sector makes up 38 percent of formal employment (World Bank 2020). As such a large employer, it plays an influential role in the wider labor market, particularly for tertiary educated workers (Somani 2021). The same can be said for the prices it pays for the goods it procures, and the stimulus it induces when it builds infrastructure or regulates business. So even if interest is solely in improving the private sector, government analytics matters.

Yet with so much at stake, and such large margins for improvement to exploit, governments everywhere are not exploiting modern analytics (World Bank 2016; World Bank 2021). Although governments worldwide have invested heavily in digitizing their administrative work—dedicated digital government institutions have been established in 154 economies—few have systematized the use of the resulting records into data to strengthen the way they work, the World Bank's GovTech Maturity Index reveals (World Bank 2022). As noted, analytics is not necessarily dependent on such digitization, but it is illustrative of wider commitments to analytics.

## THE ANALYTICS REVOLUTION

To understand the potential for government analytics, look no further than the private sector. Tech firms have generated trillions of dollars of value in part or in full by creating, managing, and providing access to diagnostic data. More broadly, the collection and analysis of microdata at previously unforeseen scale has been one of the main drivers of economic and social advancement—from machine-learning algorithms automating customer interactions to utility firms using service data to identify critical weaknesses in physical infrastructure.

Data have enabled firms to improve their own internal operations and management. Productivity is significantly higher among plants that use predictive analytics, an assessment of 30,000 US manufacturer establishments finds (Brynjolfsson, Gorodnichenko, and Kuehn 2021). Units within a business whose workers are engaged have 23 percent higher profit compared to business units with disengaged employees; they also see



significantly lower absenteeism, turnover, and accidents, and higher customer loyalty, the Gallup analytics and advisory organization found, using its database of employee surveys to identify the impact of an organization tracking and nurturing employee engagement (Gallup 2023). In other words, management without measurement—the historically dominant approach to management in firms—puts firms at a competitive disadvantage and undermines productivity. Consequently, an increasing share of private sector companies base their entire business model on analytics (Carrera and Dunleavy 2013).

Sophisticated data collection efforts have uncovered how differences within firms—for instance, across business units—account for the largest share in productivity differences between countries (Cusolito and Maloney 2018). Microdata on firm productivity showcases that variation in firm performance is often due to differences within the firm, such as leadership and management practice (Bloom, Sadun, and Van Reenen 2010; Bloom and Van Reenen 2010). The difference between productive and unproductive firms, it seems, is that unproductive firms allow unproductive units to persist. Or conversely, they do not learn the lessons they could from the most successful units. The most productive firms identify laggards through the use of administrative records or primary data, and target them for improvements or let them go—allowing the firm as a whole to flourish. The private sector, particularly in competitive markets, is disinclined to leave money on the table.

This data revolution in firms *could* be paralleled by one inside government, for at least two reasons. First, government already does analytics on the rest of society. Governments have made significant investments to strengthen the quality of data systems toward better policy making and service delivery, by heavily expanding their measurement of their citizens, firms, and the environment they govern. In most countries, household and firm surveys have become central policy tools to avoid making policy decisions in a data vacuum (Deaton 2003; Kraay 2006). The centrality of such data for state efficacy was striking during the COVID-19 pandemic, for instance, when governments' ability to create effective track-and-trace systems to isolate COVID cases varied widely (Fetzer 2021). In other words, governments have been developing the capabilities for doing analytics on everyone else, but have not turned that lens on their own administrations.

Second, governments also sit on substantial volumes of data that could be used for government analytics. Public administration is frequently symbolized by stacks of government files packed with information on individuals, projects, or tasks. As the use of information and communication technology (ICT) for government operations increases, these records are ever more digitized (Ugale, Zhao, and Fazekas 2019). The digitization of these government records in turn creates a “big data trail,” as a by-product of people's digital behavior. For instance, a microdata point is created every time a firm submits a public procurement bid, a judge makes a court ruling, or a human resource department makes a monthly pay transfer to a public employee. This makes creating the raw material for government analytics far easier than before.

In short, now more than ever, governments are equipped to undertake the government analytics necessary to understand and improve the machinery of public administration. Yet, despite having developed analytic capacity for service delivery by measuring *others* (such as firms and households), many governments are yet to harness the full power of data to measure *themselves* and their own operations. In other words, though government is increasingly developing an evidence base for its services, much further behind is the practice and evidence base for measuring the public administration that actually generates those services.

Existing country-level governance indicators—such as the Worldwide Governance Indicators (Kaufmann, Kraay, and Mastruzzi 2022), the Transparency International (2021) Corruption Perceptions Index, and the Freedom House (2021) Freedom in the World Index—have provided a window into the global distribution of government functioning. However, they are insufficiently granular to inform specific government improvements. Understanding, in broad terms, how effective a government is perceived to be by citizens and firms is helpful, but does not provide guidance on what specific actions governments could undertake to improve effectiveness. Government analytics does just that. Utilizing microdata, it zeroes in on specific inputs, management practices, outputs, and outcomes in specific government organizations, units

within them, and/or public administration as a whole. As such, it puts a more refined sense of reality into each official, manager, and leader's hands. Much as better data to understand what society looks like has transformed public policy for the better, a better sense of what government administration looks like will eventually transform the public administration for the better.

## HOW GOVERNMENT ANALYTICS CAN PROVIDE A STRONGER EVIDENCE BASE FOR IMPROVING GOVERNMENT

Governments across the world make hundreds of thousands of personnel management decisions, undertake millions of procurements, and execute billions of processes each day. The public servants responsible for these activities possess extensive experience and innate knowledge of their respective administrations, which measurement and analytics—no matter how advanced—cannot replace. As one observer noted, “You will never be able to measure away the public sector manager.”

Yet government analytics can provide a stronger evidence base to improve how public officials understand government administration. Rather than substituting for the knowledge of and conversations about the public service, analytics are a strong complement to them. For example, in an experiment with the rollout of a monitoring technology to support agricultural extension workers in Paraguay, researchers found that managers were able to predict which of their staff would benefit most from the program, strengthening its impacts (Dal Bó et al. 2021). As in the private sector and the rest of the economy, analytics and technology are great complements to those who capitalize on them. And without analytics, valuable data trails are being left unexplored, which could lead to missed opportunities for improved decision-making and service delivery.

For example, by utilizing the data sets discussed in the *Handbook*, government officials can recognize the strengths of staff in similar organizations, gain valuable insights to help identify excellent staff in their own organizations, and better allocate their own staff across offices or tasks. This does not mean that managers must abandon their own style; rather, they can learn from the best practices of others. Furthermore, the use of survey data can help governments meet increasing employee expectations for evidence-based staff management, as already practiced by private firms. As in the example that opened this chapter, surveys give employees a voice they would not otherwise have, enriching conversations and providing valuable insights to managers across the service.

Analytics makes internal accountability more effective. When numbers refer to the productivity of an entire sector or a large network of organizations, it can be difficult to hold relevant parties accountable for their performance. However, the microdata embedded in government analytics can enable heads of state to hold the relevant heads of organizations accountable, and heads of organizations in turn can hold the relevant managers of units within the organization accountable. For example, the head of a social security agency can hold directors of regional offices accountable when the speed and quality of social security claims processing at their office falls well below that of other offices. Ultimately, the use of government analytics can enrich conversations in government so that it is more targeted, more engaged with best practices from within the public service, and more grounded in reality.

Government analytics can also enhance the accountability of the government to the public, and improve public conversations about the machinery of government more broadly. By making analytics public, citizens, civil society, and the media can hold the government accountable for how it manages public administration. This can be particularly important when analytics reveal that the machinery of government is not being administered in the public interest. For example, citizens may be interested in knowing how procurement contracts are awarded and whether public sector jobs are based on merit or political grounds. As is the case internally, analytics ensures that public accountability is targeted to organizations where improvements are needed most. That can help avoid unfairly spreading blame across the entire government and reducing the likelihood that any single organization will change (Dunleavy 2017).

## THE OBSTACLES TO ADOPTING GOVERNMENT ANALYTICS

Government analytics thus promises a transformative shift toward evidence-based and continuous improvement of government administration. Why then have many governments not adopted analytics of their administration as proactively as their private sector counterparts?

One reason lies in the lack of systematic evidence on how to do government analytics, and the lack of a systematic compilation of the methods and data available to governments to this end. This *Handbook* is motivated by the need to fill this gap, and hopes to provide a first step toward addressing it.

A second reason is skill shortages for government analytics inside many public sector organizations—both to undertake analytics and to use analytics to improve management. Chapter 3 provides a road map toward addressing these skill shortages, for instance by creating government analytics units inside public sector organizations.

Third, digital records are often primarily designed to enable government operations—such as the award of a contract tender—rather than the analytics of such operations. Governments need to make investments to repurpose government records for analytics—for instance, by connecting, storing, and analyzing data in a cost-effective and secure manner (de Mello and Ter-Minassian 2020). Similarly, online employee surveys are another data source for government analytics. Digitization has made it much cheaper, quicker, and easier for governments to obtain in-depth feedback from employees at scale and to do so more frequently. Again, however, investments in analytics are needed to design, implement, interpret, and make use of employee survey results to improve public administration.

Beyond these obstacles are, however, at least four thornier limits inherent to the analytics of public administration: (1) measurability of outputs and outcomes; (2) institutional complexity; (3) political disincentives to measure; and (4) ethical constraints on measurement. Government analytics requires careful navigation of each of them.

First, not all outputs or outcomes of public administration can be measured. Unlike a private sector firm with a bottom line, public administration organizations have multidimensional missions and tasks whose outcomes and associated processes are often challenging to measure accurately. For instance, how can the quality of policy advice by civil servants to ministers or the quality of a budget a ministry of finance prepares for submission to parliament be measured at scale? Not everything that matters in public administration can be measured. This has made constituencies wary of investing heavily in measurement innovations in the public administration. This is also why analytics templates from private sector firms cannot be indiscriminately copied and applied in public sector organizations.

The response of the *Handbook* to these limits of observability in public administration is to improve, extend, and expand measurement while respecting its limits. This allows public managers to have the best possible knowledge available for administrative improvement. As there are inherent limits to what can be measured in public administration, even with better measurement, public managers need to keep in mind the limits of measurement and triangulate data with other forms of knowledge when devising administrative improvements. Otherwise, as a large literature on performance management in public sectors has found (see, for example, Hood 2006), imperfect measurement can generate a series of unintended and adverse consequences for public management—from gaming performance to effort substitution (expending more effort on measurable metrics of performance at the expense of important but unmeasured outputs and outcomes).

A second challenge is institutional. Public sector reforms struggle with path dependencies (the tendency to become committed to develop in certain ways as a result of structural properties or embedded beliefs and values) (Gains and Stokes 2005). Overhauling data infrastructures and monitoring structures is difficult when organizations fall under different mandates and jurisdictions. Implementation times might extend far into or even span several legislative periods, impairing political incentives for change.

Third, government analytics generates numbers and data that were previously not available. The creation of such data—much like greater transparency in government generally—will sometimes generate political winners and losers (Dargent et al. 2018). To illustrate, creating data on the scale of recruitment

into government based on personal connections rather than merit can generate an avalanche of media reports (Schuster et al. 2020). Government analytics is thus not apolitical. Analysts need to understand what different government actors want to know and what they want to *not* know about the machinery of public administration. Individual analysts and officials will have to negotiate the politics of their setting. Within those constraints, we believe that producing and publishing a broader and more accurate set of government analytics will eventually lead to better conversations within government, and a better government, for society at large.

Last, collecting data on public servants raises ethical concerns, which may limit the scope of analytics. In public administration, ethical considerations are further complicated by trade-offs between individual employee demands for privacy and public demands for values such as productivity, innovation, and accountability of public administrations. Balancing these considerations appropriately is thus central to an ethical pursuit of government analytics, but also implies limits: from an ethics perspective, not everything that can be measured should be measured (see chapter 6).

For understandable reasons, then, the public sector is slower to capitalize on the data revolution than private firms. Much of what it must focus on is harder to measure and requires balancing a greater number of considerations than most private sector activity, whether those considerations are institutional, political, or ethical. Governments have thus fallen behind in their use of modern analytic methods to design and manage public administration. But as the opening stories of how data were successfully used for management improvement in the United States, Brazil, and Nigeria show, this need not be the case. Government analytics can be a powerful tool to improve government if leveraged in the right way.

## GOVERNMENT ANALYTICS AS AN ALTERNATIVE APPROACH TO PUBLIC ADMINISTRATION REFORM

How should governments go about the business of improving government administration? And how should others—such as international development organizations—support these improvement processes? The answer in many countries has been better laws that align governments with “best practices” for the administration of government. However, implementing best-practice public financial management legislation alone has had limited impact in many contexts (Andrews 2013). Similarly, while the number of countries with best-practice merit-based civil service legislation on the books has multiplied, little improvement in merit-based civil service practices has resulted (Schuster 2017). Global best practice may not be an appropriate fit for many country contexts (Grindle 2004, 2007).

Legislation requires a catalyst to make it the reality of government practice. Some observers have urged practitioners to focus on identifying local administrative problems facing particular organizations and attempt solutions in an iterative manner (Andrews 2013; Pritchett, Andrews, and Woolcock 2017). While focusing on addressing specific organizational problems is central to management improvement in public administration, this approach begs immediate questions: What helps officials and governments know which administrative problems are faced by which teams? How can officials know how isolated (or not) those problems are in public sector organizations? And how can we know whether solutions—developed locally or borrowed from other countries or global best practice—have effectively resolved the targeted problems?

This information gap can be addressed at scale through government analytics that leverage governmentwide microdata to diagnose every team in every public organization—for instance, through the data sources outlined in this *Handbook*. This approach can help the government as a whole, specific organizations within it, and individual teams better understand the specific management problems they are facing. Through benchmarking with other teams, organizations, or governments, data analytics can also shed light on opportunities for improvement and who might already be practicing useful approaches. And after governments have undertaken actions to bring about improvements, analytics can help practitioners understand

whether those actions were effective—for instance, in terms of lower procurement costs, lower staff turnover, or better ratings of the quality of managerial leadership according to employee surveys.

Managerial action should, of course, not be taken in isolation based solely on analytics data. As noted, governments will always face measurement challenges and boundaries. Analytical findings must be supplemented with practical and tacit knowledge. Public sector decision-makers thus need to triangulate analytics findings with practical and tacit knowledge. This puts public managers at the heart of interpreting and making use of government analytics findings.

What government analytics can do is strengthen the quality of conversations about how to improve public administration, rather than dictating managerial responses to specific analytics findings. Those conversations about improvement may be in a department of local government, a ministry, or may even span countries and the international community. They may involve government employees, service users, and others, extending beyond managers who are making solitary interpretations of analytics findings. In short, government analytics generates evidence for better conversations—and thus decisions—about how to improve the public administration.

This cycle is, of course, not necessarily linear; it is iterative. At times, government analytics can shed light on management problems that managers were unaware of—as with the opening example of the former marine managing a team in the US federal government. At times, managers may sense a potential problem—as with Brazil’s wage bill—that motivates analytics to better estimate the scale and nature of the problem.

The intention of this *Handbook* is not to dictate a unitary approach to measurement. Instead, the chapters that follow describe many of the most common approaches to government analytics, and present some evidence from across the world for them. This information is offered to provide a framework that government officials and analysts can draw on to select particular analytical approaches of particular use to them. Greater use of government analytics in turn will further the evidence base on how to best undertake government analytics.

This matters because effectively measuring the state allows us to manage it better. The best version of government arises from basing its design and management on the best data achievable as to how government is functioning. Having good policies on the books matters, but much less so without an effective machine to implement them. In other realms, data have revolutionized productivity. It is time to turn the lens on public administration.

## REFERENCES

- Acemoglu, D., and J. A. Robinson. 2019. *The Narrow Corridor: States, Societies, and the Fate of Liberty*. New York: Penguin Press.
- Akhari, M., D. Moreira, and L. Trucco. 2022. “Political Turnover, Bureaucratic Turnover, and the Quality of Public Services.” *American Economic Review* 112 (2): 442–93.
- Andrews, M. 2013. *The Limits of Institutional Reform in Development: Changing Rules for Realistic Solutions*. New York: Cambridge University Press.
- Besley, T., R. Burgess, A. Khan, and G. Xu. 2022. “Bureaucracy and Development.” *Annual Review of Economics* 14 (1): 397–424.
- Best, M. C., J. Hjort, and D. Szakonyi. 2017. “Individuals and Organizations as Sources of State Effectiveness.” NBER Working Paper 23350, National Bureau of Economic Research, Cambridge, MA.
- Bloom, N., R. Sadun, and J. Van Reenen. 2010. “Recent Advances in the Empirics of Organizational Economics.” *Annual Review of Economics* 2 (1): 105–37.
- Bloom, N., and J. Van Reenen. 2010. “New Approaches to Surveying Organizations.” *American Economic Review* 100 (2): 105–09.
- Bosio, E., and S. Djankov. 2020. “How Large Is Public Procurement?” *Let’s Talk Development*. World Bank blogs, February 5, 2020. <https://blogs.worldbank.org/developmenttalk/how-large-public-procurement>.
- Brynjolfsson, E., Y. Gorodnichenko, and R. Kuehn. 2021. “Predictive Analytics and the Changing Nature of Productivity in Manufacturing.” *Proceedings of the National Academy of Sciences* 118 (13): e2017984118.
- Carrera, L., and P. Dunleavy. 2013. *Growing the Productivity of Government Services*. Cheltenham, UK: Edward Elgar.
- Cusolito, A. P., and W. F. Maloney. 2018. *Productivity Revisited: Shifting Paradigms in Analysis and Policy*. Washington, DC: World Bank.



- Dahlström, C., and V. Lapuente. 2022. "Comparative Bureaucratic Politics." *Annual Review of Political Science* 25 (1): 43–63.
- Dal Bó, E., F. Finan, N. Y. Li, and L. Schechter. 2021. "Information Technology and Government Decentralization: Experimental Evidence from Paraguay." *Econometrica* 89 (2): 677–701.
- Dargent, E., G. Lotta, J. A. Mejía-Guerra, and G. Moncada. 2018. *Who Wants to Know? The Political Economy of Statistical Capacity*. Washington, DC: Inter-American Development Bank.
- Deaton, A. 2003. "Household Surveys, Consumption, and the Measurement of Poverty." *Economic Systems Research* 15 (2): 135–59.
- de Mello, L., and T. Ter-Minassian. 2020. "Digitalisation Challenges and Opportunities for Subnational Governments." OECD Working Papers on Fiscal Federalism 31. Organisation for Economic Co-operation and Development (OECD) Publishing, Paris.
- Dunleavy, P. 2017. "Public Sector Productivity." *OECD Journal on Budgeting* 17 (1): 153–70.
- Fenzia, A. 2022. "Managers and Productivity in the Public Sector." *Econometrica* 90 (3): 1063–84.
- Ferraz, C., F. Finan, and D. B. Moreira. 2012. "Corrupting Learning: Evidence from Missing Federal Education Funds in Brazil." *Journal of Public Economics* 96 (9): 712–26.
- Fetzer, T. 2021. "Did Covid-19 Improve Global Health? A Mixed-Methods Analysis of the Global Health Effect of the Covid-19 Pandemic." *The Lancet* 1 (1): e10–e18.
- Finan, F., B. A. Olken, and R. Pande. 2015. "The Personnel Economics of the Developing State." Chapter 6 in *Handbook of Economic Field Experiments* 2 (2017): 467–514.
- Freedom House. 2021. Freedom in the World. Data file. <https://freedomhouse.org/report-types/freedom-world>.
- Gains, J., and S. Stokes. 2005. "Path Dependencies and Policy Learning in Comparative Public Policy." *Governance* 18 (3): 475–98.
- Gallup. 2023. *State of the Global Workplace: 2022 Report*. <https://www.gallup.com/workplace/349484/state-of-the-global-workplace-2022-report.aspx>.
- Grindle, M. S. 2004. "Good Enough Governance Revisited." *Development Policy Review* 22 (6): 695–703.
- Grindle, M. S. 2007. *Going Local: Decentralization, Democratization, and the Promise of Good Governance*. Princeton, NJ: Princeton University Press.
- Hood, C. 2006. "Gaming in Targetworld: The Targets Approach to Managing British Public Services." *Public Administration Review* 66 (4): 515–21. <http://www.jstor.org/stable/3843937>.
- Kaufmann, D., A. Kraay, and M. Mastruzzi. 2022. Worldwide Governance Indicators. World Bank. Data file. <https://datacatalog.worldbank.org/dataset/worldwide-governance-indicators>.
- Khan, A. Q., A. I. Khwaja, and B. A. Olken. 2019. "Making Moves Matter: Experimental Evidence on Incentivizing Bureaucrats through Performance-Based Postings." *American Economic Review* 109 (1): 237–70.
- Kraay, A. 2006. "When Is Growth Pro-Poor? Evidence from a Panel of Countries." *Journal of Development Economics* 80 (1): 198–227.
- Pepinsky, T. B., J. H. Pierskalla, and A. Sacks. 2017. "Bureaucracy and Service Delivery." *Annual Review of Political Science* 20 (1): 249–68.
- Pritchett, L., M. Andrews, and M. Woolcock. 2017. "Escaping Capability Traps through Problem-Driven Iterative Adaptation (PDIA)." *World Development* 99: 74–84.
- Rasul, I., and D. Rogger. 2018. "Management of Bureaucrats and Public Service Delivery: Evidence from the Nigerian Civil Service." *Economic Journal* 128 (608): 413–46.
- Schuster, C. 2017. "Legal Reform Need Not Come First: Merit-Based Civil Service Management in Law and Practice." *Public Administration* 95 (3): 571–88.
- Schuster, C., J. Fuenzalida, J. Meyer-Sahling, K. Sass Mikkelsen, and N. Titelman. 2020. *Encuesta Nacional de Funcionarios en Chile: Evidencia para un Servicio Público Más Motivado, Satisfecho, Comprometido y Ético*. <https://www.serviciocivil.cl/wp-content/uploads/2020/01/Encuesta-Nacional-de-Funcionarios-Informe-General-FINAL-15ene2020-1.pdf>.
- Somani, R. 2021. "The Returns to Higher Education and Public Employment." *World Development* 144: 105471.
- Transparency International. 2021. Corruption Perceptions Index. <https://www.transparency.org/en/cpi/2021>.
- Ugale, G., A. Zhao, and M. Fazekas. 2019. *Analytics for Integrity: Data-Driven Approaches for Enhancing Corruption and Fraud Risk Assessment*. OECD (Organisation for Economic Co-operation and Development), Paris.
- Wilson, J. Q. 1989. *Bureaucracy: What Government Agencies Do and Why They Do It*. New York: Basic Books.
- World Bank. 2016. *World Development Report 2016: Digital Dividends*. Washington, DC: World Bank.
- World Bank. 2020. Worldwide Bureaucracy Indicators. <https://www.worldbank.org/en/data/interactive/2019/05/21/worldwide-bureaucracy-indicators-dashboard>.
- World Bank. 2021. *World Development Report 2021: Data for Better Lives*. Washington, DC: World Bank.
- World Bank. 2022. "GovTech Maturity Index, 2022 Update: Trends in Public Sector Digital Transformation." Equitable Growth, Finance and Institutions Insight–Governance. World Bank, Washington, DC.
- World Bank Group. 2019. *Innovating Bureaucracy for a More Capable Government*. Washington, DC: World Bank Group.

## CHAPTER 2

# How to Do Government Analytics

## Lessons from the *Handbook*

*Daniel Rogger and Christian Schuster*

### SUMMARY

How can practitioners and researchers undertake government analytics effectively? This chapter summarizes lessons from *The Government Analytics Handbook*. The chapter begins by introducing a public administration production function, which illustrates how different data sources, such as procurement data or public servant survey data, shed light on different parts of the machinery of government. The chapter then highlights lessons to keep in mind when undertaking any measurement and analysis of government administration. In the chapter-by-chapter summary that follows, lessons on how to generate and analyze a range of core data sources for government analytics are presented. These are data from administrative records (such as human resources and payroll, budget, procurement, administrative cases, task completion and projection completion); data from surveys of public servants; and external assessments of government, such as household or citizen surveys. The chapter concludes by showcasing how different data sources can be integrated to understand core challenges in the administration of government, such as personnel management.

### ANALYTICS IN PRACTICE

- This *Handbook* presents a wealth of approaches and data sources available to governments to improve the analytics of the machinery of government and identify evidence-based improvements. Many of these approaches rely on data that governments already collect as part of their day-to-day operations.
- By conceiving of government's electronic records as data in themselves, existing government records, and in particular the vast troves of data now being produced, can be repurposed as a means of diagnosing and strengthening government administration.
- Government analytics can be undertaken with at least three types of data: administrative data collected or published by government entities (such as payroll data); surveys of public servants; and external assessments (such as household surveys or anthropological assessments).

- To organize the various data sources assessed throughout the *Handbook*, this chapter introduces a public administration production function. A production function relates input factors of production through processes (such as management practices) to the output of an organization, and their eventual outcomes.
- Some data sources are better suited to assessing inputs into public administration, such as payroll data assessing the costs of different personnel. Some data sources are better suited to assessing the processes, practices, and cultures that convert inputs into outputs, such as surveys of public servants that assess how they are being managed. And some data sources are better suited to assessing the outputs and outcomes of public administration, such as citizen satisfaction surveys. What type of data source is appropriate for analytics depends on what aspect of public administration the analyst is seeking to diagnose and improve. Data sources can also be powerfully combined to understand overarching themes in how government is functioning, such as corruption.
- Frontier government analytics would integrate the analytics of these distinct data sources. It would generate them at a scale sufficient to inform the decision-making of individual managers. And it would make them easily accessible to those managers across government organizations and departments. For instance, dashboards integrating data sources and updating in real time would provide managers with comparisons for their staffing issues, process quality, the perceived quality of management practices, and so on. They could track outputs and outcomes, from task completion and case productivity to external assessments from citizens. Comparative data would allow them to benchmark themselves against other government organizations, or where appropriate, other countries. Managers would be capable of understanding the limitations and strengths of different analytics. The result would be a transformational change toward leveraging data to strengthen public administration.

## INTRODUCTION: AN UNPRECEDENTED EXPLOSION OF DATA

A common perception of government is that it is a place that no one truly understands: an incomprehensible maze, filled with a procedural fog. Actually, public administration is brimming with information. The everyday business of government has generated an abundance of records on who does what where. Each of these records has its origin in a contemporary administrative need of public officials. With the right approach, these records also provide a window into the workings of the administrations that created them—a way to clear the fog and provide a map through the maze.

For example, as a social security officer decides how to organize their work, they are making decisions that will affect the delivery of cases they are working on. These choices are reflected in the time stamps that accompany their cases, the degree to which they follow procedure, and the feedback they receive from those they serve. On its own, the record of their work is interesting, but combined with the records of their colleagues from their own organization and across the country it becomes a means to understanding the functioning of a public sector organization. Do all the officers working for a particular manager perform well on some aspects of their work but less so on others? Are some offices able to regularly process retirement benefits faster than others? The fog of uncertainty about how social security officers perform across their community can be cleared by turning records into analytics of the public service.

The data trail of government records like these has expanded greatly recently. The move toward digital government has multiplied the scale of electronic transactions between public administrations and citizens. From citizens paying taxes electronically to registering online for driver's licenses, every interaction is now recorded electronically in government data systems. Digital government has also multiplied electronic transactions within public administrations. Payrolls, procurement records, and budget disbursement records are some examples. Digital government has also facilitated the collection of other data, for instance by enabling governments to survey their employees at scale online. All this makes data to understand public administration easier to access than ever before.



This trove of data is critically important to the productive functioning of government, much in the same way as it has improved firm performance (Bakshi, Bravo-Biosca, and Mateos-Garcia 2014). “Without high-quality data providing the right information on the right things at the right time, designing, monitoring and evaluating effective policies becomes almost impossible,” a United Nations report notes (UN 2014, 2).

While the use of data for public *policy* has exploded, far less emphasis has been placed on how government records might be repurposed to understand how effectively the administrative machinery of government is functioning. By conceiving of government’s electronic records as data in themselves, existing government records can be seen as a means of diagnosing and strengthening government administration.

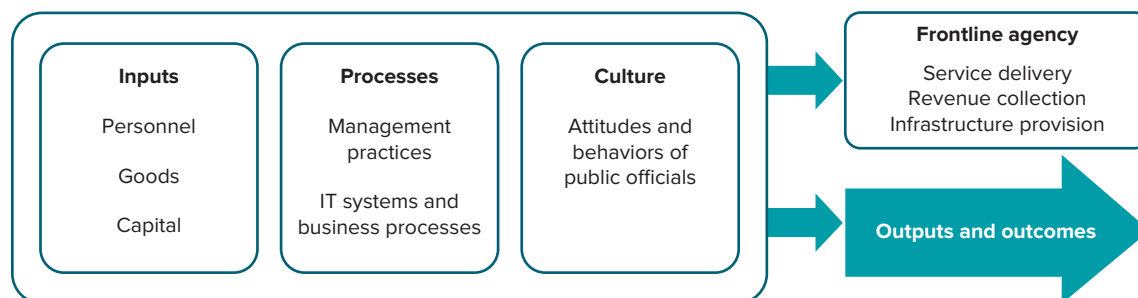
As discussed in chapter 1, public administration presents particular measurement challenges and opportunities. This *Handbook* focuses on sources and methods of measurement to meet those challenges, and on the need to create a system of measurements that is fit for public administration, rather than repurposed from other settings. Although current government records are an excellent foundation for analysis, their suitability for generating knowledge requires work. Once analytics becomes a goal, the range and nature of the public record is likely to change, to optimize government data for both the task at hand and the broader assessment of government functioning. This chapter summarizes lessons to that end—and on how to do government analytics—from across the *Handbook*.

## WHAT GOVERNMENT ANALYTICS CAN ANALYZE: UNDERTAKING ANALYTICS ALONG A PUBLIC ADMINISTRATION PRODUCTION FUNCTION

Government analytics refers to the use of data to diagnose and improve the machinery of government, or public administration. This chapter introduces a public administration production function to provide an overarching framework to organize the various data sources assessed in different chapters.<sup>1</sup> A production function relates input factors of production to the output of deliverables of an organization, and their eventual outcomes. The productivity of an organization thus depends on the quality and quantity of outputs relative to inputs. Figure 2.1 visualizes the different components of our production function for public administration (Meyer-Sahling et al. 2021; World Bank Group 2019). While many core elements coincide with typical private sector production functions (Mas-Colell, Whinston, and Green 1995), the functioning of government administration has been characterized as distinct from that of private firms due to the multiplicity of principals, the ambiguity of tasks, and the presence of principals with political incentives, among other features.

In public administration, inputs include personnel (public employees), goods (such as computers), and capital (such as office space). Outputs refer to, first, the deliverables produced by public

**FIGURE 2.1 The Public Administration Production Function**



Source: Adapted from World Bank Group 2019.  
Note: IT = information technology.

administration organizations themselves. For instance, a ministry of finance might issue public sector debt at a certain interest rate. Further, public administration organizations produce outputs (activities) that enable frontline agencies in the public sector—such as hospitals, schools, or police forces—to deliver services and goods to citizens. The outcomes in these examples are better health, education, or public safety, respectively. To fund the outputs, a ministry of finance may oversee budgets that frontline agencies then disburse to deliver their services.

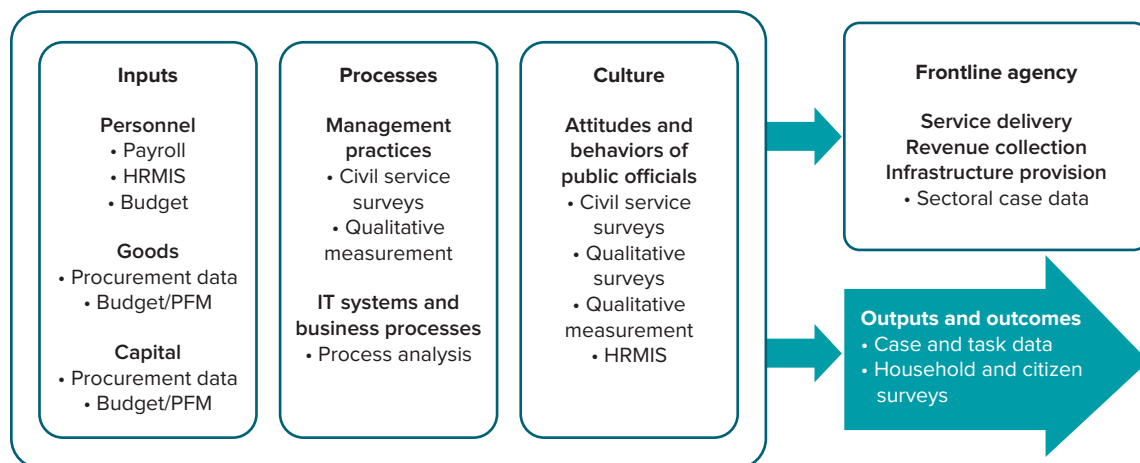
How do public administrations convert inputs (such as personnel) into outputs and outcomes? In our production function, this conversion is enabled by policies (organizational objectives and work procedures), systems, and management practices, and mediated by norms and behaviors inside public administration. For instance, a ministry of finance may have a policy in place to review a budget for an organization by a certain date. A team lead inside the ministry then manages employees to ensure the task is completed well and on time—such as through effective performance management practices. Those practices and organizational policies shape the norms and behaviors of the ministry’s employees—such as their motivation to work hard—which in turn then allows the ministry to produce outputs (such as a budget review).<sup>2</sup>

By utilizing different data sources and different methods, government analytics can shed light on all parts of this production function and identify bottlenecks, whether overpriced input goods, ghost workers on the payroll, high staff turnover, or slow processing of administrative cases, to name just a few. Contemplating government analytics along the production function enables analysts to diagnose public administration challenges holistically, and to understand how different data and approaches to government analytics relate.

To illustrate, figure 2.2 maps a number of different data sources analyzed in various chapters to their respective components in the production function. Several types of administrative data have particular strengths in diagnosing inputs into the public administration production function. For instance, payroll data and human resources management information system (HRMIS) data can help governments understand personnel as an input into the public administration production function, such as whether pay of public servants is competitive and fiscally sustainable, or whether staffing levels are adequate (see chapters 9 and 10). Budget data and procurement data can help governments understand spending on goods and capital as inputs into public administration—for instance, whether public administrations acquire similar goods cost-effectively across organizations in the public administration (see chapters 11 and 12).

Government analytics can also shed light on the processes and practices that convert inputs into outputs and outcomes. Surveys of public servants and qualitative measurement have particular strengths

**FIGURE 2.2 Mapping Different Government Analytics Data in the Public Administration Production Function**



Source: Original figure for this publication adapted from World Bank Group 2019.

Note: HRMIS = human resources management information system; IT = information technology; PFM = public financial management.

at diagnosing management practices. Management quality is fundamentally experienced by employees and a result from the interaction between managers and employees. Surveys can, for instance, ask public servants how they perceive the leadership of their superior or the quality of their performance feedback (see chapter 18). Government analytics can also shed light on the quality of processes inside public administration, such as whether these processes adhere to government procedure or meet deadlines (see chapter 13).

Whether practices and processes effectively turn inputs into outputs and outcomes is, as noted, mediated by the norms, attitudes, and behaviors of public administrators. Surveys of public servants and qualitative measurement are standard practice in many governments to evaluate this component of public administration production—for instance, to understand how engaged, committed, and ethical in their behavior public administrators are (see, for example, chapter 18). HRMIS data often complement rich survey data by providing insights into specific behaviors of public employees that are digitally recorded, such as whether public servants leave the organization, work overtime, or take sick leave (see chapter 9).

Last, public administrations produce outputs and outcomes both of their own (such as a ministry of finance issuing debt), and to enable outputs and outcomes of frontline providers. The productivity of frontline, service delivery agencies such as hospitals, schools, and police forces has been extensively measured, not least as direct contact with citizens enables more direct measurement of service delivery outcomes (such as patient outcomes in hospitals or learning outcomes in schools) (see chapter 29).

This *Handbook*, instead, focuses on the analytics of *administrative* outputs and outcomes. Administrative case data are one important source for measurement in such contexts. Such data are often routinely collected by organizations (for instance, the number of tax or social security cases processed) and can be repurposed by organizations to measure outputs and outcomes (such as the amount of tax revenue raised), and thus gauge productivity (see chapters 14 and 15). Beyond administrative data, surveying households and citizens (such as by asking citizens about their trust in public administration organizations) can be an important data source to understand outcomes of public administration (see chapter 28). What will be of most significance to measure and analyze will depend on a specific organizational setting and topic of interest to decision-makers.

The various chapters of this *Handbook* provide insights into how to use these different data sources to understand how government is functioning and improve the management of public administration. For a brief overview, table 2A.1 in annex 2A at the end of this chapter presents a chapter-by-chapter mapping of data sources to the topics covered in chapters 10 to 30 of the *Handbook*, to help readers pick and choose the chapters most of interest to them. The remainder of this chapter thus summarizes their lessons on how to do government analytics well.

## PART 2: FOUNDATIONAL THEMES IN GOVERNMENT ANALYTICS

The second part of the *Handbook* focuses on four cross-cutting challenges in the measurement and analysis of public administration: how to ensure that government analytics are undertaken (1) with due recognition and management of the risks involved (from managing political pressure to protecting key indicators from manipulation); (2) in a way that ensures analytical practices accord with current best practice in social science; (3) in an ethical manner; and (4) with measures to determine whether the government analytics generated are, in fact, used.

### Measuring What Matters: Principles for a Balanced Data Suite That Prioritizes Problem Solving and Learning

Chapter 4, by Kate Bridges and Michael Woolcock, tackles the first of these four challenges. The chapter notes that overreliance on quantitative government analytics comes with risks. For example, as with any performance targeting, hitting easy-to-measure targets risks becoming a false standard of broader success. Overemphasis on measurement also risks neglect of other important forms of knowledge—such as

qualitative work or practical knowledge—and thus may curtail a deeper understanding of key public administration problems. Moreover, political pressures, if undetected, can lead to falsification of data—for instance, to cover up problems such as corruption—and thus undermine the integrity of government analytics. Left unattended, these risks may mean that government analytics may curtail the problem-solving and implementation capabilities of public sector organizations, rather than strengthening them.

In light of these risks, chapter 4 offers four principles for government analytics based on a balanced data suite that strengthens the diagnosis and solving of problems in public administration:

1. Identify and manage the organizational capacity and power relations that shape data management: for instance, by defining and communicating to all staff professional standards for collecting, curating, analyzing, and interpreting government analytics data.
2. Focus quantitative measures of success on those aspects that are close to the problem: for instance, by targeting measurement of public administration problems prioritized by government.
3. Embrace a role for qualitative data, especially for those aspects that require in-depth, context-specific knowledge.
4. Protect space for judgment, discretion, and deliberation in those (many) decision-making domains that inherently cannot be quantified.

### **Practical Tools for Effective Measurement and Analytics**

Chapter 5, by Maria Ruth Jones and Benjamin Daniels, turns to a second cross-cutting challenge: how to employ the best practices of modern social science when utilizing statistical tools for government analytics. The chapter discusses important underlying statistical principles for government analytics to ensure that government analytics are credible, including the transparency of analysis and reproducibility of results. Producing analysis that accords with best practice requires considering the full life cycle of data work, such that each stage of handling data can be designed to support the next stages. The chapter introduces the suite of tools and resources made available for free by the World Bank's Development Impact Evaluation (DIME) Department to support the achievement of best-practice statistical analysis, such as research-cycle frameworks, extensive training tools, detailed archives of process and technical guidance, and a collaborative approach to data and analytics. The chapter also links to online tools, such as an online training hub, available to help implement these practices.

### **The Ethics of Measuring Public Administration**

Chapter 6, by Annabelle Wittels, discusses a third cross-cutting challenge in government analytics: how to undertake ethical measurement and analysis in public administration. While guides for the ethical collection of data on citizens exist, there is a dearth of practical guides on the ethics of government analytics, particularly with respect to data collection by governments on their own employees. Chapter 6 introduces a heuristic to encourage ethical government analytics, which balances three, at times competing, ethical considerations: (1) an individual dimension that encompasses demands by public employees for dignity and privacy; (2) a group dimension that relates to allowing for voice and dissent of public employees; and (3) a public-facing dimension that ensures that analytics enable public administrators to deliver on public sector values—accountability, productivity, and innovation. These three considerations can be in tension. For instance, data diagnostics on public personnel can inform better management practices and enable greater productivity but impinge on the privacy of employees, whose data are required for the diagnostic.

To guide practitioners, chapter 6 presents a 10-point framework. For instance, the ethics of government analytics requires consideration of the granularity of data required for the diagnostic. Are data that identify individuals required for the diagnostic, or could group-level or anonymous individual-level data enable a

similar diagnosis? In a survey of public servants, do demographic questions need to identify the exact age of a respondent in an organization (which risks identifying the respondent), or are broad age bands (with a lower risk of identification) sufficient? As a second example, ethical government analytics requires consideration of who has a say in what gets measured. In particular, are those who are being measured consulted in advance and given an opportunity to provide inputs? Questions in the framework like these can guide ethical government analytics.

### Measuring and Encouraging Performance Information Use in Government

Chapter 7, by Donald Moynihan, discusses efforts to address a fourth challenge in government analytics: measuring whether government analytics measures are, in fact, being used. If costly data collection or analysis is undertaken on public administration but the resulting analytics are not used to improve public administration, government analytics harms rather than advances better government. Chapter 7 illustrates how survey and administrative data can be drawn on to measure use of government analytics data (or, to use the term in chapter 7, performance information). In particular, the chapter recounts the experience of the US Government Accountability Office (GAO), which periodically surveys public employees on their use of performance information. The chapter describes the survey measures deployed to this end. The chapter also discusses administrative data that governments can utilize, such as data that track the use of dashboards displaying government analytics data. The use of government analytics can thus be subjected to the same evaluative measurement rigor that government analytics applies to public administration.

### Understanding Corruption through Government Analytics

Chapter 8, by James Anderson, David S. Bernstein, Galileu Kim, Francesca Recanatini, and Christian Schuster, illustrates how the approaches presented in this *Handbook* can be combined and leveraged to holistically diagnose a major obstacle to more effective public administrations. Chapter 8 is discussed further at the end of this chapter.

## PART 3: GOVERNMENT ANALYTICS USING ADMINISTRATIVE DATA

The nine chapters in part 3 discuss the range of administrative data sources that governments can draw on to undertake government analytics. Each main type of administrative data is covered in a different chapter. Chapters 9 and 14 contextualize these discussions, showcasing how to create underlying data infrastructures (chapter 9) and how to combine data sources to measure the performance of public organizations with multidimensional missions (chapter 14).

### Creating Data Infrastructures for Government Analytics

Chapter 9, by Khuram Farooq and Galileu Kim, focuses on how to create data infrastructures—or management information systems (MIS)—that are well suited for government analytics. Using the case of human resources management information systems (HRMIS), the chapter provides a road map to guide the development of data infrastructures that enable government analytics. The road map emphasizes the importance of first getting in place high-quality foundational data modules in the MIS, and only then transitioning to developing more complex analytical modules. In the case of HRMIS, for instance, ensuring high-quality foundational modules including basic information on personnel and payroll compliance should take precedence over more advanced modules such as talent management and analytics. Without quality foundations,

other modules will produce imprecise or inaccurate analytics. Analytical modules that include dashboards and reports require that foundational modules are set in place and their data are accurate.

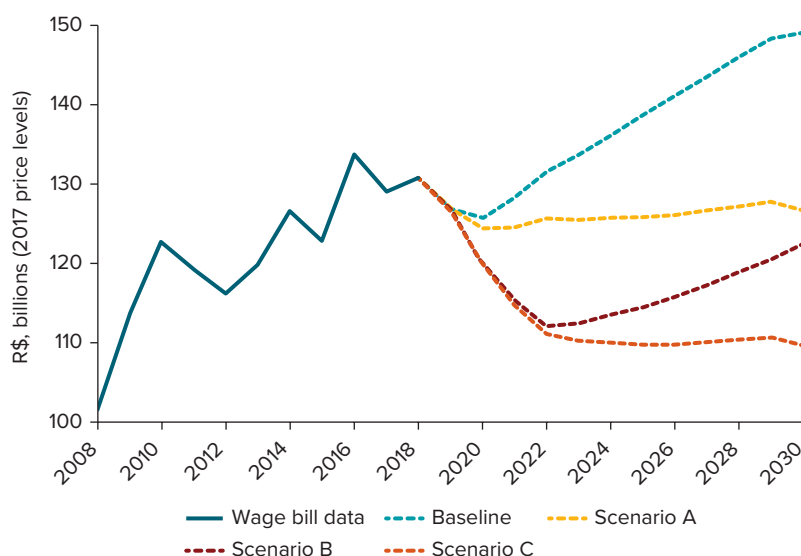
The road map thus emphasizes a sequential approach to creating data infrastructures, coupled with repeated testing of new infrastructures, accessible technical support for users of new MIS systems, tracking of usage, and building in-house capacity to maintain the system. To illustrate potential applications, the chapter is complemented by three case studies of analytical transformations in HRMIS systems in Luxembourg, Brazil, and the United States. Ludwig Balmer, Marc Blau, and Danielle Bossaert discuss how the government of Luxembourg introduced a Business Intelligence Center for human resources (HR) analytics, which transformed HR operations. Luciana Andrade, Galileu Kim, and Matheus Soldi Hardt showcase the development of a system to detect irregularities in Brazil's federal payroll, using machine learning. Robin Klevins and Camille Hoover illustrate how a US federal government agency developed a simple but effective dashboard to extract insights from a federal public employee engagement survey.

The quality of analytics arising from any MIS is critically dependant on the quality of the underlying data, and the analysts' understanding of their origins and limitations. The rest of part 3 discusses how to ensure quality measurement and analytics with a range of administrative data sources that measure aspects of the production function of public administration.

## Government Analytics Using Human Resources and Payroll Data

In chapter 10, Rafael Alves de Albuquerque Tavares, Daniel Ortega Nieto, and Eleanor Florence Woodhouse illustrate how to use payroll and HRMIS data for government analytics. Using a series of examples from Latin American countries, the chapter underscores how the analytics of such data can enable more evidence-based decisions around both fiscal planning and personnel policy. For instance, payroll data can be drawn on to better understand the likely future dynamics of wages and retirement based on the modeling of individual career trajectories. Figure 2.3 illustrates how analytical use of payroll data allowed Brazilian policy makers to simulate the financial implications of a set of policies related to pay and employment. The difference between the wage bill costs under the least and most expensive options amounted to nearly 50 percent

**FIGURE 2.3 Wage Bill Projection and Policy Scenarios, Brazil, 2008–30**



Source: Original figure for this publication (see chapter 10).



of total wage bill expenditures. Such a granular lens on employment trends enables governments to better plan for the longer-term fiscal implications of public employment and understand its drivers.

The analytics of payroll and HRMIS data can also enable governments to improve personnel policy. For instance, data on workforce allocation across different regions of a country (such as those of a tax administration) matched to data on the number of service users (such as taxpayers) in each region can help governments understand under- or overstaffing and improve workforce allocation by prioritizing new recruitment in understaffed regional offices. Payroll data also enable governments to measure turnover of employees working in different organizations or for different managers, helping governments pinpoint retention problems and their origins. By comparing pay for similar positions across government institutions, payroll analysis identifies potential salary inequities. Chapter 10 recommends centralizing payroll and HR data collection systems to render such data accessible and provides a road map to this end that complements the discussions and case studies in chapter 9.

### **Government Analytics Using Expenditure Data**

Chapter 11, by Moritz Piatti-Fünfkirchen, James Brumby, and Ali Hashim, discusses government analytics using expenditure data. Budget and government expenditure data are already widely used by governments to understand whether resources are used for budgeted priorities; whether spending is sustainable, efficient, effective, and equitable; and which government transactions (typically, large-value ones) might have high fiduciary risks. The World Bank, for instance, has a long-standing practice of Public Expenditure Reviews. Such reviews are often accompanied by benefit incidence analysis to orient spending toward those in need, by linking data on spending distribution with who receive services. The chapter briefly reviews these well-established uses of expenditure government analytics, and then delves into an aspect missing in much expenditure analytics: how to ensure high-quality data for expenditure analytics.

The chapter underscores the need to periodically review government expenditure microdata for five features: (1) data provenance and integrity; (2) comprehensiveness; (3) usefulness; (4) consistency; and (5) stability. This requires a prior, clear definition of what counts as a government expenditure, as well as a clear understanding and documentation of how transactions across spending items in government are created, what control protocols they are subject to, how this information is stored, and how microdata are aggregated for analysis (such as by classifying government transactions by function, to be able to analyze aggregate spending data by function). The chapter provides practitioners with seven questions to include in expenditure analytics to ensure that the data are high quality, and the resulting analytics are as informative as possible to improve government spending.

### **Government Analytics Using Procurement Data**

Chapter 12, by Serena Cocciolo, Sushmita Samaddar, and Mihaly Fazekas, discusses how to use procurement records as data for government analytics. The digitization of national public procurement systems across the world has multiplied opportunities for procurement data analytics. Such analytics allow governments to strategically monitor procurement markets and trends, to improve the procurement and contracting process through data-driven policy making—for instance, by identifying overpriced goods or corruption risks in procurement—and to assess the potential trade-offs of distinct procurement strategies or reforms. The chapter explores the range of procurement indicators that can serve these purposes. For instance, indicators to measure the economy and efficiency of procurement include the time needed for contracting and the final price paid. Indicators to proxy transparency and integrity include the share of single bidder tenders and the share of excluded bids. Indicators to measure competition include the number of bidders. Indicators of inclusiveness and sustainability include the share of bids coming from small and medium enterprises.

When e-procurement systems are integrated with other e-government systems—such as systems generating firm registries or tax data—analytics can go even further: for instance, by allowing governments

to detect potential family relations (and thus collusion risk) between owners of firms bidding for government contracts and procurement officials. Chapter 12 also showcases how governments can use interactive dashboards to track, analyze, and display key procurement indicators through customizable and user-friendly visualizations. All this requires that procuring entities record procurement transactions consistently, that such records are then centralized, and that (automated) data quality checks and periodic data audits ensure the data are accurate.

### Government Analytics Using Data on the Quality of Administrative Processes

The eventual value of the core inputs to the effective functioning of government (personnel, goods, and capital) is determined by how they are processed and managed. Chapter 13—by Jane Adjabeng, Eugenia Adomako-Gyasi, Moses Akrofi, Maxwell Ampofo, Margherita Fornasari, Ignatius Geegbae, Allan Kasapa, Jennifer Ljungqvist, Wilson Metronao Amevor, Felix Nyarko Ampong, Josiah Okyere Gyimah, Daniel Rogger, Nicholas Sampah, and Martin Williams—presents approaches to assessing the quality of *administrative* processes. Applying proper processes and procedure to a project, file, or case is core to the work of public administrators. The chapter presents a range of indicators to this end, such as the share of processes undertaken by public administrators that are timely with respect to deadlines, adhere to government procedure, and are logical in flow. Such data can be collected automatically as part of digitized government work, or manually by assessors employed to judge the quality of process in the physical records of projects, files, or cases. Chapter 13 showcases two applications of this approach. The example from Liberia highlights adherence to new processes for performance appraisal. The example from Ghana highlights the quality of process in core office duties, such as project planning, budgeting, and monitoring.

### Government Analytics Using Customs Data

Chapter 14, by Alice Duhaut, provides an overview of government analytics using customs data. Customs agencies typically have three core objectives: facilitating trade, collecting revenue, and ensuring the security and safety of the goods entering or exiting the country. As in many other government agencies with multidimensional missions, attaining one objective (such as greater safety of traded goods) can come at the expense of another (such as facilitating trade). Incomplete measurement of objectives risks encouraging attainment of measured objectives while unknowingly impairing other objectives. This puts a premium on effective measurement of all dimensions of a customs mission, which often requires triangulating different data sources. Chapter 14 showcases how this can be done, deriving indicators for trade facilitation (such as costs of the process, particularly in terms of delays); revenue collection (such as trade volume and revenue collected based on the assessed value); and safety (such as number of goods in infraction seized). Collecting these indicators requires integrating multiple data sources. The chapter thus discusses several data sources. These include the Automated System for Customs Data, used by 100 countries and designed by the United Nations Conference on Trade and Development (UNCTAD), which captures items declared, excise, and duties; as well as complementary data sources, such as time-release studies and GPS data on the time spent at borders. The chapter also illustrates how such data can be used not only for risk management *ex ante*, but also to assess customs performance *ex post*.

### Government Analytics Using Administrative Case Data

Chapter 15, by Michael Carlos Best, Alessandra Fenizia, and Adnan Qadir Khan, provides insights into the analytics of administrative case data in government more broadly. A case file is typically a collection of records regarding, for instance, an application for government payments (such as social security) or access to services (such as government-sponsored child care); to obtain licenses and permits; or to bid on a government contract. The chapter draws on three example types of administrative cases: social security programs,



tax collection, and public procurement. In all three examples, governments routinely collect case data as part of their day-to-day operations. The chapter shows how these case data can be repurposed to construct objective measures of performance. These include measures to benchmark productivity across, say, regional tax offices, such as the average time to complete similar tax cases or the number of cases completed per officer. Map 2.1 presents an index of the productivity of social security case processing in Italy. It illustrates how substantial the variability in government productivity can be across a single country, with some offices taking 2.5 times as long as others to process similar cases. Case data also enable governments to understand differences in quality, such as comparing the share of tax cases leading to appeals by taxpayers or being identified as erroneous by audits.

Chapter 15 also emphasizes the importance of accounting for the complexity of each case. For example, a social security claim that clearly meets the requirements of regulation is less complicated to process than a case in which there are ambiguities in eligibility and external validation is required. The chapter provides guidance on how to adjust case data for complexity. When accounting for complexity and quality, the chapter concludes that the case data governments already collect provide a wealth of performance information to make such adjustments.

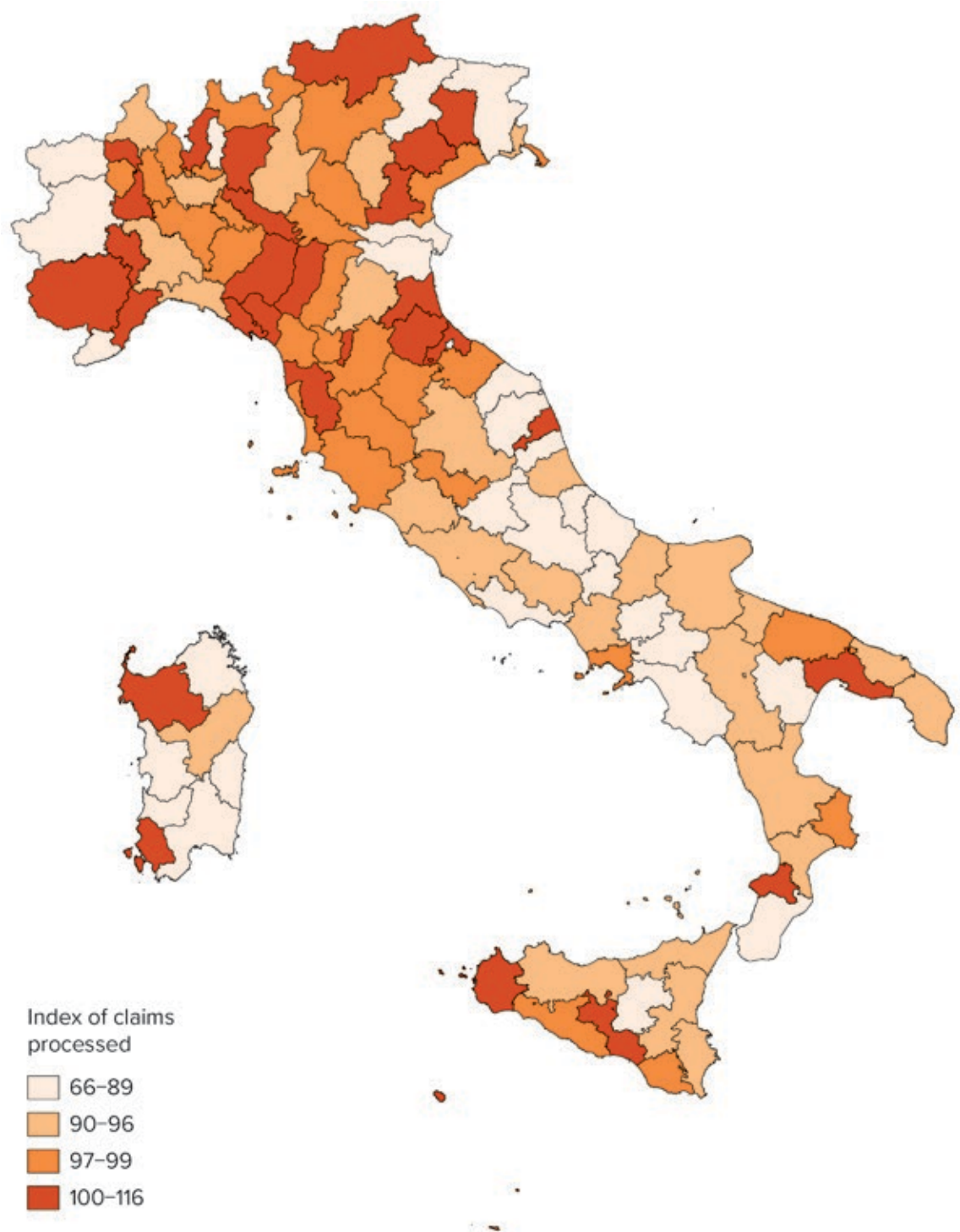
### Government Analytics Using Machine Learning

Chapter 16, by Sandeep Bhupatiraju, Daniel Chen, Slava Jankin, Galileu Kim, Maximilian Kupi, and Manuel Ramos Maqueda, shifts the focus to a different data source—text-as-data—and a different methodological approach, the use of machine learning (ML) and artificial intelligence (AI) for government analytics. Machine learning is fundamentally a methodological approach: it defines a performance indicator and trains an algorithm to improve this indicator, using the data collected. Such data can include text, allowing government to classify quantitatively the “big data” of texts it produces in records or communications. As a result, ML can be applied in a range of government analytics domains, from detecting payroll fraud to understanding bias in welfare appeal decisions, to name a few. The chapter illustrates the use of ML and AI for government analytics in the case of the judiciary. In the justice system, the increasing digitization of legal documents and court sentences, and the development of new techniques in natural language processing, enable analytics to improve judicial decision-making. India, for instance, has 27 million pending court cases; the sheer number of cases precludes manual classification of (often inconsistent) records and legal texts. ML algorithms can be trained to classify such records. This enables, for instance, analytics of bias and discrimination in courts (such as where judges with certain characteristics are associated with certain judicial outcomes in similar cases), or evaluations of how judicial reforms shape judicial productivity and bias. The chapter also describes the enabling environment for ML application—including how to build ML human capital and data infrastructure; the ethical considerations to keep in mind; and the importance of collaboration between ML engineers, domain experts, and the agencies that will use the technology to develop effective ML-based government analytics.

### Government Analytics Using Data on Task and Project Completion

Chapter 17, by Imran Rasul, Daniel Rogger, Martin Williams, and Eleanor Florence Woodhouse, discusses government analytics using task completion data. Much government work consists of the completion of tasks, from creating major reports to undertaking training programs and building infrastructure. A task completion approach allows for the investigation of which units and organizations are most likely to initiate, make progress on, and complete tasks—particularly where organizations complete similar tasks (such as preparing budgets). A task completion approach is particularly important to understand performance in administrative organizations in which the productivity data discussed in previous chapters (such as case data or frontline service delivery indicators) are not available.

**MAP 2.1** Variations in Productivity of Processing Social Security Cases, Subregions of Italy



Source: Fenizia 2022, using Italian Social Security Agency data.  
Note: The key refers to the number of social security claims of a particular type that are processed by an office in a particular time period divided by the full-time equivalent of workers of that office during that time.

Data for task completion can be extracted from a variety of sources. In Ghana, for instance, government units complete quarterly progress reports on all activities, which can be digitized and repurposed into task completion data. In the United Kingdom, the National Audit Office completes major project reports that assess the progress of large infrastructure projects against corresponding planning documents. By subsetting—that is, culling—these data to similar tasks undertaken by all government units, units can be benchmarked on the average time it takes them to complete tasks and the share of uncompleted tasks, for instance. Matching these data to other data about these units—for instance, on budget disbursements or management practices—can help governments understand why some government units are more effective at completing tasks and identify drivers to improve productivity in lagging government units. Chapter 17 underscores the importance of—and provides guidance for—classifying task characteristics correctly (such as in terms of their complexity) to ensure that cross-unit benchmarking is valid.

### Cross-Cutting Insights from Part 3

The chapters in part 3 reveal some cross-cutting insights to keep in mind when undertaking government analytics using administrative data. To begin with, high-quality administrative data are not a foregone conclusion. The infrastructure for—and thus quality of—the underlying measurement and resulting data is paramount and requires careful consideration. The first challenge is data coverage. A central payroll system, for instance, might cover only part of the public employment workforce (chapter 10); and a central financial management information system might cover only part of government expenditures (chapter 11). A second challenge is data validity. To illustrate, case data in tax or social security are often not recorded or designed to measure performance. Careful thought is needed to repurpose such data for performance measurement. Self-reported data—as in the case of some task completion data—may also suffer from inaccurate or manipulated data entry by the units being evaluated, putting a premium on independent, third-party data collection or validation (chapter 7). A third challenge for performance data in particular is completeness relative to the mission of an organization. As chapter 14 notes, missions and goals of public sector organizations are typically multidimensional (and at times contradictory). This requires the triangulation of multiple data sources to ensure performance is measured holistically, so that government analytics does not incentivize the attainment of one goal of an organization at the expense of another (chapter 4).

The chapters also emphasize the human capital requirements for government analytics using administrative data. Creating the information technology (IT) systems that underlie centralized data infrastructures requires IT and data science skills (chapters 9 and 16). Processing records into usable data, analyzing the data, and making the data accessible for analysis and management improvements (such as through dashboards) similarly require data science and visualization skills. In some governments, advanced data science skills (such as for machine learning) might be in short supply. Upskilling data scientists in government, or creating data science teams for government analytics, is thus important to make the most of government analytics opportunities.

Lastly, the chapters also emphasize the importance of data accessibility for decision-makers. Creating dashboards to allow users to explore key insights from the data—for instance, on procurement indicators or HR—facilitates such accessibility. These can be complemented by briefings and slide decks with key diagnostic findings and insights and data storytelling for senior policy makers to take action. In other words, government analytics data do not speak for themselves, but need to be made understandable to support government action.

## PART 4: GOVERNMENT ANALYTICS USING PUBLIC SERVANT SURVEYS

Part 4 focuses on a single data source for government analytics: surveys of public servants. There are three reasons for dedicating an entire section to one data source.

First, surveys of public servants are one of the most widely used data sources for government analytics. The review of the global landscape in chapter 18 finds that the number of governments implementing governmentwide surveys of public servants repeatedly every year or two has increased continuously over the last two decades. At least nine governments of member countries of the Organisation for Economic Co-operation and Development (OECD) were conducting annual or biannual surveys as of 2021, and many others are surveying their public servants on a more ad hoc basis.

Second, surveys of public servants can be costly in terms of staff time in a way that repurposing administrative data is not. Such surveys often sample a census of (that is, all) government employees. The staff time cost of completing the US federal government employee survey reaches US\$30 million annually (as cited in chapter 20). It is therefore important to design such surveys to be as efficient and effective as possible.

Third and more important, many key features of public administration production cannot be measured efficiently through other data (such as administrative data or citizen surveys). For example, understanding how public servants are managed, their motivations, and their behaviors are all internal to the official's lived experience, yet matter for public sector productivity. Public employees' motivations are difficult to observe outside of their own expressions of their motives. Thus, self-reporting through surveys becomes the primary means of measurement for many aspects of the public administration production function, and serves as a lever for improving public sector productivity.

For all these reasons, effectively designing, implementing, and making the most of surveys of public servants for government improvement is crucial.

## Surveys of Public Servants: The Global Landscape

In chapter 18, Ayesha Khurshid and Christian Schuster review current government practices in surveys of public servants. The chapter finds that governments undertake such surveys with relatively similar objectives, and thus most governments tend to measure similar concepts in their surveys. These include, on the one hand, measures of core employee attitudes correlated with productivity, such as job satisfaction and engagement, commitment to the organization, and intent to remain working for the organization. On the other hand, governments measure a relatively consistent set of management practices as antecedents of these employee attitudes, such as the quality of leadership, performance management, and training.

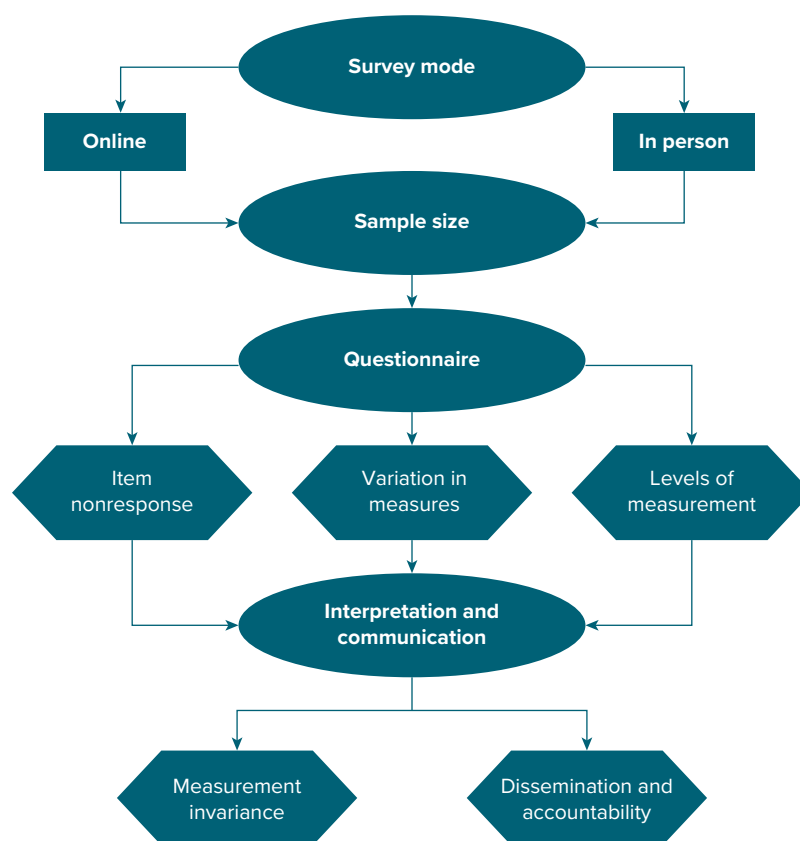
Yet chapter 18 finds that governments differ in how they design, implement, and report on surveys of public servants. For instance, the wording of survey questions differs, even when similar concepts are being measured. Approaches to sampling public servants differ, as do survey modes or approaches to dealing with nonresponse. Governments also differ widely in how they report survey results: for instance, in terms of what kind of benchmarks are reported or what levels of hierarchy inside organizations are measured and provided with results reports. Given that governments undertake surveys with similar objectives, why is there such diversity in how they design, implement, and report on surveys?

The answer, arguably, lies in part in the limited evidence available to governments that could guide choices about design, implementation, and reporting of surveys of public servants. The chapters in part 4 thus provide novel empirical evidence to enable governments and practitioners to make more evidence-based choices in response to some of these and other methodological questions in public servant surveys.

The decision tree pictured in figure 2.4 structures the choices facing governments in surveying public servants. This decision tree should *not* be read as a linear set of steps: there are interdependencies between choices. For instance, how many respondents need to be sampled depends on the expected variation in survey measures, which in turn is a function of questionnaire design.

Nonetheless, a common first choice concerns a survey mode: Are surveys conducted online, on paper, in person, by phone, or through a combination of these modes? Governments then need to determine the appropriate survey population, and an approach to sampling respondents, including determining the desired sample size given the purpose of the survey. Subsequently, questionnaires need to be designed. While measures may vary across concepts, several general concerns apply across them. For instance, how can measures be designed so that public servants are willing to answer questions (thus avoiding item nonresponse)?

**FIGURE 2.4** Decision Tree for Surveys of Public Servants



Source: Original figure for this publication.

How can measures be designed that vary sufficiently, so that comparisons between organizations or groups of public servants on these measures become meaningful? And should survey measures inquire about the individual experience of public servants themselves or ask them about their perceptions of practices in the organization as a whole?

Finally, governments need to decide how to interpret and report results. For instance, can responses from different groups—such as public servants in different countries, organizations inside a country, or demographic groups inside a country—be meaningfully compared? Or might concepts such as job engagement mean different things to different public servants (even when answering the same question), so benchmarking is not valid? Once decisions about who to benchmark are made, how can results be reported effectively to enable action? For instance, how are survey results presented to decision-makers, and who receives results? How is capacity built to enable decision-makers to take action based on survey results, and how are they incentivized to do so? The chapters in part 4 provide novel evidence on each of these key questions.

### **Determining Survey Modes and Response Rates: Do Public Officials Respond Differently to Online and In-Person Surveys?**

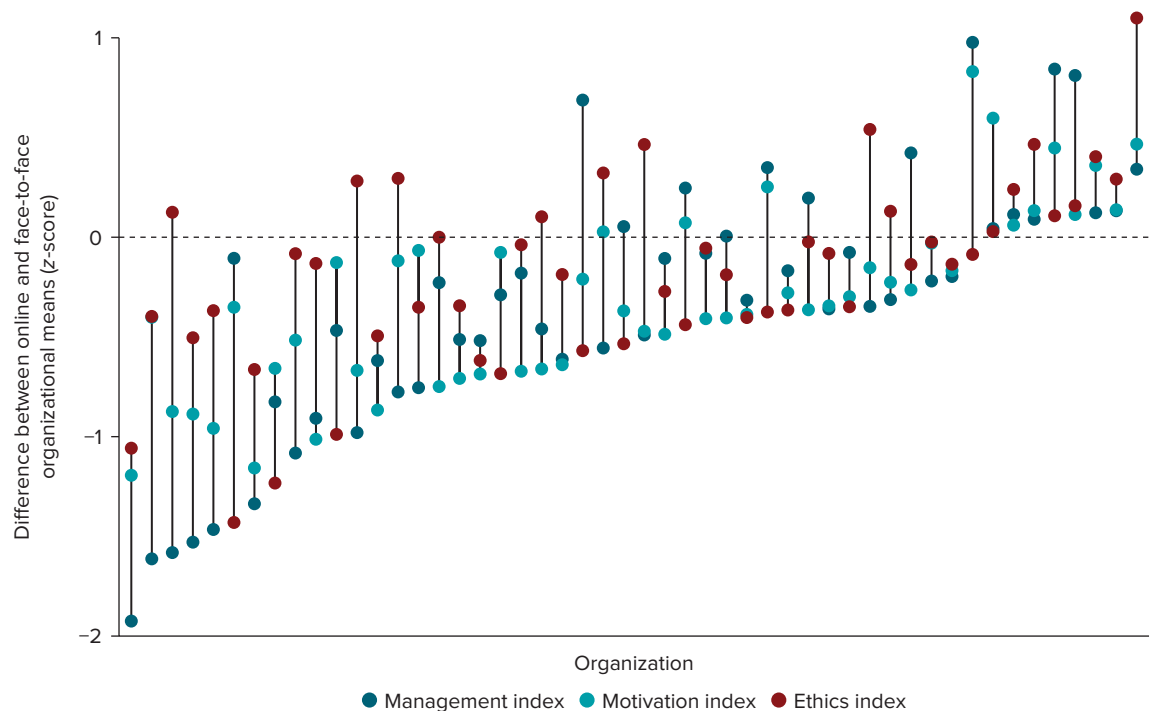
Chapter 19, by Xu Han, Camille Parker, Daniel Rogger, and Christian Schuster, assesses the first methodological choice in the decision tree: which enumeration method or survey mode to choose. This matters because different survey modes may come with different response biases to questions and different overall response rates. In OECD governments, surveys of public servants are typically conducted online, though not exclusively so. All nine government surveys reviewed in chapter 18 are implemented online, although,

to enhance accessibility (for instance, for staff who have difficulty accessing or completing an online survey), Colombia, Switzerland, the United Kingdom, and a few agencies in Australia also offer their survey in a paper format, while New Zealand offers its survey through paper and telephone upon request. The advantage of a predominantly online approach to surveying public servants across governments is clear: it reduces costs and may reduce biases, such as those induced by respondents' notions of socially desirable answers when faced with an in-person or phone interviewer.

Online surveys, however, also tend to have lower response rates than other survey modes, such as in-person surveys. For instance, the last US Federal Employee Viewpoint Survey had a response rate of 44 percent. This raises a concern that the data resulting from an online survey are not a valid representation of the population—in the case of the United States, the entire federal public administration. In some instances, these concerns about the validity of online surveys of public servants become severe. In the United Kingdom, the validity and quality of the underlying data of the Civil Service People Survey was questioned in a parliamentary inquiry, in part motivated by low response rates in some government organizations (UK Parliament 2022).

To what extent are low response rates in online surveys a concern (thus putting a premium on survey modes with higher response rates, such as in-person surveys)? To find out, chapter 19 presents evidence from a randomized control trial that compares face-to-face and online survey responses from Romanian public servants. The face-to-face surveys had consistently high response rates across Romanian government organizations, while the response rates for the online surveys varied across organizations, as is typical in other governments. The results suggest that these diverging survey modes do not substantially affect aggregate estimates at the national level. They do, however, affect the comparability of findings across organizations. Figure 2.5, reproduced from chapter 19, shows how big of a difference the mode of survey makes for indexes of survey topics. For some organizations, the impact of the survey mode is substantial.

**FIGURE 2.5 Average Difference between Survey Modes for Different Topics across Romanian Government Organizations**



Source: Original figure for this publication (see chapter 19).

Note: The figure shows, by organization in the Romanian government, the difference in the management index, motivation index, and ethics index between online and face-to-face survey respondents.



Basic organizational and demographic characteristics explain little of the variation in these effects. In other words, survey weights are not effective in addressing these effects.

Governments that offer varying survey modes should thus be careful when comparing the scores of organizations if some implement the survey primarily online while others implement it primarily based on pen and paper. Rankings of organizations in such instances do not appear to be valid. Nonetheless, chapter 19 does not find evidence to suggest that the (lower-response) online survey mode biases national-level inferences from the survey of public servants in Romania. More research is required to confirm the external validity of this finding in other countries.

In a second step, governments need to define survey populations and their sampling approach. Who the appropriate survey population is, of course, depends on the government's measurement objectives. The global review in chapter 18 shows that the survey population generally consists of central government civil servants, albeit with variations in the extent to which public sector organizations and types of employee contracts outside the (legally defined) civil service are also covered—for instance, in other branches of government or frontline services. To cite just one example, for the United Kingdom's government employee survey, all public servants from 101 agencies are eligible, excluding the Northern Ireland Civil Service, the National Health Service (NHS) (which conducts its own survey), and frontline officials (such as police officers and teachers) (UK Cabinet Office 2022).

Who, then, should be sampled within this survey population? As chapter 18 shows, approaches to sampling vary across governments, ranging from census to random, ad hoc, and stratified sampling. Australia, Canada, Ireland, New Zealand, and the United Kingdom adopt a census approach where all eligible public sector employees may participate in the survey. Switzerland and the United States use (stratified) randomized sampling approaches for most years but conduct a census every few years. Colombia's public servant survey uses a mixed approach: for larger organizations, a stratified sampling approach is used, while for smaller organizations a census is taken to protect anonymity. The Republic of Korea adopts a sampling approach for all annual surveys.

## **Determining Sample Sizes: How Many Public Officials Should Be Surveyed?**

As noted in chapter 20, by Robert Lipinski, Daniel Rogger, Christian Schuster, and Annabelle Wittels, determining the appropriate sample of a public administration survey is often a trade-off between increasing the precision of survey estimates through greater sample sizes and the high costs of surveying a larger number of civil servants. Greater precision enables both more precise benchmarking (such as between organizations) and survey result reports at lower levels of hierarchy in an organization. Less precision reduces the staff time lost responding to the survey.

How can this trade-off be resolved? Chapter 20 shows that, ultimately, survey administrators must decide on the sample size based on the type of inferences they want the survey to yield and the staff time they can claim for the survey. By employing Monte Carlo simulations on survey data from Chile, Guatemala, Romania, and the United States, chapter 20 shows that governmentwide averages can be reliably derived using sample sizes considerably smaller than those used by governments currently. On the other hand, detecting differences between demographic groups (such as gender and rank) and, in particular, ranking individual public administration organizations precisely requires larger sample sizes than are collected in many existing surveys.

These results underscore, on the one hand, the importance of not overinterpreting the substantive significance of small differences between organizations in public servant survey results (or individual ranks of organizations in results). On the other hand, the results emphasize that governments should determine sample sizes based on the type of inferences and benchmarking exercises they wish to make with the data. Existing governmental surveys of public servants do not seem to be based on such a data-driven approach to sampling. Chapter 20 addresses this gap and offers an online sampling tool to enable such sampling.

In a third step, surveys of public servants require the definition of a questionnaire. Part 4 sheds light on three cross-cutting dimensions of questionnaire design: How can measures be designed that vary sufficiently



so that comparisons between organizations or groups of public servants become meaningful? How can measures be designed so that public servants are willing to answer? And should survey measures inquire about the individual experience of public servants or instead ask them about their perceptions of practices in the organization as a whole?

### **Designing Survey Questionnaires: Which Survey Measures Vary and for Whom?**

A first prerequisite for effective question design is variation: Survey measures should provide a sufficient degree of discriminating variation across respondents to be useful—or, in other words, sufficient variation to understand differences between key comparators, such as organizations or demographic groups. Without discriminating variation across organizations, demographic groups, or countries, survey measures cannot inform governments about strengths and areas for improvement. With this in mind, chapter 21, by Robert Lipinski, Daniel Rogger, Christian Schuster, and Annabelle Wittels, assesses variation in a set of typical indicators derived from data sets of public service surveys from 10 administrations in Africa, Asia, Europe, North America, and South America.

The results show that measures related to personal characteristics such as motivation do not vary as much as those relating to management practices such as leadership. When respondents are asked to assess practices of others, such as their superior or their organization, survey responses exhibit significant discriminant variation across organizations and groups. By contrast, when respondents are asked to self-assess whether they possess desirable characteristics such as work motivation, survey responses across countries tend to be heavily skewed toward favorable answers, and variation is so compressed that meaningful differences between organizations or demographic groups are difficult to detect. Standard measures for desirable attitudes, such as motivation, may therefore need to be redesigned in surveys of public servants to better discriminate between values at the top end of indexes.

### **Designing Survey Questionnaires: To What Types of Survey Questions Do Public Servants Not Respond?**

Chapter 22, by Robert Lipinski, Daniel Rogger, and Christian Schuster, shows that surveys of public servants differ sharply in the extent to which respondents skip or refuse to respond to questions. So-called item nonresponse can affect the legitimacy and quality of public servant survey data. Survey results may be biased, for instance, if those least satisfied with their jobs are also most prone to skipping survey questions. Understanding why public servants respond to some survey questions but not others is thus important.

The chapter offers a conceptual framework and empirical evidence to further this understanding. Drawing on other survey methodology research, the chapter theorizes that public servants are less likely to respond to questions that are too complex (because they are unable to answer them) or sensitive (because they are unwilling to respond). Coding the complexity and sensitivity of public servant survey questions in Guatemala, Romania, and the United States, chapter 22 finds one indicator of complexity to be the most robust predictor of item nonresponse across countries: respondents' lack of familiarity with the information or topic examined by a survey question. By contrast, other indicators of complexity or sensitivity or machine-coded algorithms of textual complexity do not predict item nonresponse. The implication for survey design is clear: Avoid questions that require public servants to speculate about topics with which they are less familiar.

### **Designing Survey Questionnaires: Should Surveys Ask about Public Servants' Perceptions of Their Organization or Their Individual Experience?**

A third prerequisite for effective questionnaire design is valid measurement of organizational aggregates, such as which surveyed organization has the highest level of job satisfaction among its employees or

which organization has the highest quality of leadership practices of superiors. This raises the issue of whether respondents should be asked about their perceptions of organizational practice overall (so-called organizational referents) or whether questions should ask about the respondent's own experience, such as the quality of their superior's leadership practices or their own job satisfaction (so-called individual referents). In chapter 23, Kim Sass Mikkelsen and Camille Mercedes Parker examine this question using survey experiments with public servants in Guatemala and Romania. The survey experiments randomly assign public servants to respond to questions about a topic with phrasing using either an organizational referent or an individual referent.

The chapter finds that, while there are no strong conceptual grounds to prefer either organizational or individual referents—both have advantages and disadvantages—the choice matters to responses and alters response means (such as the average job satisfaction of employees in an organization). Organizational questions are particularly useful when questions are very sensitive (such as on corruption) because respondents may skew their response more strongly toward a socially desirable response option when asked about their own individual experience rather than practices in the organization. Individual questions are particularly useful when the attitudes or practices being measured are rare in the organization. In such cases, many respondents may lack the information to accurately assess the prevalence of a practice in the organization, risking that they rely instead on unrepresentative information or stories, for instance, rather than actual organizational characteristics. In short, whether survey questions should ask about the individual's own experience or the individual's perception of organizational practice depends on the characteristics of the question and the organization it seeks to measure.

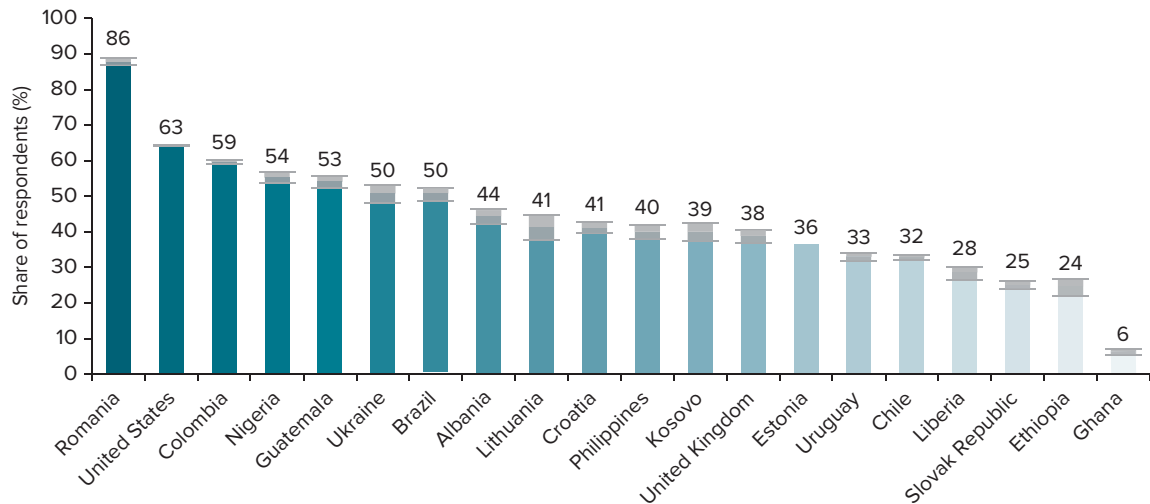
### **Interpreting Survey Findings: Can Survey Results Be Compared across Organizations and Countries?**

Chapters 24 to 26 turn to the interpretation and reporting of survey results. In chapter 24, Robert Lipinski, Jan-Hinrik Meyer-Sahling, Kim Sass Mikkelsen, and Christian Schuster focus on interpretation and in particular the question of benchmarking: Can survey results be compared across organizations and countries? This matters because survey results can rarely be understood in a void. Rather, they require benchmarks and points of reference. If, for instance, 80 percent of public servants are satisfied with their jobs, should a manager interpret this as a high or low level? Historical comparisons provide a sense of dynamics over time, but not a sense of degree. The availability of similar statistics from comparator organizations in the public sector or other countries is a potentially valuable complement to a manager's own results. However, such benchmarking requires that survey questions measure the same concept in the same way, making meaningful comparisons possible. Even when questions are phrased in the exact same way, however, the validity of comparison is not obvious. For multiple reasons, including work environment, adaptive expectations, and cultural factors, different people might understand the same question in distinct ways and adjust their answers accordingly. This might make survey results incomparable not only across countries but also across different groups of public servants within a national public administration.

To assess this concern empirically, chapter 24 investigates to what extent the same survey questions measure the same concept similarly—that is, questions are measurement invariant—using questions related to “transformational leadership” and data from seven public service surveys from Europe, Latin America, and South Asia. The chapter finds support for so-called scalar invariance: the topic (in this case, means of transformational leadership) can be compared *within* countries across organizations and demographic groups (the chapter authors test for gender and educational levels). Across countries, the chapter finds tentative evidence for scalar invariance, and stronger evidence when countries are grouped by regions and income.

The findings—although tentative and requiring further confirmatory evidence from other settings—thus underscore the utility of global benchmarking of countries in surveys of public servants.<sup>3</sup> As chapter 18 explains, the Global Survey of Public Servants offers a tool to harmonize questions across governments. In conjunction with the freely accessible Global Survey Indicators dashboard, the Global Survey of Public

**FIGURE 2.6** Share of Public Servants Satisfied with Their Pay and/or Total Benefits, Various Countries



Source: Fukuyama et al. 2022, figure 9.

Note: Years of measurement vary by country. Colors denote the extent of job satisfaction, with darker shades signifying greater job satisfaction. The gray vertical bars denote 95% confidence intervals.

Servants thus enables governments to understand strengths and areas for development of their public administration in global comparative terms. For example, the Global Survey provides comparative data on pay satisfaction from public administrations around the world. As can be seen from figure 2.6, this varies greatly across countries.

Once decisions are made about whom to benchmark against, consideration turns to how to report and disseminate survey results—that is, how to make the most of survey results. Chapters 25 and 26 provide two complementary perspectives on this challenge.

### Making the Most of Public Servant Survey Results: Lessons from Six Governments

Chapter 25—by Christian Schuster, Annabelle Wittels, Nathan Borgelt, Horacio Coral, Matt Kerlogue, Conall Mac Michael, Alejandro Ramos, Nicole Steele, and David Widlake—presents a self-assessment tool that lays out low-cost actions governments can take to support evidence-based reforms based on the insights from public servant surveys. The chapter applies this tool to governments of six countries (Australia, Canada, Colombia, Ireland, the United Kingdom, and the United States) to assess the comprehensiveness of their ecosystem to turn survey results into management improvements.

The self-assessment tool focuses on three main components of an ecosystem to turn survey results into management improvements: (1) information, (2) capacity, and (3) incentives to take action. For the first component (information), public servant survey results can improve public administration by providing tailored survey results to four main types of users: the government as a whole; individual public sector organizations; individual units or departments within a public sector organization; and the public, including public sector unions. Results reporting should identify key takeaways about the strengths and weaknesses of particular organizations and enable users to explore aggregate survey results in a customized manner, such as through dashboards.

For the second component (capacity to take action), reporting of results can become more effective when it includes (automated) recommendations to users—such as managers of units or organizations—on how to best address survey findings, as well as action plans for users to develop their own actions. Where more

resources are available, tailored technical assistance—or a human resources management (HRM) consultancy, provided either by a central HR unit or an external provider—can further help managers turn survey findings into improvements.

For the third component (incentives to take action), accountability mechanisms are key. For instance, governments can introduce central oversight of actions taken in response to survey findings by government organizations and units; can construct and publicize “best place to work” in government indexes to foster external oversight; or measure employee perceptions of the extent to which government organizations take action in response to survey findings.

Applying this self-assessment framework to the six governments, chapter 25 finds that many governments could undertake a range of additional low-cost actions to enhance the benefits they derive from public servant surveys to improve public administration.

## Using Survey Findings for Public Action: The Experience of the US Federal Government

Chapter 26—by Camille Hoover, Robin Klevins, Rosemary Miller, Maria Raviele, Daniel Rogger, Robert Seidner, and Kimberly Wells—complements chapter 25 by delving into the experience of the United States, the country with the longest-standing governmentwide employee survey. The chapter emphasizes, first, the importance of considering action in response to survey results at the time the survey is being designed. In particular, questions should focus on topics that staff and senior leaders find most important to achieve their mission. Second, the chapter describes the critical architecture necessary in each public sector organization to translate survey results into management improvements. This includes, for instance, a technical expert in the organization capable of interpreting survey data; a strong relationship between that expert and a senior manager in the organization who acts as a “change champion”; and the development of a culture for initiatives for improvements informed by the survey.

The chapter also provides guidance on how to develop a culture of responsiveness to surveys of public servants. It emphasizes the importance of leaders in an organization being transparent in sharing and discussing the survey results with their workforce, codeveloping action plans with staff, and coproducing improvements in response to survey results.

Part 4 thus provides evidence-based good practice on a range of choices involved in designing, implementing, interpreting, and reporting on surveys of public servants. This evidence can inform actions by governments seeking to improve their existing regular governmentwide employee surveys, as is the case in many OECD countries. It can enable governments that have yet to introduce surveys of public servants to leapfrog to best practice from the start. Such governments are encouraged to consult the range of toolkits on the Global Survey of Public Servants toolkit site.<sup>4</sup>

## PART 5: GOVERNMENT ANALYTICS USING EXTERNAL ASSESSMENTS

Part 5 turns to select data sources available to undertake government analytics through external assessments: that is, assessments conducted on or by those outside government (rather than data on public servants or administrative data collected by government organizations themselves). Chapters 27 through 30 provide guidance on four external data sources: household survey data, citizen survey data, service delivery indicators, and anthropological methods. These data sources illustrate the possibility of government analytics through external assessments but do not cover the full range of microdata for external assessments. Enterprise surveys of businesses, for instance, can provide insights on topics such as bribery or government regulation (World Bank 2023). With that caveat in mind, the part 5 chapters provide important insights about how to do government analytics using external assessments.

## Government Analytics Using Household Surveys

Chapter 27, by Faisal Ali Baig, Zahid Hasnain, Turkan Mukhtarova, and Daniel Rogger, describes how to use household survey data for government analytics. Such data are readily available in many countries. In particular, national statistical authorities frequently collect labor force (and related household) surveys that are broadly consistent across time and developed using globally standardized definitions and classification nomenclatures. Governments can leverage these data to gain insights into the public sector workforce that administrative data or public servant survey data do not provide. In particular, labor force survey data allow governments to explore and compare public and private sector labor markets (because labor surveys cover both populations), as well as labor markets in different regions of a country or over time. Map 2.2, reproduced from chapter 27, presents differences in labor market features for Indonesia. The role of public sector employment in the formal sector varies from 15 percent to 60 percent of paid employment, implying substantial economic vulnerability of some regions to changes in public employment practices and policies.

Chapter 27 shows how such comparisons provide a wealth of insights into the input side of the public administration production function. To cite just three examples: Labor force data analytics enable governments to understand gender pay and employment differences between the public and private sectors, and whether the public sector promotes gender equality in employment in both absolute terms and relative to the private sector. The analytics help governments assess pay competitiveness, providing answers to whether the public sector pays competitive wages compared to the private sector to attract talent while not crowding out private sector jobs. Household and labor force survey data can also shed light on the skills composition of the public sector workforce with respect to the private sector and identify in what areas the government is competing most intensively for skills with private sector actors.

In short, such data can complement payroll, HRMIS data, and public servant survey data to provide a more complete diagnosis of public pay and employment. To facilitate access to these analytics, chapter 27 also highlights the freely available Worldwide Bureaucracy Indicators (WWBI), a set of indicators based on labor force survey data from more than 200 countries compiled by the World Bank to assess public and private labor markets and their interaction across the world.

**MAP 2.2 Subnational Patterns in Public Sector Employment, Indonesia, 2018**



Source: Original map for this publication, based on data from the World Bank Worldwide Bureaucracy Indicators database v 1.1, <https://www.worldbank.org/en/data/interactive/2019/05/21/worldwide-bureaucracy-indicators-dashboard>.

## Government Analytics Using Citizen Surveys: Lessons from the OECD Trust Survey

Chapter 28, by Monica Brezzi and Santiago González, examines government analytics using citizen surveys. These surveys can help governments shed light on certain outcomes in the public administration function. In particular, they can capture the outcomes of public governance as perceived and experienced by people, through nationally representative population samples. For instance, citizen surveys are used in many countries to measure satisfaction with widely used public services (such as tax administrations, schools, or hospitals).

As chapter 28 shows, they can also be used to understand broader government outcomes. In particular, the chapter illustrates the potential of such surveys for government analytics using the example of the OECD's Survey on the Drivers of Trust in Public Institutions (OECD Trust Survey). The survey measures trust of citizens in government and captures their expectations of and experiences with public institutions around key drivers of trust. Measures of trust and its drivers include the competence of public institutions—including access to public services and their quality and reliability—as well as the perceived values of public institutions, notably in terms of integrity, openness to involving citizens, and fairness in the treatment of citizens.

Chapter 28 showcases how governments have used evidence resulting from the survey to develop concrete actions to strengthen institutional trust. The chapter provides guidance for other governments wishing to apply this international benchmark on measuring trust in public institutions, following the OECD Guidelines on Measuring Trust.

## Government Analytics Using Measures of Service Delivery

In chapter 29, Kathryn Andrews, Galileu Kim, Halsey Rogers, Jigyasa Sharma, and Sergio Venegas Marin go beyond this book's core focus on public administration to the frontline of service delivery and introduce measures of service delivery (MSDs). Currently, MSDs are focused on education and health. Mirroring this *Handbook's* approach to conceptualize government analytics along the public administration production function from inputs to outcomes, MSDs provide objective measurements not only of service quality (such as absenteeism of medical doctors, and test results of students in school) but the entire process involved in delivering frontline public services (including input and process measures such as availability of medicine, and management practices in schools).

MSDs offer a granular view of the service delivery system, providing actionable insights on different parts of the delivery chain, from the physical infrastructure to the knowledge of frontline providers. These can be usefully viewed as outputs and “outcomes” of the administrative environment embedded in government agencies under which these frontline providers fall. Measurement of these different factors of production allows practitioners to map out, conceptually, how each part of the production chain is faring, and where improvements can be made, at the individual provider level as well as part of the wider production function for government outlined in figure 2.1. MSDs also provide action-oriented visualizations of these indicators, enabling practitioners to design their service delivery policies in an intuitive and evidence-based approach.

Chapter 29 provides a road map to interested practitioners to produce MSDs, from design, implementation, and analysis to dissemination. The chapter emphasizes that developing service delivery indicators requires considering and defining relevant dimensions of quality in a public service, along with relevant inputs, with a premium on defining indicators according to policy objectives and resource constraints. Akin to the discussions in chapters 25 and 26, chapter 29 also underscores the importance of linking MSDs directly to stakeholders who have the ability to enact change in the delivery system. Many of the steps involved in the analytics of core public administration are thus mirrored in the analytics of service delivery.



## Government Analytics Using Anthropological Methods

Finally, chapter 30, by Colin Hoag, Josiah Heyman, Kristin Asdal, Hilde Reinertsen, and Matthew Hull, returns to chapter 4's call for embedding qualitative studies in government analytics. It considers how government analytics can be undertaken through an anthropological approach, a powerful means of collecting qualitative data. Anthropologists are most commonly associated with immersive, ethnographic methods such as participatory observation. Chapter 30 applies that lens to studying public administration. As the chapter authors emphasize, "Anthropologists are motivated by an abiding concern with empirical rigor—a refusal to ignore any sort of data or to content oneself with a single view of such a multifarious thing as bureaucracy." Doing so risks overlooking factors that shape organizations.

Anthropological methods suggest that data collection should approach government analytics by engaging with the staff who are involved at every level of the organization, from senior officers to low-level staff and contractors, and across different demographic groups; studying everyday documents; and watching how officials interact. By observing every part of what public officials do at work in a holistic way, from their interactions in corridors and meetings to the protocols they observe in their relationships, the analyst undertakes the most holistic data collection strategy feasible.

Such an approach requires analysts to develop relationships with a variety of types of people in an organization and have open-ended conversations about their work and unrelated issues to understand their values and perspectives. It also requires analysts to engage in participant observation to capture activities that may be so routine they go unnoticed by public officials and are not self-reported in surveys. Moreover, it requires analysts to collect the widest practical range and amount of qualitative and quantitative data, even if such data cannot be easily standardized. Finally, it requires analysts to study not only data but also the interactions and microscopic decisions that affect the gap between stated policy goals and the actual work being carried out by public officials—for instance, by studying what public officials say and do, including the rationales they draw on for their decisions.

Chapter 30 thus emphasizes that government analytics can incorporate methods that provide insights into aspects of the public administration function that quantitative microdata cannot. It also brings the *Handbook* back to the first cross-cutting chapter (chapter 4), which emphasizes the importance of a holistic and "balanced data suite." Part of this "suite" is ensuring that problem analysis in public administration is holistic: that important parts of a problem are not neglected due to the absence of quantitative data and measurement. This, in turn, puts a premium on utilizing qualitative and anthropological methods to complement insights gleaned from microdata.

Holistic analytics benefit not only from triangulating different quantitative and qualitative methods, but also from triangulating and integrating the analytics of different components of the public administration production function. The chapter concludes with a challenge: How can the different approaches and data sources in the government analytics toolbox be integrated effectively to diagnose major challenges in public administration holistically?

## CONCLUSION: TOWARD A HOLISTIC ANALYTICS OF CORE CHALLENGES IN THE MACHINERY OF GOVERNMENT

How can practitioners take the distinct data sources detailed in the *Handbook* to the frontier? Frontier government analytics would integrate the analytics of the data sources described across the *Handbook* into standard government practice. It would generate them at a scale sufficient to inform the decision-making of individual managers. And it would make them easily accessible to those managers across government organizations and departments. For instance, dashboards integrating data sources and updating in real time would provide managers with comparisons for their staffing issues, process quality, the perceived quality of



management practices, and so on. They could keep tabs on outputs and outcomes, from task completion and case productivity to external assessments from citizens. Comparative data would allow them to benchmark themselves against other government organizations, or where appropriate, other countries. Managers would be capable of understanding the limitations and strengths of different analytics. The result would be a transformational change toward leveraging data to strengthen public administration.

Where should the journey toward this frontier begin? As a first step, the government analytics of the individual data sources explored in detail in various chapters can provide important insights into different components of the public administration production function. Once governments have analytical tools for several components in place (such as payroll diagnostics, and surveys of public servants), the possibilities for government analytics further expand. The integration of multiple government analytics data sources enables governments to diagnose and address major government challenges holistically (chapter 9). In some cases, such as customs (chapter 14), such integration is a vital part of measuring attainment of the organization's mission. In others, the insights that can be gleaned go beyond those available from individual data sources and thus enable a holistic perspective on public administration.

To illustrate, consider how the integration of government analytics data sources described in this *Handbook* can shed light on key challenges in public administration: corruption and personnel management.

Corruption is a multidimensional phenomenon, affecting public administration across its production function. As chapter 8 discusses, corruption can first be detected with input data. For example, payroll data can be drawn on to detect ghost workers. HRMIS data can be drawn on to detect instances of nepotism in recruitment, such as when family members with similar last names are hired. Procurement data can be drawn on to detect procurement fraud and collusion risks (such as when organizations grant contracts without competitive bidding). Expenditure data can be drawn on to detect off-budget spending at risk of embezzlement.

Second, corruption can be detected in the processes and practices that convert inputs into outputs and generate norms and behaviors in public administration. For instance, surveys of public servants can measure unethical leadership by superiors (such as pressures on subordinates to collude in corruption), as well as the ethical norms and integrity of public servants themselves (such as their perceptions of corruption of colleagues).

Third, corruption can be detected in output and outcome data (such as in tax audit case data, or through surveys of citizens querying them about bribery requests from public officials). Understanding at a granular level where corruption occurs—in a specific public administration production function, or in a particular department or organization—enables governments to identify comprehensive but tailored and evidence-based solutions to curb corruption.

Consider, as a second example, civil service management—and how the integration of analytics across the public administration production function can aid better personnel management in government. As chapters 9 and 10 emphasize, data analytics of payroll and HRMIS systems can diagnose personnel as an input into the production function. To cite two examples: pay equity can be assessed across groups (such as by gender) and institutions; and retention challenges can be pinpointed by comparing turnover rates for public servants at different ranks or in different regions of the country. Such data can be complemented by labor force and household survey data (chapter 27)—for instance, to understand whether public sector pay is competitive relative to the private sector, or whether the workforce allocation across the territory is appropriate given the respective number of service users. Other input data, such as on budgets, can help understand whether productivity problems might arise from the lack of complementary inputs to personnel in the production function.

Data on processes and practices can then shed light on whether government is effectively converting personnel inputs into outputs and outcomes. Surveys of public servants, as explored in part 4, can measure the experience of public servants with management practices, as well as the culture, attitudes, and behaviors in public administration that mediate the conversion of inputs into outputs. For instance, is leadership by line managers effective? Are public servants motivated to work hard? Anthropological and qualitative

methods, as discussed in chapter 30, can enrich and contextualize survey responses through participant observation and detailed case study work. As chapter 13 shows, government analytics can also assess processes converting inputs into outputs through administrative data—for instance, to assess whether institutions follow due procedure when evaluating the performance of public servants.

Last, data on outputs and outcomes can help diagnose and improve personnel management. For instance, as chapter 17 explains, data on task completion can help understand where in government public servants are effectively completing tasks and where they are not. For public sector organizations that complete cases, administrative case data, as explored in chapters 14 and 15, can help understand differences in the productivity of public servants more broadly, while surveys of citizens can help understand differential citizen satisfaction and trust with public services across institutions and regions in a country.

## Strengthening Public Sector Management through Government Analytics

At the center of government are the public officials who navigate the strategic and daily decisions that determine public policy and the effectiveness of public administration. The information public managers have, and the extent to which they use it for running public organizations, drives the efficacy of government. An effective system of government analytics empowers officials to manage government based on evidence of the administration's current realities.

Empowering public officials with government analytics can transform the basis for personnel management—and public administration more generally. It enables government decision-makers to complement their tacit and practical knowledge about public administration with evidence-based and data-informed insights to improve how the machinery of government is run.

Almost all officials manage, organize, or process in ways that could usefully be analyzed as an input to strengthening government functioning. Thus, communicating analytical insights to government officials at the time that they are making a decision, as discussed in chapters 25 and 26, is a critical link in making analytics effective. One approach is to support decision-making through a human resources management dashboard that brings together the distinct data sources a government has available to visualize and benchmark the strengths and areas for improvement in personnel management to decision-makers in each government institution. By implementing any of the approaches in this *Handbook*, government or its stakeholders are building a platform for managers and decision-makers to do more with the public sector's finite resources.

Even with 30 chapters, this book does not cover all potential data sources and approaches to government analytics. There will always be new analytical opportunities on the horizon. Chapter 3 discusses how government can continuously prepare for a transition to new analytical approaches. Above all, building the culture that binds public administrators and public administration together requires a commitment from senior management and individual officials to use evidence about their own administration in their decision-making. We invite all public servants and related stakeholders to capitalize on the insights summarized in this chapter and this *Handbook* and push forward the quality of management of their administration. The quality of government around the world will be shaped by how its decision-makers leverage data to strengthen their administration.

## ANNEX 2A MAPPING GOVERNMENT ANALYTICS DATA SOURCES TO CHAPTERS IN THE *HANDBOOK*

**TABLE 2A.1 Mapping Government Analytics Data Sources to Chapters 10–30**

Data source	Chap.	Examples of uses of analytics	Examples of indicators
<i>Part 3. Administrative data</i>			
Payroll and HRMIS	10	<ul style="list-style-type: none"> <li>Examine fiscal planning and sustainability of the wage bill.</li> <li>Allocate the workforce across government departments and territories.</li> <li>Set pay.</li> </ul>	<ul style="list-style-type: none"> <li>Wage bill by sector and years.</li> <li>Distribution of civil servants by rank.</li> <li>Turnover of civil servants.</li> <li>Pay inequity between ministries.</li> <li>Retirement projections.</li> </ul>
Expenditure	11	<ul style="list-style-type: none"> <li>Assess efficiency, equity, and effectiveness of government spending.</li> <li>Assess whether government spending corresponds to budget priorities.</li> <li>Identify large transactions with high fiduciary risk.</li> </ul>	<ul style="list-style-type: none"> <li>Budget execution rates.</li> <li>Share of government expenditures covered by the FMIS.</li> <li>Share of total expenditures by transaction value.</li> <li>Overdue accounts payable (payment arrears).</li> </ul>
Procurement	12	<ul style="list-style-type: none"> <li>Monitor procurement markets and trends.</li> <li>Improve procurement and contracting (for instance, by identifying goods organizations overpay for, or organizations with high corruption risks in procurement).</li> <li>Assess effects of distinct procurement strategies or reforms.</li> </ul>	<ul style="list-style-type: none"> <li>Time needed for contracting.</li> <li>Number of bidders.</li> <li>Share of contracts with single bidder.</li> <li>Final price paid for a good or service.</li> <li>Share of contracts with time overruns.</li> <li>Share of bidders that are small and medium enterprises.</li> </ul>
Administrative processes	13	<ul style="list-style-type: none"> <li>Assess the speed and quality of administrative back-office processes and process implementation (for instance, for project planning, budget monitoring, or performance appraisals).</li> </ul>	<ul style="list-style-type: none"> <li>Adherence of administrator's process work to accepted government procedure.</li> <li>Timeliness of administrator's process work with respect to deadlines.</li> </ul>
Customs	14	<ul style="list-style-type: none"> <li>Assess customs revenue collection.</li> <li>Assess trade facilitation (flow of goods) through customs across borders.</li> <li>Assess whether customs safeguards safety of goods and protects people (for instance, prevents dangerous goods from crossing).</li> </ul>	<ul style="list-style-type: none"> <li>Time delays in customs clearances.</li> <li>Cost of customs process.</li> <li>Total customs revenue collected.</li> <li>Number of goods in infraction seized in customs.</li> </ul>
Case data	15	<ul style="list-style-type: none"> <li>Assess productivity of organizations and individuals processing cases (such as tax claims).</li> <li>Assess the quality of case processing.</li> <li>Identify “best performing” offices and transfer best practices to other offices.</li> </ul>	<ul style="list-style-type: none"> <li>Total number of cases completed.</li> <li>Average time to complete one case.</li> <li>Error rate in completing cases.</li> <li>Timeliness of case completion.</li> <li>Share of cases with complaints.</li> </ul>

(continues on next page)

**TABLE 2A.1 Mapping Government Analytics Data Sources to Chapters 10–30**  
(continued)

Data source	Chap.	Examples of uses of analytics	Examples of indicators
Text-as-data (machine learning)	16	<ul style="list-style-type: none"> <li>Analyze “big data” and convert unstructured text from government records into data for analytics.</li> <li>Examples of big data include payroll disbursements to civil servants or tax filings by citizens and firms.</li> <li>Examples of text data include court rulings, procurement tenders, or policy documents.</li> </ul>	<ul style="list-style-type: none"> <li>Risk score for a procurement tender or payroll disbursement based on a predictive algorithm.</li> <li>Bias score in court ruling, based on textual analysis of wording in the document (sexist or racial-profiling terms).</li> </ul>
Task and project completion data	17	<ul style="list-style-type: none"> <li>Examine the frequency of completion of tasks on schedule (such as delivery of training program, preparation of budget).</li> <li>Understand quality of administrative task completion.</li> <li>Understand drivers of differences in task completion across government units.</li> </ul>	<ul style="list-style-type: none"> <li>Number of tasks completed.</li> <li>Share of tasks completed on time.</li> <li>Share of unfinished tasks.</li> <li>Share of tasks completed as planned.</li> </ul>
<i>Part 4. Surveys of public servants</i>			
Civil service survey data	18–26	<ul style="list-style-type: none"> <li>Assess the quality of management practices (such as performance management, recruitment).</li> <li>Assess norms, attitudes, and cultures in public administration (such as work motivation).</li> </ul>	<ul style="list-style-type: none"> <li>Share of public servants hired through merit examinations.</li> <li>Share public servants wishing to leave the public sector in the next year.</li> <li>Share of public servants favorably evaluating the leadership practices of their superior.</li> </ul>
<i>Part 5. External assessments</i>			
Household survey data	27	<ul style="list-style-type: none"> <li>Assess wage differences between the public and the private sector.</li> <li>Identify regions with lower public sector wage premiums.</li> <li>Assess gender representation in the public sector.</li> </ul>	<ul style="list-style-type: none"> <li>Average public sector wage premium.</li> <li>Share of women in public employment.</li> <li>Share of public employment in total employment.</li> </ul>
Citizen survey data	28	<ul style="list-style-type: none"> <li>Assess satisfaction of citizens with public services.</li> <li>Assess trust of citizens in government institutions.</li> <li>Assess interactions of citizens with public administration (such as bribery requests).</li> </ul>	<ul style="list-style-type: none"> <li>Share of citizens satisfied with health services.</li> <li>Share of citizens trusting the civil service.</li> <li>Share of citizens asked for a bribe by a government official in the last year.</li> </ul>
Service delivery indicators	29	<ul style="list-style-type: none"> <li>Assess the quality of education and health service delivery.</li> <li>Assess facility-level characteristics (such as teacher presence).</li> </ul>	<ul style="list-style-type: none"> <li>Share of days during which teachers are not present in school.</li> <li>Availability of resources in hospital to treat patients.</li> </ul>
Anthropological analytics	30	<ul style="list-style-type: none"> <li>Observe everyday practices in public administration, to capture routine but unnoticed parts of administration.</li> <li>Observe social engagement to understand formal and informal rules, relationships, and interactions between public servants.</li> </ul>	<ul style="list-style-type: none"> <li>Holistic participant observation of everyday life inside particular public administrations.</li> <li>Understand how public servants interpret broader policy goals.</li> </ul>

Source: Original table for this publication.

Note: FMIS = financial management information system; HRMIS = human resources management information system.

## NOTES

1. In contrast to the relatively coherent consensus of functions of private sector production (Mas-Colell, Whinston, and Green 1995), no consensus has formed around an integrated model of a production function for public administration. This is due in part to the limited use of microlevel data on the workings of different components of public administration—the very gap that motivated this *Handbook*.
2. Whether inputs effectively convert into outputs is also moderated by exogenous factors, such as the political environment. This *Handbook* does not discuss the analysis of microdata to assess these exogenous factors. Instead, readers are encouraged to consult the many existing excellent resources to understand the exogenous and political environment of public administrations (see, for example, Moore 1995).
3. Read in conjunction with other research cited in chapter 24, this conclusion holds particularly for questions that are more factual and less culturally specific. For instance, questions on specific management practices (such as around the presence of certain performance evaluation practices) can be more plausibly benchmarked across countries without measurement invariance concerns than attitudinal questions (such as on how engaged employees are).
4. The Global Survey of Public Servants was cofounded by the two editors of this *Handbook*, along with a range of practitioners and academic colleagues. The toolkits on the website build upon and incorporate much of the evidence reviewed in this *Handbook*. See <https://www.globalsurveyofpublicservants.org>.

## REFERENCES

- Bakshi, H., A. Bravo-Biosca, and J. Mateos-Garcia. 2014. “The Analytical Firm: Estimating the Effect of Data and Online Analytics on Firm Performance.” Nesta Working Paper 14/05, Nesta, London.
- Fenizia, A. 2022. “Managers and Productivity in the Public Sector.” *Econometrica* 90 (3): 1063–84.
- Fukuyama, F., D. Rogger, Z. Husnain, K. Bersch, D. Mistree, C. Schuster, K. Sass Mikkelsen, K. Kay, and J.-H. Meyer-Sahling. 2022. Global Survey of Public Servants. <https://www.globalsurveyofpublicservants.org/>.
- Mas-Colell, A., M. D. Whinston, and J. R. Green. 1995. *Microeconomic Theory*. Oxford, UK: Oxford University Press.
- Meyer-Sahling, J., D. Mistree, K. Mikkelsen, K. Bersch, F. Fukayama, K. Kay, C. Schuster, Z. Hasnain, and D. Rogger. 2021. *The Global Survey of Public Servants: Approach and Conceptual Framework*. [https://www.globalsurveyofpublicservants.org/\\_files/ugd/acee03\\_91fc43ed92774eb88704717d6c7c80a2.pdf](https://www.globalsurveyofpublicservants.org/_files/ugd/acee03_91fc43ed92774eb88704717d6c7c80a2.pdf).
- Moore, M. 1995. *Creating Public Value: Strategic Management in Government*. Cambridge, MA: Harvard University Press.
- UK Cabinet Office. 2022. “UK Civil Service People Survey: Technical Summary.” <https://www.gov.uk/government/publications/civil-service-people-survey-2022-results/civil-service-people-survey-2022-technical-guide>.
- UK Parliament. 2022. “MPs to Probe Quality of Civil Service People Survey.” <https://committees.parliament.uk/committee/327/public-administration-and-constitutional-affairs-committee/news/173010/mps-to-probe-quality-of-civil-service-people-survey/>.
- UN (United Nations). 2014. *A World That Counts: Mobilising the Data Revolution for Sustainable Development*. Report prepared at the request of the United Nations Secretary-General by the Independent Expert Advisory Group on a Data Revolution for Sustainable Development. <https://www.undatarevolution.org/wp-content/uploads/2014/12/A-World-That-Counts2.pdf>.
- World Bank. 2023. Enterprise Surveys. <https://www.enterprisesurveys.org/en/enterprisesurveys>.
- World Bank Group. 2019. “Innovating Bureaucracy for a More Capable Government.” <https://openknowledge.worldbank.org/handle/10986/31284?show=full>.



## CHAPTER 3

# Government Analytics of the Future

*Daniel Rogger and Christian Schuster*

### SUMMARY

The investments governments make in measurement today will determine what they know tomorrow. Building an analytics system in government has long-term benefits for our ability to manage scarce public resources and detect unforeseen risks. This chapter provides guidance on what public organizations can do today to become more analytical tomorrow. Government institutions themselves require reshaping: by enhancing structures for planning; by equipping public sector managers with a greater ability to consume and interpret analytics; and by developing new architecture for analytical units. Assisting each public sector organization to develop its own analytics agenda induces cultural change and targets the analytics to the requirements of its specific mission. Rewarding experimentation with novel data sources improves government's capacity to innovate more generally. Each of these changes helps chart a course to the government analytics of the future.

### ANALYTICS IN PRACTICE

The guidance that follows aims to facilitate the transition process to build an environment for analytical insights across government:

1. Continuously plan to capitalize on the opportunities afforded by innovations in measurement and analysis of government functioning.
2. Develop units of government analytics at the center of government and within each major organizational unit, and embed them in a community of practice. Centralized units enable economies of scale in both the implementation of analytics and the breadth of comparable data created, as well as network economies from users investing in a common data architecture. Units within organizations can complement central analytics by helping interpret analytics for their organization, and adapting analytics tools to particular organizational needs.



3. Build a public sector management cadre able to undertake and interact with frontier measurement and government analytics. Technological advances in measurement and analysis reinforce the importance of capable public sector managers. A cadre of public managers literate in government analytics is aware of the boundaries and assumptions of government measurement and analysis and adept in using analytical results to complement a broader understanding of public service.
4. Pursue a centralized analytical agenda that harmonizes common variables and conducts joint analysis of them, with specific agencies supporting this “public good.” The lack of objective benchmarks in many areas of government work puts a premium on harmonization and benchmarking through common variables across organizations. Similarly, governments should invest in measures that allow international comparisons.
5. Incentivize experimentation and an innovation culture in analytics through institutions that take responsibility for failures. Cultural shifts that reward smart experimentation irrespective of the outcome often come from the explicit support of senior leadership and political actors who endorse the process of innovation—and corresponding success and failure. To reinforce that cultural shift, actors across government—from senior managers through unit supervisors to individual employees—should define analytics agendas for their areas of responsibility.

## A GOVERNMENT COMMITMENT TO CONTINUOUSLY SEEK INNOVATION

The choices governments make today on what aspects of their machinery to measure and how to do so will determine what governments know tomorrow. Reviewing the analytical status quo, and planning for its future, should be an integral part of organizational and governmentwide strategies. This chapter provides guidance on how to build a government analytics of the future based on lessons in the chapters of *The Government Analytics Handbook*. Its starting point is recognition that measurement challenges are a defining feature of the public sector.

Outcomes of government intervention in many areas of the economy and society are often hard to observe. Inside government, measurement challenges pervade public management, with managers engaging their staff in tasks that cannot be fully defined in a manual or contract and that can change rapidly in response to a societal or political shock. Management in government is thus anchored in ambiguity and uncertainty, rather than measurement and measurability. This is the environment that public sector managers must make their decisions in every day.

This has always been true. Early governments grappled with measuring the scale of the economy and its taxable component. As governments have scaled up their activities to take on an increasing role in society, they have had to measure an increasingly broad range of activities. As the complexity of society grows, so does the complexity of the organizations government must build to manage its interactions with society, and the corresponding measurement tasks. Conversely, given the centrality of the government budget process, public sector managers have had to collapse much of their measurement and activity back to cash terms that can be put into a centralized budget. Thus, public officials have always faced the tension between the incompleteness of what they know and the practical requirement to make policy decisions.

Take the case of regulation. The performance of public sector regulators will be judged by regulators’ capacity to understand and make effective decisions about the sectors that they regulate. As society’s economic environment becomes more complex, it naturally yields more complex issues for regulators to comprehend. In response, governments may hire more specialized public sector professionals to undertake required analysis of the complexity—which in turn increases the complexity of the public sector itself. Government must then determine the performance of those professionals, how they should be managed, and what might increase their capacity to undertake their job now and in the future.

Thus, in the future, government will struggle with measurement issues, just as its predecessors have. But there is a qualitative difference: in terms of both its understanding of the world it must govern, and in comprehending its own structures, capabilities, and challenges, the future of government will be far more complex as the world grows more complex.

Fortunately, however, future efforts will also benefit from technical advances in three aspects that facilitate the measurement and management of government, as discussed in chapters 1 and 2. The first is the digitization of many government services, activities, and records. The second is improvements in the analytical technology for understanding government functioning. The third is the increasing redesign of government as an analytical organization. An organization that is not designed for the collection and use of analysis about its functioning will simply not be able to use evidence, however rich. Such a redesign is multifaceted, starting with increased recognition of the need for more and better analytics, continuing through building the mindset and skills of public officials to support an analytical approach, and gaining momentum by setting up institutions that can undertake analytics and embed it in public administration.

In other words, the future of government is characterized by tension. On the one hand, governments can capitalize on the opportunities that innovations in digitization and analytics afford. On the other hand, they face increasing complexity in both society and in the organization of government. Managing the interplay between these two forces will be a central challenge. Government must build its capacity to engage with greater complexity in the world and within its own architecture.

How can this be done? This chapter provides answers. It begins by describing the institutional architecture that has been shown in the cases covered by the *Handbook* to strengthen government analytics. It continues by outlining what an effective agenda might look like to capitalize on the analytics that are available today and what may be available in the future. It concludes with a discussion of how a government's analytical architecture and analytical agenda might prepare for novel data sources that are not yet part of the administration's strategy but could be a useful addition in the future. These transformations require an informed conversation throughout the public service that drives the requisite cultural change. This *Handbook* aims to improve that conversation.

## **BUILD A GOVERNMENT OF THE FUTURE**

### **A Vision of the Future**

Like any forward-planning activity in the public sector, government should continuously plan to capitalize on the opportunities afforded by innovations in measurement and analysis of government functioning. Given the speed at which analytics is evolving, this should be a constant task. Each new innovation in the measurement of government functioning is a new opportunity for improved public sector performance.

The future of government analytics thus begins with an approach by government to continuously support innovation. This should become a strategic goal of politicians in their political agendas; of senior management of the public service in their guidelines for public service action, formulation of appraisals, and circulars to staff; and of middle management in their prioritization of analytical activities and use of corresponding results in their decision-making. Implementing these commitments in planning documents and work plans presents stakeholders with a credible signal of cultural change. These efforts are catalyzed by the formation of coalitions of political and technical officials interested in developing or innovating the analytics agenda. Chapter 26 describes how such coalitions around the Federal Employee Viewpoint Survey have substantively improved the US federal government.

Planning that includes a review of analytical opportunities should capitalize on the best evidence available in public decision-making. For example, in workforce planning, basic analytics would regularly monitor the likely shortfalls in staffing as current employees leave or retire. A higher level of analytics would aim to

predict what new roles might be necessary and which might become redundant. An even higher level would assess trends in the productivity of distinct roles, enabling adjustments to be made for shifting burdens of work. Attaining each of these levels entails strategic choices in bringing empirical evidence to the management of the public service. It also requires resources to be allocated for analytical purposes, and necessitates technical staff to work closely with senior managers to articulate the implications for personnel management. Chapter 10, for instance, zeroes in on how investments in the use of personnel management data have allowed future wage costs to be predicted. The chapter describes a ladder of quality of data, with each higher layer enabling an increase in the depth of analytical insights. The analytics provided a platform for the corresponding governments to head off fiscal shortfalls and avert a major wage bill crisis. But it was the strategic choices key decision-makers made to request the analysis and proactively respond to the results that were the key to success.

Public service cultures often guard against rapid innovation and technological change. To enable cultural change, a multitude of approaches can be taken. For instance, to signal high-level support, senior management can convey a vision of a government managed based on evidence. Senior managers and middle managers can celebrate the implementation of analytical approaches. Publicizing and explaining how government functioning in a specific agency has improved can help shift service norms toward acceptance. Hiring employees trained in data analytics and upskilling existing employees in data analytics can increase the interest in adopting innovations in analytics, and reduce the cost of making those changes. Evidence of how quickly public service culture can accept—and come to expect—novel measurement of administrative functioning can be seen in the rapid adoption of surveys of public servants in many countries in recent years, documented in chapter 18.

### Creating or Expanding Management Capabilities

As the task of integrating precision analytics with less measurable aspects of government work becomes more sophisticated, the need will grow for decision-makers capable of interpreting and integrating analytical insights with tacit managerial knowledge. For example, in the case of machine learning (ML), chapter 16 notes that “continuous collaboration between the ML implementation team and policy colleagues who will use its insights ensures that applications are adapted to and stay relevant to public administration’s needs.”

Ethical considerations are also paramount. For instance, chapter 16 emphasizes the important role public managers must play in assessing the ethics of machine-learning approaches in specific cases. Balancing the need for innovation to collect and analyze more and better data and safeguarding the public good will always be a fundamental aspect of public managers’ application and oversight of analytics. This is particularly true when data concern government itself because managers and organizations are the ultimate safeguards of their employees’ rights. Yet, as chapter 6 notes, “there is a dearth of discussion and practical guides on the ethics of data collection by government on its own employees.” The chapter provides a framework for public sector managers to judge the ethical use of government analytics in particular cases.

Another important foundation is to build what chapter 4 calls a balanced data suite to inform decision-making. As the chapter warns, “an overreliance on quantitative data comes with its own risks, of which public sector managers should be keenly aware.” While embracing a role for qualitative data, especially for those aspects that require in-depth, context-specific knowledge, analytics should focus quantitative measures of success on those aspects that are close to the problem. Analytics also needs to protect space for judgment, discretion, and deliberation in those (many) decision-making domains that inherently cannot be quantified. One way to attain a balanced suite is through the use of external assessments, as discussed in part 5, such as anthropological methods (chapter 30). To attain balance in the data suite, public managers need to identify and manage the organizational capacity and power relations that shape data management.

Managers of the public service also need to be able to have an informed discussion about when measurement is of the right nature and accuracy to make a particular claim. For example, chapter 20 shows that public servant surveys frequently do not have a sufficiently large sample to make valid comparisons across organizations about employee attitudes. An informed public official could therefore disregard such comparisons

when there is not a statistical basis to make them. The more profound the understanding of public sector decision-makers as to how measurement should be undertaken and how related analysis should optimally be mapped into decisions, the more useful government analytics will be.

All this implies that along with a commitment to analytical work, building a government of the future requires building a public sector management cadre capable of directing and interacting with frontier measurement and government analytics. This cadre should be aware of the boundaries and assumptions of that measurement and analysis and be capable of using analytical results in the context of their broader tacit understanding of the public service. Such managers should also be continuously aware of what the frontier of good practice looks like in undertaking analytical work. As chapter 5 shows, the range of freely available resources to support achieving this awareness is expanding rapidly.

Managers in individual organizations need to link to a community of practice, where they can combine learning from their own organization—and from their specific set of tasks—with learning from others. Embedding public managers in a community of practice for government analytics across the relevant administration, or across government, bolsters opportunities for learning and motivating officials, rather than leaving them as independent analysts who could be subsumed within the wider institutional environment. The network effects that arise from such a community underlie the rationale for central offices of government analytics. To encourage network effects, for instance, the US federal government holds workshops to build communities and connect with staff (see chapters 9 and 26).

## Analytics Architectures

Centralized units of analytics enable economies of scale in both the implementation of analytics and the breadth of comparable data created, as well as facilitating network economies from users investing in a common data architecture. For example, by mainstreaming public servant surveys into an integrated data system, a single entity can brand, market, and implement the survey (chapter 25); the statistical rigor of question design and analysis can be improved (chapters 19 through 23); and all agencies and units can compare their results to similar entities across the service (chapter 24). As more agencies use the survey for personnel management, the cultural norms around acceptability of the usefulness of the survey results shift and favor adoption (chapter 26).

Such benefits might be realized by mainstreaming government analytics into Integrated National Data Systems, typically managed by national statistical agencies. Locating analytics teams in statistical authorities may improve the statistical rigor and country ownership of the corresponding analysis. Such agencies provide a solid foundation for achieving scale and sustainability in the collection of data on public service. They also offer a platform for integrating data on government functioning with broader data on the targets of government action, such as the Sustainable Development Goals.

However, locating analytics teams in national statistical agencies outside of management agencies risks isolating analytics teams from decision-makers. In particular, these teams may not be responsive to the requirements of specific managers. To address that issue, the UK Cabinet Office and the US Office of Personnel Management have created centralized units of government analytics and located them in central management authorities rather than statistical agencies. Centralized delivery approaches have typically housed analytics teams within the heart of government, either in the presidency/prime minister's cabinet office or in a ministry of finance or public administration. The evidence leans toward developing units of government analytics at the center of government and within each major organizational unit, though whether this holds in a given government depends on its institutional context.

There may be ways to share analytical responsibilities across national statistical authorities and implementing agencies, but at this nascent stage in government analytics, few examples of such relationships exist. One example is Colombia's National Statistical Office (DANE), which conducts the country's governmentwide employee survey (chapter 18). Statistical agencies can also use existing data, such as household surveys, to provide insights into the functioning of public administration (see chapters 27 and 28). Chapter 29 provides examples of how some sectors and associated line ministries can use service

delivery assessments as diagnostic tools, particularly when combined with other forms of data on public administration, as discussed in chapter 8.

Having distinct analytics teams spread across a public service carries its own risk: fragmentation in analytical approaches, limiting comparability. Such a risk can be mitigated by building servicewide management information systems (MIS) to harmonize and aggregate measurements across government, and by embedding local analytics teams in governmentwide communities of practice. As the case studies in chapter 9 focusing on human resources management information systems (HRMIS) show, integrating different data sources can enhance analytical impacts. Chapter 9 describes the stages in developing harmonized management information systems focused on public administration, and outlines key decision points and trade-offs involved in building public sector data architectures. It warns against constraints in existing legislative institutional environments impeding experimentation with integrating data sources. Such experimentation enables the costs and benefits of integration to be clearly identified, providing important inputs into any scale-up decision. Introducing legislation to allow for small-scale experimentation in measurement and data integration can generate precise inputs and demonstration effects to inform discussions about how to push forward a government's analytical agenda.

Even within existing legislative and institutional constraints, a range of actions can be taken to strengthen analytical integration. A basic activity is the recording of common metadata for all data and existing integration in a centralized repository. This can promote clear communication across government and facilitate public scrutiny. Another action is to monitor what analytics are being used and by whom (see chapter 7). By taking this step, analysts can turn the lens on themselves, and assess how well the architecture of analytics they have developed is holding up, and whether analytics are being used purposefully rather than abused.

## **BUILD AN ANALYTICS AGENDA OF THE FUTURE**

### **Develop Analytical Agendas Everywhere**

Every institution comes with its own staff, tasks, and culture. Thus, the specific analytical requirements of any unit, organization, or public service will vary over a particular task, space, and time. At the same time, for each activity and employee, the questions of what success looks like and how to measure it remain relevant. As such, an agenda for government analytics can be defined at a very granular “micro” level. Every member of the public service can apply an analytical lens of measurement and analysis to their activities, and as such can define an analytical agenda for themselves.

Yet what success looks like and how it can be measured are not central concerns in many government agencies. A first step in resolving this is for actors across government—from senior managers, through unit supervisors, to individual employees—to articulate their commitment to using government analytics where beneficial in their work and to define associated analytics agendas for their areas of responsibility. To institutionalize this approach, performance appraisals could include a compulsory component on the curation of an analytics agenda for an official's areas of responsibility. Organizations could be required to publish an annual update on their analytical agenda. And government could have a central strategy for adopting or expanding government analytics (chapter 26).

None of the discussions in the *Handbook*, or in this chapter, insist that everything in government that can be measured should be measured. Measurement and analysis are costly, and have opportunity costs. Part of an analytics agenda should be to identify an optimal level of analytical investment. Such a level will be iterative, to be updated by the results from previous rounds of analytical investments.

The investment in analytics has potentially high returns. Chapters 12 and 15 show how administrative case data can be used to identify public agencies or individuals who overpay to procure goods for government or take considerably longer than their peers to complete tasks, for instance. Supporting these agencies



or individuals to harmonize their practices with those closer to the average can yield substantial cost savings. Although not all analytics will yield such a return on investment, searching for and prioritizing those that do is a shrewd financial investment. In this vein, evidence from the private sector suggests that data analytics can drive organizational productivity and profitability.

Although the discussion in the *Handbook* has been framed in terms of central government, much of what is discussed applies to any public administrative environment, including subnational public administrations. Subnational entities face some distinct management challenges, such as the management of jurisdictional boundaries. However, subnational analysis can capitalize on the fact that many subnational entities within a country have comparable units with the same functions that can vary considerably in performance and quality. Coordination of analytical agendas across subnational entities has analogous benefits to the centralized analytical units discussed. Institutions that can help coordinate government analytics across subnational entities as part of a community of practice will capitalize on those benefits.

A strength of anchoring analytics agendas in the public administration is that they have a greater chance of continuing beyond any single government administration. By making an analytics agenda a basic part of the administration of government, it can take a longer-term perspective than political agendas can (see chapter 18). This is important because the credibility of measurement and its use for strengthening public service matters for the quality of that measurement. If public servants do not believe that survey results will be used for action and as a management tool, response rates fade, for instance (chapter 26). By clearly signaling that analytics will only be molded but not undermined by political actors, it is likely to be of higher quality. Analytics can build credibility over time, and with stability gain a degree of familiarity.

Similarly, by embedding analytical agendas in public administration, many political leaders are more likely to accept the preexisting analytical architecture as a foundation for their own efforts to strengthen the administration. This may shift political actors toward evidence-based management of the public service.

## Build Comparable Measurement

Many elements of government functioning can be usefully measured across sectors and organizational units. For instance, many features of budget, payroll, human resources management, and process quality have commonalities across all of government (chapters 10 to 13). Thus, centralized analytical agendas should push for the harmonization and joint analysis of these features, and agencies should be open to supporting these public goods. Other measures—such as those related to budget utilization and task completion—are more challenging to compare across tasks, but are central to organizational decision-making in areas such as budget allocations (chapter 17). Thus, making explicit the assumptions of such comparisons, and then refining them, is better than skewing measurement toward the most measurable and comparable areas.

Similarly, individual governments should invest in measures that allow international comparisons, such as internationally standardized modules in public servant surveys. Within suitable comparison groups, such harmonization does not substitute for, but powerfully complements, internal measurement. Concerns regarding comparisons of officialdom across sectors or tasks within a single public service can be balanced against concerns regarding comparisons across countries. Having access to measurement from multiple ministries of health around the world will support a health minister's understanding of their own organization's particular strengths and weaknesses in a way that is complementary to their comparison to ministries of agriculture and education in their own countries (chapter 24). To this end, the Global Survey of Public Servants (Fukuyama et al. 2022) aims to increase the volume, quality, and coherence of survey data on public administration over time (see chapter 18). It presents both harmonized data from existing surveys and a suggested common set of questions to be included in surveys of public servants to aid comparability of data from any specific setting. It also recognizes the boundaries of such comparisons and provides access to data (chapter 24).

Comparisons across and within governments can also be made based on administrative data. When such comparisons are made, a frequent challenge is that different organizations in government complete different tasks. One approach the analyst can take is to focus on homogeneous units that do very similar work, such as procurement units across agencies. This is particularly useful for analysts focusing on a specific sector, such

as those described in chapters 14 and 15. As a more general principle, however, such an approach is liable to skew the analytics of government toward areas that are easier to measure (chapter 4). An analyst will gain a more comprehensive picture by defining a holistic agenda for understanding public administration and defining areas where comparability is useful. This relates back to the capacity for public servants to discriminate between when analytics rests on assumptions that fit their setting and when they do not.

With this in mind, when comparing organizations based on administrative data, the analyst should address three questions: (1) Is such a comparison being made implicitly somewhere in government, such as the ministry of finance (for example, when it compares task completion or budget execution rates across organizations)? (2) Can adjustments be made that will make the comparisons more valid (such as by measuring the complexity of the underlying task)? (3) Are there subgroups of comparators for which comparison is more reasonable? As these questions suggest, taking an analytical lens to the issue of comparability itself sometimes allows unspoken assumptions to be surfaced and discussed. Much comparison occurs in public administration without proper care being taken that the underlying issues surrounding comparability are understood and factored into decision-making based on the comparison.

## Use Experimentation Broadly

Faced with uncertainty or ambiguity, how should decision-makers proceed? As tech firms have realized, the optimal choice is to experiment with multiple sensible choices and measure which one works in what environments. For example, according to the company's "rigorous testing" blog, Google "ran over 700,000 experiments that resulted in more than 4,000 improvements to Search" in 2021.<sup>1</sup>

Experimentation in the field of government analytics allows the analyst to trial distinct approaches to measurement of public administration, or the design of government itself, and assess, through proper measurement, the advantages of each. The use of experimentation in the design of public administration is growing in policy circles, and a complementary academic literature is burgeoning in public administration, economics, political science, and beyond. Within this *Handbook*, chapters use experimentation to shed light on methodological questions, such as how the mode of a public servants survey affects responses (chapter 19) and how responses change when questions focus on organizational-level or individual-level referents (chapter 23).

The overlap in work programs across the public sector, and across public sector organizations around the world, presents an opportunity for the repeated testing of successful approaches from other settings, both in measurement and policy. This *Handbook* has illustrated key ideas in government analytics from specific governments. The lessons provide a starting point for undertaking methodological experiments (akin to the A-B testing of large technology firms) in the use of government analytics in a particular setting. In their specific analytics agenda, one question an analyst should aim to answer is: "Does what worked elsewhere work here?" In addition, there is a global benefit to repeated replication of any approach to measurement. Repeated testing of measurement approaches in different settings allows extensions of the findings in this *Handbook* and advances global knowledge toward "stylized facts" about what works where. This will also enhance the quality of conversation on how government functions, grounded in the empirical realities of the service, rather than only perceptions and tacit knowledge.

## PREPARE FOR NOVEL DATA SOURCES

### Search and Testing

The speed at which analytics is evolving requires a constant perspective on novel data sources. A new way of measuring and assessing government functioning could appear at any time. Thus, governments should set



themselves up to capitalize on novel data sources. This requires an approach to analytics that experiments with new approaches to measurement and analysis without the need for wholesale change. Analytics agendas should include an approach to searching for and testing innovations. Individual analysts can assist the search process by publicizing their experiments, by collaborating with others on testing, and by being open to the insights presented by others in the public and private sectors. Centralized analytics units are perhaps the most common way for an organization to engage with new approaches to government analytics.

Setting up an analytics agenda that has a component of search and testing requires a legislative environment that allows public officials a space for experimentation (chapters 9 and 26). Thus, how a government is built will affect its ability to experiment (chapter 16). Institutions that can take the responsibility for the failures that naturally come from testing innovations increase the incentives to experiment. Complementary cultural shifts that reward smart experimentation irrespective of the outcome often require support from senior leadership and political actors. Political actors who can articulate the case for experimentation to their peers and the public buy senior administrative officials space to improve the quality of government administration.

### The Limits of the *Handbook*

There are areas that this *Handbook* has not covered where some governments are starting to make inroads in their analytical efforts. In the area of recruitment, sentiment analysis toward public sector jobs and a wide range of recruitment analytics—for instance, on the diversity of the application pool or the extent of competition for different public sector jobs—can be drawn on by government to improve its quality of personnel. Analysis of communications between government employees, enabled by off-the-shelf solutions from large technology firms, is being experimented with; properly managed and with due care for employee privacy, it promises an understanding of how organizational structure affects team dynamics. Connecting tax records with procurement and customs data can enable an understanding of how government procurement policy affects private businesses and international trade. Machine-learning approaches to images can allow governments to automatically cross-check progress records in infrastructure projects with photos of those infrastructure projects to detect anomalies. And so on.

The *Handbook* has limited the data sources it presents to those of widest use to the largest number of public service organizations. All such entities must deal with payroll and budget, processes, and measures of task completion. Yet this focus on the most standard data sources in government has meant that the *Handbook* has not included some innovative approaches to assessing government functioning.

For example, substantial efforts have been made in geospatial analysis of the impacts of public policy, but there is little evidence that this has been applied to the public administration of the state beyond simple geographic comparisons. Matching personnel with geolocated project data will allow analytics to shed light on whether managers are better or worse at managing projects closer to their homes, or whether there are strong links between characteristics of local labor markets and the quality of recruitment into the public administration in that area. As the world shifts further toward remote work, the utility of tracking exactly where a public official is working and how this affects their productivity may allow for more sophisticated telework policies.

The potential for applying machine learning to text analysis of the vast quantities of documents produced by the public sector is in its infancy (chapter 16). Given that much public service communication is now online, such text analysis and machine learning might be applied to the communications of public officials in real time, and provide automated interventions when there is evidence of a personnel, management, or public policy issue arising.

As governments become more capable of integrating their electronic data systems, the capacity to build maps of the networks of government officials and firms will increase, and it will be easier to assess how personnel who move across different tasks (such as from managing procurement to budget) prosper in different environments and with different colleagues. Overall, gaining a greater sense of what the informal coalitions

in public administration are that facilitate strengthening of government may require triangulation between different data sources.

All these examples underscore the point that a comprehensive analytical agenda is forward-looking, capitalizing on what is available today and readying itself for what might be useful tomorrow.

### The Continuing Validity of the *Handbook*

A number of the foundational themes highlighted in the *Handbook* will continue to be of relevance to any innovations in the field. These include a robust discussion of the ethical implications of government analytics, the boundaries of measurement, and the rigor of analysis.

In terms of ethical issues, the use of data by governments on their own employees has received very little attention, as chapter 6 notes. Although checks and balances exist in public service, these will not always be well-equipped to deal with the pivot to government analytics. Where governments have begun to undertake government analytics, efforts have often not been complemented by a corresponding discussion of the ethical issues involved. For instance, it is important to have robust, servicewide debates about questions such as the extent to which analytics on public officials' remote work communications be undertaken at the level of anonymized individual email or message exchanges, and the ways in which this influences officials' behavior and the capacity to have wide-ranging and honest discussions about public policy.

It is key that such debates are undertaken both sectorwide and within specific organizations because what is considered as ethical and morally right can be very dependent on context (chapter 6). For example, what obligations of transparency around individual activities come with seniority in public service, and how much should officials be actively involved in this debate as they rise up the ranks? Chapter 6 presents a framework for evaluating the ethics of measuring and tracking public sector workers that will continue to be useful to evaluate the impact of innovations in measurement and analysis.

Similarly, the framework presented in chapter 4 will facilitate discussions around the relationship new measurements have to a holistic investigation of the environment being examined. Every new measurement or piece of analysis should come with a "health warning" regarding the boundaries of what it measures, and what it is likely missing. The principles outlined in chapter 5 serve as benchmarks by which new methods can be assessed for their credibility and transparency. Chapter 7 reminds us to turn the analytical lens on analytics themselves and continuously monitor what and how analytics are being (mis)used. And the principles of holistic measurement illustrated in chapter 14 push us to question the extent to which we have "triangulated" any specific measurement with others as a means of capturing distinct dimensions of a variable.

The insights offered in the *Handbook* can strengthen some innovations in measurement and analytics. Better measures of budget or task completion will still rely on the principles outlined in chapters 11 and 17. Innovations focused on improving data quality, availability, regularity, and connectedness will all need to implement the basics outlined in this *Handbook*. Chapters 10, 11, and 12 explicitly discuss layers of data quality that innovations in local settings will help achieve. Similarly, some innovations will build infrastructures that enable more regular, secure, and timely data collection (chapter 9).

## GOVERNMENT ANALYTICS IN AN INCREASINGLY COMPLEX WORLD

As measurement, data, and analysis become the central mediators of decision-making, government must build its capacity to engage with greater complexity in the world and in its own architecture. The question is whether public organizations will reform themselves sufficiently fast so that they can keep up. A solid machinery for government analytics can help empower government organizations to do so.

This chapter lays out the key components of a strategic review process for government actors to think through how they are building a government analytics system that responds not only to today's demands,

but also those of the future. Such thinking is useful at every level of government, from a project manager assessing how they are using administrative diagnostics in their project to the most senior management of the public service thinking through how they might optimally manage service staff.

The lessons presented in this chapter are drawn from across the chapters of the *Handbook*.

The *Handbook's* inability to cover all potential sources of government analytics mirrors the fact that governments will have to prioritize their investments in measurement and analysis. To make those choices strategically, a governmentwide vision of the future, linked to diverse analytical agendas of officials across government, will define the objectives of analytics. Managers who are aware of the trade-offs involved, and supported by specialized offices, will balance investments in basic measurement and the testing of innovations.

As the world gets more complex, the demands on public managers and decision-makers will increase as they manage a more complex government in response. Making the public administration fit-for-purpose will require an informed conversation throughout public service that drives the requisite cultural change. This *Handbook* hopes to inform that conversation. Important public sector conversations regarding reform may occur in a department of local government, a ministry of civil service, or even span countries and the international community. It is therefore important for all government actors to make an informed choice today about how they are setting up a system of analytics that will define what they will know tomorrow.

## HOW TO USE THE *HANDBOOK*

The chapters in the *Handbook* aim to be freestanding overviews of a particular topic in government analytics and can be read independently. The book is accompanied by a website with annexes and tools for analytics that enable readers to immediately apply insights from the *Handbook* in their own work ([www.worldbank.org/governmentanalytics](http://www.worldbank.org/governmentanalytics)).

To make the best use of the *Handbook*, readers are encouraged to choose the chapters that provide guidance on the data sources most relevant to the management challenges they are facing. For instance, if fiscal sustainability is the core challenge, consider focusing on chapters related to data sources that can yield solutions, such as chapter 10 on the payroll and chapter 11 on budget data. Table 2A.1 at the end of chapter 2 provides a tool to map areas of interest and data sources to the content of the chapters.

The *Handbook* aims at three main external audiences: government analytics practitioners (in governments, international development organizations, and elsewhere); educators; and researchers.

## GOVERNMENT ANALYTICS PRACTITIONERS

The *Handbook* has been designed to make use of the most widespread sources of data on public administration and to address some of the most pressing problems in managing government. As such, our hope is that government analytics practitioners will be able to find inspiration and useful advice in each of the chapters. We also hope that they will see the connections between their immediate interest and other data sources that might enrich the analysis they originally envisaged.

For readers interested in building the analytical capabilities of their organization, this chapter provides a vision of how government might move itself toward being more able to undertake analytics. Chapter 9 describes how to generate an integrated management information system for government. Chapter 26 provides a case study of the US government that presents the complementary management infrastructure that catalyzes any physical data system to become a platform for action.

For readers interested in making the most of their analytics, consider chapter 7 on how to measure whether government analytics are being used and chapter 25 on how to use results from surveys of public servants to strengthen public administration.

For those interested in how different data sources fit together, consider chapter 4 on holistic measurement, and chapter 8, showcasing how analytics can be combined to understand corruption holistically.

Readers looking for practical general statistics tools should go to chapter 5.

For those seeking guidance to think through the underlying ethical considerations of any government analytics effort, turn to chapter 6.

## EDUCATORS

Instructors in a school of public administration or public service training center, or in an academic institution, for instance, should pick and choose areas of particular interest and adapt lessons to the time available.

A single session could provide a useful overview of government analytics. Beginning with the motivation for government analytics (chapter 1), the class could then review a summary of approaches available outlined in chapter 2, and then focus on one particular data source of interest to the use (such as how to use procurement analytics).

A potential module on government analytics could proceed as follows. After an introductory session discussing chapters 1 and 2, consider a class summarizing chapters 4 to 6, to give students a sense of foundational considerations in government analytics. Students could be asked to consider the right ways to apply and manage statistical tools, the ethical considerations particular to studying public administration, and ways to measure holistically in public administration. Perhaps, students could design their own analytics study of public administration that has a pre-analysis and ethics plan that accords to the messages in these chapters.

The third session could focus on chapters 18 and 27, to give students a sense of comparative public administration around the world, and how to diagnose them. The discussion of these chapters could act as an introduction to what data sources are available for government analytics.

Chapter 27 introduces methods for using household surveys to understand public administration, which are the foundations of the World Bank's Worldwide Bureaucracy Indicators. Using the indicators in conjunction with the reading in chapter 27 allows students to understand the global footprint of the public administration, and its relationship to the private sector.<sup>2</sup>

Similarly, chapter 18 outlines the surveys of public administrators undertaken on a regular basis around the world. This chapter complements the data provided by the Global Survey of Public Servants initiative so as to provide the most comprehensive window into the public administration available to date based on surveys of public servants (Fukuyama et al. 2022).

For those students interested in undertaking their own surveys of public officials, the methodological lessons in chapters 19 to 25 provide useful inputs to their design process. These methodological considerations could be covered in a further teaching session on how to do surveys of public servants.

In subsequent sessions, instructors could cover different data sources introduced in parts 3 and 5, focused on the data sources of greatest interest to students. For instance, sessions could cover how to use payroll data, procurement data, and citizen survey data. These sessions should make use of publicly available data sources for students to practice analyzing these data sources.<sup>3</sup>

A teaching module could conclude with a discussion of how to build the analytical capability for government analytics (chapter 3), and how to integrate different analytics sources to assess management challenges holistically (chapter 8).

## RESEARCHERS

Overall, the *Handbook* discusses how to repurpose or construct a range of data sources that are rarely used by scholars, yet provide a fascinating window into public administration and government productivity. For many of the data sources discussed, the *Handbook* is the first consolidated attempt at discussing appropriate measurement. It is one of the goals of the *Handbook* to encourage researchers to expand and improve on the measurement of data sources for government analytics through their work. These researchers—in the fields of public administration, economics, management, political science, or elsewhere—may be in traditional research centers, or from inside government itself, perhaps in an analytics unit focused on improving their part of the public service.

A key early consideration of any research project is what the ethical framework is in which research questions and designs are produced. Chapter 6 provides a useful lens for a researcher to evaluate the ethical implications of their research approach.

Given the weight placed on the rigor and reproducibility of any data analysis, chapter 5 provides a reminder of the principles of good data analysis, and links to a set of resources to make those good practices straightforward to apply. Similarly, given the importance of understanding the limits of interpretation of any single data source or study, chapter 4 provides important reminders as to the validity of any single empirical study or approach.

Part 3 on administrative data can help researchers gain insights into how to construct a broader range of data to better understand the state. Some data sources have featured centrally in recent scholarly work, such as procurement data (chapter 12).<sup>4</sup> Other data sources explored in the *Handbook*—such as payroll data (chapter 10), task completion data (chapter 17), or process data (chapter 13)—have been seldom studied.<sup>5</sup>

Part 4 on survey data presents a range of methodological work related to investigations by the editors and others into how to undertake public servant surveys. Although, as outlined in chapter 18, surveys play an increasingly important part in managing the public sector in a number of countries, rigorous research on how to navigate the decision points that arise in designing, implementing, and interpreting surveys of public servants is limited. Chapter 2 presents a decision tree (figure 2.4) that might be useful to arrange thoughts on factors to be addressed in the survey approach chosen.

Research on the public service is not contingent on having access to proprietary government data. Though some public institutions are making their administrative data publicly available in one form or another, this is the exception rather than the rule. Part 5 presents four approaches that researchers have undertaken to understand features of the public administration using assessments that can be undertaken “external” to the public administration. Each of these data sources can be analyzed by researchers independent of government partnership.

We hope future research on public administration, whether in the fields of public administration, economics, management, political science, or elsewhere, will further capitalize on the data sources outlined in the *Handbook*. With the intention of the *Handbook* evolving in response to new methodological insights in government analytics, we look forward to reading your work or hearing from you.

## NOTES

In formulating this chapter, the authors benefited from discussions with Donna Andrews, Pierre Bachas, Jurgen Blum, Gero Carletto, Verena Fritz, Galileu Kim, Florence Kondylis, Tracey Lane, Arianna Legovini, and Daniel Ortega.

1. See <https://www.google.com/search/howsearchworks/how-search-works/rigorous-testing/>.
2. All of the code associated with chapter 27 is available online. Thus, students can extend the methods presented in the chapter to a country and corresponding household survey of their choice. Such an extension provides an opportunity to work directly with household survey data and learn about what underlies the comparisons made in the indicators, as well as get to study a particular public administration in detail.

3. Payroll data, for instance, are made public by governments such as Brazil (<https://portal.datatransparencia.gov.br/servidores/orgao?>). Similarly, citizen survey data are available on topics such as satisfaction with public services (see, for example, <https://www.gu.se/en/quality-government/qog-data/data-downloads/european-quality-of-government-index>).
4. See, for example, Bandiera et al. (2020); Dahlström, Fazekas, and Lewis (2021).
5. There are exceptions. See, for example, Rasul, Rogger, and Williams (2021).

## REFERENCES

- Bandiera, Oriana, Michael Carols Best, Adnan Qadir Khan, and Andrea Prat. 2020. "The Allocation of Authority in Organizations: A Field Experiment with Bureaucrats." NBER Working Paper 26733, National Bureau of Economic Research, Cambridge, MA.
- Dahlström, Carl, Mihály Fazekas, and David E. Lewis. 2021. "Partisan Procurement: Contracting with the United States Federal Government, 2003–2015." *American Journal of Political Science* 65 (3): 652–69.
- Fukuyama, Francis, Daniel Rogger, Zahid Husnain, Katherine Bersch, Dinsha Mistree, Christian Schuster, Kim Sass Mikkelsen, Kerenssa Kay, and Jan-Hinrik Meyer-Sahling. 2022. Global Survey of Public Servants. <https://www.globalsurveyofpublicservants.org/>.
- Google. No date. "Improving Search with Rigorous Testing." <https://www.google.com/search/howsearchworks/how-search-works/rigorous-testing/>.
- Rasul, Imran, Daniel Rogger, and Martin J. Williams. 2021. "Management, Organizational Performance, and Task Clarity: Evidence from Ghana's Civil Service." *Journal of Public Administration Research and Theory* 31 (2, April): 259–77.
- University of Gothenburg. No date. European Quality of Government Index. <https://www.gu.se/en/quality-government/qog-data/data-downloads/european-quality-of-government-index>.

The background features a series of wavy, horizontal lines in a light red color, composed of small squares. Overlaid on these are various strings of binary code (0s and 1s) in a light red color, scattered across the upper half of the page.

## **PART 2**

# Foundational Themes in Government Analytics





## CHAPTER 4

# Measuring What Matters

## Principles for a Balanced Data Suite That Prioritizes Problem Solving and Learning

*Kate Bridges and Michael Woolcock*

### SUMMARY

Responding effectively and with professional integrity to public administration's many challenges requires recognizing that access to more and better quantitative data is necessary but insufficient. An overreliance on quantitative data comes with risks, of which public sector managers should be keenly aware. We focus on four such risks: first, that attaining easy-to-measure targets becomes a false standard of broader success; second, that measurement becomes conflated with what management is and does; third, that an emphasis on data inhibits a deeper understanding of the key policy problems and their constituent parts; and fourth, that political pressure to manipulate key indicators, if undetected, leads to falsification and unwarranted impact claims or, if exposed, jeopardizes the perceived integrity of many related (and otherwise worthy) measurement efforts. The cumulative concern is that these risks, if unattended to, will inhibit rather than promote public sector organizations' core problem-solving and implementation capabilities, an issue of high importance everywhere but especially in developing countries. We offer four cross-cutting principles for building an approach to the use of quantitative data—a “balanced data suite”—that strengthens problem solving and learning in public administration: (1) identify and manage the organizational capacity and power relations that shape data management; (2) focus quantitative measures of success on those aspects that are close to the problem; (3) embrace a role for qualitative data and a theory of change, especially for those aspects which require in-depth, context-specific knowledge; and (4) protect space for judgment, discretion, and deliberation because not everything that matters can be measured.

---

Kate Bridges is an independent consultant. Michael Woolcock is the lead social scientist in the World Bank's Development Research Department.

## ANALYTICS IN PRACTICE

- Identify and manage the organizational capacity and power relations that shape data management. Make professional principles and standards for collecting, curating, analyzing, and interpreting data clear to all staff—from external consultants to senior managers—in order to affirm and enforce commitments to ensuring the integrity of the data themselves and the conclusions drawn from them. Make measurement accountable to advisory boards with relevant external members. Communicate measurement results to the public in a clear and compelling way, especially on contentious, complex issues.
- Focus quantitative measures of success on those aspects that are close to the problem. Ensure that the measurement approach itself is anchored to a specific performance problem. Target measurement investments at those performance problems that are prioritized by the administration. Ensure that any judgments on an intervention's success or failure are based on credible measures of the problem being fixed and not simply on output or process metrics. Where measures of success relate to whether the intervention is functioning, allow flexibility in the implementation of the intervention (where possible) and in the related measurement of its functioning. In this way, implementation strategies can shift if it becomes clear from the collected data that they are not making progress toward fixing the problem.
- Embrace an active role for qualitative data and a theory of change. Include qualitative data collection as a complement to quantitative data. This may be a prelude to future large-scale quantitative instruments or perhaps the only available data option for some aspects of public administration in some settings (such as those experiencing sustained violence or natural disasters). Draw on qualitative methods as a basis for eliciting novel or “unobserved” factors driving variation in outcomes. Tie measurement (both qualitative and quantitative) back to a theory of change. If the implementation of an intervention is not having its intended impact on the problem, assess whether there are mistaken assumptions regarding the theory of change.
- Protect space for judgment, discretion, and deliberation because not everything that matters can be measured. Consider carefully what you choose to measure, recognizing that whatever you choose will inevitably create incentives to neglect processes and outcomes that cannot be measured. Actively identify what you cannot (readily) measure that matters and take it seriously, developing strategies to manage that as well. Identify those aspects of implementation in the public sector that require inherently discretionary decisions. Employ strategies that value reasoned judgment and allow meaningful space for qualitative data inputs and the practical experience of embedded individuals, treating such inputs as having value alongside more quantitative ones.
- In the longer term, develop organizational systems that foster “navigation by judgment”—for example, a management structure that delegates high levels of discretion to allow those on the ground the space to navigate complex situations, recruitment strategies that foster high numbers of staff with extensive context-specific knowledge, and systems of monitoring and learning that encourage the routine evaluation of theory against practice.

## INTRODUCTION

“What gets measured gets managed, and what gets measured gets done” is one of those ubiquitous (even clichéd) management phrases that hardly require explanation; it seems immediately obvious that the data generated by regular measurement and monitoring make possible the improvement of results. Less well known than the phrase itself is the fact that, although it is commonly attributed to the acclaimed management

theorist Peter Drucker, Drucker himself never actually said it (Zak 2013). In fact, Drucker's views on the subject were reportedly far more nuanced, along the lines of those of V. F. Ridgway, who argued over 65 years ago that not everything that matters can be measured and that not everything that can be measured matters (Ridgway 1956). Simon Caulkin (2008), a contemporary business management columnist, neatly summarizes Ridgway's argument, in the process expanding the truncated "to measure is to manage" phrase to "What gets measured gets managed—even when it's pointless to measure and manage it, and even if it harms the purpose of the organisation to do so."

Ridgway's and Caulkin's warnings—repeated in various guises by many since—remind us that the indiscriminate use of quantitative measures and undue confidence in what they can tell us may be highly problematic in certain situations, sometimes derailing the very performance improvements that data are intended to support (Merry, Davis, and Kingsbury 2015).<sup>1</sup> We hasten to add, of course, that seeking more and better quantitative data is a worthy aim in public administration (and elsewhere). Many important gains in human welfare (for example, recognizing and responding to learning disabilities) can be directly attributed to interventions conceived of and prioritized on the basis of empirical documentation of the reality, scale, and consequences of the underlying problem. The wonders of modern insurance are possible because actuaries can quantify all manner of risks over time, space, and groups. What we will argue in the following sections, however, is that access to quantitative data alone is not a sufficient condition for achieving many of the objectives that are central to public administration and economic development.

This chapter has five sections. Following this introduction, we lay out in section two how the collection, curation, analysis, and interpretation of data are embedded in contexts: no aspect takes place on a blank slate. On one hand, the institutional embeddedness of the data collection and usage cycle—in rich and poor countries alike—leaves subsequent delivery efforts susceptible to a host of possible compromises, stemming from an organization's inability to manage and deploy data in a consistently professional manner. At the same time, the task's inherent political and social embeddedness ensures it will be susceptible to influence by existing power dynamics and the normative expectations of those leading and conducting the work, especially when the political and financial stakes are high. In contexts where much of everyday life transpires in the informal sector—rendering it "illegible" to, or enabling it to actively avoid engagement with, most standard measurement tools deployed by public administrators—sole reliance on formal quantitative measures will inherently only capture a slice of the full picture.

In section 3, we highlight four specific ways in which an indiscriminate increase in the collection of what is thought to be "good data" can lead to unintended and unwanted (potentially even harmful) consequences. The risks are that (1) the easy-to-measure can become a misleading or false measure of broader reality, (2) measurement can become conflated with what management is and does, (3) an emphasis on what is readily quantified can inhibit a fuller and more accurate understanding of the underlying policy problem(s) and their constituent elements, and (4) political pressure to manipulate selected indicators, if undetected, can lead to falsification and unwarranted expectations—or, if exposed, can compromise the perceived integrity of otherwise worthy measurement endeavors.

Thankfully, there are ways to anticipate and mitigate these risks and their unintended consequences. Having flagged how unwanted outcomes can emerge, we proceed to highlight, in section 4, some practical ways in which public administrators might thoughtfully anticipate, identify, and guard against them. We discuss what a balanced suite of data tools might look like in public administration and suggest four principles that can help us apply these tools to the greatest effect, thereby enabling the important larger purposes of data to be served. For further methodological guidance, practitioners should consult appendix A, which provides a checklist titled "Using Expansive and Qualified Measurement for Informed Problem Solving and Learning in Public Administration." We stress from the outset that our concerns are not with methodological issues per se, or with the quality or comprehensiveness of quantitative data; these concerns are addressed elsewhere in *The Government Analytics Handbook* and in every econometrics textbook, and they should always be considered as part of doing "normal social science." The concerns we articulate are salient even in a best-case scenario, in which analysts have access to great data acquired from a robust methodology, although they are obviously compounded when the available data are of poor quality—as is often the case, especially in low-income countries—and when too much is asked of them.

## HOW DATA ARE IMPACTED BY THE INSTITUTIONAL AND SOCIOPOLITICAL ENVIRONMENT IN WHICH THEY ARE COLLECTED

For all administrative tasks, but especially those entailing high-stakes decision-making, the collection and use of data is a human process inherently subject to human foibles (Porter 1995). This is widely accepted and understood: for example, key conceptual constructs in development (such as “exclusion,” “household,” and “fairness”) can mean different things to different people and translate awkwardly into different languages. With this in mind, professional data collectors will always give serious attention to “construct validity” concerns to ensure there is close alignment between the questions they ask and the questions their informants hear.<sup>2</sup> For present purposes, we draw attention to issues given less attention, but which are critical nonetheless—namely, the institutional and political factors that comprise the context shaping which data are (and are not) collected, how and from whom they are collected, how well they are curated over time, and how carefully conclusions and policy implications are drawn from analyses of them. We briefly address each item in turn.

## INSTITUTIONAL EMBEDDEDNESS OF DATA

Beyond the purposes to which they are put, the careful collection, curation, analysis, and interpretation of public data are themselves complex technical and administrative tasks, requiring broad, deep, and sustained levels of organizational capability. In this section, we briefly explore three institutional considerations shaping these factors: the dynamics shaping the (limited) “supply” and refinement of technical skills, the forging of a professional culture that is a credible mediator of complex (and potentially heated) policy issues yet sufficiently robust to political pressure, and the related capacity to infer what even the best data analysis “means” for policy, practice, and problem solving.

These issues apply in every country but are especially salient in low-income countries, where the prevailing level of implementation capability in the public sector is likely to be low, and where the corresponding expectations of those seeking to improve it by expanding the collection and use of quantitative data may be high. At the individual level, staff with the requisite quantitative analytical skills are likely to be in short supply because acquiring such skills requires considerable training, while those who do have them are likely to be offered much higher pay in the private sector. (One could in principle outsource some data collection and analysis tasks to external consultants, but doing so would be enormously expensive and potentially compromise the integrity and privacy of unique public data.)

So understood, it would be unreasonable to expect the performance of data-centric public agencies to be superior to other service delivery agencies in the same context (for example, public health). Numerous studies suggest the prevailing levels of implementation capability in many (if not most) low-income countries are far from stellar (Andrews, Pritchett, and Woolcock 2017).<sup>3</sup> For example, Jerven’s (2013) important work in Africa on the numerous challenges associated with maintaining the System of National Accounts—the longest-standing economic data collection task asked of all countries, from which their respective gross domestic products (GDPs) are determined—portends the difficulties facing less high-profile metrics (see also Sandefur and Glassman 2015).<sup>4</sup> Put differently: if many developing countries struggle to curate the single longest-standing, universally endorsed, most important measure asked of them, on what basis do we expect these countries to manage lesser, lower-stakes measures?

To be sure, building quantitative analytical skills in public agencies is highly desirable; for present purposes, our initial point is a slight variation on the old adage that the quality of outcomes derived from quantitative data is only as good as the quality of the “raw material” and the competence with which it is analyzed and interpreted.<sup>5</sup> Fulfilling an otherwise noble ambition to build a professional public sector whose decisions are informed by evidence requires a prior and companion effort to build the requisite

skills and sensibilities. Put differently, precisely because effective data management is itself such a complex and difficult task, in contexts where agencies struggle to implement even basic policy measures at a satisfactory level (for example, delivering mail and ensuring attendance at work), it is unlikely that, *ceteris paribus*, asking these agencies to also take a more “data-driven” approach will elicit substantive improvement. More and better “data” will not fix a problem if the absence of data is not itself the key problem or the “binding constraint”; the priority issue is discerning what *is* the key policy problem and its constituent elements. From this starting point, more and better data can be part of, but not a substitute for, strategies for enhancing the effectiveness of public sector agencies.

Even if both data management and broad institutional capability are functioning at high and complementary levels, there remains the structural necessity of interpreting what the data *mean*. Policy inference from even the best data and most rigorous methodology is never self-evident; it must always be undertaken in light of theory. This might sound like an abstract academic concern, but it is especially important when seeking to draw lessons from, or to make big decisions regarding the fate of, complex interventions. This is so because a defining characteristic of a complex problem is that it generates highly variable outcomes across time, space, and groups.

Promoting gender equality, for example, is a task that rarely generates rapid change: it can take a generation (or several, or centuries) for rules requiring equal participation in community meetings, or equal pay for equal work, to become the “new normal.”<sup>6</sup> So, assessed over a five-year time frame, a “rigorous” methodology and detailed data may yield the empirical finding that a given gender empowerment project (GEP) has had “no impact”; taken at face value, this is precisely what “the data” would show and is the type of policy conclusion (“the GEP doesn’t work”) that would be drawn. However, interpreted in the light of a general theory of change incorporating the likely impact trajectory that GEP-type interventions follow—that is, a long period of stasis eventually leading to a gradual but sustained takeoff—a “doesn’t work” conclusion would be unwarranted; five years is simply too soon to draw a firm conclusion (Woolcock 2018).<sup>7</sup> High-quality data and a sound methodology alone cannot solve this problem: a GEP may well be fabulous, neutral, useless, or a mixture of all three, but discerning which of these it is—and why, where, and for whom it functions in the way it does—will require the incorporation of different kinds of data into a close dialogue with a practical theory of change fitted for the sector, the context, and the development problem being addressed.

## SOCIOPOLITICAL EMBEDDEDNESS OF DATA

Beyond these institutional concerns, a second important form of embeddedness shaping data collection, curation, and interpretation is the manner in which all three are shaped by sociopolitical processes and imperatives. All data are compiled for a purpose; in public administration, the scale and sophistication of the required data are costly and complex (requiring significant financial outlay and, thus, competition with rival claimants). Data are frequently called upon to adjudicate both the merits of policy proposals *ex ante* (for example, the Congressional Budget Office in the United States) and the effectiveness of programmatic achievements *ex post* (for example, the World Bank’s Independent Evaluation Group), which frequently entails entering into high-stakes political gambits—for example, achieving signature campaign proposals in the early days of an administration and proclaiming their subsequent widespread success (or failure) as election time beckons again. (See more on this below.)

Beyond the intense political pressure “data” are asked to bear in such situations, a broader institutional consideration is the role large-scale numerical information plays in “rendering legible” (Scott 1998) complex and inherently heterogeneous realities, such that they can be managed, mapped, and manipulated for explicit policy purposes. We hasten to add that such “thin simplifications” (Scott’s term) of reality can be both benign and widely beneficial: comprehensive health insurance programs and pension systems have largely tamed the otherwise debilitating historical risks of, respectively, disease and old age by generating premiums based on

general demographic characteristics and the likelihood of experiencing different kinds of risks (for example, injuries or cancer) over the course of one's life.

A less happy aspect of apprehending deep contextual variation via simplified (often categorical) data, however, is the corresponding shift it can generate in the political status and salience of social groups. The deployment of the census in colonial India, for example, is one graphic demonstration of how the very act of “counting” certain social characteristics—such as the incidence of caste, ethnicity, and religion—can end up changing these characteristics themselves, rendering what had heretofore been relatively fluid and continuous categories as fixed and discrete. In the case of India, this massive exercise in data collection on identity led to “caste” being created, targeted, and mobilized as a politically salient characteristic that had (and continues to have) deep repercussions (for example, at independence, when Pakistan split from India, and more recently within the rise of Hindu nationalism; see Dirks 2011).<sup>8</sup> Similarly, influential scholars have argued that the infamous Hutu/Tutsi massacre in Rwanda was possible at the scale at which it was enacted because ethnic categories were formalized and fixed via public documents whose origins lie in colonial rule (for example, Mamdani 2002).

For Scott (1998), public administration can only function to the extent its measurement tools successfully turn widespread anthropological variation, such as in languages spoken, into singular modern categories and policy responses—for instance, to ensure that education is conducted in one national language, in a school, and on the basis of a single curriculum.<sup>9</sup> The net welfare gains to society might be unambiguous, but poorer, isolated, marginalized, and less numerous groups are likely to bear disproportionately the costs of this trade-off. If official “data” themselves constitute an alien or distrusted medium by which certain citizens are asked to discern the performance of public agencies, merely providing (or requiring) “more” is unlikely to bring about positive change. In such circumstances, much antecedent work may need to be undertaken to earn the trust of citizens and to help them more confidently engage with their administrative systems.<sup>10</sup> By way of reciprocity, it may also require such systems to interact with citizens themselves in ways that more readily comport with citizens' own everyday (but probably rather different) vernacular for apprehending the world and interpreting and responding to events. Either way, it is critical that officials be wary of the potentially negative or unintended effects of data collection, even when it may begin with a benign intention to facilitate social inclusion and more equitable policy “targeting.”<sup>11</sup>

## THE UNINTENDED CONSEQUENCES OF AN INDISCRIMINATE PURSUIT OF “MORE DATA”

There is a sense in which it is axiomatic that more and better data are always a good thing. But the institutional and sociopolitical embeddedness of data generation and the use of data in public administration (as discussed in the preceding section) means we need to qualify this assertion by focusing on where and how challenges arise. With this in mind, we turn our attention to instances where the increased collection of what is thought to be “good data” leads to perverse outcomes. Here, we highlight four such outcomes that may materialize as the result of an undue focus on issues, concepts, inputs, or outcomes that happen to be most amenable to being quantified.

### Outcome 1: The Easy-to-Measure Becomes a False Standard of Success

What may start as a well-intentioned managerial effort to better quantify meaningful success can instead generate a blinkered emphasis on that which is simply easiest to quantify. The result can be a skewed or false sense of what a project has (or has not) achieved, and how, where, and for whom outcomes have been achieved.

In a recent study, we demonstrate how a variety of institutional incentives align across the government of Malawi and the World Bank in such a way that both government and World Bank officials consistently favor



easy-to-measure indicators (inputs and outputs, or what we refer to as “changes in form rather than function”) as the yardstick of project success (Bridges and Woolcock 2017). This is a quintessential example of what strategy writer Igor Ansoff describes as a situation in which “managers start off trying to manage what they want, and finish up wanting what they can measure” (quoted in Cahill 2017, 152). As a result of evaluating public financial management (PFM) projects that were implemented over the course of 20 years in Malawi, we show that almost 70 percent of what projects measure or aim for is “change in terms of whether institutions look like their functioning counterparts (that is, have the requisite structures, policies, systems, and laws in place),” whereas only 30 percent of what is measured can be said to be “functional”—that is, focused on “purposeful changes to budget institutions aimed at improving their quality and outcomes” (Andrews 2013, 7). What is more, we find that World Bank PFM projects have considerably more success in achieving “formal” results than “functional” ones. Unsurprisingly, demonstrable improvement in actual performance is far harder to achieve than change that is primarily regulative, procedural, or systems oriented. Unfortunately, an emphasis on what is easy-to-measure obfuscates this reality and allows reform “success” to be claimed.

In practice, Malawi’s history of PFM reform is littered with projects that claim “success” based on hardware procured, software installed, legislation developed, and people trained, whereas even a basic analysis reveals stagnation or even regression in terms of more affordable spending decisions, spending that reflects budgeted promises, greater ability to track the flow of funds, or reduction in corruption. As long as the World Bank and the Malawian government focus on “formal” measures, they are able to maintain the illusion of success: that is, until something like Malawi’s 2013 “Cashgate” crisis—in which it was revealed that about US\$32 million in government funds had been misappropriated between April and September 2013—lifts the lid on the deep-rooted financial management problems that have remained largely unaffected by millions of dollars of reform efforts. In this sense, Malawi is a microcosm of many institutional reform efforts globally. Although similar financial reforms have been globally implemented in a manner that suggests some level of consensus about “what works,” the outcomes of those reforms are varied at best and often considerably lower than anticipated (Andrews 2013).

In the same way that an emphasis on the easy-to-measure can lead to overestimation of success, it can also contribute to underestimation. Reforms can sometimes yield meaningful change via what McDonnell (2017) calls “the animating spirit of daily practice” but end up being missed because managers do not have good means of measuring, attributing, and enhancing these *kinds* of shifts. For example, when researching the impact of technical assistance on a large government health program in Nigeria, we found that there were strong indications that important innovations and shifts took place at the local level, including in aspects as difficult to shift as cultural practices regarding contraceptives (Bridges and Woolcock 2019). These shifts in practice and their impact on contraceptive uptake could not be apprehended by aggregated statewide indicators, however, and since no measurement was being done below this level, the progress and valuable lessons of such interventions were being missed.

Another example of the importance of having access to a broader suite of data comes from an assessment of a program in rural India seeking to promote participatory democracy in poor communities, where the curation of such a data suite enabled more nuanced and constructive lessons to be drawn (see Rao, Ananthpur, and Malik 2017). The results of the initial randomized controlled trial (RCT) deemed the program to have had no mean impact—and if these were the only data available, that would have been the sole conclusion reached.<sup>12</sup> Upon closer inspection, however, it was learned that there was considerable variation in the program’s impact. The average of this variation may have been close to zero, but for certain groups, the program had worked quite well, for others it had had no impact, while for still others it had been detrimental. Who were these different groups, and what was it about them that led to such variable outcomes? A companion qualitative process evaluation was able to discern that the key differences were the quality of implementation received by different groups, the level of support provided to them by managers and political leaders, and variations in the nature and extent of local-level inequalities (which, in turn, shaped which groups were able to participate and on what terms).<sup>13</sup> The administrative rules and implementation guidelines provided to all groups were identical, but in this case, a qualitative process evaluation was able to document the ways

and places in which variable fidelity to them yielded widely different outcomes (albeit with no net impact). Moreover, the qualitative data were able to discern subtle positive effects from the program that the quantitative survey instrument alone would have missed.

## Outcome 2: Measurement Becomes Conflated with Management

An extension of the above point is that an undue emphasis on quantitative data can lead measurement to become a *substitute for* rather than a *complement to* management. This is evident when only that which is quantifiable receives any significant form of managerial attention, an outcome made possible when the easily quantifiable becomes the measure of success, becoming, in turn, the object of management's focus, typically to the exclusion of all else. As Wilson (1989, 161) famously intoned in a classic study of bureaucratic life, "Work that produces measurable outcomes tends to drive out work that produces immeasurable outcomes."

In one sense this is hardly surprising: the need for managers to make decisions on the basis of partial information is difficult and feels risky, so anything that claims to fill that gap and bypass the perceived uncertainty of subjective judgment will be readily welcomed. "The result," Simon Caulkin (2016) argues, "both practically and theoretically, is to turn today's management into a technology of control that attempts to minimise rather than capitalise on the pesky human element." And in public administration, a managing-it-by-measuring-it bias can mean that, over time, the bulk of organizational resources end up neglecting the "pesky human element" of change processes, even though it is this element that is often central to attaining the transformational outcomes managers are seeking.

This dynamic characterizes key aspects of the Saving One Million Lives (SOML) initiative, an ambitious health sector reform program launched by the government of Nigeria. The original goal of SOML was to save the lives of one million mothers and children by 2015; to this end, SOML gave priority to a package of health interventions known as "the six pillars."<sup>14</sup> The World Bank actively supported SOML, using its Program-for-Results (PforR) instrument to reward Nigerian states financially based on improvements from their previous best performance on the six key indicators.<sup>15</sup> Improvements were to be measured through yearly household surveys providing robust estimates at the state level.

In practice, of course, these six pillars (or intervention areas) were wildly different in their drivers and complexity; improvement within them was therefore destined to move at different trajectories and different speeds for different groups in different places. State actors, keen to raise their aggregate measure of success and get paid for it, soon realized that there was some gaming to be done. Our field research documents how the emphasis on singular measures of success introduced a perverse incentive for states to focus on the easier metrics at the expense of the harder ones (Bridges and Woolcock 2019). Interviews with state officials revealed that frontline staff increasingly focused their time and energies on those constituent variables that they discerned were easiest to accomplish (for example, dispensing vitamin supplements) over those that were harder or slower—typically those that involved a plethora of "pesky human elements," such as lowering maternal mortality or increasing contraceptive use. In selecting certain outcomes for measurement and managing these alone, others inevitably end up being sidelined.

Likewise, a recent report on results-based financing (RBF) in the education sector (Dom et al. 2020) finds evidence of a "diversion risk" associated with the signposting effect of certain reward indicators, with important areas deprioritized because of the RBF incentive. For example, in Mozambique, they find that an emphasis on simple process indicators and a focus on targets appears to have led officials to divert resources and attention away from "more fundamental and complex issues," such as power dynamics in the school council, the political appointment of school directors, and teachers' use of training. Dom et al. also report evidence of "cherry-picking risks," in which less costly or politically favored subgroups or regions see greater resources, in part because they are more likely to reach a target. For example, in Tanzania, they find evidence that the implementation of school rankings based on exam results was correlated with weaker students not sitting, presumably in an effort by the schools to raise average exam pass rates.

This tendency becomes a particular issue when the sidelined outcomes end up being the ones we care most about. Andrew Natsios (2011), the former administrator of the United States Agency for International

Development (USAID, an organization charged with “demonstrating the impact of every aid cent that Congress approves”), argues compellingly that the tendency in aid and development toward what he calls “obsessive measurement disorder” is a manifestation of a core dictum among field-based development practitioners—namely, “that those development programs that are most precisely and easily measured are the least transformational, and those programs that are most transformational are the least measurable.” The change we often desire most is in very difficult-to-measure aspects, such as people’s habits, cultural norms, leadership characteristics, and mindsets.

This reality is also aptly illustrated in many anticorruption efforts, where imported solutions have managed to change the easy-to-measure—new legislation approved, more cases brought, new financial systems installed, more training sessions held—but have failed to shift cultural norms regarding the unacceptability of whistleblowing or the social pressures for nepotism (Andrews 2013). Failure to measure and therefore manage these informal drivers of the problem ensures that any apparent reduction in fund abuses tends to be short-lived or illusory.

This phenomenon is hardly limited to poor countries. A more brutal example of how what cannot be measured does not get managed, with disastrous results, can be found in the United Kingdom’s National Health System (NHS). While investigating the effects of competition in the NHS, Propper, Burgess, and Gossage (2008) discovered that the introduction of interhospital competition improved waiting times while also substantially *increasing* the death rate following emergency heart attacks. The reason for this was that waiting times were being measured (and therefore managed), while emergency heart-attack deaths were not tracked and were thus neglected by management. The result was shorter waiting times but more deaths as a result of the choice of measure. The authors note that the issue here was not intent but the extent to which one target consumed managerial attention to the detriment of all else; as they note, it “seems unlikely that hospitals deliberately set out to decrease survival rates. What is more likely is that in response to competitive pressures on costs, hospitals cut services that affected [heart-attack] mortality rates, which were unobserved, in order to increase other activities which buyers could better observe” (Propper, Burgess, and Gossage 2008).

More recently, in October 2019, the Global Health Security Index sought to assess which countries were “most prepared” for a pandemic, using a model that gave the highest ranking to the United States and the United Kingdom, largely on the basis of these countries’ venerable medical expertise and technical infrastructure, factors which are readily measurable (McCarthy 2019). Alas, the model did not fare so well when an actual pandemic arrived soon thereafter: a subsequent analysis, published in *The Lancet* on the basis of pandemic data from 177 countries between January 2020 and September 2022, found that “pandemic-preparedness indices . . . were not meaningfully associated with standardised infection rates or IFRs [infection/fatality ratios]. Measures of trust in the government and interpersonal trust, as well as less government corruption, had larger, statistically significant associations with lower standardised infection rates” (Bollyky et al. 2022, 1).

Needless to say, variables such as “trust” and “government corruption” are hard to measure, are hard to incorporate into a single theory anticipating or informing a response to a pandemic, and map awkwardly onto any corresponding policy instrument. For present purposes, the inference we draw from these findings is not that global indexes have no place; rather, they suggest the need, from the outset, for curating a broad suite of data when anticipating and responding to complex policy challenges, the better to promote real-time learning. Doubling down on what can be readily measured limits the space for eliciting those “unobserved” (and perhaps unobservable) factors that may turn out to be deeply consequential.

### **Outcome 3: An Undue Emphasis on Data Inhibits Understanding of the Foundational Problem(s)**

An indiscriminate emphasis on aggregated, quantitative data can erode important nuances about the root causes of the problems we want to fix, thereby hampering our ability to craft appropriate solutions and

undermining the longer-term problem-solving capabilities of an organization. All too often, the designation of indicators and targets has the effect of causing people to become highly simplistic about the problems they are trying to address. In such circumstances, what should be organizational meetings held to promote learning and reflection on what is working and what is not instead become efforts in accounting and compliance (Honig and Pritchett 2019). Reporting, rather than learning, is incentivized, and management increasingly focuses on meeting target numbers rather than solving the problem. Our concern here is that, over time, this tendency progressively erodes an organization's problem-solving capabilities.

The education sector is perhaps the best illustration of this: time and again practitioners have sought to codify “learning,” and time and again this has resulted in an obfuscation of the actual causes underlying the problem. In a well-intentioned effort to raise academic performance, “the standards movement” in education promoted efforts hinged on quantitative measurement, as reported in the league tables of the Program for International Student Assessment (PISA).<sup>16</sup> PISA runs tests in mathematics, reading, and science every three years with groups of 15-year-olds in countries around the world. Testing on such a scale requires a level of simplicity and “standardization,” thus the emphasis is on written examinations and the extensive use of multiple-choice tests so that students’ answers can be easily codified and processed (Robinson and Aronica 2015). Demonstrating competence in fundamental learning tasks certainly has its place, but critics have increasingly argued that such tests are based on the incorrect assumption that what drives successful career and life outcomes is the kind of learning that is capable of being codified via a standardized test (Claxton and Lucas 2015; Khan 2012).

In reality, the gap between the skills that children learn and are tested for and the skills that they need to excel in the 21st century is becoming more obvious. The World Economic Forum noted in 2016 that the traditional learning captured by standardized tests falls short of equipping students with the knowledge they need to thrive.<sup>17</sup> Yong Zhao (2012), the presidential chair and director of the Institute for Global and Online Education in the College of Education at the University of Oregon, points out that there is an inverse relationship between those countries that excel in PISA tests and those that excel in aspects like entrepreneurship, for example (see figure 4.1).

While a focus on assessing learning is laudable—and a vast improvement over past practices (for example, in the Millennium Development Goals) of merely measuring attendance (World Bank 2018)—for present purposes the issue is that the drivers of learning outcomes are far more complex than a quantifiable content deficit in a set of subjects. This is increasingly the case in the 21st century, which has brought a need for new skills and mindsets that go well beyond the foundational numeracy and literacy skills required during the Industrial Revolution (Robinson and Aronica 2015). A survey of chief human resources and strategy officers by the World Economic Forum (2016) finds a significant shift between 2015 and 2020 in the top skills future workers will need, with “habits of mind” like critical thinking, creativity, emotional intelligence, and problem solving ranking well ahead of any specific content acquisition. None of this is to say that data do not have a role to play in measuring the success of an educational endeavor. Rather, the data task in this case needs to be informed by the complexity of the problem and the extent to which holistic learning resists easy quantification.<sup>18</sup>

Finally, relying exclusively on high-level aggregate data can result in presuming uniformity in underlying problems and thus lead to the promotion of simplistic and correspondingly generic solutions. McDonnell (2020) notes, for example, that because many developing countries have relatively high corruption scores, an unwelcome outcome has been that *all* the institutions in these countries tend to be regarded by would-be reformers as similarly corrupt and uniformly ineffectual. In her impressive research on “clusters of effectiveness,” however, she offers evidence of the variation in public-sector performance within states, noting how the aggregated data on “corruption” masks the fact that the difference in corruption scores between Ghana’s best- and worst-rated state agencies approximates the difference between Belgium (WGI = 1.50) and Mozambique (WGI = −0.396), in effect “spanning the chasm of so-called developed and developing worlds.” The tendency of reform actors to be guided by simplistic aggregate indicators—such as those that are used to determine a poor country’s “fragility” status and eligibility for International Development Association (IDA) funding—has prevented a more

**FIGURE 4.1 Country Scores on Program for International Student Assessment Tests and Perceived Entrepreneurial Capability**



Source: Based on Zhao 2012.

Note: PISA = Program for International Student Assessment.

detailed and context-specific understanding of the lessons that could be drawn from positive outlier cases, or what McDonnell refers to as “the thousand small revolutions quietly blooming in rugged and unruly meadows.”<sup>19</sup>

#### Outcome 4: Pressure to Manipulate Key Indicators Leads to Falsification and Unwarranted Impact Claims

As an extension of our previous point regarding how the easy-to-measure can become the yardstick for success, it is important to acknowledge that public officials are often under extreme pressure to demonstrate success in selected indicators. Once data themselves, rather than the more complex underlying reality, become the primary objective by which governments publicly assess (and manage) their “progress,” it is inevitable that vast political pressure will be placed on these numbers to bring them into alignment with expectations, imperatives, and interests. Similar logic can be expected at lower units of analysis (for example, field offices), where it tends to be even more straightforward to manipulate data entry and analysis. This, in turn, contributes to a perverse incentive to falsify or skew data, to aggregate numbers across wildly different variables into single indexes, and to draw unwarranted inferences from them.

This risk is particularly acute, for instance, when annual global rankings are publicly released (assessing, for example, a country’s “investment climate,” “governance,” and gender equity), thereby shaping major investment decisions, credit ratings, eligibility for funding from international agencies, and the fate of senior officials charged with “improving” their country’s place in these global league tables. Readers will surely be aware of the case at the World Bank in September 2021, when an external review revealed that the *Doing Business* indicators had been subject to such pressure, with alterations made to certain indicators from certain countries (WilmerHale 2021). Such rankings are now

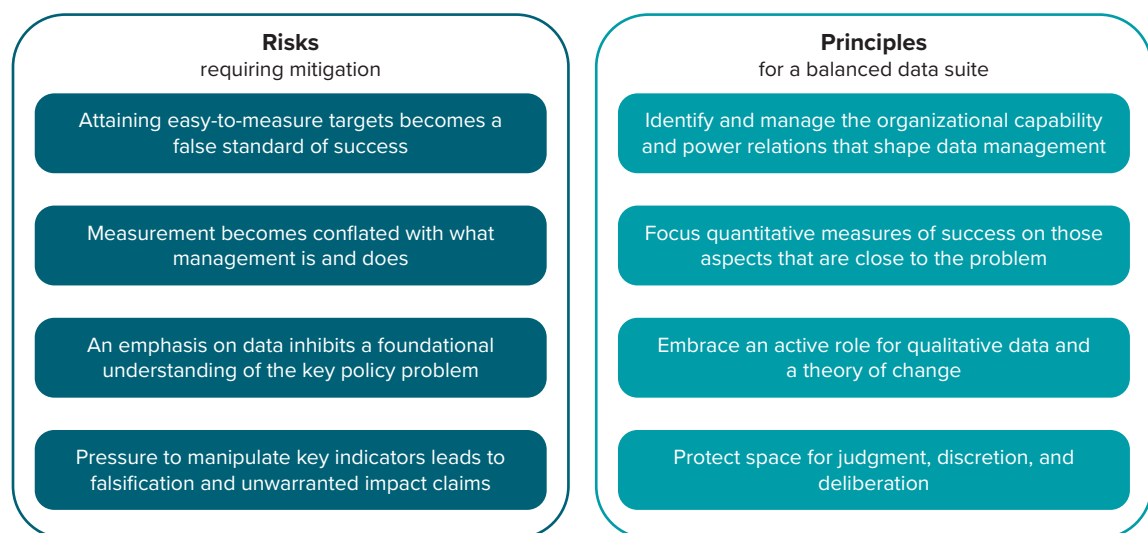
omnipresent, and if they are not done by one organization, they will inevitably be done by another. Even so, as the *Economist* (2021) magazine concluded, some might regard the *Doing Business* episode as “proof of ‘Goodhart’s law,’ which states that when a measure becomes a target, it ceases to be a good measure.” At the same time, it pointed out that there is a delicate dance to be done here, since “the *Doing Business* rankings were always intended to motivate as well as measure, to change the world, not merely describe it,” and “if these rankings had never captured the imagination of world leaders, if they had remained an obscure technical exercise, they might have been better as measures of red tape. But they would have been worse at cutting it.”

Such are the wrenching trade-offs at stake in such exercises, and astute public administrators need to engage in them with their eyes wide open. Even (or especially) at lower units of analysis, where there are perhaps fewer prying eyes or quality-control checks, the potential is rife for undue influence to be exerted on data used for political and budgetary-allocation purposes. Fully protecting the integrity of data collection, collation, and curation (in all its forms) should be a first-order priority, but so, too, is the need for deploying what should be standard “risk diversification” strategies on the part of managers—namely, not relying on single numbers or methods to assess inherently complex realities.

## PRINCIPLES FOR AN EXPANSIVE, QUALIFIED DATA SUITE THAT FOSTERS PROBLEM SOLVING AND ORGANIZATIONAL LEARNING

In response to the four risks identified above, we offer a corresponding set of cross-cutting principles for addressing them. Figure 4.2 summarizes the four risks in the left-hand column and presents the principles as vertical text on the right, illustrating the extent to which the principles, when applied in combination, can serve to produce a more balanced data suite that prioritizes problem solving and learning.

**FIGURE 4.2 Four Risks with Corresponding Principles for Mitigating Them to Ensure a Balanced Data Suite**



Source: Original figure for this publication.

Note: The principles are “cross-cutting,” in the sense that they apply in some measure to all the risks; they are not one-to-one.



## **Principle 1: Identify and Manage the Organizational Capacity and Power Relations That Shape Data Management**

The data collection and curation process takes place not in isolation but in a densely populated political and institutional ecosystem. It is difficult, expensive, and fraught work; building a professional team capable of reliably and consistently doing this work—from field-level collection and curation at headquarters to technical analysis and policy interpretation—will be as challenging as it is in every other public sector organization. Breakdowns can happen at any point, potentially compromising the integrity of the entire endeavor. For this reason, it is important for managers not just to hire those with the requisite skills but to cultivate, recognize, and reward a professional ethos wherein staff members can do their work in good faith, shielded from political pressure. Such practices, in turn, need to be protected by clear, open, and safe procedures for staff to report undue pressure, complemented by accountability to oversight or advisory boards comprising external members selected for their technical expertise and professional integrity. In the absence of such mechanisms, noble aspirations for pursuing an “evidence-based policy” agenda risk being perceived as means of providing merely “policy-based evidence.”

The contexts within or from which data are collected are also likely to be infused with their own socio-political characteristics. Collecting data on the incidence of crime and violence, for example, requires police to faithfully record such matters and their response to them, but they must do so in an environment where there may be strong pressure to underreport, whether because of personal safety concerns, lack of administrative resources, or pressure to show that a given unit’s performance is improving (where this is measured by showing a “lower incidence” of crime). In this respect, good diagnostic work will reveal the contours of the institutional and political ecosystem wherein the data work will be conducted and the necessary authorization, financing, and protection sought; it will also help managers learn how to understand and successfully navigate this space.<sup>20</sup> The inherent challenges of engaging with such issues might be eased somewhat if those closest to them see data deployment not as an end in itself or an instrument of compliance but as a means to higher ends—namely, learning, practical problem solving, and enhancing the quality of policy options, choices, and implementation capability.<sup>21</sup>

A related point is that corresponding efforts need to be made to clearly and accurately communicate to the general public those findings that are derived from data managed by public administrators, especially when these findings are contentious or speak to inherently complex issues. This has been readily apparent during the COVID-19 pandemic, with high-stakes policy decisions (for example, requiring vulnerable populations to forgo income) needing to be made on the basis of limited but evolving evidence. Countries such as Vietnam have been praised for the clear and consistent manner in which they issued COVID-19 response guidelines to citizens (Ravallion 2020), but the broader point is that even when the most supported decisions are based on the best evidence generated by the most committed work environments, it remains important for administrators to appreciate that the very act of large-scale measurement and empirical interpretation, especially when enacted by large public organizations, can potentially be threatening to or misunderstood by the very populations they are seeking to assist.

## **Principle 2: Focus Quantitative Measures of Success on Those Aspects That Are Close to the Problem**

If we wish to guard against the tendency to falsely ascribe success based on the achievement of poorly selected indicators, then we should ensure that any indicators used to claim or deny reform success are as readily operational and close to the service delivery problem as possible. Output and process indicators are useful in their own ways, but we should not make the mistake of conflating their achievement with “problem fixed.” There are often strong institutional incentives to claim reform success based on whether a new mechanism or oversight structure has been created, a new law has been passed, or a percentage of participation has been achieved, but if meaningful change is sought, these incentives need to be countered. All of these



measures are changes in *form* that, while useful as indicators of outputs being met, can be achieved (and have been in the past) without any attendant functional shifts in the underlying quality of service delivery.

Officials can guard against this tendency by taking time to ensure that an intervention is focused on specific problems, including those that matter at a local level, and that the intervention's success and its attendant metrics are accurate measures of the problems being fixed. Tools like the Problem-Driven Iterative Adaptation (PDIA) Toolkit, designed by members of the Building State Capability program at the Center for International Development (CID) at Harvard University, can help guide practitioners in this process.<sup>22</sup> The PDIA approach is designed to help practitioners break down their problems into root causes, identify entry points, search for possible solutions, take action, reflect upon what they have learned, adapt, and then act again. By embedding any intervention in such a framework, practitioners can ensure that success metrics are well linked to functional responses to locally felt problems.

Whatever tool is applied, the goal should be to arrive at metrics of success that represent a compelling picture of the root performance problem being addressed (and hopefully solved). So, for example, in our education example, metrics such as the number of teachers hired, the percentage of the budget dedicated to education, and the number of schools built are all output measures that say nothing about actual learning. Of course, there are assumptions that these outputs *lead* to children's learning, but as many recent studies now show, such assumptions are routinely mistaken; these indicators can be achieved even as actual learning regresses (Pritchett 2013; World Bank 2018). By contrast, when a robust measure of learning—in this case, literacy acquisition—was applied in India, it allowed implementers to gain valuable insights about which interventions actually made a difference, revealing that teaching to a child's actual level of learning, not their age or grade, led to marked and sustained improvements. Crucially, such outcomes are the result of carefully integrated qualitative and quantitative approaches to measurement (Banerjee et al. 2016).

Going further, various cross-national assessments around the world are trying to tackle the complex challenge of finding indicators that measure learning not just in the acquisition of numeracy, science, and literacy skills but in competencies that are increasingly valuable in the 21st century: grit, curiosity, communication, leadership, and compassion. PISA, for example, has included an “innovative domain” in each of its recent rounds, including creative problem solving in 2012, collaborative problem solving in 2015, and global competence in 2018. In Latin America, the Latin American Laboratory for Assessment of the Quality of Education (LLECE) included a module on socioemotional skills for the first time in its assessment of sixth-grade students in 2019, focusing on the concepts of conscience, valuing others, self-regulation, and self-management (Global Partnership for Education 2020). Much tinkering remains to be done, but the increase in assessments that include skills and competencies such as citizenship (local and global), socioemotional skills, information and communication technology literacy, and problem solving is a clear indication of willingness to have functional measures of success, capturing outcomes that matter.

In summary, then, public administrators who wish to guard against unwarranted impact claims and ensure metrics of success are credible can begin by making sure that an intervention itself is focused on a specific performance problem that is locally prioritized and thereafter ensure that any judgment of that intervention's success or failure is based not on output or process metrics but on measures of the problems being fixed. And having ensured that measures of success are functional, practitioners must allow for flexibility of implementation where possible so strategies can shift if it becomes clear from the collected data that they are not making progress toward fixing the problem, possibly due to mistaken assumptions regarding their theory of change.

### **Principle 3: Embrace an Active Role for Qualitative Data and a Theory of Change**

The issues we have raised thus far, we argue, imply that public administrators should adopt a far more expansive concept of what constitutes “good data”—namely, one that includes insights from theory and qualitative research. Apprehending complex problems requires different forms and sources of data; correctly interpreting empirical findings requires active dialogue with reasoned expectations about what outcomes should be attained by when. Doing so helps avoid creating distortions that can generate (potentially wildly) misleading claims regarding “what's going on and why” and “what should be done.”

Specifically, we advocate for the adoption of a complementary *suite* of data forms and sources that favors flexibility, is focused on problem solving (as opposed to being an end in itself), and values insights derived from seasoned experience. In the examples we have explored above, reliance on a single form of data (sometimes even a single number) rendered projects vulnerable to political manipulation, unwarranted conclusions, and an inability to bear the decision-making burdens thrust upon them. More constructively, it was the incorporation of alternative methods and data in dialogue with a reasoned theory of change that enabled decision-makers to anticipate and address many of these same concerns.

To this end, we have sought to get beyond the familiar presumption that the primary role of qualitative data and methods in public administration research (and elsewhere) is to provide distinctive insights into the idiosyncrasies of an organization's "context" and "culture" (and thus infuse some "color" and "anecdotes" for accompanying boxes).<sup>23</sup> Qualitative approaches can potentially yield unique and useful material that contributes to claims about *whether* policy goals are being met and delivery processes duly upheld (Cartwright 2017); they can be especially helpful when the realization of policy goals requires integrating both adaptive and technical approaches to implementation—for example, responding to COVID-19. But perhaps the more salient contributions of qualitative approaches, we suggest, are to explore *how*, *for whom*, and *from whom* data of all kinds are deployed as part of broader imperatives to meet political requirements and administrative logics in a professional manner and to elicit novel or previously "unobserved" variables shaping policy outcomes.

#### Principle 4: Protect Space for Judgment, Discretion, and Deliberation

Our caution is against using data reductively: as a replacement or substitute for managing. Management *must* be about more than measuring. A good manager needs to be able to accommodate the immeasurable because so much that is important to human thriving is in this category; dashboards etc. certainly have their place, but if these were all that was needed, then "managing" could be conducted by machines. We all know from personal experience that the best managers and leaders take a holistic interest in their staff, taking the time and making the effort to understand the subtle, often intangible processes that connect their respective talents. As organizational management theorist Henry Mintzberg (2015) wisely puts it,

Measuring as a complement to managing is a fine idea: measure what you can; take seriously what you can't; and manage both thoughtfully. In other words: If you can't measure it, you'll have to manage it. If you can measure it, you'll especially have to manage it. Have we not had enough of leadership by remote control: sitting in executive offices and running the numbers—all that deeming and downsizing?<sup>24</sup>

Contrary to the "what can't be measured can't be managed" idea, we *can* manage the less measurable if we embrace a wider set of tools and leave space for judgment. The key for practitioners is to begin with a recognition that measurability is not an indicator of significance and that professional management involves far more than simply "running the numbers," as Mintzberg puts it. Perhaps the most compelling empirical case for the importance of "navigating by judgment" in public administration has been made by Honig (2018), in which he shows—using a mix of quantitative data and case study analysis—that the more complex the policy intervention, the more necessary it becomes to grant discretionary space to frontline managers, and the more necessary such discretion is to achieving project success. Having ready access to relevant, high-quality quantitative data can aid in this "navigation," but true navigation requires access to a broader suite of empirical inputs.

In a similar vein, Ladner (2015, 3) points out that "standard performance monitoring tools are not suitable for highly flexible, entrepreneurial programs as they assume that how a program will be implemented follows its original design." To avoid "locking in" a theory of change that prevents exploration or responsive adaptation, some practitioners have provided helpful suggestions for how to use various planning frameworks in ways that support program learning.<sup>25</sup> The Building State Capability team highlights lighter-touch methods, such as their PDIA "check-ins," which include a series of probing questions

to assist teams in capturing learning and maximizing adaptation. Teskey and Tyrrel (2017) recommend participating in regularized formal and informal Review and Reflection (R&R) points, during which a contractor can demonstrate how politics, interests, incentives, and institutions were systematically considered in problem selection and design and, in turn, justify why certain choices were made to stop, drop, halt, or expand any activity or budget during implementation. The common connection across all these tools is that they seek to carve out meaningful space for qualitative data and the hard-won insights born out of practical experience.

In summary, then, public administrators can embed the recognition that management must be about more than measuring by first recognizing that whatever they choose to measure will inevitably create incentives to neglect processes and outcomes that cannot be measured (or are hard to measure) but are nonetheless crucial for discerning whether, how, where, and for whom policies are working. Following that recognition, they need to be very careful about what they choose to measure. Second, they can actively identify what they cannot (readily) measure that matters and take it seriously, developing strategies to manage that as well. A key part of those strategies will be that they create space for judgment, qualitative data inputs, and the practical experience of embedded individuals (focus group discussions, case studies, semi-structured interviews, review and reflection points, etc.) and treat these inputs as equally valid alongside more quantitative ones. As far as longer-term strategies to manage the immeasurable, administrations can work toward developing organizational systems that foster navigation. Such systems might include, for example, a management structure that delegates high levels of discretion to allow those on the ground the ability to navigate complex situations, recruitment strategies that foster high numbers of staff with extensive context-specific knowledge, and systems of monitoring and learning that encourage the routine evaluation of theory with practice.

## CONCLUSION

Quantitative measurement in public administration is undoubtedly a critical arrow in the quiver of any attempt to improve the delivery of public services. And yet, since not everything that matters can be measured and not everything that can be measured matters, a managerial emphasis on measurement alone can quickly and inadvertently generate unwanted outcomes and unwarranted conclusions. In the everyday practices of public administration, effective and professional action requires forging greater complementarity between different epistemological approaches to collecting, curating, analyzing, and interpreting data. We fully recognize that this is easier said than done. The risks of reductive approaches to measurement are not unknown, and yet simplified appeals to “what gets measured gets managed” persist because they offer managers a form of escape from those “pesky human elements” that are difficult to understand and even more so to shift.

Most public administrators might agree in principle that a more balanced data suite is necessary to navigate their professional terrain, yet such aspirations are too often honored in the breach: under sufficient pressure to “deliver results,” staff from the top to the bottom of an organization are readily tempted to reverse engineer their behavior in accordance with what “the data” say (or can be made to say). Management as measurement is tempting for individuals and organizations that fear the vulnerability of their domain to unfavorable comparison with other (more readily measurable and “legible”) domains, as well as the complexity of problem solving and the necessity of subjective navigation that it often entails. But given how heavily institutional and sociopolitical factors shape how data are collected, how well they are collected and curated, and how they can be manipulated for unwarranted purposes, a simplistic approach to data as an easy fix is virtually guaranteed to obscure learning and hamper change efforts. If administrations genuinely wish to build their problem-solving capabilities, then access to more and better quantitative data will be necessary, but it will not be sufficient.

Beginning with an appreciation that much of what matters cannot be (formally) measured, public administration must routinely remind itself that promoting and accessing data is not an end in itself: data's primary purpose is not just monitoring processes, compliance, and outcomes, but contributing to problem solving and organizational learning. More and better data will not fix a problem if the absence of such data is not itself the key problem or the "binding constraint." Administrations that are committed to problem solving, therefore, will need to embed their measurement task in a broader problem-driven framework, integrate complementary qualitative data, and value embedded experience in order to apprehend and interpret complex realities more accurately. Their priority in undertaking good diagnostic work should be to identify and deconstruct key problems, using varied sources of data, and then to track and learn from potential solutions authorized and enacted in response to the diagnosis. Accurate inferences for policy and practice are not derived from data alone; close interaction is required between data (in various forms), theory, and experience. In doing all this, public administrators will help mitigate the distortionary (and ultimately self-defeating) effects of managing only that which is measured.

## NOTES

Our thanks to Galileu Kim, Daniel Rogger, Christian Schuster, and participants at an authors' workshop for helpful comments and constructive suggestions. More than 20 years of collaboration with Vijayendra Rao have also deeply shaped the views expressed herein. Remaining errors of fact or interpretation are solely ours.

1. See, for example, former USAID Administrator Andrew Natsios (2011), citing Lord Wellington in 1812 on the insidious manner in which measures of "accountability" can compromise rather than enable central policy objectives (in Wellington's case, winning a war). For his part, Stiglitz has argued that "what you measure affects what you do. If you don't measure the right thing, you don't do the right thing" (quoted in Goodman 2009). Pritchett (2014), exemplifying this point, notes (at least at the time of his writing) that the Indian state of Tamil Nadu had 817 indicators for measuring the delivery of public education but none that actually assessed whether students were learning. In this instance, an abundance of "measurement" and "data" was entirely disconnected from (what should have been) the policy's central objective. In many cases, however, it is not always obvious, especially *ex ante*, what constitutes the "right thing" to measure—hence the need for alternative methodological entry points to elicit what this might be.
2. Social science methodology courses classically distinguish between four key issues that are at the heart of efforts to make empirical claims in applied research: *construct validity* (the extent to which any concept, such as "corruption" or "poverty," matches particular indicators), *internal validity* (the extent to which causal claims have controlled for potential confounding factors, such as sample selection bias), *external validity* (the likelihood that claims are generalizable at larger scales and to more diverse populations or novel contexts), and *reliability* (the extent to which similar findings would be reported if repeated or replicated by others). See, among many others, Johnson, Reynolds, and Mycoff (2019). Of these four issues, qualitative methods are especially helpful in ensuring construct validity, since certain terms may mean different things to different people in different places, complicating matters if one seeks to draw comparisons across different linguistic, cultural, or national contexts. In survey research, for example, it is increasingly common to include what is called an "anchoring vignette"—a short, real-world example of the phenomenon in question, such as an instance of corruption by a government official at a port—before asking the formal survey question so that cross-context variations in interpretation can be calibrated accordingly (see, among others, King and Wand 2007). Qualitative methods can also contribute to considerations pertaining to internal validity (Cartwright 2017) and external validity—helping to identify the conditions under which findings "there" might apply "here" (Woolcock 2018; see also Cartwright and Hardie 2012).
3. If such agencies or departments do in fact happen to perform especially strongly—in the spirit of the "positive deviance" cases of government performance in Ghana provided in McDonnell (2020)—then it would be useful to understand how and why this has been attained. For present purposes, our point is that, perhaps paradoxically, we should not expect, *ex ante*, that agencies or departments in the business of collecting and curating data for guiding policy and performance should themselves be exemplary exponents of the deployment of that data to guide their *own* performance—because doing this is a separate ontological task, requiring distinct professional capabilities. Like the proverbial doctors, if data analysts cannot "heal themselves," we should not expect other public agencies to be able to do so merely by "infusing them with more and better data."
4. A special issue of *The Journal of Development Studies*, 51.2, was dedicated to this problem. For example, on the enduring challenges associated with agricultural data—another sector with a long history of data collection experience—see Carletto, Jolliffe, and Banerjee (2015).

5. The adage is popularly known as GIGO: garbage in, garbage out.
6. See the evolution in early work on gender inclusion in rural India and subsequent work (Ban and Rao 2008; Duflo 2012; Sanyal and Rao 2018).
7. This does not mean, of course, that nothing can be said about GEPs after five years—managers and funders would surely want to know by this point whether the apparent “no net impact” claim is a result of poor technical design, weak implementation, contextual incompatibility, countervailing political pressures, or insufficient time having elapsed. Moreover, they would likely be interested in learning whether the GEP’s zero “average treatment effect” is nonetheless a process of offsetting outcomes manifest in a high standard deviation (meaning the GEP works wonderfully for some groups in some places but disastrously for others) and/or is yielding unanticipated or unmeasured outcomes (whether positive or negative). For present purposes, our point is that reliance on a single form and methodological source of data is unlikely to be able to answer these crucial administrative questions; with a diverse suite of methods and data, however, such questions become both askable and answerable. (See Rao, Ananthpur, and Malik 2017 for an instructive example, discussed below.)
8. One could say that this is a social scientific version of the Heisenberg uncertainty principle, in which the very act of measuring something changes it. See also Breckenridge (2014) on the politics and legacy of identity measurement in pre- and postcolonial South Africa and Hostetler (2021) on the broader manner in which imposing singular (but often alien) measures of time, space, and knowledge enabled colonial administration. More generally, Sheila Jasanoff’s voluminous scholarship shows how science is a powerful representation of reality, which, when harnessed to technology, can reduce “individuals to standard classifications that demarcate the normal from the deviant and authorize varieties of social control” (Jasanoff 2004, 13).
9. Among the classic historical texts on this issue are *Peasants into Frenchmen* (Weber 1976) and *Imagined Communities* (Anderson 1983). For more recent discussions, see Lewis (2015) on “the politics and consequences of performance measurement” and Beraldo and Milan (2019) on the politics of big data.
10. This is the finding, for example, from a major empirical assessment of cross-country differences regarding COVID-19 (Bollyky et al. 2022), wherein—controlling for a host of potential confounding variables—those countries with both high infections and high fatalities are characterized by low levels of trust between citizens and their governments and between each other. See further discussion of this study and its implications below.
11. The British movie *I, Daniel Blake* provides a compelling example of how even the literate in rich countries can be excluded by administrative systems and procedures that are completely alien to them—for example, filling out forms for unemployment benefits on the internet that require users to first “log on” and then “upload” a “CV.” The limits of formal measurement to bring about positive policy change has long been recognized; when the Victorian-era writer George Eliot was asked why she wrote novels about the lives of the downtrodden rather than contributing to official government reports more formally documenting their plight, she astutely explained that “appeals founded on generalizations and statistics require a sympathy ready-made, a moral sentiment already in activity” (quoted in Gill 1970, 10). Forging such Smithian “sympathy” and “moral sentiment” is part of the important antecedent work that renders “generalizations and statistics” legible and credible to those who might otherwise have no reason for engaging with, or experience interpreting, such encapsulations of reality.
12. We fully recognize that, in principle, econometricians have methods available to identify both outcome heterogeneity and the factors driving it. Even so, if local average treatment effects are reported as zero, the “no impact” conclusion is highly likely to be the (only) key takeaway message. The primary benefit of incorporating both qualitative and econometric methods is the capacity of the former to identify factors that were not anticipated in the original design (see Rao 2022). In either case, Ravallion’s (2001) injunction to “look beyond averages” when engaging with complex phenomena is worth being heeded by all researchers (and those that interpret researchers’ findings), no matter their disciplinary or methodological orientations.
13. On the use of mixed methods in process evaluations, see Rogers and Woolcock (2023).
14. The six pillars were: maternal, newborn, and child health; childhood essential medicines and increasing treatment of important childhood diseases; improving child nutrition; immunization; malaria control; and the elimination of mother-to-child transmission of human immunodeficiency virus (HIV).
15. A PforR is one of the World Bank’s three financing instruments. Its unique features are that it uses a country’s own institutions and processes and links disbursement of funds directly to the achievement of specific program results. Where “traditional” development interventions proceed on the basis of ex ante commitments (for example, to designated “policy reforms” or to the adoption of procedures compliant with international standards), PforR-type interventions instead reward the attainment of predetermined targets, typically set by extrapolating from what recent historical trajectories have attained. According to the Project Appraisal Document for SOML, “each state would be eligible for a grant worth \$325,000 per the percentage point gain they made above average annual gain in the sum of six indicators of health service coverage.” The six indicators were: vitamin A, Pentavalent3 immunization, use of insecticide-treated nets (ITNs) by children under five, skilled birth attendance, contraceptive prevalence rate, and the prevention of mother-to-child transmission of HIV.
16. These tables are based on student performance in standardized tests in mathematics, reading, and science, which are administered by the Paris-based Organisation for Economic Co-operation and Development (OECD).



17. A summary of the report explains that

whereas negotiation and flexibility are high on the list of skills for 2015, in 2020 they will begin to drop from the top 10 as machines, using masses of data, begin to make our decisions for us. A survey done by the World Economic Forum's Global Agenda Council on the Future of Software and Society shows people expect artificial intelligence machines to be part of a company's board of directors by 2026. Similarly, active listening, considered a core skill today, will disappear completely from the top 10. Emotional intelligence, which doesn't feature in the top 10 today, will become one of the top skills needed by all. (Gray 2016)

See also Soffel (2016).

18. Many companies and tertiary institutions are ahead of the curve in this regard. Recently, over 150 of the top private high schools in the US, including Phillips Exeter Academy and the Dalton School—storied institutions that have long relied on the status conveyed by student ranking—have pledged to shift to new transcripts that provide more comprehensive, qualitative feedback on students while ruling out any mention of credit hours, GPAs, or A–F grades. And colleges—the final arbiters of high school performance—are signaling a surprising willingness to depart from traditional assessments that have been in place since the early 19th century. From Harvard and Dartmouth to small community colleges, more than 70 US institutions of higher learning have weighed in, signing formal statements asserting that competency-based transcripts will not hurt students in the admissions process. See the “College Admissions” page on the New England Secondary School Consortium website: <http://www.newenglandssc.org/resources/college-admissions/>.
19. See Milante and Woolcock (2017) for a complementary set of dynamic quantitative and qualitative measures by which a given country might be declared a “fragile” state.
20. For development-oriented organizations, a set of tools and guidelines for guiding this initial assessment according to a political economy analysis (PEA) framework—crafted by USAID and ODI (London) and adopted by certain parts of the World Bank—is *Thinking and Working Politically through Applied Political Economy Analysis: A Guide for Practitioners* (Rocha Menocal et al. 2018). Its key observations include the following. First, a well-designed process of policy implementation should answer not only the technical question of what needs to be done but also how it should be done. Second, in-depth understanding of the political, economic, social, and cultural forces needs to supplement technical analysis to achieve successful policy implementation. Third, PEA should incorporate three pillars: the foundational factors (geography, natural resource occurrence, national borders), the “rules of the game” (institutions at the formal [political system, administrative structure, and law] and the informal [social and cultural norms] levels), and the “here and now” (current leaders, geopolitical situation, and natural hazards). Fourth, it is crucial to pay attention to the institutions, the structure of incentives, and the constraints, as well as the gains and losses of all the actors involved in policy implementation, including those outside of the traditional purview of development organizations. Fifth, policy solutions should be adjusted to political realities encountered on the ground in an iterative and incremental fashion. And finally, the evaluation of policy success should be extended to incorporate “process-based indicators,” including trust and quality of relationship. Hudson, Marquette, and Waldo (2016) offer a guide for “everyday political analysis,” which introduces a stripped-back political-analysis framework designed to help frontline practitioners make quick but politically informed decisions. It aims to complement more in-depth political analysis by helping programming staff to develop the “craft” of political thinking in a way that fits their everyday working practices.
21. On the application of such efforts to the case of policing in particular, see Sparrow (2018).
22. The *PDI Toolkit: A DIY Approach to Solving Complex Problems* (Samji et al. 2018) was designed by members of Harvard's Building State Capability program to guide government teams through the process of identifying, deconstructing, and solving complex problems. See in particular the section “Constructing your problem,” which guides practitioners through the process of defining a problem that matters and building a credible, measurable vision of what success would look like.
23. As anthropologist Mike McGovern (2011, 353) powerfully argues, taking context seriously  
is neither a luxury nor the result of a kind of methodological altruism to be extended by the soft-hearted. It is, in purely positivist terms, the epistemological due diligence work required before one can talk meaningfully about other people's intentions, motivations, or desires. The risk in foregoing it is not simply that one might miss some of the local color of individual “cases.” It is one of misrecognition. Analysis based on such misrecognition may mistake symptoms for causes, or two formally similar situations as being comparable despite their different etiologies. To extend the medical metaphor one step further, misdiagnosis is unfortunate, but a flawed prescription based on such a misrecognition can be deadly.

More generally, see Hoag and Hull (2017) for a summary of the anthropological literature on the civil service. Bailey (2017) provides a compelling example of how insights from qualitative fieldwork help explain the strong preference among civil servants in Tanzania for providing new water infrastructure projects over maintaining existing ones. Though a basic benefit-cost analysis favored prioritizing maintenance, collective action problems among civil servants themselves, the prosaic challenges of mediating local water management disputes overseen by customary institutions, and the performance targets set by the government all conspired to create suboptimal outcomes.

24. Says Mintzberg (2015): “Someone I know once asked a most senior British civil servant why his department had to do so much measuring. His reply: ‘What else can we do when we don’t know what’s going on?’ Did he ever try getting on the ground to find out what’s going on? And then using judgment to assess that?”
25. Teskey (2017) and Wild, Booth, and Valters (2017) give examples of an adaptive logframe, drawn from Department for International Development experiences, that sets out clear objectives at the outcome level and focuses monitoring of outputs on the quality of the agreed rapid-cycle learning process. Strategy Testing (ST) is a monitoring system that the Asia Foundation developed specifically to track programs that are addressing complex development problems through a highly iterative, adaptive approach.

## REFERENCES

- Anderson, Benedict. 1983. *Imagined Communities: Reflections on the Origins and Spread of Nationalism*. London: Verso.
- Andrews, Matt. 2013. *The Limits of Institutional Reform in Development: Changing Rules for Realistic Solutions*. New York: Cambridge University Press.
- Andrews, Matt, Lant Pritchett, and Michael Woolcock. 2017. *Building State Capability: Evidence, Analysis, Action*. New York: Oxford University Press.
- Bailey, Julia. 2017. “Bureaucratic Blockages: Water, Civil Servants and Community in Tanzania.” Policy Research Working Paper 8101, World Bank, Washington, DC.
- Ban, Radu, and Vijayendra Rao. 2008. “Tokenism or Agency? The Impact of Women’s Reservations on Village Democracies in South India.” *Economic Development and Cultural Change* 56 (3): 501–30. <https://doi.org/10.1086/533551>.
- Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukherji, Marc Shotland, and Michael Walton. 2016. “Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of ‘Teaching at the Right Level’ in India.” NBER Working Paper 22746, National Bureau of Economic Research, Cambridge, MA.
- Beraldo, Davide, and Stefania Milan. 2019. “From Data Politics to the Contentious Politics of Data.” *Big Data & Society* 6 (2): 1–11. <https://doi.org/10.1177/2053951719885967>.
- Bollyky, Thomas J., Erin N. Hulland, Ryan M. Barber, James K. Collins, Samantha Kiernan, Mark Moses, David M. Pigott, et al. 2022. “Pandemic Preparedness and Covid-19: An Exploratory Analysis of Infection and Fatality Rates, and Contextual Factors Associated with Preparedness in 177 Countries, from Jan 1, 2020, to Sept 30, 2021.” *The Lancet* 399 (10334): 1489–512. [https://doi.org/10.1016/S0140-6736\(22\)00172-6](https://doi.org/10.1016/S0140-6736(22)00172-6).
- Breckenridge, Keith. 2014. *Biometric State: The Global Politics of Identification and Surveillance in South Africa, 1850 to the Present*. Cambridge, UK: Cambridge University Press.
- Bridges, Kate, and Michael Woolcock. 2017. “How (Not) to Fix Problems That Matter: Assessing and Responding to Malawi’s History of Institutional Reform.” Policy Research Working Paper 8289, World Bank, Washington, DC.
- Bridges, Kate, and Michael Woolcock. 2019. “Implementing Adaptive Approaches in Real World Scenarios: A Nigeria Case Study, with Lessons for Theory and Practice.” Policy Research Working Paper 8904, World Bank, Washington, DC.
- Cahill, Jonathan. 2017. *Making a Difference in Marketing: The Foundation of Competitive Advantage*. London: Routledge.
- Carletto, Calogero, Dean Jolliffe, and Raka Banerjee. 2015. “From Tragedy to Renaissance: Improving Agricultural Data for Better Policies.” *The Journal of Development Studies* 51 (2): 133–48. <https://doi.org/10.1080/00220388.2014.968140>.
- Cartwright, Nancy. 2017. “Single Case Causes: What Is Evidence and Why.” In *Philosophy of Science in Practice*, edited by Hsiang-Ke Chao and Julian Reiss, 11–24. New York: Springer.
- Cartwright, Nancy, and Jeremy Hardie. 2012. *Evidence-Based Policy: A Practical Guide to Doing It Better*. New York: Oxford University Press.
- Caulkin, Simon. 2008. “The Rule Is Simple: Be Careful What You Measure.” *Guardian* (US edition), February 9, 2008. <https://www.theguardian.com/business/2008/feb/10/businesscomment1>.
- Caulkin, Simon. 2016. “Decision-Making: How to Make the Most of Data.” *The Treasurer*, January 4, 2016. <https://www.treasurers.org/hub/treasurer-magazine/decision-making-how-make-most-data>.
- Claxton, Guy, and Bill Lucas. 2015. *Educating Ruby: What Our Children Really Need to Learn*. New York: Crown House Publishing.
- Dirks, Nicholas. 2011. *Castes of Mind*. Princeton, NJ: Princeton University Press.
- Dom, Catherine, Alasdair Fraser, Joseph Holden, and John Patch. 2020. “Results-Based Financing in the Education Sector: Country-Level Analysis. Final Synthesis Report.” Report submitted to the REACH Program at the World Bank by Mokoro Ltd.



- Duflo, Esther. 2012. "Women Empowerment and Economic Development." *Journal of Economic Literature* 50 (4): 1051–79. <https://doi.org/10.1257/jel.50.4.1051>.
- Economist*. 2021. "How World Bank Leaders Put Pressure on Staff to Alter a Global Index." September 17, 2021. <https://www.economist.com/finance-and-economics/2021/09/17/how-world-bank-leaders-put-pressure-on-staff-to-alter-a-global-index>.
- Gill, Stephen. 1970. "Introduction to Elizabeth Gaskell." In *Mary Barton: A Tale of Manchester Life*. London: Penguin.
- Global Partnership for Education. 2020. *21st-Century Skills: What Potential Role for the Global Partnership for Education?* Washington, DC: Global Partnership for Education Secretariat. <https://www.globalpartnership.org/sites/default/files/document/file/2020-01-GPE-21-century-skills-report.pdf>.
- Goodman, Peter S. 2009. "Emphasis on Growth Is Called Misguided." *New York Times*, October 4, 2009. <https://www.nytimes.com/2009/09/23/business/economy/23gdp.html>.
- Gray, Alex. 2016. "The 10 Skills You Need to Thrive in the Fourth Industrial Revolution." World Economic Forum. <https://www.weforum.org/agenda/2016/01/the-10-skills-you-need-to-thrive-in-the-fourth-industrial-revolution>.
- Hoag, Colin, and Matthew Hull. 2017. "A Review of the Anthropological Literature on the Civil Service." Policy Research Working Paper 8081, World Bank, Washington, DC.
- Honig, Dan. 2018. *Navigation by Judgment: Why and When Top-Down Management of Foreign Aid Doesn't Work*. New York: Oxford University Press.
- Honig, Dan, and Lant Pritchett. 2019. "The Limits of Accounting-based Accountability in Education (and Far Beyond): Why More Accounting Will Rarely Solve Accountability Problems." Working Paper No. 510, Center for Global Development, Washington, DC.
- Hostetler, Laura. 2021. "Mapping, Registering, and Ordering: Time, Space and Knowledge." In *The Oxford World History of Empire: Volume One: The Imperial Experience*, edited by Peter Fibiger Bang, C. A. Bayly, and Walter Scheidel, 288–317. New York: Oxford University Press.
- Hudson, David, Heather Marquette, and Sam Waldo. 2016. "Everyday Political Analysis." Working paper, Developmental Leadership Program, University of Birmingham, Birmingham, UK.
- Jasanoff, Sheila. 2004. "Ordering Knowledge, Ordering Society." In *States of Knowledge: The Co-Production of Science and the Social Order*, edited by Sheila Jasanoff, 13–45. London: Routledge.
- Jerven, Morten. 2013. *Poor Numbers: How We Are Misled by African Development Statistics and What to Do About It*. Ithaca, NY: Cornell University Press.
- Johnson, Janet Buttolph, Henry T. Reynolds, and Jason D. Mycoff. 2019. *Political Science Research Methods*. 9th ed. Thousand Oaks, CA: Sage.
- Khan, Salman. 2012. *The One World Schoolhouse: Education Reimagined*. London: Hodder & Stoughton.
- King, Gary, and Jonathan Wand. 2007. "Comparing Incomparable Survey Responses: Evaluating and Selecting Anchoring Vignettes." *Political Analysis* 15 (1): 46–66. <https://doi.org/10.1093/pan/mpl011>.
- Ladner, Debra. 2015. *Strategy Testing: An Innovative Approach to Monitoring Highly Flexible Aid Programs*. Working Politically in Practice Case Study 3. San Francisco: The Asia Foundation.
- Lewis, Jenny M. 2015. "The Politics and Consequences of Performance Measurement." *Policy and Society* 34 (1): 1–12. <https://doi.org/10.1016/j.polsoc.2015.03.001>.
- Mamdani, Mahmood. 2002. *When Victims Become Killers: Colonialism, Nativism, and the Genocide in Rwanda*. Princeton, NJ: Princeton University Press.
- McCarthy, Niall. 2019. "The Countries Best Prepared to Deal With a Pandemic." *Statista*, October 28, 2019. <https://www.statista.com/chart/19790/index-scores-by-level-of-preparation-to-respond-to-an-epidemic/>.
- McDonnell, Erin. 2017. "Patchwork Leviathan: How Pockets of Bureaucratic Governance Flourish within Institutionally Diverse Developing States." *American Sociological Review* 82 (3): 476–510. <https://doi.org/10.1177/0003122417705874>.
- McDonnell, Erin. 2020. *Patchwork Leviathan: Pockets of Bureaucratic Effectiveness in Developing States*. Princeton, NJ: Princeton University Press.
- McGovern, Mike. 2011. "Popular Development Economics: An Anthropologist among the Mandarins." *Perspectives on Politics* 9 (2): 345–55. <https://doi.org/10.1017/S1537592711000594>.
- Merry, Sally Engle, Kevin E. Davis, and Benedict Kingsbury, eds. 2015. *The Quiet Power of Indicators: Measuring Governance, Corruption, and Rule of Law*. New York: Cambridge University Press.
- Milante, Gary, and Michael Woolcock. 2017. "New Approaches to Identifying State Fragility." *Journal of Globalization and Development* 8 (1): 20170008. <https://doi.org/10.1515/jgd-2017-0008>.
- Mintzberg, Henry. 2015. "If You Can't Measure It, You'd Better Manage It." *Henry Mintzberg* (blog). May 28, 2015. <https://mintzberg.org/blog/measure-it-manage-it>.
- Natsios, Andrew. 2011. "The Clash of the Counter-Bureaucracy and Development." Essay, Center for Global Development, Washington, DC. [https://www.cgdev.org/sites/default/files/1424271\\_file\\_Natsios\\_Counterbureaucracy.pdf](https://www.cgdev.org/sites/default/files/1424271_file_Natsios_Counterbureaucracy.pdf).

- Porter, Theodore M. 1995. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.
- Pritchett, Lant. 2013. *The Rebirth of Education: Schooling Ain't Learning*. Washington, DC: Center for Global Development.
- Pritchett, Lant. 2014. "The Risks to Education Systems from Design Mismatch and Global Isomorphism: Concepts, with Examples from India." WIDER Working Paper 2014/039, United Nations University World Institute for Development Economics Research, Helsinki.
- Propper, Carol, Simon Burgess, and Denise Gossage. 2008. "Competition and Quality: Evidence from the NHS Internal Market 1991–9." *The Economic Journal* 118 (525): 138–70. <https://doi.org/10.1111/j.1468-0297.2007.02107.x>.
- Rao, Vijayendra. 2022. "Can Economics Become More Reflexive? Exploring the Potential of Mixed-Methods." Policy Research Working Paper 9918, World Bank, Washington, DC.
- Rao, Vijayendra, Kripa Ananthpur, and Kabir Malik. 2017. "The Anatomy of Failure: An Ethnography of a Randomized Trial to Deepen Democracy in Rural India." *World Development* 99 (11): 481–97. <https://doi.org/10.1016/j.worlddev.2017.05.037>.
- Ravallion, Martin. 2001. "Growth, Inequality and Poverty: Looking Beyond Averages." *World Development* 29 (11): 1803–15. [https://doi.org/10.1016/S0305-750X\(01\)00072-9](https://doi.org/10.1016/S0305-750X(01)00072-9).
- Ravallion, Martin. 2020. "Pandemic Policies in Poor Places." CGD Note, April 24, Center for Global Development, Washington, DC.
- Ridgway, V. F. 1956. "Dysfunctional Consequences of Performance Measurements." *Administrative Science Quarterly* 1 (2): 240–7. <https://doi.org/10.2307/2390989>.
- Robinson, Ken, and Lou Aronica. 2015. *Creative Schools: Revolutionizing Education from the Ground Up*. London: Penguin UK.
- Rocha Menocal, Alina, Marc Cassidy, Sarah Swift, David Jacobstein, Corinne Rothblum, and Ilona Tservil. 2018. *Thinking and Working Politically through Applied Political Economy Analysis: A Guide for Practitioners*. Washington, DC: USAID, DCHA Bureau Center of Excellence on Democracy, Human Rights, and Governance. [https://usaidlearninglab.org/sites/default/files/resource/files/pea\\_guide\\_final.pdf](https://usaidlearninglab.org/sites/default/files/resource/files/pea_guide_final.pdf).
- Rogers, Patricia, and Michael Woolcock. 2023. "Process and Implementation Evaluation Methods." In *Oxford Handbook of Program Design and Implementation*, edited by Anu Rangarajan, 294–316. New York: Oxford University Press.
- Samji, Salimah, Matt Andrews, Lant Pritchett, and Michael Woolcock. 2018. *PDIAtoolkit: A DIY Approach to Solving Complex Problems*. Cambridge, MA: Building State Capability, Center for International Development, Harvard University. <https://bsc.cid.harvard.edu/PDIAtoolkit>.
- Sandefur, Justin, and Amanda Glassman. 2015. "The Political Economy of Bad Data: Evidence from African Survey and Administrative Statistics." *The Journal of Development Studies* 51 (2): 116–32. <https://doi.org/10.1080/00220388.2014.968138>.
- Sanyal, Paromita, and Vijayendra Rao. 2018. *Oral Democracy: Deliberation in Indian Village Assemblies*. New York: Cambridge University Press.
- Scott, James C. 1998. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven, CT: Yale University Press.
- Soffel, Jenny. 2016. "The 21st-Century Skills Every Student Needs." World Economic Forum. <https://www.weforum.org/agenda/2016/03/21st-century-skills-future-jobs-students/>.
- Sparrow, Malcolm. 2018. "Problem-Oriented Policing: Matching the Science to the Art." *Crime Science* 7 (1): 1–10. <https://doi.org/10.1186/s40163-018-0088-2>.
- Teskey, Graham. 2017. "Thinking and Working Politically: Are We Seeing the Emergence of a Second Orthodoxy?" Governance Working Paper Series 1, ABT Associates, Canberra.
- Teskey, Graham, and Lavinia Tyrrel. 2017. "Thinking and Working Politically in Large, Multi-Sector Facilities: Lessons to Date." Governance Working Paper Series 2, ABT Associates, Canberra.
- Weber, Eugen. 1976. *Peasants into Frenchmen: The Modernization of Rural France, 1870–1914*. Palo Alto, CA: Stanford University Press.
- Wild, Leni, David Booth, and Craig Valters. 2017. *Putting Theory into Practice: How DFID Is Doing Development Differently*. London: ODI.
- WilmerHale. 2021. *Investigation of Data Irregularities in "Doing Business 2018" and "Doing Business 2020": Investigation Findings and Report to the Board of Executive Directors*. <https://thedocs.worldbank.org/en/doc/84a922cc9273b7b120d49ad3b9e9d3f9-0090012021/original/DB-Investigation-Findings-and-Report-to-the-Board-of-Executive-Directors-September-15-2021.pdf>.
- Wilson, James Q. 1989. *Bureaucracy: What Government Agencies Do and Why They Do It*. New York: Basic Books.

- Woolcock, Michael. 2018. "Reasons for Using Mixed Methods in the Evaluation of Complex Projects." In *Contemporary Philosophy and Social Science: An Interdisciplinary Dialogue*, edited by Michiru Nagatsu and Attilia Ruzzene, 149–71. London: Bloomsbury Academic.
- World Bank. 2018. *World Development Report 2018: Learning to Realize Education's Promise*. Washington, DC: World Bank.
- World Economic Forum. 2016. *New Vision for Education: Fostering Social and Emotional Learning through Technology*. Industry Agenda. Geneva: World Economic Forum.
- Zak, Paul. 2013. "Measurement Myopia." Drucker Institute, Claremont, CA. December 6, 2021. <https://www.drucker.institute/thedx/measurement-myopia/>.
- Zhao, Yong. 2012. "Test Scores vs. Entrepreneurship: PISA, TIMSS, and Confidence." *Yong Zhao* (blog). <http://zhaolearning.com/2012/06/06/test-scores-vs-entrepreneurship-pisa-timss-and-confidence/>.



## CHAPTER 5

# Practical Tools for Effective Measurement and Analytics

*Maria Ruth Jones and Benjamin Daniels*

### SUMMARY

Increasingly important features of good data analysis are the transparency with which the analysis is undertaken and the reproducibility of its results. These features ensure the credibility of analytical outputs and policy guidance. The World Bank's Development Impact Evaluation (DIME) Department has developed freely available tools and processes to support the achievement of these best practices by analysts across the world. These resources include research-cycle frameworks, extensive training tools, detailed archives of process and technical guidance, and a collaborative approach to data and analytics. The DIME Analytics team continuously updates many of these resources and makes them available globally as a free knowledge product. This chapter describes the frameworks, the approach, and the products that are available to bring these best practices into any organization that relies on data analytics for decision-making. The chapter provides a discussion of how to apply these elements to public administration, thus ensuring that analytics of government accord with international best practices.

### ANALYTICS IN PRACTICE

- For credibility, modern methods for data analysis rely on analysts undertaking their work in a transparent way that ensures their results can be replicated.
- Best practices for reproducibility and transparency assure the internal quality and organization of data work and provide a template for publishing materials externally when appropriate.
- Producing analysis that accords with best practices requires considering the full life cycle of data work, such that each stage of handling data can be designed to support the next stages.

---

Maria Ruth Jones is a senior economist for the World Bank's Development Impact Evaluation (DIME) Department, where she coordinates DIME Analytics. Benjamin Daniels is a research fellow in the gui<sup>2</sup>de group at the McCourt School of Public Policy at Georgetown University, who works with DIME Analytics.

- Development Impact Evaluation (DIME) Analytics, a unit of the DIME Department at the World Bank, has developed extensive, publicly available tools and trainings to standardize and promulgate these best practices among analysts across the world.
- This chapter describes these resources and links to their locations online. It provides a discussion of how to apply these elements to public administration, thus ensuring government analytics accord with international best practices.

## INTRODUCTION

The principles of good data analysis and empirical research have been refined over centuries. Today's analysts and researchers work in an environment in which adherence to modern standards of analysis is an essential part of the credibility of their results. As such standards become more abundant and layered, it can be a challenge to keep up to date on best practices. These considerations go beyond the internal analysis of government data and are of general concern to anyone attempting to undertake rigorous analysis.

To support analysts and researchers across the world in understanding and implementing modern approaches to data analysis, the World Bank's Development Impact Evaluation (DIME) Department has created a series of resources to support the adoption of best practices in data collection and analysis. The DIME Analytics unit aims to help data analysts identify inefficiencies and practices that compromise the quality of analysis and to develop a workflow that strengthens their work. DIME Analytics creates and tests tools that support these practices and provides the training and technical support necessary to sustain their adoption (illustrated in figure 5.1). This chapter aims to provide an introduction to the resources (typically free) DIME Analytics provides, which are in themselves an introduction to the global frontier of best practices in data analytics.

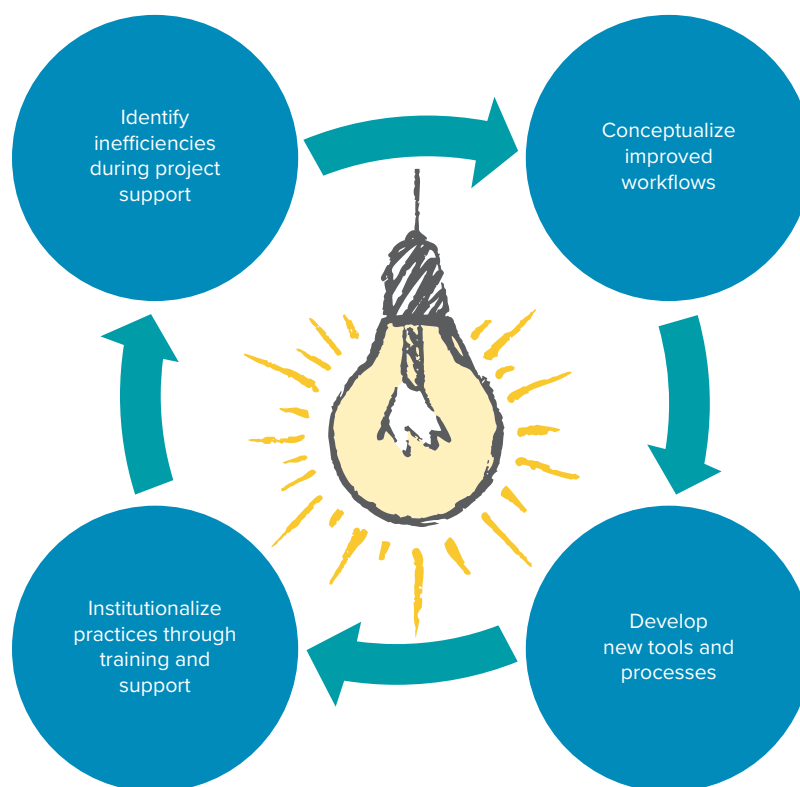
Creating high-quality data and research outputs is central to generating useful insights to inform public administration. DIME Analytics' resources are well suited to those interested in updating their administration's capacity to collect high-quality data and evaluate the impact of public administration reforms. This chapter therefore describes protocols for high-quality data collection relevant to all public administration data collection efforts and discusses how best practices for reproducible and transparent data collection and analysis can establish the credibility of public administration analysis and research.

## ESTABLISHING THE CREDIBILITY OF PUBLIC ADMINISTRATION MEASUREMENT

The framework of reproducibility, transparency, and credibility has become the basis for empirical research and data analysis worldwide. There are several core reasons for the adoption of these principles in the academic world, and most of them apply equally to the sphere of policy analysis, design, and governance. In academic research, collaborators are often not employed in the same organization, or direct collaboration may never occur. These standards have been adopted so that disparate teams and individuals can access, reuse, assess, and build upon the work done by other researchers in the ecosystem, even when they do not have institutional or personal connections to one another. A similar argument can be made of the many distinct units of government administrations.

Reproducible data work is designed so that the materials and processes from a task can be easily reused and adapted within and across contexts. Such data work is done with the goals of collaboration and the handoff of materials in mind, even if the research materials are not intended for public consumption. When analytical processes are designed for collaboration and reuse, they become a form of knowledge

**FIGURE 5.1** DIME Analytics Uses an Iterative Process to Expand Technical Capacity throughout the Research Cycle



Source: DIME (Development Impact Evaluation) Analytics, World Bank.

accumulation within an organization. By contrast, data processes that are completed as “one-off” tasks—nonreproducible processes that cannot be transferred to other people after completion—cannot transfer workflow improvements and efficiencies to other tasks. Reproducible data work also enables more people to work on a task because the tools can be handed off among individuals on the team; it also enables more people to conduct quality control over the work of others, which is a built-in feature of reusing others’ work.

Transparent work is designed so the materials and processes from a task can be understood and used by individuals not involved in the original task. In research, this often means publishing or releasing code, data, and documentation to the general public (or under license to other researchers). However, publication is not the most important aspect of transparent work. The essential characteristic of transparent work is that the documentation of all steps is complete, the organization of subtasks is easily understood, and the complete materials are archived in a known location. In this way, a transparent process is one that can be accessed and replicated by a new team without contacting the original analysts. This approach allows knowledge transfer to happen rapidly both across an organization, from team to team, and within an organization as materials are reused and improved upon progressively over time.

Together, these two approaches produce a foundation of credibility for conclusions and outputs within an organization. Reproducible work usually demands the adoption of standardized processes for subtasks. Processes that are reused frequently and by many people will attract higher levels of investment and quality assurance. Similarly, when data, code, and documentation are intended to be seen and reused by others, authors are incentivized to improve their own organization and description of materials, particularly if they or their own team will be the ones reusing them in the future. When these materials are archived and are visible to other members

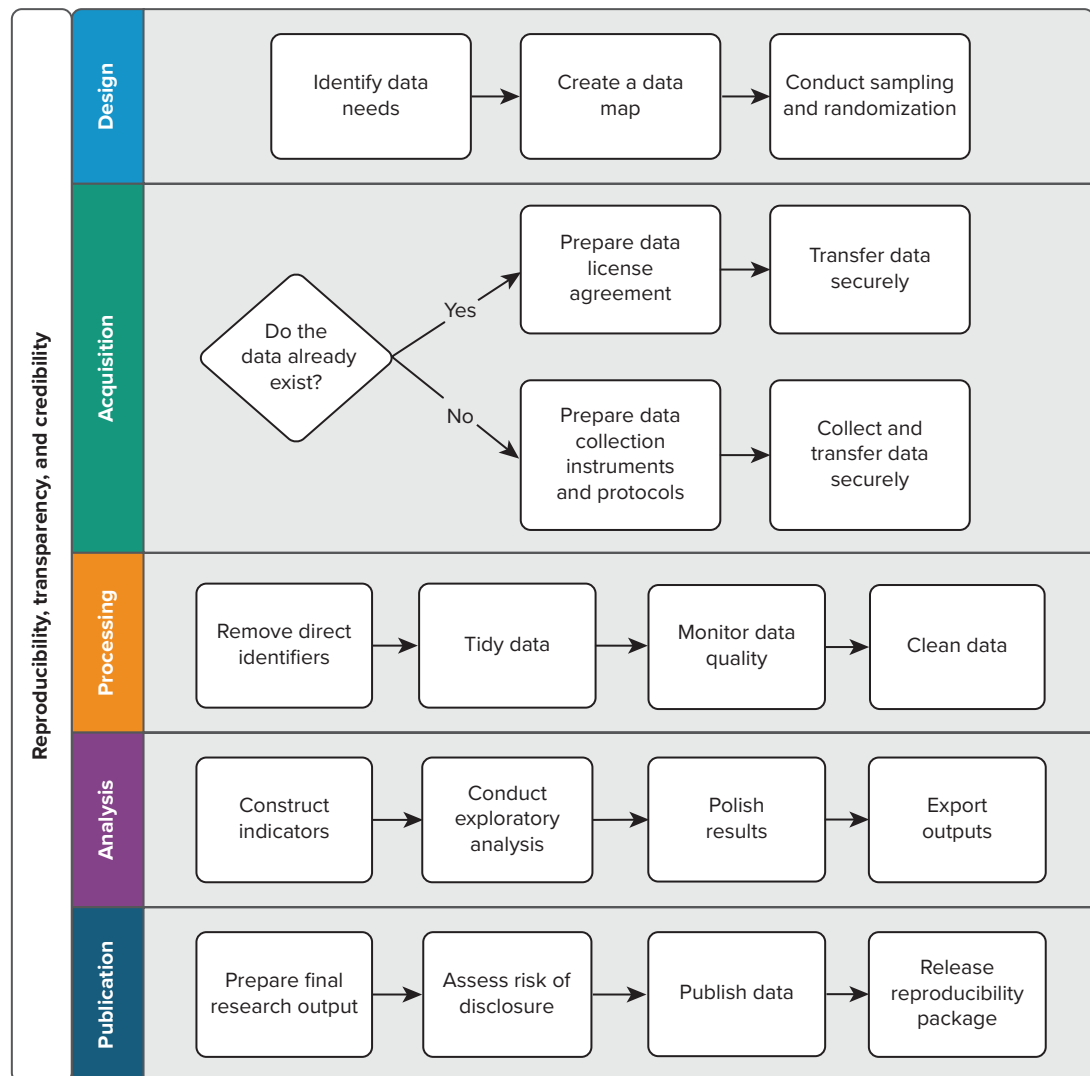


of an organization, they can be quality-controlled, and, over time, the best work becomes widely improved and adopted, establishing the reputation of high-quality tools and processes. Whether or not analytical materials are ever released to the public is secondary to adopting a mindset of collaboration and reuse across a research organization.

## PROTOCOLS FOR HIGH-QUALITY DATA COLLECTION

The conventions, standards, and best practices that are fast becoming a necessity for high-quality empirical research affect most elements of data collection and analysis. Figure 5.2 describes the standard workflow of analysis from design to reporting or publication. Some of these steps may be redundant, such as when using administrative data, but considering the full workflow allows analysts to imagine what the optimal process would have been if they had been in control of each stage of data collection and analysis.

**FIGURE 5.2** Overview of the Tasks Involved in Development Research Data Work



Source: Bjärkefur et al. 2021.

DIME Analytics has described best practices at each of these stages in its published handbook, *Development Research in Practice: The DIME Analytics Data Handbook* (Björkefur et al. 2021). *The DIME Analytics Data Handbook* is intended to train users of data in effective, efficient, and ethical data handling. It covers the full data workflow for an empirical project, compiling all the lessons and tools developed by DIME Analytics into a single narrative of best practices. It provides a step-by-step guide to high-quality, reproducible data work at each stage of the data workflow, from design to publication, as visualized in figure 5.2. Each chapter contains boxes with examples of how the practices and workflows described in that chapter were applied in a real-life case. The handbook is available for free download through the World Bank's Open Knowledge Repository or for purchase from Amazon. The handbook and related resources will be updated over time as best practices evolve, and feedback can always be provided through the handbook's GitHub repository.<sup>1</sup>

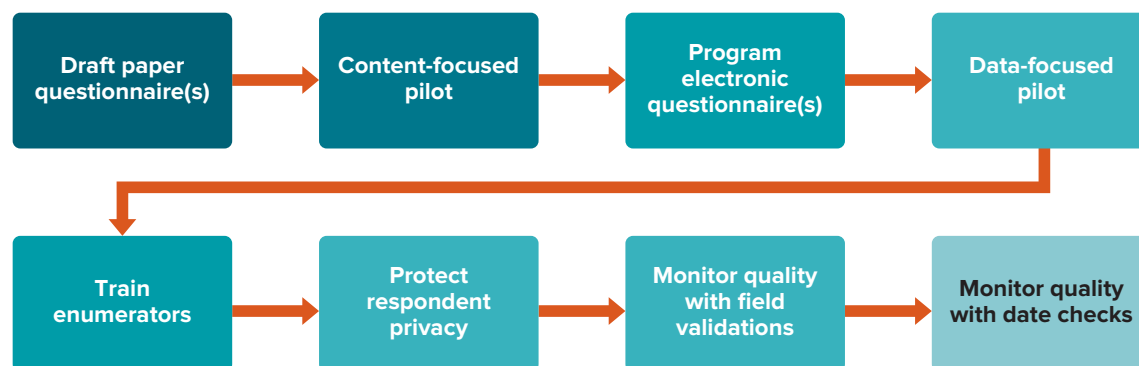
Let us take as an example the element “Prepare data collection instruments and protocols” in the “Acquisition” pillar in figure 5.2. This activity can be further broken down into constituent parts. The handbook provides narrative descriptions and best practices for each stage of the activity. It also provides links to corresponding entries on the DIME Wiki, which contain more specific technical details and are kept up to date as these details change and evolve.<sup>2</sup> Figure 5.3 illustrates the full workflow for an electronic survey.

Each of these stages can then be broken down further into constituent activities. Figure 5.4 summarizes the key protocols for ensuring data quality at every stage of the survey workflow. In this way, for each activity related to data collection, DIME Analytics has outlined the activities required to undertake rigorous data collection and analysis and has provided a knowledge network for analysts to obtain the appropriate level of specific detail about each task and subtask for their current needs. Adhering to such standardized and organized protocols, even when analysis is not intended to be made public, ensures that work is organized and internally consistent. In this way, the DIME Analytics knowledge products reflect the world's frontier knowledge about how to obtain credible empirical results.

As an example of the application of these principles in public administration, consider an assessment of the impact of in-service training. Much effort is made to keep the skills of public servants current with new procedures within government and innovations in wider society. However, there is frequently little evaluation of whether such training actually has an impact on the quality of government processes or productivity. An effective measurement framework might include an immediate assessment of public servants' skills upon entry to and exit from the training, as well as a follow-up with objective assessments of the quality of procedure and productivity.

To assess the broad impact of an in-service training intervention, we could imagine returning to the unit of the trainee after some time had passed and assessing the impact of the training through a survey

**FIGURE 5.3 Workflow for High-Quality Survey Data Collection**



Source: DIME (Development Impact Evaluation) Analytics, World Bank.

**FIGURE 5.4** Summary Protocols for High-Quality Data Collection at Every Stage of the Survey Workflow



Source: DIME (Development Impact Evaluation) Analytics, World Bank.

of the unit's public servants. Working through the elements of figure 5.4, this would require the following steps:

- **Prepare for a survey:** Ensure that we have a sufficiently large sample of trainees to detect a quantitative increase in their knowledge. Determine which organizations or units we want to survey and any sub-groups of respondents we want to focus on (managers vs. nonmanagers, contract workers vs. staff, etc.). Obtain organizational structure charts if necessary.
- **Draft survey instrument:** Develop questions sufficiently close to the concepts of interest to provide us with measures we can act on. We might ask the trainees questions about how they have used the new

procedures in their daily work, and we might ask their colleagues whether this has made them easier to work with on related tasks. Basing our survey on existing surveys of these trainees would allow us to track the impacts of training across time.

- **Pilot:** Test intended questions with the audience of interest so the trainees understand what we are asking about. Though we might be asking about the actual practice of using the new procedures, the wrong question will make trainees respond about the change in rules rather than whether they are actually being followed.
- **Train enumerators:** If data collection is managed by enumerators, ensure they understand the appropriate protocols of the public service and take a consistent approach to measurement. If trainees differ in seniority, how will an enumerator ensure a common approach to surveying across the hierarchy?
- **Protect respondents' privacy:** Create protocols to ensure the process protects respondents' privacy. Individual public servants' capability at aspects of their jobs is highly sensitive information. If officials believe that their data will not be completely private, they may refuse to cooperate with the data collection exercise.
- **Monitor quality with field validations and/or data checks:** Implement a system that monitors survey data as they come in, and monitor whether a specific enumerator, department, or agency is presenting unusual data.

For a more detailed explanation of the protocols associated with surveys, please refer to chapter 4 of *The DIME Analytics Data Handbook* (Bjärkefur et al. 2021) and the videos and lectures from the Manage Successful Impact Evaluation Surveys course (DIME Analytics 2020a, 2021b).<sup>3</sup> To learn how to implement high-quality surveys in practice, please refer to the DIME Wiki articles on Primary Data Collection, Field Surveys, Survey Protocols, and Remote Surveys.<sup>4</sup> For specific considerations regarding phone surveys, please refer to chapters 1 and 2 of *Mobile Phone Panel Surveys in Developing Countries: A Practical Guide for Microdata Collection* (Dabalen et al. 2016).

## PUBLIC RESOURCES AND TOOLS

To support the implementation of a rigorous research process, DIME Analytics has made a range of resources, technical solutions, and research protocols available through open-access trainings and open-source tools. We have found that there is significant unmet demand for these public goods, demonstrated by fast-growing and widespread global interest in our offerings. This section describes DIME Analytics' flagship resources.

### Development Research in Practice: The Course

*The DIME Analytics Data Handbook* is accompanied by the course Development Research in Practice.<sup>5</sup> This free and fully virtual course lasts 8 weeks, with seven lecture weeks each corresponding to one of the chapters from the handbook. This course provides attendees with a high-level overview of the entire process of empirical research so that they understand how each stage of the research workflow fits among the others and how the framework of transparency, reproducibility, and credibility informs the entire structure. Each week presents a motivational video with paired readings from the handbook and the DIME Wiki, a detailed lecture and Q&A session with DIME

Analytics team members on the topic of the week, and a knowledge assessment for attendees. The course will be offered annually, and all course materials are publicly available for self-study (DIME Analytics 2021a).

### Manage Successful Impact Evaluation Surveys

Manage Successful Impact Evaluation Surveys is a free, virtual course in which participants learn and practice the workflows involved in primary data collection. It acts as a complement to the Development Research in Practice course, providing a deep dive into the processes that are described at a high level in chapter 4 of *The DIME Analytics Data Handbook*. The course covers best practices at all stages of the survey workflow, from planning to piloting instruments and monitoring data quality once fieldwork begins. There is a strong emphasis throughout on research ethics and reproducible workflows. During the global pandemic, a special module focused on adapting surveys to remote data collection. Participants learn to plan for and prepare successful surveys, design high-quality survey instruments, effectively train surveyors (including remote training), monitor survey implementation, ensure high-quality data, and handle confidential data securely, among other topics. The course uses a combination of virtual lectures, readings, and hands-on exercises. It is offered annually, and all course materials are available online for self-study (DIME Analytics 2020a, 2021b).

### Measuring Development Conference

DIME Analytics annually invites a diverse group of attendees to become part of a community of analysts interested in innovations in measurement. Measuring Development is an annual conference organized jointly by DIME, the World Bank's Development Economics Data Group, and the Center for Effective Global Action at the University of California, Berkeley. The conference focuses on data and measurement innovations across different sectors and themes. It was held virtually and was open to the public in 2020 and 2021. The focus for 2021 was "Emerging Data and Methods in Global Health Research" (a summary blog can be found in Jones, Fishman, and Reschechko 2021). Previous topics have included "Data Integration and Data Fusion" (2020), "Crisis Preparedness and Response" (2019), and "Artificial Intelligence and Economic Development" (2018).<sup>6</sup>

### DIME Wiki

The DIME Wiki is a one-stop shop for resources on the best practices and resources across all phases of an impact evaluation (IE): design, fieldwork, data, and analysis. It focuses on practical implementation guidelines rather than theory. With over 200 content pages, the DIME Wiki is open to the public, easily searchable, and suitable for users of varying levels of expertise. The DIME Wiki is closely linked to *The DIME Analytics Data Handbook*; the handbook links to the DIME Wiki for implementation details and specific examples of the best practices it outlines.<sup>7</sup>

### Stata Packages

If you use Stata (a statistical package many researchers use for their analysis), you may be interested in the DIME Analytics Stata software packages, *ietoolkit* and *iefieldkit*. These packages are a direct result of DIME Analytics' efforts to identify inefficiencies and common sources of error in data workflows and to create tools that routinize best practices. The *ietoolkit* package includes standardized commands for data analysis tasks that are common in DIME work. The *iefieldkit* package includes commands related to primary data

collection that create rapid, standardized, and well-documented data acquisition workflows. These statistical packages can be installed through the Statistical Software Components (SSC) archive. The source code for `ietoolkit` and `iefieldkit` is available for public review and contribution via GitHub.<sup>8</sup> By using these standardized commands, users avoid repeating common mistakes and produce more efficient code as well as share common workflows for major tasks.

## Data Visualization Libraries

If you are producing graphics from your data, you may be interested in the DIME Analytics visual libraries for Stata and R users (Andrade et al. 2020).<sup>9</sup> The libraries contain example code for model data visualizations in an easy-to-browse format. These libraries help researchers reduce the time they spend creating professional-quality, reproducible graphs and maps. Both the Stata library and the R library are open to contributions through GitHub.<sup>10</sup>

## Technical Trainings

DIME Analytics provides regular technical trainings to World Bank staff and consultants. All training materials are shared with the public through the Open Science Framework platform; self-study is encouraged, and DIME Analytics provides support to independent learners by answering questions and responding to feedback in the relevant GitHub repositories. To access an index of these trainings and browse all materials, visit the DIME Analytics profile page on the Open Science Framework website.<sup>11</sup>

## Research Assistant Onboarding Course

The Research Assistant Onboarding Course is designed to familiarize research assistants (or, in a public administration setting, junior analysts) with best practices for reproducible research.<sup>12</sup> By the end of the course's six sessions, participants have the tools and knowledge to implement these best practices and to set up a collaborative workflow for code, data sets, and research outputs. Most content is platform independent and software agnostic, though participants are expected to be familiar with statistical software. The course is offered twice yearly to World Bank staff and consultants; course materials are available to the public (DIME Analytics 2020b).

## R for Advanced Stata Users

The R for Advanced Stata Users course provides an introduction to the R programming language, building on knowledge of Stata.<sup>13</sup> The course focuses on common tasks in development research related to descriptive analysis, data visualization, data processing, and geospatial data work. The course is offered twice yearly to World Bank staff and consultants; course materials are available to the public (DIME Analytics 2019).

## Continuing Education Series

DIME Analytics offers a biweekly Continuing Education Series in the fall and spring semesters. The trainings are typically hands-on, and the topics are decided based on a review of common issues faced by DIME project teams. For example, in 2020, DIME Analytics delivered 11 Continuing Education sessions on topics including “Data Quality Monitoring,” “Working with Spatial Data in Stata,” “Optimizing Survey Length,” “GitHub Pull Requests,” and “Introduction to Python for Stata Users” (DIME Analytics 2020–2023).

## CONCLUSION

The World Bank's DIME Analytics team has made a sustained, years-long effort to implement the principles of transparency, reproducibility, and credibility across the cycle of data work. It has taken an iterative approach. First, through direct engagement with analytical teams, DIME Analytics identifies processes that are both common and ad hoc, such that the whole organization would benefit from standardization. Then, the DIME Analytics team works to understand the essential needs of the analytics teams for the process or subtask, mapping each process as part of a research flowchart and defining the inputs and outputs that are desired for the process. Next, the team either identifies an external tool or process that can be utilized or develops its own tools or guidelines that are appropriate to data work. Finally, the materials are documented, archived, and disseminated into team workflows through the frequent training and support sessions the team organizes.

In this way, DIME as an organization has accumulated a knowledge base of high-quality analytical tools and standardized processes and best practices that are used across almost all its projects. For the DIME Department, these resources form a visible foundation that makes DIME research well known for its attention to quality and underscores the reliability of DIME research products. Additionally, an essential element of the mission of DIME and the World Bank is to produce global public resources for high-quality analytical evidence. In response to this mission, DIME Analytics makes its tools, processes, and trainings publicly available whenever possible and enables them to be self-paced and remote to the greatest practical extent. This chapter has summarized some of the key features and philosophies of the DIME Analytics approach and has offered these resources to readers so they may use and access whatever materials are helpful to them.

## NOTES

We acknowledge the initiative of Arianna Legovini (alegovini@worldbank.org) and Florence Kondylis (fkondylis@worldbank.org) for their creation of and ongoing support for the DIME Analytics team, and Vincenzo di Maro (dimaro@worldbank.org) for his leadership and management. We especially acknowledge the work of Kristoffer Bjärkefur (kbjarkefur@worldbank.org), Luíza Cardoso de Andrade (lcardoso@worldbank.org), and Roshni Khincha (rkhincha@worldbank.org) as members of that team, as well as the support and assistance of Avnish Singh (asingh42@worldbank.org), Patricia Paskov (ppaskov@worldbank.org), Radhika Kaul (rkaul1@worldbank.org), Mizuhiro Suzuki (msuzuki1@worldbank.org), Yifan Powers (ypowers@worldbank.org), and Maria Arnal Canudo (marnalcanudo@worldbank.org). We further recognize all members of the DIME team, past and present, for their contributions to this portfolio and all the World Bank teams and external support we have graciously received. DIME Analytics has been financially supported by the United Kingdom Foreign, Commonwealth and Development Office (FCDO) through the DIME i2i Umbrella Facility for Impact Evaluation at the World Bank.

1. *The DIME Analytics Data Handbook* GitHub repository can be found at <https://worldbank.github.io/dime-data-handbook/>.
2. The DIME Wiki is maintained by DIME Analytics and can be found at <https://dimewiki.worldbank.org/>.
3. More information about the Manage Successful Impact Evaluation Surveys course is available on the World Bank website at <https://www.worldbank.org/en/events/2021/05/24/manage-successful-impact-evaluation-surveys>.
4. DIME Wiki, s.v. "Primary Data Collection," last modified May 6, 2022, 17:02, [https://dimewiki.worldbank.org/Primary\\_Data\\_Collection](https://dimewiki.worldbank.org/Primary_Data_Collection); DIME Wiki, s.v. "Field Surveys," last modified May 6, 2021, 17:27, [https://dimewiki.worldbank.org/Field\\_Surveys](https://dimewiki.worldbank.org/Field_Surveys); DIME Wiki, s.v. "Survey Protocols," last modified July 11, 2022, 16:18, [https://dimewiki.worldbank.org/Survey\\_Protocols](https://dimewiki.worldbank.org/Survey_Protocols); DIME Wiki, s.v. "Remote Surveys," last modified April 7, 2022, 17:35, [https://dimewiki.worldbank.org/Remote\\_Surveys](https://dimewiki.worldbank.org/Remote_Surveys).
5. More information about the Development Research in Practice course is available on the World Bank website at <https://www.worldbank.org/en/events/2021/07/12/development-research-in-practice>.
6. More information about the 2018 and 2019 conferences can be found on the World Bank website at <https://www.worldbank.org/en/events/2018/01/29/artificial-intelligence-for-economic-development> and <https://www.worldbank.org/en/events/2019/03/27/crisis-preparedness-and-response>. More information about the 2020 and 2021 conferences can be found on the website of the Center for Effective Global Action at <https://cega.berkeley.edu/event/annual-conference-on-measuring-development-vi-data-integration-and-data-fusion/> and <https://cega.berkeley.edu/event/measuring-development-2021-emerging-data-and-methods-in-global-health-research/>.



7. The DIME Wiki can be found at <https://dimewiki.worldbank.org/>.
8. The GitHub repository for ietoolkit is available in the World Bank's repository at <https://github.com/worldbank/ietoolkit>. The GitHub repository for iefieldkit is available in the World Bank's repository at <https://github.com/worldbank/iefieldkit>.
9. The Stata visual library is available at <https://worldbank.github.io/stata-visual-library/>. The R visual library is available at <https://worldbank.github.io/r-econ-visual-library/>.
10. The GitHub repository for the Stata visual library is available in the World Bank's repository at <https://github.com/worldbank/stata-visual-library>. The GitHub repository for the R visual library is available on the World Bank's repository at <https://github.com/worldbank/r-econ-visual-library>.
11. The DIME Analytics profile page on the Open Science Framework website can be found at <https://osf.io/wzjtk>.
12. More information about the Research Assistant Onboarding Course is available at <https://www.worldbank.org/en/events/2020/01/30/ra-onboarding-course>.
13. More information about the R for Advanced Stata Users course can be found on the World Bank website at <https://www.worldbank.org/en/events/2019/04/16/R-for-advanced-stata-Users>.

## REFERENCES

- Andrade, Luíza, Maria Jones, Florence Kondylis, and Luis Eduardo (Luise) San Martin. 2020. "New Visual Libraries for R and Stata Users." *Development Impact* (blog). *World Bank Blogs*, October 15, 2020. <https://blogs.worldbank.org/impacitevaluations/new-visual-libraries-r-and-stata-users>.
- Björkefur, Kristoffer, Luíza Cardoso de Andrade, Benjamin Daniels, and Maria Ruth Jones. 2021. *Development Research in Practice: The DIME Analytics Data Handbook*. Washington, DC: World Bank. <https://openknowledge.worldbank.org/handle/10986/35594>.
- Dabalen, Andrew, Alvin Etang, Johannes Hoogeveen, Elvis Mushi, Youdi Schipper, and Johannes von Engelhardt. 2016. *Mobile Phone Panel Surveys in Developing Countries A Practical Guide for Microdata Collection*. Directions in Development. Washington, DC: World Bank. <http://documents.worldbank.org/curated/en/877231468391801912/Mobile-phone-panel-surveys-in-developing-countries-a-practical-guide-for-microdata-collection>.
- DIME Analytics. 2019. "R for Advanced Stata Users." Recordings and Course Materials from Virtual Course Held April 16, 2019, World Bank, Washington, DC. Created March 17, 2020, and last updated May 31, 2022. Open Science Framework. <https://osf.io/nj6bf/>.
- DIME Analytics. 2020a. "Manage Successful Impact Evaluation Surveys." Recordings and Course Materials from Virtual Course Held July 6–24, 2020, World Bank, Washington, DC. Created June 3, 2020, and last updated May 31, 2022. Open Science Framework. <https://osf.io/resya/>.
- DIME Analytics. 2020b. "Research Assistant Onboarding Course." Recordings and Course Materials from Virtual Course Held January 30–March 5, 2020, World Bank, Washington, DC. Created March 31, 2020, and last updated May 31, 2022. Open Science Framework. <https://osf.io/qtmdb/>.
- DIME Analytics. 2020–23. "DIME Continuing Education." Recordings and Course Materials for Biweekly Virtual Course Series, World Bank, Washington, DC. Created March 20, 2020, and last updated February 17, 2023. Open Science Framework. <https://osf.io/8sgrh/>.
- DIME Analytics. 2021a. "Development Research in Practice." Recordings and Course Materials from Virtual Course Held July 12–September 3, 2021, World Bank, Washington, DC. Created June 24, 2021, and last updated May 31, 2022. Open Science Framework. <https://osf.io/6fsz3/>.
- DIME Analytics. 2021b. "Manage Successful Impact Evaluation Surveys." Recordings and Course Materials from Virtual Course Held May 24–June 11, 2021, World Bank, Washington, DC. Created April 30, 2021, and last updated May 31, 2022. Open Science Framework. <https://osf.io/672ej/>.
- Jones, Maria, Sam Fishman, and Yevanit Reschekko. 2021. "Data-Driven Global Health Research in the Time of COVID." *Data Blog*. *World Bank Blogs*, April 26, 2021. <https://blogs.worldbank.org/opendata/data-driven-global-health-research-time-covid>.



## CHAPTER 6

# The Ethics of Measuring Public Administration

*Annabelle Wittels*

### SUMMARY

As data collection *on* government *within* government becomes more prevalent, a review of research on data ethics fit for use within public administration is needed. While guides on data ethics exist for public sector employees, as well as guides on the use of data about citizens, there is a dearth of discussion and few practical guides on the ethics of data collection by governments about their own employees. When collecting data about their employees, public administrations face ethical considerations that balance three dimensions: an individual dimension, a group dimension, and a public-facing dimension. The individual dimension comprises demands for dignity and privacy. The group dimension allows for voice and dissent. The public-facing dimension ensures that analytics enable public servants to deliver on public sector values: accountability, productivity, and innovation. The chapter uses this heuristic to investigate ethical questions and provide a tool (in appendix B) with a 10-point framework for governments to guide the creation of fair and equitable data collection approaches.

### ANALYTICS IN PRACTICE

- Collecting data on government employees involves a different set of challenges from collecting data on service users and private sector employees.
- Depriving public sector employees of privacy can erode democratic principles because employees may lose spaces to dissent and counteract malpractice pursued by powerful colleagues, managers, or political principals.
- Designing data collection in a way that enhances dignity serves several functions. For one, if we understand the problem of privacy in terms of disruptions of autonomy, collecting data in ways that enhance dignity can help to move away from zero-sum thinking. If public sector employees gain dignity from data collection, they are unlikely to see it as an unwanted intrusion. For instance, data could be collected that celebrate personal initiative, good management practices, and outstanding achievements, as opposed to disciplinary uses of data, such as identifying malpractice or inefficiencies. In some cases, employers

---

Annabelle Wittels is an independent researcher and former consultant in the World Bank's Development Impact Evaluation (DIME) Department.

might also consider disclosing the identities of participating individuals, if they consent, to highlight valuable contributions and give credit.

- Despite improved data quality, which helps to produce efficiencies, value-based decision-making will not become obsolete. Negotiating values is required as much as ever to produce evidence-based and ethically sound policy (Athey 2017).
- Development practitioners and donor countries working on data strategies for public sector reform in countries where political, religious, or other basic human freedoms are not guaranteed must thus tread carefully to guard against setting up data infrastructures that can be used to the detriment of public sector employees.
- Navigating ethical dilemmas sustainably requires that, when individuals join social groups in which different norms on data privacy are applied, they do so knowingly and are provided with the opportunity to advocate to change these norms. In practice, this can mean giving public sector staff unions a voice in what data are made available about their members and what guarantees of accountability they can offer the public in place of such data.
- Creating data approaches for public sector innovation thus requires that time and resources be set aside to make the process explicable to those affected. This is no simple task because governments still face skills gaps in cutting-edge areas of information technology. In many instances, governments will need to rely on external expertise to develop and maintain the skills of their staff to implement data solutions that are ethically sound and secure.
- What is considered ethical and morally right can depend on context. There are, however, questions that provide general guidance for how measurement can be conducted in an ethically sound manner, if they are asked regularly at key junctures of data collection, analysis, and reporting.
- It is important to construct objective measures of organizational and/or individual performance rather than relying only on subjective evaluations, such as performance appraisals.
- To construct an objective measure of performance using case data, one should ensure that cases are comparable to one another. This could entail comparing cases only within a homogeneous category or constructing a metric that captures the complexity of the case.
- Measures of performance for public sector organizations will depend on the specific context of study and data availability, but they should reflect both the *volume* of services provided as well as their *quality*.

## INTRODUCTION

Data, from the Latin *datum*, meaning “given,” are the embodiment of something factual, technical, value-free. Yet, data are more than that. They have and create monetary worth: “Data is the new capital” has become the catchphrase of the 21st century as their importance in value generation has increased. Data are central to the business models of most leading Fortune 500 companies (MIT Technology Review and Oracle 2016; Wang 2012). Their potential for poverty alleviation, growth, and development has been recognized. For instance, the World Bank’s *Europe and Central Asia Economic Update, Spring 2021: Data, Digitalization, and Governance* places “data” center stage (World Bank 2021a). Several governments have already demonstrated how they can use data to provide better services and protections to their citizens (UN Data Revolution Group 2014; World Bank 2021a). Collecting data on people who work in government has become part of using metrics to improve government service delivery (see table 6.1 for an overview of the types of data that governments can collect on their workforce and examples of how they can contribute to a mission of service improvement).

**TABLE 6.1** Types of Data Collected on Public Administration, with Examples

Type of data collected on public sector employees	Examples
Prehiring metrics	Qualifications, work experience, gender, age, ethnicity, sexual orientation, disability
Recruitment metrics	Test and application scores, background checks
Performance metrics	Output rate, user feedback, supervisor reviews
Learning and development metrics	Rate of promotion, courses taken
Incentives	Disclosing salaries, disclosing tax returns, pay scales, bonuses
Survey data	Attitudes, self-reported behaviors, practices
Linked data	Survey + administrative data, survey + geospatial data, administrative + geospatial data

Source: Original table for this publication.

Parallel to increased awareness of data as a source of value creation, greater attention is being paid to how rendering observations into data relies on structures of power and value trade-offs. Over the last decade, public debate on the ethics of data use has become a fixture of global news, with most articles focusing on the use of consumer data (Bala 2021; BBC News 2019; Fung 2019; Pamuk and Lewis 2021), some on governments' use of data on citizens (for example, Beioley 2022a; Williams and Mao 2022; World Bank 2021a), and some on companies' use of data on their employees (Beioley 2022b; Clark 2021; Hunter 2021; Reuters 2022). The intersection of the last two arenas—the use of data by governments on their own employees—has received next to no attention. This is likely because governments are only starting to catch up in the use of employment and within-government metrics and because claims to privacy are more complicated in the case of government employees.

This chapter tries to address this gap by providing a thorough, albeit not exhaustive, discussion of ethical issues specific to data collection on government employees. It regards *data* as a multifaceted construct: an amalgamation of points of information that—depending on how they are processed and analyzed—can become capital, commodity, truth, or all at once. Data's function as truth, in particular, distinguishes them from other resources that have shaped world economies over the last centuries. Neither gold, oil, nor 5G has inherent value because of what it says about human behavior and the world we live in. In that sense, data and their use raise ethical conundrums not seen before in other phases of technological adoption.

Data linkage and triangulation offer the best chance for constructing measures of public sector productivity that are meaningful and provide an acceptable level of accuracy. Such developments have brought their own problems. Greater triangulation can lead to greater reliance on indicators. See the discussion in chapter 4 of this *Handbook*.

Choosing these indicators means choosing what to make salient in an official's work environment.

Guides to ethical data use for both the private and public sectors now exist in many jurisdictions (Mehr 2017; Morley et al. 2020a; OECD 2019; Office for Science 2020; Open Data Institute 2021). Guidelines for the public sector tend to be heavily modeled on those for the private sector, meaning that they mostly speak to data interactions between the government and public service users. There is, however, a significant amount of data concerning processes and people within government. Collecting data on government employees involves a different set of challenges from collecting data on service users and private sector employees (see table 6.1 for an overview). The ethics of collecting data on government employees thus merits a separate discussion.

This chapter uses the terms “government employee,” “public sector employee,” and “civil servant” interchangeably. The majority of the concerns discussed here are relevant to people employed in central government functions as much as those employed in frontline services (“street-level bureaucrats”). The proportionate relevance of concerns will differ by the type of public sector employment and the regime type. As the aim here is to provide a general framework for data collection on employees in the public sector,

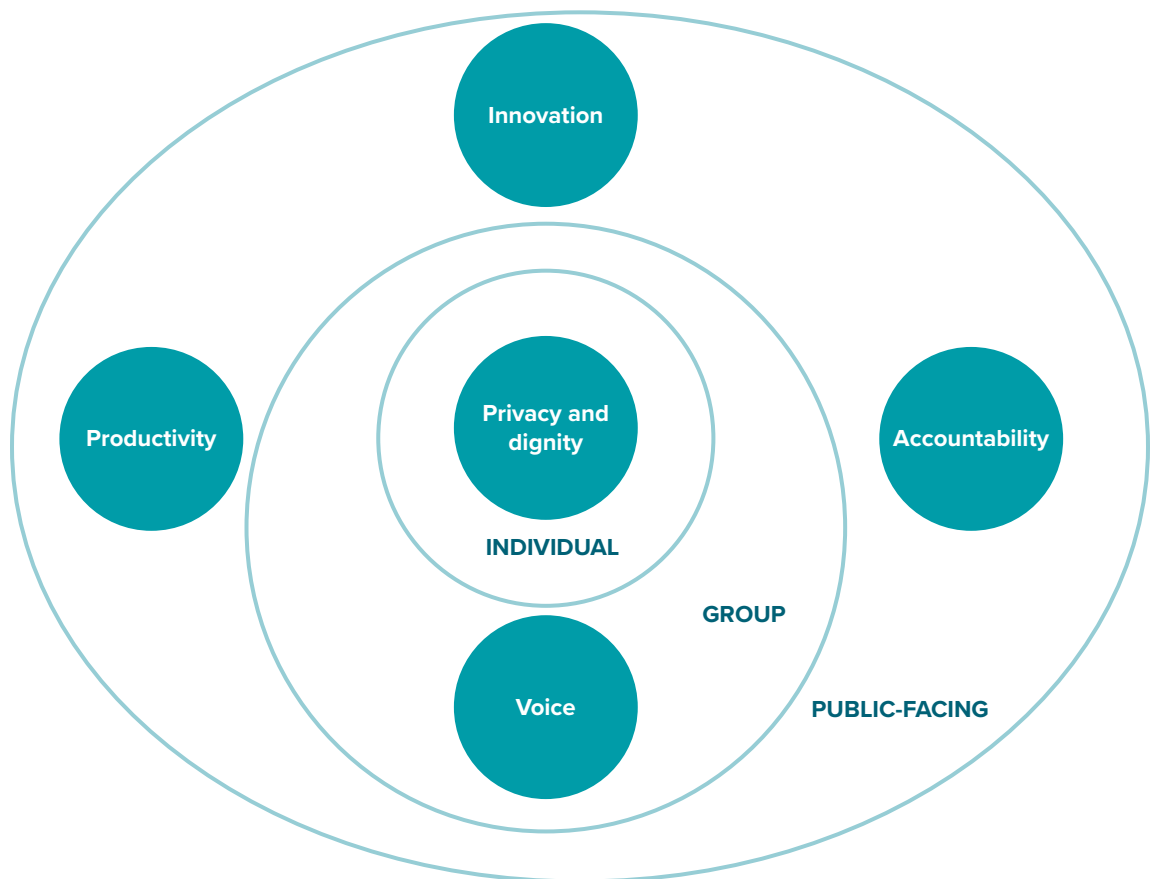
differences will not be discussed in detail, in favor of providing more space for discussion of how ethical data collection can be put into practice in the public sector.

### Data and Public Value

What public value *is* in and of itself subject to debate. Seminal scholarship on the topic, like Mark Moore's work on public value (Moore 1995), eschews defining public value by ascribing the task to public managers and their interactions—perhaps better called “negotiations”—with political and societal agents (Rhodes and Wanna 2007). Deborah Stone's (2002) work on public value highlights the ideals of equality, liberty, and security core to the idea of statehood and the role of government but stresses that none of these ideals can be fully catered to in most policy settings.

For a working definition of public value that serves the debate about the use of data on government employees, this chapter will focus on three broad aspects of public sector endeavors that can produce public value by producing goods and supporting equitable information exchange between the governed and the governing: accountability, productivity, and innovation (see figure 6.1).

**FIGURE 6.1** Ethical Dimensions That Require Balancing in Data Collection Efforts on Public Sector Employees



Source: Original figure for this publication.

Note: As a heuristic, one can imagine key questions about the ethics of collecting data on public sector employees falling into three circles. The innermost circle describes the individual dimension. These questions mainly concern the privacy and dignity of the individual public sector employee. The middle circle signifies the group dimension. These questions concern voice and dissent, which are central to many functions that public sector employees carry out and the tensions that arise when collecting data on them as groups. The third, outermost circle encapsulates questions related to the qualities that define public sector value creation in relation to stakeholders: political principals, public service users, and society at large.

The chapter discusses accountability first because it is central to ethical exchanges between citizens and governments and employers and employees. It next focuses on productivity because it considers the creation of public value in the sense of government output. Finally, it turns to innovation because concerns about public value creation are not limited to what is produced but, particularly in the context of big data and artificial intelligence, include whether data used for public value creation can also produce innovation in the form of improvements or altogether new outputs. The chapter discusses all three concepts with a view to the wide-ranging types of data that can be collected to inform public administration, such as pre-hiring, recruitment, performance, and learning and development metrics (see table 6.1).

### **The Three Dimensions of Public Sector Data Ethics**

As a heuristic, the ethical considerations facing public administrations when collecting data on their employees comprise three dimensions: an individual dimension, which comprises demands for dignity and privacy; a group dimension, which relates to voice and dissent; and a public-facing dimension, which ensures that data enable public administrators to deliver on public sector values.

#### ***The Individual Dimension: Individual Dignity and Privacy***

Individual demands for dignity and privacy—the first dimension of ethical concern for employees—have been discussed widely in standard volumes on research ethics (Carpenter 2017; Macdonald 2017). Although the other two dimensions have received less attention, the first still merits discussion because, owing to their unique position in political-economic systems, public sector employees face a different set of demands, incentives, threats, and opportunities than private sector employees. For instance, because of the public sector's dominance in providing jobs and offering better wages in many countries (Düwell et al. 2014), exit options for public sector employees are more limited than for those employed in the private or nonprofit sector. This has implications for informed consent. Employees might accept trade-offs because of the constrained choice sets they face rather than satisfying their needs for privacy and dignity.

The temporal and spatial dimensions of how data on employees are safe-guarded also differ from the private sector. What is considered “good public service” and service “in the national interest”—and thus what types of measurement and data use are justified—can shift as governments and their view of the nation-state and societal ideals change. Such shifts in value positions affect both individual freedoms and those of groups within the public sector.

Another difference is created by the pressure to reflect political shifts in how government-owned organizations are run. For example, state-owned enterprises not only deliver services where market failures exist but also serve as model employers (for example, by pioneering inclusive hiring practices), act as symbols of national identity (for example, airlines and national health services), and can play a redistributive function (for example, by providing subsidized goods or privileged access to select groups of citizens and stakeholders; see Heath and Norman [2004] for a more extensive discussion of state-owned enterprises). The aim of data collection on individuals and groups in the public sector might thus change over time compared with the private sector. Some political factions believe many services and functions should not be performed by the public sector at all, or if they do not go this far, they have a deep-seated mistrust of civil servants. The threat of surveillance and a push to replace workers with compliant and efficient machines thus might be even more acute for the public sector than the private sector, depending on the political leaning of the government in power. As a case in point, remote and flexible work has become standard in many industries. Because of competition among companies and sometimes even industries (think aerospace engineers joining tech companies or banks), these standards are unlikely to be reversed. Data on sick days and other types of leave taken by employees in the United Kingdom suggest that private and public sector workers are absent for fewer days in a year than public sector workers. These data, and the supposedly empty seats (due to remote work) in government halls, led leaders of the UK Conservative Party to campaign for remote-work policies to be curtailed



(BBC News 2021; Cabinet Office 2022; Lynn 2022; Office for National Statistics 2022). How data on public sector workers are collected and used might change more with the political flavor of the day than with the industry standard or best practice in other fields.

### **The Group Dimension: Voice and Representation**

The second, group dimension becomes relevant in a setting where data relate to existing or ad hoc groups of employees. The group dimension recognizes that employees as a group have a right to actively shape what is done to them: they have a right to exercise *voice*.

Hirschman (1970) first introduced the concept of *voice*, alongside *exit* and *loyalty*, to define the dynamics of engagement with institutions. *Voice*, in this sense, is defined as the expression of opinions and thoughts in a manner that is impactful or at least has the potential to be impactful. Denying stakeholders the option of voice makes exit more likely, which, in an organizational setting, means disengaging from the cause or leaving the organization. Voice is thus a moral imperative as much as a practical necessity.

The need for voice in public service creates many ethical conundrums. Voice is necessary because “freedom defined as noninterference or the absence of burdensome restraints clearly will not do” (Preston 1987, 776). Civil servants need space to speak up, take initiative, disagree, and innovate.

On the other hand, for organizational goals to be attained, the expression of voice needs to be bounded. It requires agreement on a narrative about what is good, achievable, and desirable. This is particularly true in many public service organizations, where adherence to mission is an important guide to action and a motivator for public servants as a whole. Voice may place individuals or groups in the position of identifying themselves as in defiance of, or as distinct from, the prevailing public service culture.

Boundaries to voice are also necessary because of the mechanics of representative democracy: central to the role of civil servants is that they implement directives from democratically elected or appointed political leaders. Civil servants thus need to subordinate some of their own concerns to the policies elected representatives choose. Such subordination can be demanded more easily of individual civil servants. However, when voice is exercised on behalf of groups—for example, women, ethnic minorities, and people with disabilities in the civil service—boundaries are much more difficult to draw. Civil servant groups, then, are both implementers of democratic will and constituent groups with a right to voice at the same time.

### **The Public-Facing Dimension: Data to Operate in Service of the Public**

The third, public-facing dimension is particular to the public sector because it concerns data collected with a view to serving public service functions. It highlights the ethical challenges linked to creating public value and the organizational capabilities required to do so (Panagiotopoulos, Klievink, and Cordella 2019). Data collection and use must be designed in a way that enables government employees to operate effectively and efficiently, to collaborate and innovate.

Challenges located within this third dimension include balancing data security requirements, openness for innovation, and room for experimentation, as well as the certainty, consistency, accountability, and reliability of public service. While not necessarily mutually exclusive, these demands create tricky trade-offs. For instance, making data on what public sector workers do available can help the population to monitor them and call out malpractice; it can also help others spot ways of innovating, doing things better or differently; however, it can also create fear and political suppression and encourage inflexibility.

The following sections will discuss each ethical dimension in turn. Each section first outlines specific demands and incentives facing public sector employees before discussing how data can help to address ethical challenges and where they introduce new challenges that require closer scrutiny. Appendix B includes a framework for evaluating the ethics of measuring and tracking public sector workers. Practitioners can use this framework to think through the key tensions laid out in this chapter.

## INDIVIDUAL LEVEL: DIGNITY AND PRIVACY OF PUBLIC SECTOR EMPLOYEES

Dignity and, by extension, privacy are two values central to the ethical handling of data. Dignity (Schroeder and Bani-Sadr 2017) and privacy (Solove 2008, 1–25) are concepts that have caused controversy because they are complex and, at times, tautological. This chapter employs *dignity* to mean that humans—in this case, public sector employees—have worth in and of themselves that *they themselves* can define. This stands in contrast to seeing people as *means* to effect government outputs and *ascribing to them the essence of their ends*. Practically, dignity requires respect for the individual as a human being with free will, thoughts, and feelings. This means that the options available to people on whom data are collected should be central to the design of research efforts. For instance, if the designers of employee web pages decide it is nice to show pictures of staff, individual employees should be allowed to have their pictures removed if they do not want others to know what they look like, regardless of whether the employer agrees that this concern is central to what it means to respect their employees.<sup>1</sup>

*Privacy*, as it is used in this chapter, does not mean *anonymity*. Information might be disclosed anonymously—for instance, via a survey where no names are provided and IP addresses or similar identifying characteristics of the respondent cannot be linked to responses. However, this act still entails giving up privacy because what was internal and possibly unknown to others is now known to others. The reasons why such a strict definition of privacy is adopted in this chapter become clear when discussing the group-level dimension of data ethics concerning public sector employees: even when information cannot be linked to an individual, as soon as information can be linked to a group—public sector employees—their thoughts, behaviors, and environs become known to people other than the members of this group. Discussions of *individual* privacy, the focal point of this section, must therefore be separated from discussions of *collective* privacy, which will appear in later sections.

Relatedly, *privacy* as it is used here is understood in terms of a “typology of disruptions” (Solove 2008), which acknowledges that the definition of privacy is highly contextual. Just as quirks are distinguished from mental health disorders, disclosure, transparency, and openness are distinguished from infringements on privacy by the disruption they cause. Does a person only occasionally need to return home to check whether the stove was left on, or is this a daily occurrence that interferes with a person’s life? Is it a minor issue that a public servant’s address is publicly available online, since everyone’s address is publicly available online, or is it a danger to the public servant’s safety and right to conduct personal matters in private? Until the 1980s and into the late 1990s, it was common in Western European countries for the phone numbers of the inhabitants of entire cities to be listed, publicly available in white pages or equivalent phone and address directories. In several Scandinavian countries, it is still the case that every resident’s social security number, address, and taxable income is made publicly available. The key differences are the extent to which something is experienced as a disruption, as opposed to the norm, and the extent to which people can stop a practice if they start to experience it as a disruption. In the case of telephone and address registries in democracies, residents can use their voting powers to change the laws surrounding the publication of personal details. It is less clear how employees—particularly state employees, of whom transparency is expected—can demand change when they find practices intrusive. Privacy, as defined in this chapter, is thus closely linked to the idea of control: the extent to which civil servants control how much is known about them when they experience it as intruding on realms they perceive as reserved for their private as opposed to their professional (work) persona.

Commonly, informed consent procedures serve to preserve dignity by affording individuals the opportunity to ascertain what they see as acceptable in how they or their data are handled. Informed consent means that the individuals on whom data are collected or who are asked to divulge information are fully informed about the purpose of the research, how their data are handled, and how they will be used. Even if they agreed at an earlier stage, individuals ought to be given the right to withdraw their consent at any stage, which requires the secure deletion of any data collected on them. Typically, there are few options for public officials to opt out of servicewide databases.

The extent to which informed consent is de facto voluntary is important. In situations of near employer monopoly, in which exiting the public sector is not a viable option, and in situations of suppression, in which exercising voice is not possible, employees might consent to data collection because they see little other choice.<sup>2</sup> This mirrors problems with consent ubiquitous in today's highly oligopolistic landscape of online service providers: if one is pressed to quickly find a parking spot and only Google Maps can pinpoint one with timeliness and accuracy, one is likely to accept the data use conditions that come with this service. The viability of informed consent is intricately bound to the ability to exercise exit and voice.

Valuing dignity also extends to claims to privacy. People ought to be allowed to keep certain details of their lives and personhood protected from the scrutiny of others—even their human resources (HR) managers. Any form of data collection on individuals will require some abnegation of privacy. Research ethics typically tries to acknowledge this right by providing confidentiality: a declaration that the information disclosed to the data collector will be known to and used by a limited, preidentified set of people. In the public sector, this is the case, for example, when an internal research or strategy unit collects survey responses from civil servants with a view to sharing them with HR and finance operations to improve planning, staffing, and training.

The principle of confidentiality in research brings to mind that research ethics has its foundation in the Hippocratic Oath and in bioethics. Patients trust their doctors to share their medical information and concerns only with relevant personnel, and always with the intention to help the patient. In a medical setting and in most everyday circumstances, we tend to assume confidentiality and do not give explicit consent to use otherwise private details divulged as a part of interactions. We do so willingly because disclosing information encourages reciprocity, builds trust, and helps shape the world around us to better meet our needs. Reductions in privacy are not a zero-sum game but can offer substantial welfare gains for the individual and society at large when negotiated carefully. The individual is, however, poorly positioned to negotiate these tradeoffs against the interests of large corporations or the government. As discussed above, this is particularly true if an individual would like to exercise exit or voice when the options presented to them do not inspire trust.

In response to this problem, legal protections have been put in place to guard against the worst misuses of data. Data regulations such as the European Union's General Data Protection Regulation (GDPR) require that entities collecting data clearly lay out which data are collected, how they are handled, and who has access to them. California, Virginia, New Zealand, Brazil, India, Singapore, and Thailand have all implemented legislation similar to the GDPR in recent years. However, detailing how data are used typically leads to documents that require, on average, 10 minutes to read (Madrigal 2012). As data collection has become ubiquitous, the time burden that consent processes introduce implies that most of us have become accustomed to quickly clicking through consent forms, terms and conditions, and other common digital consent procedures. This means that in practice, consent is either based on trust rather than complete knowledge or, in the face of a lack of exit and voice options, is coerced.

Following legal guidelines is thus not enough to ensure that dignity and privacy concerns are adequately addressed. Those planning to collect data on public sector employees must take into account what ethical challenges could arise, how to offer exit and voice options, and how to foster trust. This is not a simple feat. The discussion will thus next turn to how three common dilemmas concerning data privacy and dignity in the public sector can be addressed.

### **Government Employee Data, Dignity, and Privacy**

Discussions of data ethics concerning the dignity and privacy of civil servants could fill volumes. This chapter, therefore, cannot offer a comprehensive account of the debate. Instead, it focuses on three likely areas of concern for someone wanting to collect data on civil servants: trade-offs surrounding access, transparency, and privacy; how data collection can be designed to enhance dignity; and how data transfer and storage should be managed to safeguard privacy and dignity.

## Trade-Offs Surrounding Access, Transparency, and Privacy

When collecting data on public sector employees, it can be argued that knowing about their behaviors and attitudes is in the public interest. For example, using data to identify inefficiencies associated with public employees is only possible without their active consent, but is certainly in the public interest. In many jurisdictions, public sector workers do not have to be asked for their consent for research to be conducted on them as long as it can be shown that the research is in the public interest. In most countries, where these provisions are not made explicit in the law, data ethics codes include clauses that allow the requirement for consent to be lifted if there is a strong case that research is in the public interest. These waiver clauses tend to use unequal power relations as grounds for the lifting of consent requirements: researchers might be prevented from gaining access to public institutions if they require explicit consent from people in positions of power who are hostile to the idea of research (Israel 2015). Based on the public-interest argument, consent procedures can, in many instances, be circumvented when researching public sector employees.

A reduction in privacy and an overruling of informed consent can thus promote *accountability*. They can also enhance *transparency*. Having more in-depth knowledge of the behavior of government employees can help to increase government transparency: just as the work of election volunteers was live streamed during the 2020 US elections (Basu 2020), civil servants could be put under constant surveillance to increase transparency and inspire trust. Evidence from the consumer context suggests that extreme levels of transparency can create a win-win situation, in which customers rate services more highly and are willing to pay more when they can monitor how a product is created or a service delivered (Buell, Kim, and Tsay 2017; Mohan, Buell, and John 2020). Radical transparency could thus inspire greater trust and mutual respect between government employees and government service users as opposed to simply reducing dignity and privacy by increasing surveillance.

However, promoting one type of public good might infringe on another (see the discussions of contradicting policy choices in Stone 2002). Privacy is instrumental to guarding collective freedom of speech and association (Regan 2000). Depriving public sector employees of privacy can erode democratic principles because employees may lose spaces to dissent and counteract malpractice pursued by powerful colleagues, managers, or political principals. To date, the ambiguity of public-interest claims has been most commonly revealed in cases of whistleblowing (Boot 2020; Wright 1987): government whistleblowers often endanger some public interests (for example, national security) in favor of others (for example, transparency). How convoluted claims to public interest can become is highlighted when whistleblowers reveal previously private information about some actors with a public-interest claim (for example, disclosing that a particular person was at a particular location at a particular time or making private communication between individuals public). The privacy of whistleblowers needs to be protected in order to shelter them from unfair prosecution and attacks so that future whistleblowers do not shy away from going public. The future public interest, then, is guarded by protecting the present privacy of the whistleblower, who might have rendered public what was previously thought to be private “in the public interest.”

In this sense, what is in the public interest and what the limits are to a utilitarian logic of increased surveillance must remain part of the public debate. For public debate to be a viable strategy for dealing with the contradictions of disclosure and the protection of privacy in the public interest, society must protect fora and institutions that publicize such issues, involve powerful stakeholders, and have tools at their disposal to enforce recommendations. To date, this often means supporting the capacities of the media, civil society, the political opposition, and the judiciary to fulfill these functions. Researchers and governments thinking about collecting data on government employees need to assess whether these institutions function sufficiently or, if not, whether actions can be taken to remedy their absence. For instance, governments could create independent review committees, actively publicize planned data collection efforts, and provide extra time for public consultations. In settings where such mechanisms lack potency, the international community, most likely in the form of intergovernmental organizations and donors, has a responsibility to monitor how changes in data regimes affect the dignity, privacy, and welfare of data subjects and the citizenry more broadly.

This is important not solely with a view to balancing trade-offs between transparency, public access, and privacy. The next section looks at how a thorough design and review of government data collection strategies could help to create regimes that enhance dignity despite entailing reductions in privacy.

### Designing Data Collection to Enhance Dignity

Designing data collection in a way that enhances dignity serves several functions. For one, if we understand the problem of privacy in terms of disruptions of autonomy, collecting data in ways that enhance dignity can help to move away from zero-sum thinking. If public sector employees gain dignity from data collection, they are unlikely to see it as an unwanted intrusion. For instance, data could be collected that celebrate personal initiative, good management practices, and outstanding achievements, as opposed to disciplinary uses of data, such as identifying malpractice or inefficiencies. In some cases, employers might also consider disclosing the identities of participating individuals, if they consent, to highlight valuable contributions and give credit. (See also Israel [2015] for case studies of how sociological and anthropological researchers typically acknowledge the role that research participants play in scientific discovery.)

Apart from a normative view of why data collection efforts should enhance dignity, there are clear utilitarian reasons: this can improve data quality, trust, and openness to data collection and data-based management practices. Public sector employees might refuse to disclose their true opinions when they feel pressured into consenting to data collection. This can, for instance, be the case when managers or political principals provide consent on behalf of their employees. Lifting consent procedures with public-interest claims can backfire in such cases. Engaging with employee representatives and living up to promises of transparency concerning the objective of data collection can help.

Processes that ensure that data collection is linked to clear action points can further help to guard the dignity of research participants. Data collectors have an additional incentive to do so because the validity of responses and response rates will likely deteriorate when staff see that data collected on them are not used to their benefit.

Staff will more likely develop trust in the process and engagement with results when they have a stake in developing research questions and action plans following up on results. Principles of action research, including building phases of review, consultation, and revision into the research process, could help to create data collection strategies on public sector employees that enhance the dignity of the individuals involved.<sup>3</sup>

### Designing Data Transfer and Storage to Guard Dignity and Privacy

Both dignity and privacy are at risk when data are not secure. Cyberattacks are becoming more common. For example, in 2021, the Washington, DC, Police Department was subject to a massive data leak following a ransomware attack. Disciplinary files and intelligence reports including names, addresses, and sensitive details about conduct were leaked into the public domain (Suderman 2021). In 2015, the Office of Personnel Management (OPM) of the US federal government was subject to a hack that led to the leaking of the personal data of millions of employees, many of whom suffered from identity theft for years following the data breach (CBS and the Associated Press 2015).

Guarding dignity and privacy in this sense is as much a technical as a moral issue. Legal frameworks such as the GDPR have been created with this in mind. Several international best-practice guides on data protection elaborate on the technicalities of such efforts. Good examples include sections on data protection in *Development Research in Practice: The DIME Analytics Data Handbook* (Bjärkefur et al. 2021) and the Organisation for Economic Co-operation and Development's (OECD) Privacy Guidelines (OECD 2022).

## GROUP LEVEL: VOICE AND DISSENT

As the chapter so far has reviewed, there are many aspects of data collection that affect government employees as individuals. These are perhaps most comparable with the concerns affecting private persons and research subjects. There is, however, a dimension that becomes particularly important when thinking of public sector employees as a group and the groups within the public sector that can be created based on observable characteristics or self-elected labels.



As described above, the concept of *voice*—alongside *exit* and *loyalty*—was coined by Hirschman (1970). In the public sector, voice amounts to the ability of employees to express opinions and the potential of these expressions to impact how public administrations are run. When exit options are limited, voice becomes a more pertinent tool for employees to exercise control over their environment. In public sector employment, voice is also conceptualized as part of the job of civil servants. For instance, the UK's Civil Service Code demands that public administrators “provide information and advice, including advice to ministers, on the basis of the evidence, and accurately present the options and facts” and that they “not ignore inconvenient facts or relevant considerations when providing advice or making decisions” (UK Government 2015). Voice in this function is mainly intended to enable elected officials to deliver on their policy programs. If, however, elected officials endanger public sector values and the integrity of rule-based and meritocratic government institutions, a professionalized bureaucracy is expected to exercise voice to counteract this erosion. This can take place within legitimate remits of voice and discretion (Miller and Whitford 2016), or it can take the form of dissent (Kenny 2019).

Demands for voice are intricately linked to those for productivity and innovation. In organizational studies, it has long been established that psychological safety—the feeling that employees can voice ideas and concerns without facing an immediate threat to their jobs or selves—is necessary for innovation and sustainable increases in organizational performance (Baer and Frese 2003; Nembhard and Edmondson 2006). Empirical research shows that more-diverse workplaces, in the private and public sector, are more-productive workplaces (Alesina and Ferrara 2005; Hjort 2014; Rasul and Rogger 2015).

Voice has also been conceptualized as civil servants' representation of the interests of the demographic groups to which they belong—a kind of “passive voice.” It is theorized that they do so through increased attention to these groups' needs and a heightened awareness of how to design and deliver public services to address them (summarized under theories of representative bureaucracy; see Kingsley [1944] 2003; Meier 2019).

Data on civil servants can help to promote or curtail voice in its active and passive forms. With regard to the passive form of voice, data can showcase how certain groups (for instance, women or bodily disabled people) are affected by or think differently about certain employment policies. With regard to its active form, data can be used by groups to raise issues themselves. For example, if the number of bullying or sexual harassment complaints filed by a department or government subdivision is made public, victims understand that their cases did not happen in isolation, and numbers can be used to pressure leadership for change.

### Government Employee Data, Voice, and Dissent

Data on groups of public sector employees raise ontological and utilitarian questions. The former questions relate to how data can define groups and how groups can use data to define their workplace, work, and position between political principals and citizens.<sup>4</sup> The latter questions concern how the risks and benefits of using data relate to attempts to improve working conditions and service delivery in a way that is responsive to the needs and wants of groups that become apparent through the group-level aggregation of data.

Assigning group labels to individuals implies that their individual rights and identities are pegged to those of a group—potentially one with which they do not identify or of which they did not consent to be part. Such *passive group* membership has typically been applied to individuals grouped together as “vulnerable” or “fragile” populations (Grantham 2020, 39). As the now-mainstream debate on gender and gender pronouns has raised, similar questions can be applied to group labels that have traditionally been considered more stable.

Consent can be a vehicle to ensure the alignment of grouping with personal preference. For instance, in surveys, civil servants can opt out of providing demographic details. However, for most administrative data collection efforts, consent is limited or wholly unfeasible. Employees might be grouped together as “fit for early retirement” or as a target group for “offering overseas posting” because of their age, gender, tenure, family situation, and other administrative data available to other administrators. These groups might not align with the desires or career ambitions of the grouped individuals. Basing planning decisions solely on results arrived at from demographic or ad hoc grouping risks wrong conclusions. The availability and increasing richness of data available on groups thus cannot substitute for meaningful engagement with them. These arguments are touched on by Bridges and Woolcock in chapter 4.

The creation of groups and the pegging of data to group labels not only creates risks; it also holds immense positive power. Data can open up avenues for voice because they showcase where differences between groups exist. For instance, in 2014, the cross-organizational initiative Data2X started a global drive to increase the collection and use of data on women and gender-disaggregated data. The initiative has helped to increase women's access to finance and meaningful work and has significantly reduced maternal mortality rates (Grantham 2020). In the context of collecting data on public sector employees, data can help practitioners better understand issues such as the proportionality of the representation of groups in different positions and sections (for example, Are women proportionally represented in leadership positions? Does the proportion of civil servants from minority groups map onto their proportion in the general population?); planning for skills gaps (for example, Are there enough people with advanced IT or data analysis skills in each department?); or spotting management problems in part of the civil service (for example, Do people staffing local government offices have the same level of goal clarity as those working in central locations?). In this sense, data can increase voice and benefit society.

Navigating the collection and use of data on public sector employees requires moving beyond acknowledging how data shape realities to discussing how the risks and benefits created by this process can be negotiated. Data can catalyze employees' voice by giving groups a platform to assess metrics pertaining to these groups compared with relevant benchmarks. For instance, detailed employee data including demographic details can help practitioners to understand what predicts turnover and whether certain staff members—ethnic minorities, women, or people with care responsibilities—can be retained as well as others (Grissom, Viano, and Selin 2016). Where performance data are available, data on staff can be linked in order to better understand how staffing decisions affect service delivery. For example, employing ethnic minority teachers in publicly funded schools in otherwise homogeneous districts has been associated with better performance for ethnic minority pupils attending these schools (Ford 2022).<sup>5</sup> Data disaggregated by groups can also help provide better access to training and career progression for public sector employees.

As Viswanath (2020) notes, data equity, in terms of knowing what ranks people from different sections of society can attain in public service, is critical to providing better service (in line with theories of representative bureaucracy) but also to providing greater equity in opportunity for public sector staff. As a case in point, public sector unions in Canada have successfully used wage data disaggregated by gender to support calls for gender-based pay parity (Card, Lemieux, and Riddell 2020). This has created more equality between men and women in the public sector and, as a consequence of the non-negligible amount of the population employed in the public service, has improved pay equality across a large section of society.

At the same time, there is no guarantee that the availability of data will safeguard the rights of minority groups and promote equity and equality of opportunity. Data on groups can also be used to curtail voice. For example, while collecting recruitment metrics could heighten the potential for governments to hire a more diverse workforce, it could equally enable them to weed out people who are deemed less desirable. Such weeding out could be based on demographic details, but it is increasingly also founded on what potential employees say online. Hiring managers can easily introduce bias into the hiring process if the recruitment process is not sufficiently anonymized. For instance, it might be important to collect data on which universities applicants attended. These data, however, can also be used by hiring managers to prescreen candidates—consciously or unconsciously—based on associations of quality and merit with these universities. In a similar vein, even though hiring managers might not get access to detailed information on routine background checks, they can use an applicant's name and previous employer or university affiliation to conduct their own online searches.

Indeed, public sector unions in Canada and Australia now actively discourage their members from posting online or having an online presence that can be linked to their identities, in fear of potential discrimination for current employment and future employment opportunities (Cooper 2020a, 2020b). In some government contexts, there is the danger that governments collect information on employee opinions systematically. This is problematic not only at the individual but also at the group level. Investigations by big data scholars have illustrated how, for instance, people using hashtags on Twitter related to the Black Lives Matter movement could be grouped together (Taylor, Floridi, and Sloot 2016, 46). In a public sector context, such information could be used to map out the political affiliations of employees.



Other administrative data could be used for targeting based on religion or personal circumstances. For instance, data scholars have shown that people can be identified as devout Muslims by mapping out their work break schedules over a year (Rodriguez 2018).

Development practitioners and donor countries working on data strategies for public sector reform in countries in which political, religious, or other basic human freedoms are not guaranteed must thus tread carefully in order to guard against setting up data infrastructures that can be used to the detriment of public sector employees.

## SYSTEM LEVEL: PRODUCTIVITY, ACCOUNTABILITY, AND INNOVATION

Moving on from the group dimension, this section discusses the most distinctive aspect of data collection on public sector employees: ethical concerns that relate to the duty of public sector employees to serve the public, which can support but also interfere with safeguards designed to protect against unethical data strategies.

Public sector values are commonly defined as the set of qualities of public sector work that make for an ideal public service. Such values typically pertain to productivity (delivering public goods and services in an efficient way), accountability (delivering in a way that is answerable to the public and key stakeholders), and, ever more commonly, innovation (constantly adapting to keep pace with and anticipate societal and economic developments). Each of these qualities can be served by data. The next sections discuss them in more detail.

Other public sector values that are often discussed include equity, transparency, impartiality, and a concern for the common good. As equity and impartiality are supported by mechanisms enforcing accountability and a degree of transparency is required for accountability to be effective, these themes will not be discussed here separately. Similarly, a concern for the common good is difficult to define. As this chapter takes a view built on economic theories that see the common good as the product of welfare-maximizing actions, the next section will discuss the ethics of data collection for the common good together with those aimed at increasing productivity.

### Productivity

Public sector workers are meant to maximize productivity in service of the public, in response to their political principals' directives and while remaining accountable to a diverse group of societal stakeholders. In the 21st century, execution is not enough; public sector employees are also expected to do their work as efficiently as possible. They are expected to maximize quality output per tax dollar contributed by each tax-paying citizen. Core to the task of a public sector employee is thus to be productive (for the common social good).

This is not guaranteed.<sup>6</sup> Public sector workers have a lot of room to diverge from the productive delivery of public service. Public sector workers have specialist knowledge and skills that make it difficult for outsiders to assess the quality and efficiency of their work. A more fast-paced economy and social changes also demand that the public sector be more flexible and responsive, which requires awarding more discretion to public sector workers.

Public sector productivity is difficult to measure because it is a collective effort. There are no market prices readily available for many of the complex goods and services the public sector provides. Efficiency is often a poor marker of success because the services are supplied by the government precisely because there are market failures.

In lieu of rule-based control, oversight in the form of monitoring metrics has become more common. Data and analytics can help overcome the feasibility of, and individual employees' proclivity for, ethical violations. They can ensure that officials are focused on the productivity of public service. Advances in the measurement of productivity have been made, in particular through combining micro- and macro-data, such as process data, project- and task-completion rates, staff and user satisfaction data, performance evaluations, and cost-weighted budget data (Somani 2021).

### Data Ethics and Public Sector Productivity

Both accountability and productivity can be promoted by making data on public works projects publicly available and easily understandable. Lauletta et al. (2019) illustrate how a website that geotags public works projects in Colombia can speed up project completion. Several administrations have started using predictive algorithms to allocate inspectors (for example, fire and food safety inspectors) more efficiently. Data can help to promote meritocracy by identifying bottlenecks and resource deficiencies. If data are used to address resource inequalities, they can help public sector workers be productive.

Nehf (2003) uses the term “mischaracterization as harm” to summarize the core of ethical problems related to measuring productivity in the public sector: imperfections in measurement can create misjudgments that are harmful to the dignity of the individuals and groups described as much as they cause more tangible social harm. For instance, when productivity is measured without appropriate awareness of epistemological challenges, it can encourage management to targets. In the context of education, this can lead to grade inflation (De Witte, Geys, and Solondz 2014; Hernández-Julián and Looney 2016). In health care, it has been linked to underinvestment in preventative care (Bevan 2009; Gubb 2009).

Problems with construct validity could unfairly sideline certain groups. For instance, the amounts of sick leave taken and overtime logged are not necessarily good measures of productivity, skill, or work effort. If employees struggle with their health, it infringes on their dignity to equate sick leave with a lack of motivation or commitment to organizational performance. In a similar vein, employees with care responsibilities might be unable or reluctant to work overtime but could nonetheless be stellar performers.

The lack of agreement about what is productive for many job roles in the public sector—or perhaps the agreement that there is no clear definition—means that measurement risks undermining principles of meritocracy. Public services whose productivity is hard to measure might be defunded relative to those whose measurement is easier. Personnel who manage to targets rather than those who create meaningful value for citizens might get promoted. It can also create imbalances in workforce planning. Specialists have been found to be disadvantaged in terms of career progress in the civil service. This seems to be connected to a lack of adequate data on skill matches and to managers’ inability to evaluate what good specialist (as opposed to generalist) performance looks like (Guerin et al. 2021).

An ethical approach to measuring productivity will entail an emphasis on understanding the values that underlie what is considered productive and how power relations shape how problems are defined and acted upon. For example, microdata can also help show where public sector employees might engage in corrupt practices. An ethical approach, however, does not guard against using data on corruption selectively (Nur-tegin and Jakee 2020). Depending on the relative political power of different governing parties and the opposition, data collection efforts might be channeled away from some activities to focus on others. Who has power over data collection efforts and the use of data is thus a question that lies between data capabilities and their possible positive effects on public sector productivity and how ethically public sector personnel are treated.

As discussed in chapter 4 of the *Handbook* effective validation and benchmarking exercises can help to create meaningful and ethically sound measurement of public administration. The chapter argued that a balanced approach to measurement ensures that measurement is meaningful. This chapter argues further that a balanced distribution of power over that measurement and corresponding data will make it more likely that measurement and data are used ethically and justly. Enabling stakeholders to provide checks and balances against data misuse and opportunities for review (see the framework proposed in appendix B) remains key.

Epistemological and practical problems are here to stay. Epistemologically, what is defined as productive is questionable. Questions range from the definitions of core work tasks and what makes an externality to what metric should be used to signal positive impact. Quicker social security claim processing times might signal productivity, or a reduction in maternal mortality at public hospitals might speak to the quality of care, but neither speaks to the dignity with which service users are treated—arguably, an attribute central to public service. Despite improved data quality, which helps to produce efficiencies, value-based decision-making will not become obsolete. Negotiating values is required as much as ever to produce evidence-based and ethically sound policy (Athey 2017).

## Accountability

We next turn to the other two key public values that elicit tough ethical challenges for data collection in the public sector: first, accountability, then, innovation. One of the defining characteristics of public sector work is the demand for accountability. Public sector workers are expected to be answerable and liable when they do wrong. The group of people to whom they are answerable is large and diverse. It includes the clients of their services, members of the communities in which their services or policy actions take effect, and organizational stakeholders, such as firms and civil society organizations, whose cooperation might be necessary to arrive at sustainable and equitable policy solutions.

Creating an accountable public administration is a challenging task. The need for delegation and specialization requires that public administrators be provided with discretion. Theory and empirical evidence suggest that political control over the bureaucracy is not a sufficient lever for accountability (Brierley 2020; Gailmard 2002; Keiser and Soss 1998; Meier, Stewart, and England 1991; Raffler 2020; Weingast and Moran 1983; White, Nathan, and Faller 2015).<sup>2</sup>

Democracy in the deep sense—one that goes beyond elections and rests upon an informed and politically active population—requires that policy and its implementation can be scrutinized by the public. The quality of a democracy hinges on the ability of its population to be informed about what state apparatuses do, to voice their opinions about them, and to enforce policy change (Dahl 1998). Bureaucratic accountability also requires transparency and explicability.<sup>8</sup> As data become more widely available, it becomes easier for experts, civil society groups, and other stakeholders to scrutinize how well the government delivers on providing social goods. Data on the public sector and the civil service thus play an important role in helping to provide accountability.

However, this does not come without challenges. Ethical questions surrounding the use of data for accountability promotion center on the difference between transparency and explicability, concerns surrounding throughput, and the risk of causing unacceptable negative outcomes unintentionally as a result of data management solely focused on external accountability. Such risks require the circumspect creation of data infrastructure.

## Data Ethics and Explicability

*Explicability*, as it is used here, means the quality of being communicated in such a way that most people understand what a thing is or what a process does, its purpose and use. For instance, an algorithm is explicable if the average person could have a heuristic understanding of what it does (for example, rank news items by relevance to the reader based on their previous online behavior, such as clicking on and viewing articles or liking posts). Explicability does not require a person to know the technical details of how this is achieved (for example, the math behind an algorithm's ranking logic). Explicability is thus different from transparency. A transparent algorithm might be open source if everyone could read and check the code that is used to create it, but it is likely not explicable to most people.

A case in point relevant to public services concerns the fact that governments across the globe are increasingly adopting data dashboards that summarize progress on targets for staff and the public. They exemplify how data can provide an easy mechanism that encourages transparency for accountability. For example, dashboards can be used by citizens to monitor the progress of city governments' attempts to improve transport systems by making visible live traffic data, problems, government intervention, and response times (Matheus, Janssen, and Maheshwari 2020). It is, however, important to bear in mind that, like any technological fix, dashboards are no panacea leading to increased accountability (Matheus, Janssen, and Maheshwari 2020). They need to provide information that is actionable for civil society and other stakeholders.

In the context of public service delivery, which increasingly takes a networked or matrix-like form whereby multiple agents within and outside government collaborate to deliver public services, data for accountability need to communicate who responsible parties are and how they can be held to account if they do not deliver on promises. A case in point is the Inter-American Development Bank's MapaInversiones regional initiative, which is "an online platform that allows users to monitor the physical and financial

progress of public investment projects through data visualizations and geo-referenced maps” (Kahn, Baron, and Vieyra 2018, 23). As an evaluation suggested, the project successfully decreased the time taken to start and complete infrastructure projects. Those projects that published their details and progress via the platform fared better than those that did not (Kahn, Baron, and Vieyra 2018, 16).

Another risk that comes with increased transparency in the name of accountability is a decrease in the appetite for innovation. If administrations face a hostile media landscape or political pressure to paint over the challenges they face, an increase in the availability of data on what goes on inside government might stymie innovation. Public sector workers might face a reduction in their freedoms to think creatively and be entrepreneurial. They might be increasingly incentivized to manage to targets and targets only. Citizens would then face an inflexible administration without the capacity and incentive to adapt to changing needs. Thus, we return to the example of data dashboards: their availability must not distract from usability and explicability.

This chapter highlights explicability in particular because there is more demand for transparency regarding the algorithms and unsupervised machine-learning techniques used in public administration (Morley et al. 2020b). Making algorithms and code publicly available increases transparency, but it does not necessarily help citizens understand what they are confronting and how they can call for reform.

Increased data availability that supports accountability must incorporate qualitative aspects, lived experiences, and room for deliberation about what results mean for public service. Open government initiatives (such as the United Nations’ Open Government Data partnership and the World Bank’s Anti-corruption, Openness, Transparency, and Political Economy Global Solutions Group) and unilateral government efforts (such as the Open Knowledge Foundation) to make government data more accessible, explicable, and usable for a diverse group of stakeholders are good cases in point for how accountability can be at the center of data collection efforts.

### **Throughput**

*Throughput* describes behavior that “makes something happen.” This contrasts with plans, values, or ideals that relate to action but are not the actions that create the announced change. For instance, having a team of legal staff who regularly check contracts and account details to verify that employees with comparable skills and experience levels receive the same pay is the difference between having equal pay policies and providing throughput on them.

Where data concern the attitudes, opinions, and experiences of public sector staff, using these data for accountability promotion requires throughput. Surveying service providers (public sector workers in this case) or service users can make organizations appear considerate and proactive. However, if momentum and resources are lacking to enact change based on survey results, what is intended as an accountability mechanism can soon amount to window dressing. This is problematic in terms of monetary value—the time taken from staff and clients to answer surveys goes wasted—and in terms of trust in institutions. If accountability mechanisms are not used as promised, they can backfire. They can create mistrust, disengagement, and cynicism where they intended to foster trust, engagement, and optimism.

### **Unintended Consequences**

For public accountability, a government should know how much it spends on its workforce and who gets what. Many jurisdictions make the incomes of public sector employees public. The disclosure of salaries has propelled efforts to close gender disparities in pay (Marcal 2017). It might subsequently seem innocuous to track data on pay and incentives.

However, organizations typically know more than how much an employee earns. Many employers, particularly public sector employers, offer benefits such as health insurance and pension schemes, sometimes even bursaries for the educational and medical needs of employees’ families. From the types of pension and insurance arrangements chosen by an employee, an employer can easily see what types of health issues an employee might face and how many dependents profit from the employee’s benefits. This can create unintended breaches of privacy in the name of accountability.

For instance, while it is admirable that employers subsidize additional insurance schemes designed to cover extra costs associated with mental illness or cancer treatment, typically not covered by basic insurance packages, this also means that employers hold data on the mental and physical health of staff. Holding such data increases the risk associated with data breaches and data misuse. On top of the broad types of information on health an employer might glean from insurance choices, organizations that opt to provide their workforce with fitness trackers face an even more granular level of data and associated risk.

It can be argued that public sector employees can be subjected to greater levels of scrutiny, with more of their personal data accessible to public view, because they inhabit positions of power and the trade-off of privacy for power is just. We see such a trade-off with people who are considered to be public personas—politicians and celebrities—and it might not amount to overstepping to extend this category to include public sector employees. What is considered an infringement on privacy is, however, deeply embedded in social context and norms (Regan 2018). The dilemma created in this situation is that two sets of social norms are at odds: the belief that a person's address, name, income, and work activities should not be freely available for consultation on the internet versus the belief that this is justified when the individual belongs to the category of "public persona."

Navigating these dilemmas sustainably requires that, when individuals join social groups in which different norms on data privacy are applied, they do so knowingly and are provided with the opportunity to advocate to change these norms. In practice, this can mean giving public sector staff unions a voice in what data are made available about their members and what guarantees of accountability they can offer the public in place of such data.

## Innovation

Gaining an improved understanding of how public administrations perform is central to promoting evidence-based innovation. Promises of revolutionizing government via increased data collection, analysis, and insights abound. The World Bank released a report in early 2021 focused solely on the role of data in promoting growth and international development (World Bank 2021b).

In Indonesia, for instance, survey, census, administrative, and satellite data are being combined to help to plan infrastructure and improve urban sustainability (World Bank Data Team 2019). A report published in 2021 suggests that data on postal mail could be used to help stem drug smuggling and counterfeit trade (Shivakumar 2021). Furthermore, HR data were successfully used by Colonnelli, Prem, and Teso (2020) to capture patterns of politicized hiring and firing in the Brazilian public service. Election, media, and administrative and accountability data, such as scorecards, highlight where governments deliver on their promises of responsiveness and service delivery and where they lag behind (Brierley 2020; Erlich et al. 2021; Ingrams, Piotrowski, and Berliner 2020). The *Handbook* includes examples of how data on public administrators can help to create happier and more effective workplaces (chapters 1 and 2), flag malpractice (chapter 7), better assess the quality of government processes (chapter 13), and combat systemic institutional biases (chapter 20).<sup>2</sup>

## Welfare Consequentialism

In the context of data collection for public sector innovation, with a particular focus on data on the inner workings of government, ethical problems mainly surround two sets of questions. First are those concerned with welfare consequentialism. As outlined above, most public sector discourse stresses how data can be used for welfare gains through the alleviation of poverty, energy savings, or similar large-scale changes. Using such a logic, however, raises questions about *whose* benefit is targeted, who decides what is valuable, how valuable certain ends are, and whether, for a process to be considered ethical, evaluation should focus on actual or intended and expected consequences. Such concerns apply to the measurement of public administration as well as its outputs. It is commonplace in many countries for public sector employers to garner data on their recruits' police records, previous addresses (going back many years), social media accounts and public postings, credit scores, and other data as part of routine background checks in hiring



processes (see Erlam [2008] for a review of European and US law and practices regarding background checks). Is the end of all this data collection and consolidation a more effective recruitment process? Or are data sets routinely collected that impinge on officials' privacy but do not speak to their effectiveness?

Garnering insights for innovation will often require accumulating and linking data that have previously not been accumulated, linked, or analyzed in the same manner. Risks associated with this include the potential to breach the rights to privacy and confidentiality of individuals or employee groups. This can occur through poorly designed data collection, usage, and storage protocols, data breaches, or data vending strategies. The most defensible position to take for an evaluation of the morality of such data strategies in the public sector is one that defines what is morally right according to whether its *intended* as well as its *actual* consequences create good outcomes. In other words, systems should be designed with a view to preventing potential harm, and action plans should be available to mitigate or stop harmful actions when they are underway.

A case in point that highlights the need for a systems perspective on the intended and unintended consequences of using data for public sector innovation is the UK National Health Service (NHS). The NHS decided to sell individual and linked health data to third-party companies in order to use analytical skills outside government to help plan, forecast, and reform health care services. However, there have been doubts about whether the sale of data can be conducted in a way that prevents third parties from misusing these data—for instance, for marketing purposes (Rahman 2021; Rapp and Newman 2021; Rector 2021). It is also questionable whether, despite anonymizing details of the data (such as names and addresses), individuals and groups are protected from reidentification via triangulation with other data sources. Another example is provided by the departments of motor vehicles (DMVs) in the US states of Texas and Colorado, which sold service user data to commercial companies (New 2020; Whalen 2019). While the sale happened under the terms and conditions agreed to by service users, some of the sales were used for identity theft and document forgery by third parties who had legally acquired the data (Lieber 2020).

As these examples illustrate, while data protocols for the innovative use of public sector data will involve working with third sector parties, mission drift as much as the intentional misuse of data needs to be considered when designing data protocols. The examples also illustrate that the existence of rules and, in the case of the GDPR (which was still in force when the NHS data system was set up), the threat of high legal fines are not enough to guarantee that data systems generate usage patterns that can be widely accepted as ethical.

The most appealing solution to this problem in democratic contexts would be to involve service users and interested parties in working groups that are used to adapt systems. Apart from the hurdles inherent to all participatory approaches, such a solution faces the challenge that innovative uses of data more often than not involve new technologies, pregnant with new jargon to decipher. Creating data approaches for public sector innovation thus requires that time and resources be set aside to make the process explicable to those affected by it. This is no simple task because governments still face skills gaps in cutting-edge areas of information technology. In many instances, governments will need to rely on external expertise to develop and maintain the skills of their staff to implement data solutions that are ethically sound and secure.<sup>10</sup>

### ***Innovation and the Core Mandates of the Public Sector***

There is a danger that a greater push for data availability for innovation's sake will conflict with other demands on the public sector. Innovation in the short run can look unproductive: there are high costs and little immediate return.

Innovation and accountability can be at odds. For example, data on innovations related to defense, energy, and telecommunications infrastructures cannot be made readily available without weighing risks to national security. It is also unclear what rate of failure is acceptable for a public sector organization. Cost-benefit arguments that downweight the ethical risks associated with increased data collection and tracking efforts based on promises associated with the potential for public sector innovation should include such limitations in order not to overstate potential benefits.

This second set of concerns relates to how the imperative to innovate interacts with other core mandates of the public sector, such as enforcing rules, protecting fundamental rights, and stepping in where there are

market failures. Jordan (2014) postulates that the moral imperative for public servants to act within their assigned responsibilities is greater than the imperative for innovation, so that even when they willingly block innovation but stay within their assigned mandate, their guilt is limited. Such a claim becomes highly problematic as soon as we move outside a context where every group in society has equal access to enforcing accountability. Even in countries where electoral democracy flourishes, state institutions can be systematically biased against certain groups. Mere rule-following can then create significant welfare costs.

Over the last decade, the use of data-driven policing has provided many examples of how a genuine conviction to render public services more efficient and automatized can negatively affect swaths of society. Profiling algorithms have been shown, in some cases, to hide systematic discrimination deep inside their processing structures. In these situations, it is arguably more ethically acceptable for public servants to go out of their way to improve algorithms and change policing rather than block changes to the status quo because they do not fit their narrow job description.

When data collection for innovation concerns data on public servants themselves—colleagues, superiors, and supervisees—resistance to innovation seems equally misplaced. For instance, if public servants willfully hinder the collection of data on gender intended to promote gender-equitable hiring and promotion practices, they harm fellow citizens (their colleagues) by denying them access to services that other parts of government granted them, and they potentially harm society by obstructing reforms of the civil service that could help to improve society as a whole.

It should be noted that what is considered a transgression and what is considered part of a public servant's sphere of discretion is subject to debate—a debate that can be highly politicized. It would hence be most appropriate to revise Jordan's (2014) idea of a moral imperative to rule-following over innovation to a moral imperative to guard public welfare over rule-following, including directives related to innovation.

The age-old problem is that in the absence of rules and moral absolutes, the state apparatus needs to rely on the moral compasses of individuals and the norms reinforced by organizational culture to navigate toward the ethical delivery of public service. Appendix B therefore tries to provide a framework for public servants to evaluate new data approaches.

## CONCLUSION

While guides on research ethics abound, there is little guidance available for how ethical data collection, analysis, and innovation with data on public sector workers should take place. This chapter has attempted to provide such guidance via a heuristic that sees ethical challenges as falling broadly within three dimensions: an individual dimension, which comprises demands for dignity and privacy; a group dimension, which relates to voice and dissent; and a public-facing dimension, which ensures that data enable public administrators to deliver on public sector values. Each of these dimensions affords similar but also different prerogatives. The individual dimension calls for dignity and personal privacy, the group dimension relates to how data create and diminish voice, and the public-facing dimension concerns how data relate to public sector values. There is a wide range of values and considerations that can be framed as “public value.” This chapter has focused on three that are central to many discussions of public administration's effectiveness, efficiency, and measurement: productivity, accountability, and innovation. The section on productivity highlighted how important it is to choose metrics well and understand their strengths and biases (see more on this in sections 2 and 3 of the *Handbook*). The discussion of accountability presented the tensions in using data to increase the accountability of the public service by emphasizing explicability over mere transparency. The discussion of the tensions inherent to using data to create and measure innovation, as well as the delicate balance between accountability and innovation, showed that dialogue and regular reviews of the data strategies adopted by governments and researchers to measure public sector work and support innovation must be baked into the process to guard against delivering more to one value than another.



To make things more practical and support the reflective approach to measuring public sector work and employee effort, in appendix B we offer a framework that practitioners can use to build and review measurement strategies.

## NOTES

This chapter is indebted to extensive comments provided by Josh Cows (University of Oxford) and Johannes Himmelreich (Syracuse University) on dimensions of big data ethics and to Dan Rogger (World Bank) and Christian Schuster (University College London) on public administration and bureaucracy.

1. In the public sector, the right to individual dignity is curtailed by the public's right to dignity. This tension is explored in more detail further along in this chapter, in the section on the public dimension of data collection ethics.
2. For instance, in Qatar, 54 percent of the population is employed by the government. While employing half of the population is rare, employing 25–30 percent of the population is the norm in most middle-income and Western European countries. Data are from the International Labour Organization ILOSTAT “ILO modelled estimates database” (accessed January 2021), [ilostat.ilo.org/data](https://ilostat.ilo.org/data), via World Bank Open Data, “Employment in industry (% of total employment) (modeled ILO estimate),” World Bank, Washington, DC, <https://data.worldbank.org/indicator/SL.IND.EMPL.ZS?>
3. Action research is a form of research that uses mostly qualitative methodologies but can also involve the creation of survey questionnaires and the scales used to quantify responses. It is defined by an emphasis on making research participatory—involving the subjects of research and data collection actively in all stages of the research process, from defining the research question and the parameters of data collection to the use of data and results (for a more in-depth explanation, see, for example, Brydon-Miller, Greenwood, and Maguire 2003).
4. Note that *citizens* here is meant also to encompass persons without formal claims to citizenship who are government service users or fall within the jurisdiction of the government.
5. The theory being that these children perform worse typically as they struggle to fit in and potentially face discrimination. Minority teachers are hypothesized to be more cognizant of these problems and to create an atmosphere that is more inclusive and nurturing for minority pupils.
6. This dilemma, dubbed the “principal-agent problem,” has been widely discussed and is still researched in a variety of ways in economics, political science, and public administration.
7. For example, in some contexts, it has been demonstrated that political control over the bureaucracy increases instances of corruption and other malpractice (Brierley 2020), while in others, it can improve practices (Raffler 2020).
8. Lavertu (2016) sees an active role for public administration scholars in contextualizing, revising, and creating metrics to provide a more holistic assessment of public sector performance, in order to prevent misguided intervention by political and citizen principals.
9. As alluded to in earlier sections, a lot of rhetoric on public sector reform invokes an imperative to innovate. The underlying assumption is usually consequentialist: innovation will create benefits; therefore, it is good. Public sector innovation's merit can also be framed in terms of the collective virtue of creativity, an endeavor that is intrinsically worth pursuing, and extending it via a hedonistic logic, for the positive life experiences it can create.
10. Some governments have already developed schemes to encourage skills transfer into the government from outside for innovative practices. For example, Latvia runs a “shadow an entrepreneur” program, as part of which civil servants learn from private and third sector entrepreneurs about new developments in tech (OECD OPSI 2020).

## REFERENCES

- Alesina, Alberto, and Eliana La Ferrara. 2005. “Ethnic Diversity and Economic Performance.” *Journal of Economic Literature* 43 (3): 762–800. <https://www.jstor.org/stable/4129475>.
- Athey, Susan. 2017. “Beyond Prediction: Using Big Data for Policy Problems.” *Science* 355 (6324): 483–85. <https://doi.org/10.1126/science.aal4321>.
- Baer, Markus, and Michael Frese. 2003. “Innovation Is Not Enough: Climates for Initiative and Psychological Safety, Process Innovations, and Firm Performance.” *Journal of Organizational Behavior* 24 (1): 45–68. <https://doi.org/10.1002/job.179>.
- Bala, Sumathi. 2021. “Rise in Online Payments Spurs Questions over Cybersecurity and Privacy.” *CNBC*, July 1, 2021. <https://www.cnbc.com/2021/07/01/newdigital-payments-spur-questions-over-consumer-privacy-security-.html>.

- Basu, Tanya. 2020. "Vote Count Livestreams Are Here to Stay: Sit Back and Enjoy the Show as You Watch Election Results Get Tallied." *MIT Technology Review*, November 4, 2020. <https://www.technologyreview.com/2020/11/04/1011648/livestream-vote-counts-are-here-to-stay/>.
- BBC News. 2019. "Wikileaks: Document Dumps That Shook the World." April 12, 2019. <https://www.bbc.com/news/technology-47907890>.
- BBC News. 2021. "Conservative Conference: Get Off Your Pelotons and Back to Work, Says Oliver Dowden." October 5, 2021. <https://www.bbc.com/news/uk-politics-58804607>.
- Beioley, Kate. 2022a. "Brexit Dividend' Rule Change Prompts Fears over Data Flow with EU." *Financial Times*, February 28, 2022. <https://www.ft.com/content/8da85688-76f0-4a97-9e58-f701f3488d78>.
- Beioley, Kate. 2022b. "Metaverse vs Employment Law: The Reality of the Virtual Workplace." *Financial Times*, February 20, 2022. <https://www.ft.com/content/9463ed05-c847-425d-9051-482bd3a1e4b1>.
- Bevan, Gwyn. 2009. "Have Targets Done More Harm Than Good in the English NHS? No." *BMJ* 338: a3129. <https://doi.org/10.1136/bmj.a3129>.
- Björknef, Kristoffer, Luíza Cardoso de Andrade, Benjamin Daniels, and Maria Ruth Jones. 2021. *Development Research in Practice: The DIME Analytics Data Handbook*. Washington, DC: World Bank. <https://openknowledge.worldbank.org/handle/10986/35594>.
- Boot, Eric R. 2020. "The Feasibility of a Public Interest Defense for Whistle-Blowing." *Law and Philosophy* 39 (1): 1–34. <https://doi.org/10.1007/s10982-019-09359-1>.
- Brierley, Sarah. 2020. "Unprincipled Principals: Co-opted Bureaucrats and Corruption in Ghana." *American Journal of Political Science* 64 (2): 209–22. <https://doi.org/10.1111/ajps.12495>.
- Brydon-Miller, Mary, Davydd Greenwood, and Patricia Maguire. 2003. "Why Action Research?" *Action Research* 1 (1): 9–28. <https://doi.org/10.1177/14767503030011002>.
- Buell, Ryan W., Tami Kim, and Chia-Jung Tsay. 2017. "Creating Reciprocal Value through Operational Transparency." *Management Science* 63 (6): 1673–95. <https://doi.org/10.1287/mnsc.2015.2411>.
- Cabinet Office. 2022. *Civil Service Sickness Absence, 2021: Report*. London: UK Government. <https://www.gov.uk/government/publications/civil-service-sickness-absence-2021/civil-service-sickness-absence-2021-report>.
- Card, David, Thomas Lemieux, and W. Craig Riddell. 2020. "Unions and Wage Inequality: The Roles of Gender, Skill and Public Sector Employment." *Canadian Journal of Economics/Revue canadienne d'économie* 53 (1): 140–73. <https://doi.org/10.1111/caje.12432>.
- Carpenter, David. 2017. "The Quest for Generic Ethics Principles in Social Science Research." In *Finding Common Ground: Consensus in Research Ethics across the Social Sciences*, edited by Ron Iphofen, 3–17. Advances in Research Ethics and Integrity 1. Bingley, UK: Emerald Publishing. <https://doi.org/10.1108/S2398-601820170000001001>.
- CBS and the Associated Press. 2015. "Millions More Government Fingerprints Deemed Stolen." *CBS News*, September 23, 2015. <https://www.cbsnews.com/news/millions-more-government-fingerprints-deemed-stolen/>.
- Clark, Pilita. 2021. "Employee Data Can Be Used for Good, but Treat It with Care." *Financial Times*, July 24, 2021. <https://www.ft.com/content/b9cf429e-5529-47cf-88e0-7122aaed0240>.
- Cofone, Ignacio N. 2020. *The Right to Be Forgotten: A Canadian and Comparative Perspective*. New York: Routledge.
- Colonnelli, Emanuele, Mounu Prem, and Edoardo Teso. 2020. "Patronage and Selection in Public Sector Organizations." *American Economic Review* 110 (10) (October): 3071–99. <https://doi.org/10.3886/E118165V1>.
- Cooper, Christopher A. 2020a. "Impartiality and Public Sector Employees' Online Political Activity: Evidence from Three Australian Elections." *Acta Politica* 57: 210–34. <https://doi.org/10.1057/s41269-020-00181-5>.
- Cooper, Christopher A. 2020b. "Public Servants, Anonymity, and Political Activity Online: Bureaucratic Neutrality in Peril?" *International Review of Administrative Sciences* 86 (3): 496–512. <https://doi.org/10.1177/0020852318780452>.
- Dahl, Robert Alan. 1998. *On Democracy*. New Haven, CT: Yale University Press.
- De Witte, Kristof, Benny Geys, and Catharina Solondz. 2014. "Public Expenditures, Educational Outcomes and Grade Inflation: Theory and Evidence from a Policy Intervention in the Netherlands." *Economics of Education Review* 40: 152–66. <https://doi.org/10.1016/j.econedurev.2014.02.003>.
- Düwell, Marcus, Jens Braarvig, Roger Brownsword, and Dietmar Mieth, eds. 2014. *The Cambridge Handbook of Human Dignity: Interdisciplinary Perspectives*. Cambridge: Cambridge University Press.
- Erlam, N. Alexander. 2008. "Understanding International Background Checks in the Age of Privacy and Data Security Concerns." *International In-House Counsel Journal* 1 (3): 489–98. <https://www.iicj.net/subscribersonly/08april/iicjapr4-data-pricay-nalexandererlam-vericalscreen.pdf>.
- Erlich, Aaron, Daniel Berliner, Brian Palmer-Rubin, and Benjamin E. Bagozzi. 2021. "Media Attention and Bureaucratic Responsiveness." *Journal of Public Administration Research and Theory* 31 (4) (February): 687–703. <https://doi.org/10.1093/jopart/muab001>.
- Ford, Michael R. 2022. "A Little Representation Goes a Long Way: Minority Teacher Representation and District Performance in a Highly Homogenous Context." *Public Organization Review* 22: 691–705. <https://doi.org/10.1007/s11115-021-00535-3>.

- Fung, Brian. 2019. "Lawmakers Want to Ban 'Dark Patterns,' the Web Designs Tech Companies Use to Manipulate You." *Washington Post*, April 9, 2019. <https://www.washingtonpost.com/technology/2019/04/09/policymakers-are-sounding-alarm-dark-patterns-manipulative-web-design-trick-youve-never-heard/>.
- Gailmard, Sean. 2002. "Expertise, Subversion, and Bureaucratic Discretion." *Journal of Law, Economics, and Organization* 18 (2): 536–55. <https://www.jstor.org/stable/3555054>.
- Grantham, Kathleen. 2020. *Mapping Gender Data Gaps: An SDG Era Update*. Washington, DC: Data2x.
- Grisson, Jason A., Samantha L. Viano, and Jennifer L. Selin. 2016. "Understanding Employee Turnover in the Public Sector: Insights from Research on Teacher Mobility." *Public Administration Review* 76 (2): 241–51. <https://doi.org/10.1111/puar.12435>.
- Gubb, James. 2009. "Have Targets Done More Harm Than Good in the English NHS? Yes." *BMJ* 338: a3130. <https://doi.org/10.1136/bmj.a3130>.
- Guerin, Benoit, Alex Thomas, Rhys Clyne, and Suhasini Vira. 2021. *Finding the Right Skills for the Civil Service*. London: Institute for Government. <https://www.instituteforgovernment.org.uk/sites/default/files/publications/civil-service-skills.pdf>.
- Heath, Joseph, and Wayne Norman. 2004. "Stakeholder Theory, Corporate Governance and Public Management: What Can the History of State-Run Enterprises Teach Us in the Post-Enron Era?" *Journal of Business Ethics* 53 (3): 247–65. <https://doi.org/10.1023/B:BUSI.0000039418.75103.ed>.
- Hernández-Julián, Rey, and Adam Looney. 2016. "Measuring Inflation in Grades: An Application of Price Indexing to Undergraduate Grades." *Economics of Education Review* 55: 220–32. <https://doi.org/10.1016/j.econedurev.2016.11.001>.
- Hirschman, Albert O. 1970. *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Cambridge, MA: Harvard University Press.
- Hjort, Jonas. 2014. "Ethnic Divisions and Production in Firms." *The Quarterly Journal of Economics* 129 (4): 1899–946. <https://doi.org/10.1093/qje/qju028>.
- Hunter, Tatum. 2021. "Here Are All the Ways Your Boss Can Legally Monitor You." *Washington Post*, September 24, 2021. <https://www.washingtonpost.com/technology/2021/08/20/work-from-home-computer-monitoring/>.
- Ingrams, Alex, Suzanne Piotrowski, and Daniel Berliner. 2020. "Learning from Our Mistakes: Public Management Reform and the Hope of Open Government." *Perspectives on Public Management and Governance* 3 (4) (December): 257–72. <https://doi.org/10.1093/ppmgov/gvaa001>.
- Israel, Mark. 2015. *Research Ethics and Integrity for Social Scientists: Beyond Regulatory Compliance*. 2nd ed. London: Sage. <https://doi.org/10.4135/9781473910096>.
- Izzo, Zachary, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. 2021. "Approximate Data Deletion from Machine Learning Models." In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, edited by Arindam Banerjee and Kenji Fukumizu, 2008–16. Proceedings of Machine Learning Research 130. <http://proceedings.mlr.press/v130/izzo21a.html>.
- Jordan, Sara R. 2014. "The Innovation Imperative: An Analysis of the Ethics of the Imperative to Innovate in Public Sector Service Delivery." *Public Management Review* 16 (1): 67–89. <https://doi.org/10.1080/14719037.2013.790274>.
- Kahn, Theodore, Alejandro Baron, and J. Cruz Vieyra. 2018. "Digital Technologies for Transparency in Public Investment: New Tools to Empower Citizens and Governments." Discussion Paper IDB-DP-634, Inter-American Development Bank, Washington, DC. <http://dx.doi.org/10.18235/0001418>.
- Keiser, Lael R., and Joe Soss. 1998. "With Good Cause: Bureaucratic Discretion and the Politics of Child Support Enforcement." *American Journal of Political Science* 42 (4): 1133–56. <https://doi.org/10.2307/2991852>.
- Kenny, Kate. 2019. *Whistleblowing: Toward a New Theory*. Cambridge, MA: Harvard University Press.
- Kingsley, J. Donald. [1944] 2003. "Representative Bureaucracy." In *Representative Bureaucracy: Classic Readings and Continuing Controversies*, edited by Julie Dolan and David H. Rosenbloom, 12–18. New York: Routledge.
- Lauletta, Maximiliano, Martín Antonio Rossi, Juan Cruz Vieyra, and Diego Arisi. 2019. "Monitoring Public Investment: The Impact of MapaRegalias in Colombia." Working Paper IDB-WP-1059, Inter-American Development Bank, Washington, DC. <https://ideas.repec.org/p/ib/brikps/9967.html>.
- Lavertu, Stéphane. 2016. "We All Need Help: 'Big Data' and the Mismeasure of Public Administration." *Public Administration Review* 76 (6): 864–72. <https://doi.org/10.1111/puar.12436>.
- Lieber, Dave. 2020. "After 27 Million Driver's License Records Are Stolen, Texans Get Angry with the Seller: The Government." *Dallas Morning News*, November 26, 2020. <https://www.dallasnews.com/news/watchdog/2020/11/26/after-27-million-drive-rs-license-records-are-stolen-texans-get-angry-with-the-seller-government/>.
- Lynn, Matthews. 2022. "If We Sacked Working-from-Home Civil Servants, Would Anyone Notice?" *Telegraph*, May 12, 2022. <https://www.telegraph.co.uk/news/2022/05/12/sacked-working-from-home-civil-servants-would-anyone-notice/>.
- Macdonald, Sharon. 2017. "Embedded Ethics and Research Integrity: A Response to 'The Quest for Generic Ethics Principles in Social Science Research' by David Carpenter." In *Finding Common Ground: Consensus in Research Ethics across the Social Sciences*, edited by Ron Iphofen, 29–35. Advances in Research Ethics and Integrity 1. Bingley, UK: Emerald Publishing. <https://doi.org/10.1108/S2398-601820170000001003>.

- Madrigal, Alexis C. 2012. "Reading the Privacy Policies You Encounter in a Year Would Take 76 Work Days." *Atlantic*, March 1, 2012. <https://www.theatlantic.com/technology/archive/2012/03/reading-the-privacy-policies-you-encounter-in-a-year-would-take-76-work-days/253851/>.
- Marcal, Katrin. 2017. "Sweden Shows That Pay Transparency Works." *Financial Times*, July 27, 2017. <https://www.ft.com/content/2a9274be-72aa-11e7-93ff-99f383b09ff9>.
- Matheus, Ricardo, Marijn Janssen, and Devender Maheshwari. 2020. "Data Science Empowering the Public: Data-Driven Dashboards for Transparent and Accountable Decision-Making in Smart Cities." *Government Information Quarterly* 37 (3): 101284. <https://doi.org/10.1016/j.giq.2018.01.006>.
- Mehr, Hila. 2017. Artificial Intelligence for Citizen Services and Government. Ash Center for Democratic Governance and Innovation, Harvard Kennedy School, Cambridge, MA. <https://ash.harvard.edu/files/ash/files/artificialintelligenceforcitizen-services.pdf>.
- Meier, Kenneth J. 2019. "Theoretical Frontiers in Representative Bureaucracy: New Directions for Research." *Perspectives on Public Management and Governance* 2 (1) (March): 39–56. <https://doi.org/10.1093/ppmgov/gvy004>.
- Meier, Kenneth J., Joseph Stewart Jr., and Robert E. England. 1991. "The Politics of Bureaucratic Discretion: Educational Access as an Urban Service." *American Journal of Political Science* 35 (1): 155–77. <https://doi.org/10.2307/2111442>.
- Miller, Gary J., and Andrew B. Whitford. 2016. *Above Politics*. Cambridge, UK: Cambridge University Press.
- MIT Technology Review and Oracle. 2016. *The Rise of Data Capital*. Cambridge, MA: MIT Technology Review Custom in Partnership with Oracle. <http://files.technologyreview.com/whitepapers/MITOracle+Report-TheRiseofDataCapital.pdf>.
- Mohan, Bhavya, Ryan W. Buell, and Leslie K. John. 2020. "Lifting the Veil: The Benefits of Cost Transparency." *Marketing Science* 39 (6): 1105–21. <https://doi.org/10.1287/mksc.2019.1200>.
- Moore, Mark H. 1995. *Creating Public Value: Strategic Management in Government*. Cambridge, MA: Harvard University Press.
- Morley, Jessica, Josh Cows, Mariarosaria Taddeo, and Luciano Floridi. 2020a. "Ethical Guidelines for COVID-19 Tracing Apps." *Nature* 582: 29–31. <https://doi.org/10.1038/d41586-020-01578-0>.
- Morley, Jessica, Luciano Floridi, Libby Kinsey, and Anat Elhalal. 2020b. "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices." *Science and Engineering Ethics* 26 (4): 2141–68. <https://doi.org/10.1007/s11948-019-00165-5>.
- Nehf, James P. 2003. "Recognizing the Societal Value in Information Privacy." *Washington Law Review* 78 (1): 1–91. <https://digitalcommons.law.uw.edu/wlr/vol78/iss1/2>.
- Nembhard, Ingrid M., and Amy C. Edmondson. 2006. "Making It Safe: The Effects of Leader Inclusiveness and Professional Status on Psychological Safety and Improvement Efforts in Health Care Teams." *Journal of Organizational Behavior* 27 (7): 941–66. <https://doi.org/10.1002/job.413>.
- New, Brian. 2020. "I-Team: Texas DMV Made USD 3M Last Year Selling Drivers' Personal Information." *CBS Dallas-Fort Worth*, February 20, 2020. <https://dfw.cbslocal.com/2020/02/20/texas-dmv-selling-drivers-personal-information/>.
- Nur-tegin, Kanybek, and Keith Jakee. 2020. "Does Corruption Grease or Sand the Wheels of Development? New Results Based on Disaggregated Data." *The Quarterly Review of Economics and Finance* 75: 19–30. <https://doi.org/10.1016/j.qref.2019.02.001>.
- OECD (Organisation for Economic Co-operation and Development). 2001. *Citizens as Partners: OECD Handbook on Information, Consultation and Public Participation in Policy-Making*. Paris: OECD Publishing. <https://doi.org/10.1787/9789264195578-en>.
- OECD (Organisation for Economic Co-operation and Development). 2019. *The Path to Becoming a Data-Driven Public Sector*. OECD Digital Government Studies. Paris: OECD Publishing. <https://www.oecd-ilibrary.org/governance/the-path-to-becoming-a-data-driven-public-sector059814a7-en>.
- OECD (Organisation for Economic Co-operation and Development). 2022. *Recommendation of the Council Concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data*. OECD/LEGAL/0188. Paris: OECD Publishing.
- OECD OPSI (Organisation for Economic Co-operation and Development Observatory of Public Sector Innovation). 2020. *Embracing Innovation in Government: Global Trends 2020: Upskilling and Investing in People*. Paris: OECD Publishing. <https://trends.oecdopsi.org/trend-reports/upskilling-and-investing-in-people/>.
- Office for National Statistics. 2022. "Sickness Absence in the UK Labour Market: 2021." UK Government, April 29, 2022. <https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/labourproductivity/articles/sicknessabsenceinthelabourmarket/2021>.
- Office for Science. 2020. *Evidence and Scenarios for Global Data Systems: The Future of Citizen Data Systems*. London: UK Government. <https://www.gov.uk/government/publications/the-future-of-citizen-data-systems>.
- Open Data Institute. 2021. *Data Ethics: How to Be More Trustworthy with Data*. <https://theodi.org/service/consultancy/data-ethics/>.
- Pamuk, Humeyra, and Simon Lewis. 2021. "U.S. State Department Names Former Ambassador Gina Abercrombie-Winstanley as First Chief Diversity Officer." *Reuters*, April 12, 2021. <https://www.reuters.com/article/us-usa-diversity-diplomacy-idUSKBN2BZ2AD>.



- Panagiotopoulos, Panos, Bram Klievink, and Antonio Cordella. 2019. "Public Value Creation in Digital Government." *Government Information Quarterly* 36 (4): 101421. <https://doi.org/10.1016/j.giq.2019.101421>.
- Preston, Larry M. 1987. "Freedom and Bureaucracy." *American Journal of Political Science* 31 (4): 773–95. <https://doi.org/10.2307/2111224>.
- Raffler, Pia. 2020. "Does Political Oversight of the Bureaucracy Increase Accountability? Field Experimental Evidence from a Dominant Party Regime." *American Political Science Review* 116 (4): 1443–59. <https://doi.org/10.1017/S0003055422000181>.
- Rahman, Grace. 2021. "What's Happening with Your NHS Data?" *Full Fact*, June 8, 2021. <https://fullfact.org/health/nhs-data/>.
- Rapp, Ben, and Sara Newman. 2021. "NHS Digital's GP Data-Scraping Plan Must Be Publicised and Delayed." *Computer Weekly*, June 7, 2021. <https://www.computerweekly.com/opinion/NHS-Digitals-GP-data-scraping-plan-must-be-publicised-and-delayed>.
- Rasul, Imran, and Daniel Rogger. 2015. "The Impact of Ethnic Diversity in Bureaucracies: Evidence from the Nigerian Civil Service." *American Economic Review* 105 (5): 457–61. <https://doi.org/10.1257/aer.p20151003>.
- Rector, Alan. 2021. "The NHS Data Grab: Why We Should Be Concerned about Plans for GPs' Records." *Guardian*, June 6, 2021. <https://www.theguardian.com/society/2021/jun/06/the-nhs-data-grab-why-we-should-be-concerned-about-plans-for-gps-records>.
- Regan, Priscilla. 2018. "Legislating Privacy: Technology, Social Values, and Public Policy." In *The Handbook of Privacy Studies: An Interdisciplinary Introduction*, edited by Bart van der Sloot and Aviva de Groot, 57–62. Amsterdam: Amsterdam University Press. <https://doi.org/10.1515/9789048540136-003>.
- Regan, Priscilla M. 2000. *Legislating Privacy: Technology, Social Values, and Public Policy*. Chapel Hill, NC: University of North Carolina Press.
- Resnik, David B. 2020. "Ethical Issues in Research Involving Employees and Students." In *The Oxford Handbook of Research Ethics* [online edition], edited by Ana S. Iltis and Douglas MacKay. Oxford: Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780190947750.013.40>.
- Reuters. 2022. "Nvidia Says Employee, Company Information Leaked Online after Cyber Attack." March 1, 2022. <https://www.reuters.com/technology/nvidia-says-employee-company-information-leaked-online-after-cyber-attack-2022-03-01/>.
- Rhodes, Rod A. W., and John Wanna. 2007. "The Limits to Public Value, or Rescuing Responsible Government from the Platonic Guardians." *Australian Journal of Public Administration* 66 (4): 406–21. <https://doi.org/10.1111/j.1467-8500.2007.00553.x>.
- Riivari, Elina, and Anna-Maija Lamsa. 2014. "Does It Pay to Be Ethical? Examining the Relationship between Organisations' Ethical Culture and Innovativeness." *Journal of Business Ethics* 124 (1): 1–17. <https://doi.org/10.1007/s10551-013-1859-z>.
- Rodriguez, N. 2018. "Expanding the Evidence Base in Criminology and Criminal Justice: Barriers and Opportunities to Bridging Research and Practice." *Justice Evaluation Journal* 1 (1): 1–14. <https://doi.org/10.1080/24751979.2018.1477525>.
- Schmutte, Ian M., and Lars Vilhuber. 2020. "Balancing Privacy and Data Usability: An Overview of Disclosure Avoidance Methods." In *Handbook on Using Administrative Data for Research and Evidence-Based Policy*. New Haven, CT: Yale University Press. <https://admindatahandbook.mit.edu/book/v1.0/discavoid.html>.
- Schroeder, Doris, and Abol-Hassan Bani-Sadr. 2017. "The Quest for Dignity." In *Dignity in the 21st Century: Middle East and West*, 1–8. Cham, Switzerland: Springer.
- Selin Davis, Lisa. 2021. "The Right to Be Forgotten: Should Teens' Social Media Posts Disappear as They Age?" *Washington Post*, June 14, 2021. <https://www.washingtonpost.com/lifestyle/2021/06/14/should-what-children-post-online-come-back-haunt-them-later-life/>.
- Shivakumar, Sujai. 2021. "Better Data on International Mail Packages Could Reduce Illegal Drugs and Counterfeit." Center for Data Innovation, May 17, 2021. <https://datainnovation.org/2021/05/better-data-on-international-mail-packages-could-reduce-illegal-drugs-and-counterfeits/>.
- Solove, Daniel J. 2008. *Understanding Privacy*. Cambridge, MA: Harvard University Press.
- Somani, Ravi. 2021. *Public-Sector Productivity (Part One): Why Is It Important and How Can We Measure It?* Washington, DC: World Bank. <https://documents.worldbank.org/en/publication/documents-reports/documentdetail/913321612847439794/public-sector-productivity-part-one-why-is-it-important-and-how-can-we-measure-it>.
- Sommer, David Marco, Liwei Song, Sameer Wagh, and Prateek Mittal. 2020. "Towards Probabilistic Verification of Machine Unlearning." arXiv:2003.04247v2 [cs.CR]. <https://doi.org/10.48550/arXiv.2003.04247>.
- Stone, Deborah A. 2002. *Policy Paradox: The Art of Political Decision Making*. New York: W. W. Norton.
- Suderman, Alan. 2021. "DC Police Victim of Massive Data Leak by Ransomware Gang." Associated Press, May 13, 2021. <https://apnews.com/article/police-technology-government-and-politics-1aedfc42a8dc2b004ef610d0b57edb9>.
- Taylor, Linnet, Luciano Floridi, and Bart van der Sloot. 2016. *Group Privacy: New Challenges of Data Technologies*. Cham, Switzerland: Springer.
- UK Government. 2015. "The Civil Service Code." March 16, 2015. <https://www.gov.uk/government/publications/civil-service-code/the-civil-service-code>.

- UN Data Revolution Group (United Nations Secretary-General's Independent Expert Advisory Group (IEAG) on a Data Revolution for Sustainable Development). 2014. *A World That Counts: Mobilising the Data Revolution for Sustainable Development*. UN Data Revolution Group. <https://www.undatarevolution.org/wp-content/uploads/2014/12/A-World-That-Counts2.pdf>.
- Vidgen, Richard, Giles Hindle, and Ian Randolph. 2020. "Exploring the Ethical Implications of Business Analytics with a Business Ethics Canvas." *European Journal of Operational Research* 281 (3): 491–501. <https://doi.org/10.1016/j.ejor.2019.04.036>.
- Viswanath, Shilpa. 2020. "Public Servants in Modern India: Who Are They?" In *The Palgrave Handbook of the Public Servant*, edited by Helen Sullivan, Helen Dickinson, and Hayley Henderson, 1–16. Cham, Switzerland: Springer. <https://doi.org/10.1007/978-3-030-03008-796-1>.
- Wang, Ray. 2012. "What a Big-Data Business Model Looks Like." *Harvard Business Review*, December 7, 2012. <https://hbr.org/2012/12/what-a-big-data-business-model>.
- Weingast, Barry R., and Mark J. Moran. 1983. "Bureaucratic Discretion or Congressional Control? Regulatory Policymaking by the Federal Trade Commission." *Journal of Political Economy* 91 (5): 765–800. <https://www.jstor.org/stable/1837369>.
- Whalen, Andrew. 2019. "DMVs across the Country Selling Your Driver's License Data for as Little as a Penny, Making Them Millions." *Newsweek*, September 6, 2019. <https://www.newsweek.com/dmv-drivers-license-data-database-integrity-department-motor-vehicles-1458141>.
- White, Ariel R., Noah L. Nathan, and Julie K. Faller. 2015. "What Do I Need to Vote? Bureaucratic Discretion and Discrimination by Local Election Officials." *American Political Science Review* 109 (1): 129–42. <https://doi.org/10.1017/S0003055414000562>.
- Williams, Sophie, and Frances Mao. 2022. "Winter Olympics: Athletes Advised to Use Burner Phones in Beijing." *BBC News*, January 18, 2022. <https://www.bbc.co.uk/news/world-asia-china-60034013>.
- World Bank. 2021a. *Europe and Central Asia Economic Update, Spring 2021: Data, Digitalization, and Governance*. Washington, DC: World Bank. <https://openknowledge.worldbank.org/handle/10986/35273>.
- World Bank. 2021b. *World Development Report 2021: Data for Better Lives*. Washington, DC: World Bank. <https://www.worldbank.org/en/publication/wdr2021>.
- World Bank Data Team. 2019. "7 Data Innovation Projects Win Funding to Tackle Local Challenges." *Data Blog. World Bank Blogs*, January 19, 2019. <https://blogs.worldbank.org/opendata/7-data-innovation-projects-win-funding-tackle-local-challenges>.
- Wright, R. George. 1987. "Speech on Matters of Public Interest and Concern." *DePaul Law Review* 37 (1): 27–52. <https://via.library.depaul.edu/law-review/vol37/iss1/3>.





## CHAPTER 7

# Measuring and Encouraging Performance Information Use in Government

Donald Moynihan

### SUMMARY

Public sector organizations can leverage government analytics to improve the quality of their services and internal functioning. But the existence of government analytics does not ensure its use. This chapter discusses how governments can measure the use of government analytics—or, in other words, conduct analytics on government analytics. The chapter assesses, in particular, one important component of government analytics: *performance management systems*. It does so by, first, contrasting different types of *performance information use*, including political and purposeful uses. The chapter next assesses different approaches to measure performance information use, including those based on surveys, administrative data, and qualitative inquiry. Finally, the chapter illustrates how the measurement of performance information use can be built over time and drawn upon to improve government analytics, using the example of the US Government Accountability Office (GAO). The chapter's findings underscore the importance of the robust measurement of government analytics use to enable governments to make the most of government analytics.

### ANALYTICS IN PRACTICE

- The existence of government analytics systems does not automatically translate to the use of government analytics. Measurement does not equate to the use of measures for management action. As a result, policy makers should pay close attention not only to whether government analytics is used but also to exactly how it is used.
- The use of government analytics should be purposeful and strategic, but it can often fall prey to passive, political, and perverse uses. Ideally, performance information should be used by government agencies to purposefully plan for achieving their goals, either through internal restructuring or decision-making.

Donald Moynihan is the McCourt Chair at the McCourt School of Public Policy, Georgetown University.

However, when the design of *performance management systems* generates distortionary incentives, these same systems can lead public officials to perversely manipulate data or use performance information for political gain. This puts a premium on measuring *performance information use* to understand such (ab)use.

- Government analytics use can be measured using administrative data, surveys, qualitative data, and decision-based inferences. Administrative data can track how frequently performance information (for example, on a dashboard) is accessed; surveys of public servants can inquire about the extent of the use of performance information and about different types of use; qualitative data (for example, through focus groups) can provide rich detail on how (elite) decision-making is shaped by performance information; and decision-based inferences can showcase how informing public servants or managers about information alters their decision-making.
- The effective measurement of government analytics use benefits from longer time horizons that enable learning from previous experiences, as well as reforms to strengthen government analytics. In the United States, the stability of the core elements of a performance management system has meant that measurement can track progress in performance information use over time in order to help reformers understand what works and what needs fixing and gradually incorporate those lessons into practice.

## INTRODUCTION

Government analytics enables governments to diagnose and improve public management across the public administration production function (see chapter 1). Yet the measurement of the internal machinery of government does not necessarily translate into the actual use of government analytics to improve public administration. This puts a premium on encouraging the use of government analytics, as a number of chapters in *The Government Analytics Handbook* discuss (see, for example, chapters 25 and 26). It also, however, underscores the importance of measuring whether—and how—government analytics is used by public sector organizations to improve their functioning. In other words, it underscores the importance of doing analytics on the use of government analytics.

This chapter discusses how this can be done, focusing on the case of *performance management systems*. Governments rely on performance management systems, sets of coordinated processes to measure, validate, disseminate, and make use of performance data within the government. Early government performance systems focused on performance measurement and have gradually transitioned, with varying degrees of success, to performance management (Moynihan and Beazley 2016). Performance management systems have often emphasized routines of data collection, including requirements to set strategic goals and short-term targets, and have measured some combination of outputs and outcomes. But the collection of performance information does not, by itself, ensure its use (Kroll 2015).

How can governments monitor progress in the use of performance measurements? This chapter provides guidance in this effort. First, it provides a conceptual grounding, discussing different types of *performance information use*, including political and purposeful uses. Subsequently, it moves to measurement: different ways of measuring performance information use, including engagement metrics based on administrative data, survey-based instruments, and qualitative inquiry. Finally, it illustrates, using the case of the US Government Accountability Office (GAO), how to construct effective governmentwide measures of performance information use and underscores their utility to improve performance management systems.

## THE (AB)USES OF PERFORMANCE INFORMATION: A TYPOLOGY

Performance management systems often struggle to demonstrate their own effectiveness. As a result, there is remarkably thin evidence on the causal effects of performance management systems on government performance

itself (Gerrish 2016). One important intermediate outcome of performance management systems, however, can be measured, and this is performance information use. By *performance information use*, I refer to the extent to which data are considered by and inform the decisions and behavior of public officials (Kroll 2015).

Performance information can be construed as an intermediate measure of the effectiveness of performance management systems. Performance measures cannot generate improved performance by themselves. To have an effect, they have to alter the judgment and decision-making of a human being. In other words, data must be effectively used for management. If data are left unused, it is impossible for them to improve public sector outcomes through changes in the behavior of civil servants.

## Types of Performance Information Use

Conceptually, it is useful to distinguish between four distinct types of performance information use (Kroll 2015; Moynihan 2009). Box 7.1 lays out four primary types of performance information use, along with the scholarly literature informing their definitions.

### BOX 7.1 Types of Performance Information Use

**Purposeful:** The central hope of performance management reformers is that public employees use data to improve program performance. Such improvements can come via goal-based learning that gives rise to efficiency improvements, better targeting of resources, and more informed strategic decisions, or by tying indicators to rewards or sanctions in contract arrangements.

**Passive:** Performance management reforms may result in passive reactions, in which officials do the minimum required to comply with requirements to create and disseminate information but do not actually use this information (Radin 2006). This approach is also more likely in hierarchical relationships, where actors often lack strong incentives to use data but are not penalized for not using them. Cynicism based on failed performance management reforms in the past increases the likelihood of a passive response because the current reform will be perceived as temporary. A passive response is also more likely if elected officials, stakeholders, and agency leaders demonstrate little real interest in implementing performance management tools. Where results-based reforms have a permanent statutory basis, it becomes more difficult for public servants to assume they can wait them out.

**Political:** Performance management reforms are grounded in a demand for public agencies to present evidence that they are performing. As a result, public employees may see performance information as a means to define their efforts and success. Performance data thereby become a means of advocacy in a political environment (Moynihan 2008). In many cases, employees have some degree of influence in selecting and measuring the performance goals by which they are judged. They are likely to select, disseminate, and interpret information that portrays them in a favorable light.

**Perverse:** In some cases, the pressures to maximize measured performance may be so great that agents will improve these measures in ways that are in conflict with the underlying goals of a program. Agents may game program indicators through a variety of tactics, including making up data, selecting easy-to-serve clients while neglecting more difficult ones (*cream-skimming*), focusing on measured goals at the expense of important unmeasured goals or values (*goal displacement*), changing performance goals to limit comparison across time, or manipulating measures (Courty and Marschke 2004). Gaming becomes more likely when strong financial, career, or reputational incentives are attached to performance indicators, as well as when measures only partially capture the underlying goal.

Source: Adapted from Moynihan (2009).

The purposeful use of performance information anchors the use of performance indicators in goal orientation, improving resource allocation and strategic decisions. As such, the purposeful use of performance indicators can improve the efficiency of public administration. There is, therefore, strong interest in creating an environment that pushes public officials from passive to purposeful performance information use, as well as in monitoring for and minimizing perverse forms of use.

Public servants are, in general, the primary users of performance data. Members of the public might support the idea of a performance-driven government in the abstract but cannot be expected to be active users of public sector performance data. The only settings in which the public may be more interested in performance data are those where there is market-based choice in the provision of services or where services have high personal salience to the public. Some examples are the choice of schools or medical providers. Elected officials may express interest in performance data and seek them out in specific domains but are generally reluctant to let them alter their ideological preferences or other factors that shape their decisions (Christensen and Moynihan 2020; Moynihan 2008).

Beyond making data available and attractive, designers of performance management systems ultimately have little control over whether members of the public or policy makers use data, but they have much more control over the environment of public servants and thus greater opportunity to shape that environment to ensure performance data are used. Public servants also enjoy extensive program-level knowledge that makes them more likely to understand how to make sense of detailed performance information that would be difficult for nonspecialist actors to understand. Thus, I focus on public servants' use of performance data.

## HOW IS PERFORMANCE INFORMATION USE MEASURED?

Performance information use is a cognitive behavior, a form of judgment and decision-making that is often difficult to observe. Methodological approaches to measuring the use of performance systems require design choices, each associated with trade-offs. I briefly consider these approaches and trade-offs before concluding that surveys are the most practical option for governments seeking to track the use of performance management systems, although administrative data and qualitative approaches can provide important complements.

### Engagement-Based Inferences on Performance Information Use, Using Administrative Data

The first—and most obvious—approach to measuring the extent of performance information use is to observe directly whether public officials engage with existing performance management systems. These engagement metrics can include, for instance, whether officials download performance data or use them in public discussions or reports.

This approach has the advantage of not relying on subjective reports and can be used in a nonintrusive way. At the same time, governments must be willing to invest in tracking real-time data on engagement by public officials with performance information. This often requires having a centralized website where information is made available. For example, Andersen and Moynihan (2016) observe that school principals are more likely to download available performance data under certain experimental conditions, including when they have been delegated more authority in hiring decisions and offered comparative information that they find more useful.

Such an approach may be especially useful to observe how public servants engage with the performance management system. There is a set of indicators available for doing so: how many times performance information has been downloaded, how much time public officials spend engaging with the data in, for example, a dashboard, and how many times performance information has been shared on social media, such as Twitter or organizational forums.

Engagement-based inference requires data quality checks and may become less reliable if employees know their behavior is being observed. Additionally, engagement measures can be biased toward frequent

but unimpactful use. After all, such measures provide information on the frequency of use but not on how performance information is used, for instance, in decision-making. A single instance of performance information use by a senior authority in an organization may result in a *low-frequency* measure in a dashboard tracking time spent on a performance dashboard. Yet this instance may be more impactful in terms of triggering organizationwide management changes than dozens of employees' consulting the dashboard.

## Survey-Based Inferences on Performance Information Use

A second popular approach to evaluate the use of performance data is through public servant surveys. To discuss survey-based inferences in practice, I highlight the example of the United States. In the US federal government, the GAO is responsible for measuring performance information use. A central part of the GAO's strategy is a series of periodic surveys of public employees about performance management. These have taken place in 1997, 2000, 2003, 2007, 2013, 2017, and 2020 (US GAO 1997, 2001, 2004, 2008, 2013, 2017, 2021). The result has been a series of insightful reports that track how well the performance management system is being implemented and offer suggestions for improvement (for example, US GAO 2005, 2008, 2013, 2014).

Over time, the GAO has used a consistent set of indicators to track different types of performance information use. This has allowed the GAO to assess variation in performance information use across agencies. The results of the surveys show variation consistent with logical sources (discussed below). Survey-based inferences are discussed below in the context of the US performance management system. Box 7.2 presents a few examples of the precise wording of the GAO survey question on the use of the performance management system, as well as potential indicators as they correspond to box 7.1 on the types of performance information use.

### BOX 7.2 US Government Accountability Office Survey Measures of Performance Information Use

For those program(s)/operation(s)/project(s) that you are involved with, to what extent, if at all, do you use the information obtained from performance measurement when participating in the following activities? Responses range from "to no extent" (0) to "to a very great extent" (4).

#### Passive performance information use

- Refining program performance measures
- Setting new or revising existing performance goals

#### Purposeful performance information use

- Developing program strategy
- Setting program priorities
- Allocating resources
- Identifying program problems to be addressed
- Taking corrective action to solve program problems
- Adopting new program approaches or changing work processes
- Coordinating program efforts with other internal or external organizations
- Identifying and sharing effective program approaches with others
- Developing and managing contracts
- Setting individual job expectations for the government employees the respondent manages or supervises
- Rewarding government employees that the respondent manages or supervises

Source: Adapted from US GAO (2021).



## Qualitative Inferences on Performance Information Use

Beyond quantitative measures of performance information use, qualitative assessments can provide contextual knowledge on how performance information is being used in government. Qualitative assessments include focus groups and structured interviews in which public servants are invited to share their experiences with performance information use. During these qualitative assessments, interviewers enable public servants and managers to share their own stories regarding performance information use.

Qualitative assessments thus provide a complementary perspective to measures of how public servants use performance information. Qualitative assessments provide a holistic assessment of how performance information use is embedded within organizational routines, public servants' perceptions, and their performance of everyday tasks. While information on organizational functioning may be collected through public servant surveys as well, interviews and focus groups provide an insider's perspective and insights into elite decision-making based on performance information within organizations, and they embed them within a narrative.

This narrative—for example, “managers do not use performance information”—may be explored using other sources of data. A narrative may also validate whether performance information is having its intended effect (purposeful use) or is being used perversely or politically within the government. For this reason—much in the spirit of Bridges and Woolcock in chapter 4 of this *Handbook*—qualitative inquiry can complement survey and administrative data to enable analysts to gain a more holistic understanding of the use of government analytics.

## Decision-Based Inferences on Performance Information Use

A final approach to measuring performance information use is to evaluate whether providing decision-makers with performance information changes observed behavior. For example, a decision-maker might be asked to make a judgment about a program or organization or a decision about budgets, management, or a contract. Researchers then observe how much this decision varies depending upon the presence or absence of performance data, variation in the content of the performance data, changes in the environment in which the data have been presented, and individual traits.

As might be apparent, the controlled nature of such studies tends to rely on hypothetical scenarios, even if researchers sometimes try to increase their generalizability by using actual policy makers or public managers as respondents and actual government analytics data. This artificiality is one constraint of this mode of study and one reason it is often not practical in trying to assess general patterns of performance information use among civil servants (though survey experiments could certainly be inserted into such surveys).

Despite these potential limitations, policy experiments introducing decision-makers to informational treatments on performance have generated a significant body of work. For example, Holm (2018) measures how performance indicators affect management decisions by analyzing how public school principals define school priorities after being informed of performance results. Indeed, he finds that school principals prioritize goals with lower performance and devote their efforts to improvement in these lagging areas.

It becomes more difficult to evaluate whether performance information affects decision-making at a governmentwide level. One reason is that the bountiful array of performance data that exists makes it difficult to isolate the influence of one piece of information. Some researchers have looked at budget recommendations by government officials and budget decisions by legislators to see how these are correlated with whether summative performance data were positive or negative.

For example, there is clear evidence that program-level performance scores issued by the George W. Bush administration in the United States affected the budget recommendations it made to Congress, with less clear evidence that Congress followed those recommendations (Moynihan 2013). In parliamentary systems, where the executive and legislature work together, there is a greater likelihood that performance-informed executive judgments match legislative decisions, according to some studies (Sohn, Han, and Bae 2022). Decision-based inference is an important tool for understanding whether performance information changes

behavior. However, it should be a complement to rather than a substitute for administrative data and survey-based or qualitative inquiries, which provide a relatively more holistic understanding of performance information use.

### Potential Biases from Performance Information Use

It is important to note that, beyond the typology of deliberate uses of performance information by decision-makers (for example, purposeful or perverse use), decision-makers may be subject to biases when consulting performance information (James et al. 2020). For example, there is ample evidence for the power of negativity bias: policy makers and members of the public pay more attention to negative rather than positive performance data (Nielsen and Moynihan 2017; Olsen 2015).

As a result, providing decision-makers with performance information and measuring their deliberate use is not enough: biases, ideological or political, may affect how decision-makers use this information (Baekgaard and Serritzlew 2016; Christensen et al. 2018; James and Van Ryzin 2017; Nielsen and Moynihan 2017). There is ample evidence that decision-makers select performance information that fits with their ideological beliefs and discount data that do not (James and Van Ryzin 2017). Policy makers are more resistant to efforts to reduce such biases than the general public, reflecting their status as more skilled and dedicated motivated reasoners (Christensen and Moynihan 2020).

## CASE APPLICATION: THE GAO'S ROLE IN THE MODERNIZATION ACT

The US federal government currently uses a performance management system designed under the Government Performance and Results Act Modernization Act (GPRAMA), enacted in 2010. The Office of Management and Budget (OMB), part of the Executive Office of the President, is charged with leading the implementation of the system, while the GAO, part of the legislative branch, is charged with monitoring it. Several lessons about how to systematically measure performance information use can be drawn from this experience—which is, arguably, the world's most sustained effort to measure performance information use over time across government.

### Lesson One: An independent agency evaluating progress through a consistent set of measurement tools enhances credibility and knowledge accumulation.

The GAO has issued reports evaluating progress in performance management and the use of performance information to Congress and made them publicly available. Some have been qualitative case analyses of performance reforms in particular agencies or focus groups of senior managers across agencies, but the primary source of information is the survey data collected from employees across the federal government over time.<sup>1</sup> The credibility of the GAO has helped ensure a high response rate and thus enhanced the credibility of the resulting information on performance information use in the US federal government.

### Lesson Two: Evaluations of progress over time in performance information use benefit from stability in performance information.

The stability of the performance management system has allowed for comparisons across time. The surveys of federal government employees and managers have asked the same core questions over time. The GAO has also added questions that reflect changes in the performance management system while keeping the basic performance information use measures intact. This has allowed for analysis of the relationship between

aspects of the performance management system that change with a standard set of items over time. The data thus show, for instance, that while early evaluations of the system were relatively critical, they have become more positive over time, reflecting a process of gradual learning (Moynihan and Kroll 2016).

### **Lesson Three: Data-sharing agreements with external researchers can provide additional insights into measures of performance information use.**

GAO reports have provided valuable insights into the progress of the performance management system. But they have also been able to generate substantial additional research and insight by soliciting suggestions from outside researchers about what questions to ask and by sharing the data when they were collected. These data have enabled researchers to generate a series of analyses. Such analyses are not constrained by the format of government reports and can ask different types of questions, combine survey with nonsurvey data, and use more advanced statistical methods. Federal government officials from both the OMB and the GAO have been responsive to this research, incorporating insights into budget and planning documents.

The GPRAMA also has lessons for encouraging performance information use more generally, complementing other chapters in the *Handbook* on this topic (for example, chapter 26). Because performance information was repeatedly measured in the US case, there is a stronger (survey) evidence base on the factors that encourage performance information use. This reinforces lesson three about sharing performance information use data with researchers. For instance, Moynihan and Kroll (2016; see also Kroll and Moynihan 2021) find that exposure to high-quality organizational routines—such as agency leaders’ identifying high-priority, agency-specific or cross-agency goals or the leadership team’s reviewing progress toward key goals every quarter, based on well-run, data-driven reviews—matters to performance information use.

Moynihan and Lavertu (2012) find that leaders who engage seriously with performance management systems encourage followers to do the same and that employees who have more discretion to make decisions are more likely to use performance data. Kroll and Moynihan (2015) also find that exposure to performance management training is associated with higher use, while Kroll and Moynihan (2018) find that managers who use program evaluations are more likely to use performance data. This has become particularly salient since the Obama administration, which built bridges between program evaluation research and performance managers in government, partly by institutionalizing performance staff, such as performance improvement officers, who view their job more broadly than reporting performance data.

As these examples underscore, analytics of repeated measurement of performance information use can help governments understand this use, the factors driving it, and how to improve performance management systems over time to further increase their use. The analytics of performance information use data can thus ensure governments make the most of their performance management systems.

## **CONCLUSION**

The existence of government analytics does not ensure its effective use. This chapter shows that government analytics data can be used in a range of ways. Data can be used purposefully for management improvements, as intended by the designers of government analytics systems. Data can also be misused politically, however, or perverted, distorting organizational incentives.

Government analytics should thus not only measure whether and how frequently data are being used but also for what purpose. Much like for government analytics generally, a range of data sources are available for such analytics, including administrative data (for example, tracking users on a dashboard), survey data (for example, surveys of public employees and managers), and qualitative inquiry, among others.

The case of the United States—arguably the world’s most sustained effort to measure performance information use over time across government—underscores that such analytics is particularly helpful with

certain design features. For instance, the US system benefits from a stable set of performance information use measures and the regular collection of data assessing progress but also from a community of practice interested in using this longer-run data on government analytics to understand which parts of performance information (or government analytics) are working and what needs fixing. This community of career officials within the executive branch, who oversee performance issues, GAO officials in Congress, who complete their analyses, and external researchers, who share their findings, has then seen the gradual incorporation of lessons into practice.

The analytics of government analytics use can help improve and institutionalize how the government uses performance data to generate better outcomes. Ultimately, government analytics can only change the way governments operate if public officials meaningfully engage with, analyze, and operationalize analytical insights from data. This chapter thus underscores the importance of robust measurement of the use of government analytics to enable governments to make the most of the tools at their disposal.

## NOTE

1. Rather than trying to survey all employees, the GAO engaged in random sampling of mid- and senior-level federal managers. In 1997 and 2000, it used mail surveys, but in 2003, it transitioned to email-based surveys linked to a web questionnaire (with follow-up phone efforts for nonrespondents).

## REFERENCES

- Andersen, S. C., and D. P. Moynihan. 2016. "Bureaucratic Investments in Expertise: Evidence from a Randomized Controlled Field Trial." *Journal of Politics* 78 (4): 1032–44.
- Baekgaard, M., and S. Serritzlew. 2016. "Interpreting Performance Information: Motivated Reasoning or Unbiased Comprehension." *Public Administration Review* 76 (1): 73–82.
- Christensen, J., C. M. Dahmann, A. H. Mathiasen, D. P. Moynihan, and N. B. G. Petersen. 2018. "How Do Elected Officials Evaluate Performance? Goal Preferences, Governance Preferences, and the Process of Goal Reprioritization." *Journal of Public Administration Research and Theory* 28 (2): 197–211.
- Christensen, J., and D. P. Moynihan. 2020. "Motivated Reasoning and Policy Information: Politicians Are More Resistant to Debiasing Interventions than the General Public." *Behavioural Public Policy*, FirstView, 1–22. <https://doi.org/10.1017/bpp.2020.50>.
- Courty, P., and G. Marschke. 2004. "An Empirical Investigation of Gaming Responses to Performance Incentives." *Journal of Labor Economics* 22 (1): 23–56.
- Gerrish, E. 2016. "The Impact of Performance Management on Performance in Public Organizations: A Meta-Analysis." *Public Administration Review* 76 (1): 48–66.
- Holm, J. M. 2018. "Successful Problem Solvers? Managerial Performance Information Use to Improve Low Organizational Performance." *Journal of Public Administration Research and Theory* 28 (3): 303–20.
- James, O., A. L. Olsen, D. P. Moynihan, and G. G. Van Ryzin. 2020. *Behavioral Public Performance: How People Make Sense of Government Metrics*. Cambridge, UK: Cambridge University Press.
- James, O., and G. G. Van Ryzin. 2017. "Motivated Reasoning about Public Performance: An Experimental Study of How Citizens Judge the Affordable Care Act." *Journal of Public Administration Research and Theory* 27 (1): 197–209.
- Kroll, A. 2015. "Drivers of Performance Information Use: Systematic Literature Review and Directions for Future Research." *Public Performance and Management Review* 38 (3): 459–86.
- Kroll, A., and D. P. Moynihan. 2015. "Does Training Matter? Evidence from Performance Management Reforms." *Public Administration Review* 75 (3): 411–20.
- Kroll, A., and D. P. Moynihan. 2018. "The Design and Practice of Integrating Evidence: Connecting Performance Management with Program Evaluation." *Public Administration Review* 78 (2): 183–94.
- Kroll, A., and D. P. Moynihan. 2021. "Tools of Control? Comparing Congressional and Presidential Performance Management Reforms." *Public Administration Review* 81 (4): 599–609.

- Moynihan, D. P. 2008. *The Dynamics of Performance Management: Constructing Information and Reform*. Washington, DC: Georgetown University Press.
- Moynihan, D. P. 2009. "Through a Glass, Darkly: Understanding the Effects of Performance Regimes." *Public Performance and Management Review* 32 (4): 592–603.
- Moynihan, D. P. 2013. "Advancing the Empirical Study of Performance Management: What We Learned from the Program Assessment Rating Tool." *American Review of Public Administration* 43 (5): 499–517.
- Moynihan, D. P., and I. Beazley. 2016. *Toward Next-Generation Performance Budgeting: Lessons from the Experiences of Seven Reforming Countries*. Directions in Development: Public Sector Governance. Washington, DC: World Bank.
- Moynihan, D. P., and A. Kroll. 2016. "Performance Management Routines That Work? An Early Assessment of the GPRA Modernization Act." *Public Administration Review* 76 (2): 314–23.
- Moynihan, D. P., and S. Lavertu. 2012. "Does Involvement in Performance Reforms Encourage Performance Information Use? Evaluating GPRA and PART." *Public Administration Review* 72 (4): 592–602.
- Nielsen, P. A., and D. P. Moynihan. 2017. "How Do Politicians Attribute Bureaucratic Responsibility for Performance? Negativity Bias and Interest Group Advocacy." *Journal of Public Administration Research and Theory* 27 (2): 269–83.
- Olsen, A. L. 2015. "Citizen (Dis)satisfaction: An Experimental Equivalence Framing Study." *Public Administration Review* 75 (3): 469–78.
- Radin, B. A. 2006. *Challenging the Performance Movement*. Washington, DC: Georgetown University Press.
- Sohn, H., D. Han, and K. B. Bae. 2022. "The Effect of Performance Information on Actual Budgeting Allocation: The Role of Policy Type." *Public Performance and Management Review* 45 (5): 1112–32. <https://doi.org/10.1080/15309576.2022.2062014>.
- US GAO (US General Accounting Office). 1997. *The Government Performance and Results Act: 1997 Governmentwide Implementation Will Be Uneven*. Report to Congressional Committees, GAO-97-109. Washington, DC: US GAO.
- US GAO (US General Accounting Office). 2001. *Managing for Results: Federal Managers' Views on Key Management Issues Vary Widely across Agencies*. Report to the Chairman, Subcommittee on Oversight of Government Management, Restructuring, and the District of Columbia, Committee on Governmental Affairs, US Senate, GAO-01-592. Washington, DC: US GAO.
- US GAO (US General Accounting Office). 2004. *Results-Oriented Government: GPRA Has Established a Solid Foundation for Achieving Greater Results*. Report to Congressional Requesters, GAO-04-38. Washington, DC: US GAO.
- US GAO (US Government Accountability Office). 2005. *Performance Budgeting: PART Focuses Attention on Program Performance, but More Can Be Done to Engage Congress*. Report to the Chairman, Subcommittee on Government Management, Finance, and Accountability, Committee on Government Reform, House of Representatives, GAO-06-28. Washington, DC: US GAO.
- US GAO (US Government Accountability Office). 2008. *Government Performance: Lessons Learned for the Next Administration on Using Performance Information to Improve Results*. Statement of Bernice Steinhardt, Director Strategic Issues. Testimony before the Subcommittee on Federal Financial Management, Federal Services, and International Security, Committee on Homeland Security and Governmental Affairs, US Senate, GAO-08-1026T. Washington, DC: US GAO.
- US GAO (US Government Accountability Office). 2013. *Managing for Results: Executive Branch Should More Fully Implement the GPRA Modernization Act to Address Pressing Governance Challenges*. Report to Congressional Committees, GAO-13-518. Washington, DC: US GAO.
- US GAO (US Government Accountability Office). 2014. *Managing for Results: Agencies' Trends in the Use of Performance Information to Make Decisions*. Report to Congressional Addressees, GAO-14-747. Washington, DC: US GAO.
- US GAO (US Government Accountability Office). 2017. *Managing for Results: Further Progress Made in Implementing the GPRA Modernization Act, but Additional Actions Needed to Address Pressing Governance Challenges*. Report to Congressional Committees, GAO-17-775. Washington, DC: US GAO.
- US GAO (US Government Accountability Office). 2021. *Evidence-Based Policymaking: Survey Results Suggest Increased Use of Performance Information across the Federal Government*. Report to Congressional Committees, GAO-22-103910. Washington, DC: US GAO.

## CHAPTER 8

# Understanding Corruption through Government Analytics

*James Anderson, David S. Bernstein, Galileu Kim, Francesca Recanatini, and Christian Schuster*

### SUMMARY

Corruption is a multidimensional phenomenon that affects governments and citizens across the world. Recent advances in data collection and analytics have generated new possibilities for both detecting and measuring corruption. This chapter illustrates how the public sector production function introduced in *The Government Analytics Handbook* helps holistically conceptualize where corruption can occur in public administration. It then outlines how corruption can be detected in its multiple dimensions using the microdata approaches outlined in the remaining chapters of this book. Along the production function, corruption can be detected with input data (for example, personnel or budget data), data about processes (for example, survey data on management practices), and output and outcome data (for example, public service delivery data). Using corruption as a thematic focus, the chapter thus showcases how the approaches presented in the *Handbook* can be combined and leveraged to holistically diagnose a key issue in public administration. The chapter complements this methodical discussion with a broader consideration of how political-economic constraints affect policy reforms to reduce corruption.

### ANALYTICS IN PRACTICE

- Corruption is a multidimensional phenomenon, affecting public administration across its parts. Corruption can affect public administration in personnel and payroll, through patronage appointments of public servants, and through the payment of ghost workers. Corruption can disrupt service delivery if public servants demand bribes in exchange for access to public services, such as health care and education. Since corruption affects public administration in various ways, a holistic assessment of corruption requires multiple data sources and methodologies.

Galileu Kim is a research analyst for the World Bank's Development Impact Evaluation (DIME) Department. James Anderson is a lead governance specialist for the World Bank's Governance Global Practice (GGP). David S. Bernstein is a lead public sector specialist for the GGP. Francesca Recanatini is a lead economist for the GGP. Christian Schuster is a professor at University College London.



- Recent advances in data collection and analytics have generated new possibilities to detect and measure corruption. The use of data and indicators on corruption is a long-standing tradition in international development. Previous efforts primarily relied on expert assessments and national-level indicators, which are susceptible to subjective bias and lack a foundation in microdata. More recently, there has been growth in administrative microdata, such as procurement, payroll, and other data sources, like surveys of public servants. These rich data environments create novel possibilities to engage in more granular analytics for measuring corruption.
- The public sector production function structures holistic conceptualization and measurement of corruption in public administration. This chapter showcases how conceptualizing public administration as a production function with distinct parts enables corruption to be unpacked into its different dimensions, such as procurement and service delivery. In procurement, for instance, corruption can take the form of the capture of bidding processes by companies or the bribery of public officials. To measure this type of corruption, procurement data can be analyzed to create red flags on bids with a single bidder.
- Since corruption cuts across the public administration production function, the integration of the different data sources and methodologies presented in other chapters in *The Government Analytics Handbook* enables a comprehensive assessment of how corruption operates. For instance, corruption by public servants can stem from different causes. Public servants who engage in corruption might be dissatisfied with their wages or pressured by their managers. Measuring this complex environment of agents and organizational dynamics requires multiple data sources. For instance, managerial pressure can be measured through public servant surveys, while payroll data can provide a sense of pay equity.
- Measurement of corruption can guide and complement public sector reforms, but it is not a substitute for the implementation of challenging policies to reduce corruption. Measuring where corruption occurs can guide public sector reform by detecting areas of vulnerability—for example, ghost workers—and informing reforms—for example, improving quality control and payroll enforcement. Measurement cannot substitute for the important step of implementing challenging policy reforms that will likely be resisted by agents who benefit from the status quo. Reformers should be cognizant of the political-economic environment, which may deter reforms from taking place.
- The multidimensional analytical approach presented in this chapter can be leveraged for other key topics in public administration. While the thematic focus of this chapter is corruption, we emphasize that other important topics in public administration can also benefit from the application of analytical approaches based on multiple data sources. For example, performance management could leverage survey data on public sector motivation and management practices, as well as administrative data on performance indicators such as delays in processing business licenses. Using multiple data sources and analytical approaches enables a more holistic understanding of how public administration maps onto these key issues.

## INTRODUCTION

Corruption in public administration has many faces (Campos and Pradhan 2007).<sup>1</sup> To name just two, corruption affects how public services work through nepotism and patronage—the use of political and personal connections for professional gain (Colonnelli, Prem, and Teso 2020; World Bank 2021). It can also shape how state resources are allocated, diverting funds from public education or health for private gain (for example, Ferraz, Finan, and Moreira 2012). This puts a premium on better understanding, measuring, and fighting against corruption. Throughout this chapter, we follow a common definition of corruption as “the use of public office for private gain” (Jain 2001; Rose-Ackerman 1978).

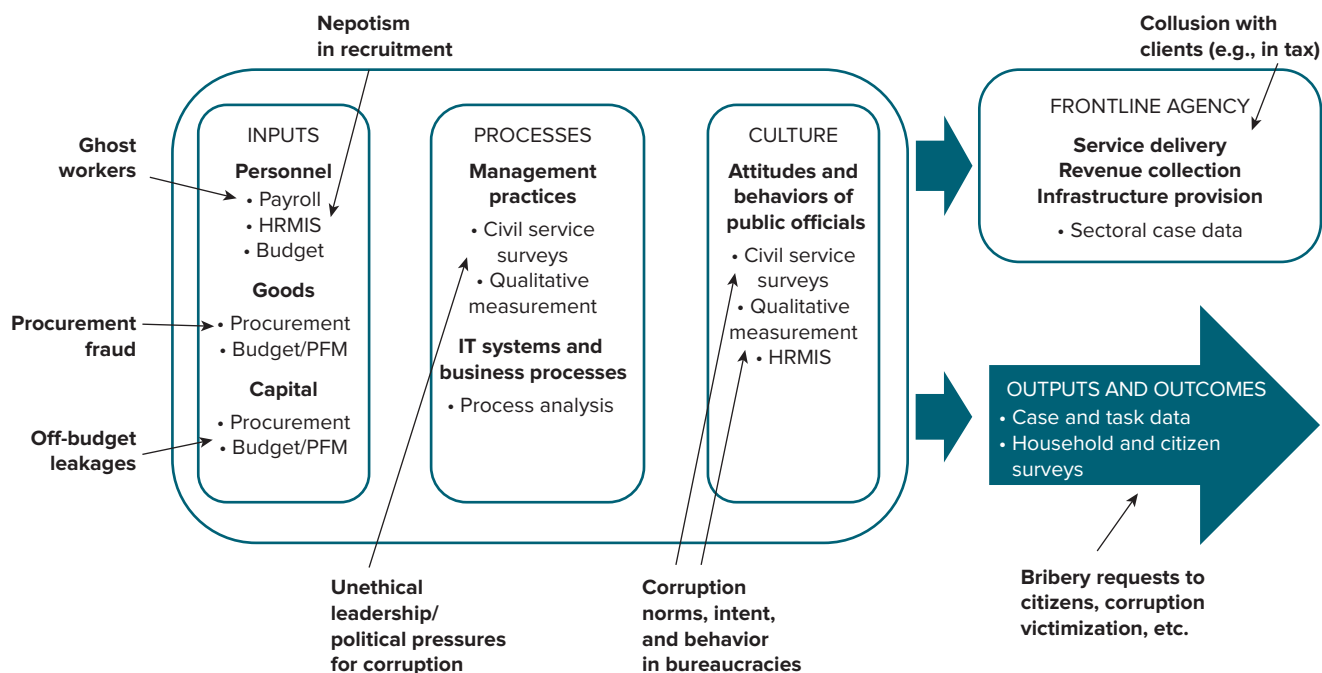
The use of government analytics—and data more broadly—to fight corruption is a long-standing tradition in the international development community (see, for example, Kaufmann, Pradhan, and Ryterman 1998).

Initiatives such as the World Bank's World Governance Indicators (WGI) and Transparency International's Corruption Perception Index (CPI) have sought to aggregate a set of indicators into a single measure to help draw attention to issues of governance and the control of corruption. These indicators often rely on expert surveys or qualitative indicators that provide national-level indicators on corruption. More recently, there has been growth in microdata analyzing corruption. For example, surveys of public servants provide staff-level perspectives on how corruption operates within countries and across sectors and government agencies (Recanatini 2011). The growing availability of microdata—administrative and survey-based—provides novel opportunities to increase and refine analytical approaches to understanding and reducing corruption.<sup>2</sup>

In this chapter, we demonstrate how to leverage the microdata sources and methodologies described in *The Government Analytics Handbook* to measure corruption. We do so through the public administration production function (figure 8.1). Corruption can be measured along the production function with input data on personnel—for example, human resources management information systems (HRMIS)—data about processes in the management of public servants—for example, surveys of public servants—and output and outcome data on the quality of service delivery—for example, service delivery measures. The following list provides a few examples of corruption along the production function:

1. **Inputs:** nepotism in hiring, procurement fraud, and off-budget leakages
2. **Processes:** unethical leadership by line managers and political pressures on public servants to act corruptly
3. **Culture and behavior in public administration:** whether public servants believe bribes are acceptable and corrupt behavior of public servants
4. **Outputs and outcomes in frontline agencies:** tax evasion in collusion with customs and tax officers
5. **Direct outputs and outcomes of public administration:** corruption in regulatory or policy decisions, bribery requests to citizens, and distorted allocation of licenses and permits.

**FIGURE 8.1** Examples of Corruption along the Public Administration Production Function



Source: Original figure for this publication.

Note: The public administration production function conceptualizes the public sector as different dimensions (personnel, management practices, and attitudes and behaviors) that connect to each other to produce outputs and outcomes. HRMIS = human resources management information systems; IT = information technology; PFM = public financial management.

While our analysis focuses on each sector of the production function individually, integrating analytics across multiple dimensions of the production function provides even greater analytical insights. For example, procurement data may be linked to personnel data to identify public servants who might benefit from discretion over contracting decisions. Management practices, such as unethical leadership, may have downstream effects on the norms and behaviors of public servants. For this reason, integrating data allows practitioners and researchers to assess risk factors associated with corruption holistically, connecting the different parts of the production function. We include in each section a brief discussion of how to integrate different data sources and methodological approaches.

Each section of this chapter focuses on a particular dimension of the production function and how to measure and identify where corruption occurs. As such, the chapter should be read as an overview of analytical approaches to measuring corruption rather than as a corruption topic chapter. To frame our discussion, in each section we provide a brief overview of relevant academic and policy literature. We then highlight how methodological tools in other chapters of the *Handbook* can be used to understand and, potentially, reduce corruption. For example, if practitioners are interested in corruption in personnel, tools outlined in chapter 9, such as compliance and control mechanisms for human resources (HR) data, may prove helpful. We provide suggestions to practitioners about how to implement the methods discussed.

Accumulated experience suggests that relying solely on data is not enough to identify and measure corruption. Public officials and private agents who stand to benefit from corruption have incentives not to disclose their corrupt behavior in both survey and administrative data. As emphasized in chapter 4, measurement efforts are subject to existing power dynamics: politicians who benefit from collusion in procurement can manipulate procurement indicators to their advantage—for instance, by misreporting the number of bidders. Beyond concerns about the integrity of measures, improving the measurement of corruption should be embedded in a wider strategy of public sector reform to reduce corruption. A reform strategy can, for example, reduce corruption through monitoring technologies (for example, audits or reporting mechanisms) and positive reinforcement (for example, ethics training).<sup>3</sup>

Finally, we highlight the importance of combining data analytics with recognition of the sensitive political-economic issues surrounding reforms to reduce corruption (Evans 1995). Resistance to reform may come from multiple stakeholders who have economic incentives to preserve the status quo, not just from public officials. Politicians, political appointees, high-ranking public servants, and large corporations may resist data collection and analytics on corruption. Politicians may collude with or pressure public officials for personal gain, derailing reforms that threaten them.<sup>4</sup> Survey data and interviews can help articulate the nature of these political dynamics and their intensity across the administration. Awareness of the institutional context within a country can guide reforms by securing buy-in from stakeholders and negotiating compromises that ensure the sustainability and effectiveness of reform efforts.

This chapter is structured as follows. Section 2 covers the input side of the production function, demonstrating how personnel, budget, and procurement data can be used to measure corruption. Section 3 dives into processes, such as management practices and business processes, that can be measured through a combination of survey and administrative data. Section 4 presents analytical approaches to measuring the norms and behaviors of public servants, particularly through surveys. Section 5 discusses corruption in frontline agencies, with a particular focus on service delivery, sectoral cases, and revenue collection. Section 6 covers the outputs and outcomes of public administration. Finally, we conclude.

## INPUTS

Inputs to public administration include personnel, goods, capital, and budgets. In this section, we provide an overview of the extant literature on personnel (HRMIS and payroll), budget, and procurement, highlighting how different types of corruption, such as patronage, fraud, and embezzlement, may impair inputs into the public administration production function. Drawing on the approaches of other chapters in the *Handbook*,

we also present indicators to measure these dimensions of corruption and discuss how to develop and implement them.

## Personnel and HRMIS Data

Personnel decisions, whether they regard selection or personnel management, have important consequences for public administration (Besley et al. 2022; Finan, Olken, and Pande 2017). Corruption may negatively affect personnel systems, particularly through patronage and nepotism, with long-term consequences (Evans and Rauch 1999; Rauch and Evans 2000). Patronage, the politically motivated selection and promotion of personnel, operates through the preferential hiring of copartisans (Brollo, Forquesato, and Gozzi 2017) or repayment for electoral contributions (Colonnelli, Prem, and Teso 2020). Patronage may adversely affect public servants' perceptions of the quality of governance and their general perceptions of corruption (Anderson, Reid, and Ryterman 2003). There can also be negative consequences for societal outcomes: the quality of government outputs such as health care and education can be compromised when people appointed based on political affiliation lack the skills or experience to perform critical functions.

It is important to consider not only how corruption occurs in personnel decisions (that is, through nepotism or patronage) but also how personnel systems affect corruption throughout public administration. A more meritocratic public service, for example, increases the opportunity cost for public servants who lose their jobs due to corruption (Cadot 1987). Conversely, if public servants believe that advancement is not based on good behavior, they may have an incentive to supplement their incomes through behavior that does not align with public policy goals (for example, bribes). Relatedly, if organizations are subject to high turnover due to political changes, officials will have a greater incentive to ensure their brief appointments pay off. Regarding the intensity of these political influences over bureaucratic careers, in Brazil's federal civil service, a quarter of civil servants believe that promotions are influenced by political connections (nepotism), and only 23.1 percent believe they are meritocratic (World Bank 2021). Such systematic prevalence of patronage may influence whether public servants engage in corruption.

Improvements in government analytics can assist in detecting and reducing corruption in personnel. Chapter 9 of the *Handbook* highlights a set of metrics that could be used to detect corruption in personnel. For example, talent management indicators that focus on recruitment—the number of applications per position or recruitment method (for example, competitive exam or political appointment)—can enable governments to better identify and measure cases of patronage. A lack of competitive exams or a low number of applicants may suggest a greater prevalence of patronage appointments. Performance indicators—the rate of performance reviews completed or employee ratings as completed by colleagues or supervisors—strengthen the measurement of meritocratic channels for promotion over political ones.

A publicly available analytics dashboard on public service can increase the transparency and accountability of personnel practices, as highlighted by the case study of Luxembourg in case study 1 of chapter 9. Moreover, HRMIS can enable governments to detect the risk of nepotism and patronage in recruitment by assessing similarity in the last names of public servants inside the same government organizations or, where such data are available, by linking name records to family records or political party membership records to understand where family or political party members are disproportionately hired into government (Bragança, Ferraz, and Rios 2015; Riaño 2021).

The implementation of analytical tools relies on robust data infrastructure, capacity, and attention to political-economic constraints. If HR data are to assist in the detection and reduction of corruption in personnel, governments need to establish processes for quality and compliance controls in HRMIS data to reduce gaps in the coverage and frequency of data. However, government agencies in which patronage is more common may resist sharing or even generating data on recruitment practices precisely to reduce this scrutiny. Additionally, there is the key issue of the sustainability of data collection. Governments often

launch new data collection efforts, but these efforts are not replicated over time. Collaborations with national statistical agencies and offices could ensure the sustainability of these efforts.<sup>5</sup>

Some countries do not have the necessary resources to implement an HRMIS. Thus, surveys and other tools are often used while an HRMIS is being designed and implemented. At the same time, it is important to complement efforts to generate HRMIS data on personnel with surveys of public servants and focus group discussions about experiences of corruption in personnel management. Political leadership from key stakeholders, such as the office of civil service, and broader institutional coordination may be necessary to reduce resistance by particular agencies.

## Payroll

Corruption in payroll occurs through irregular payments to public servants, either through undue payments to ghost workers, who do not perform their service duties (Das and Hammer 2014; La Cascia et al. 2020), or through the collection of payments that exceed established guidelines (World Bank 2019). Payroll irregularities waste valuable public resources. In the context of fiscal constraint, irregular payments in excess of existing regulations may compromise the sustainability of the wage bill and lower citizens' trust.<sup>6</sup> The irregular concentration of wages among a few public servants may lead to payroll inequalities that pose challenges to sustainability and may arise from public servants' wielding power to accumulate these resources (World Bank 2019). Reducing corruption in payroll is therefore an important policy objective for governments.

The principled use of payroll data, as well as the establishment of control and compliance mechanisms, can assist in curbing corruption in payroll. Chapter 9 outlines how indicators of payroll data can enable the accurate and timely monitoring of payroll in public administration. For example, a first exercise involves calculating the size of the wage bill and whether it complies with the actual budget allocation. A simple head count of public servants appearing in the payroll data identifies duplicated entries. A breakdown of wage-bill expenditures by sector, administrative unit, or territory enables a granular analysis of historical trends in payroll expenditure and a comparison between similar units to detect evidence of irregularity. Death and pension records can be cross-referenced with payroll data as well as attendance records to identify ghost workers. We note, however, that this exercise requires updated and reliable information systems. Compliance and control mechanisms, such as budget audits, should be set in place.

The use of digital technologies, such as machine learning, can assist in the detection of payroll irregularities, as outlined in chapter 16 and case study 2 of chapter 9. In particular, historical payroll data that are classified by payroll analysts as irregular can be used to train machine-learning algorithms to automatically classify irregularities in payroll entries. Given the large volume of payroll data being generated at any given point in public administration, these automated processes can complement and enhance the work of payroll analysts, enabling them to detect irregularities that would otherwise remain undetected. However, it is important to note that, in order to develop these advanced digital solutions, a robust payroll data infrastructure has to be set in place. Payroll data are often fragmented and decentralized. As chapter 9 outlines, a reform process may be necessary to integrate and standardize HRMIS data, which is demonstrated through its operational framework for HRMIS reforms.

The implementation of payroll reforms requires coordination with relevant stakeholders, such as the ministry of finance or the head of the civil service, and cognizance of the political-economic context. In particular, leadership support for reforms is necessary, as is navigating resistance from actors who benefit from the status quo, such as ghost workers. In contexts in which payroll data are highly decentralized, institutional coordination and data-sharing agreements are necessary to ensure that payroll data coverage improves. Additionally, an advisory rather than a punitive approach to payroll reform is recommended, in particular when justice systems are weak and unable to enforce regulations. While some duplications or excessive benefits may be intentional, they are often the result of a lack of knowledge or inadequate training, as well as legacy systems that are not regularly updated. As a result, onboarding to new control and compliance mechanisms, rather than punishing infractions outright, may reduce resistance to and ensure the uptake of policy reforms.

## Budget

Budget data measure how public resources are spent throughout the entirety of the public administration. They include multiple sectors within the government, such as procurement and payroll. Due to its cross-cutting nature, the budget is exposed to different types of corrupt behavior. Corruption may manifest itself in budget leakages: resources that are unaccounted for in the flow of public resources across administrative levels of government (Gurkan, Kaiser, and Voorbraak 2009). In extreme cases, corruption occurs through embezzlement, the diversion of public funds for personal gain (Hashim, Farooq, and Piatti-Fünfkirchen 2020). Corruption in public expenditure may also have a distortionary effect, misallocating resources to less productive sectors of the economy and ultimately inhibiting economic growth (d'Agostino, Dunne, and Pieroni 2016). Because of its potential negative effects, corruption in the budget has been the subject of extensive policy debate, particularly in public financial management (World Bank 2020). Methodologies to detect corruption in budgets include the Public Expenditure and Financial Accountability (PEFA) assessment and Public Expenditure Tracking Surveys (PETS), which are discussed in greater detail in chapter 11.<sup>7</sup>

Chapter 11 also provides guidance on how to build more robust data infrastructures for public expenditures. It outlines five guiding principles that should be respected in designing and maintaining public expenditure data: data provenance and integrity, comprehensiveness, utility, consistency, and stability. These principles ensure that the sources of expenditure data are documented and accounted for and that data are comparable and stable across public administration. One simple measure is to map out all transactions in a given fiscal year to understand what goes through the government's financial management information system (FMIS) and identify where high-value transactions are conducted. The share of the government budget transacted through the FMIS indicates the integrity of expenditure data.

## Procurement

Governments are responsible for large volumes of procurement transactions. A recent estimate places these transactions at 12 percent of the global GDP (Bosio et al. 2020). There is a growing body of academic and policy literature on how to measure and prevent corruption in procurement. A widely used definition of corruption in procurement is the violation of impartial access to public contracts—that is, the deliberate restriction of open competition to the benefit of a politically connected firm or firms (Fazekas and Kocsis 2020). Corruption in procurement can occur in many forms. A single firm may bid for a procurement contract, securing exclusive access to lucrative government contracts. Or firms may overinvoice a procured good, often in collusion with procurement officials or politicians.

**TABLE 8.1 Examples of Public Procurement Indicators**

Economy and efficiency	Transparency and integrity	Competition	Inclusiveness and sustainability
<i>Tender and bidding process</i>			
<ul style="list-style-type: none"> <li>• Total processing time</li> <li>• Evaluation time</li> <li>• Contracting time</li> </ul>	<ul style="list-style-type: none"> <li>• Time for bid preparation</li> <li>• Single-bidder tender</li> </ul>	<ul style="list-style-type: none"> <li>• Open procedure</li> <li>• Number of bidders</li> <li>• Share of new bidders</li> </ul>	<ul style="list-style-type: none"> <li>• Share of SME bidders</li> <li>• Share of WOE bidders</li> </ul>
<i>Assessment and contracting</i>			
<ul style="list-style-type: none"> <li>• Awarded unit price</li> <li>• Final unit price after renegotiation</li> </ul>	<ul style="list-style-type: none"> <li>• Share of excluded bids</li> </ul>	<ul style="list-style-type: none"> <li>• Number of bidders</li> <li>• New bidders</li> </ul>	<ul style="list-style-type: none"> <li>• Share of SME bidders</li> <li>• Share of WOE bidders</li> </ul>
<i>Contract implementation</i>			
<ul style="list-style-type: none"> <li>• Final unit price after renegotiation</li> <li>• Time overrun</li> </ul>	<ul style="list-style-type: none"> <li>• Variation orders</li> <li>• Renegotiations</li> </ul>		

Source: Original table for this publication based on chapter 12.

Note: SME = small and medium enterprise; WOE = women-owned enterprise.



Chapter 12 provides an overview of a set of indicators and data sources on public procurement and how they can be used for data-driven decision-making (table 8.1). It also provides guidance on how to build data infrastructure and capacity for procurement data analytics and emphasizes the added value of combining public procurement data with other data sources. The chapter concludes by describing how a whole-of-government approach can increase the benefits of procurement analytics, as well as the advantages of combining administrative with survey data on procurement officials.

## PROCESSES

Processes in public administration define organizational objectives and work procedures. They include management practices, which structure how managers and staff engage with each other in the performance of their duties. Processes also include business practices, which map onto the different regulations and procedures that structure how public servants should perform their duties. In this section, we provide a snapshot of the extant literature on these processes, highlighting, as illustrative examples, how unethical leadership and a lack of compliance with existing business processes may impact public administration. We also present indicators on these dimensions of corruption and discuss how to develop and implement them. Regarding data integration, management practices can affect multiple areas of public administration, including culture and behavior as well as turnover in personnel. It is therefore important to connect agency-level indicators of management practices with other administrative data sources.

### Management Practices

Corruption in management practices involves the violation of business rules that govern how public servants are managed. We follow the definition proposed in Meyer-Sahling, Schuster, and Mikkelsen (2018), focusing particularly on human resource management practices. Management practices include decisions about recruitment into the public service, compensation, and the promotion of public servants. In practice, corruption can affect these different management functions. Politicians can appoint loyalists to the public service to extract rents, while low wages may encourage public servants to request bribes. Finally, political control or pressure may be applied to promote public servants who “steal for the team.”

Surveys of public servants can help measure their experience of (corrupt) management practices (see part 3 of the *Handbook* and also Meyer-Sahling, Schuster, and Mikkelsen 2018). In identifying these practices, practitioners must choose whether to capture public servants’ perceptions at the organizational level or their individual experiences with corruption. Chapter 23 assesses the trade-offs involved in each approach, highlighting how answers may differ depending on the kind of referent used. In particular, sensitive questions about topics such as corruption in management may be better measured at the organizational rather than the individual level because organizational comparisons reduce social-desirability bias. Another key question is how to assess the degree to which management practices differ across groups within public administration, such as across genders or organizations. Therefore, the choice of referent—organizational or individual—should be considered when designing the survey.<sup>8</sup>

### Business Processes (Organizational Procedures in Government)

The *Handbook* provides tools to measure corruption in business processes, understood as the procedures that regulate how public administration is to conduct its business (for example, what kind of archives it needs to keep). Chapter 13 presents indicators that allow practitioners to measure the quality of the processes completed by public servants. Given that public administration is often regulated by a set of rules dictating which forms have to be filled out, the level of completion of these forms can be used by external evaluators to assess the quality of business processes in government and, thereby, the risk of corruption.

Indicators such as the availability of minutes, memos, and other relevant documents, as well as the proportion of incoming and outgoing correspondence with dates, stamps, and documented sources, may provide insights into the quality of business process implementation by public officials.

## CULTURE AND BEHAVIOR OF PUBLIC SERVANTS

The culture of a public service organization includes the attitudes and norms of public servants, which, in turn, shape their behavior. Attitudes and norms include the values that guide officials as they do their jobs, the level of engagement officials have with their work, and the norms that govern their behavior, among other things. This section describes methods from the *Handbook* that can be used to assess how the culture, norms, attitudes, and behaviors of public servants might reveal a propensity for corruption. We have mentioned before that management practices can shape norms and behaviors in public administration. Integrating data on cultural norms with the quality of service delivery can help identify how norms shape practice in services such as health care and education.

### Norms and Behaviors

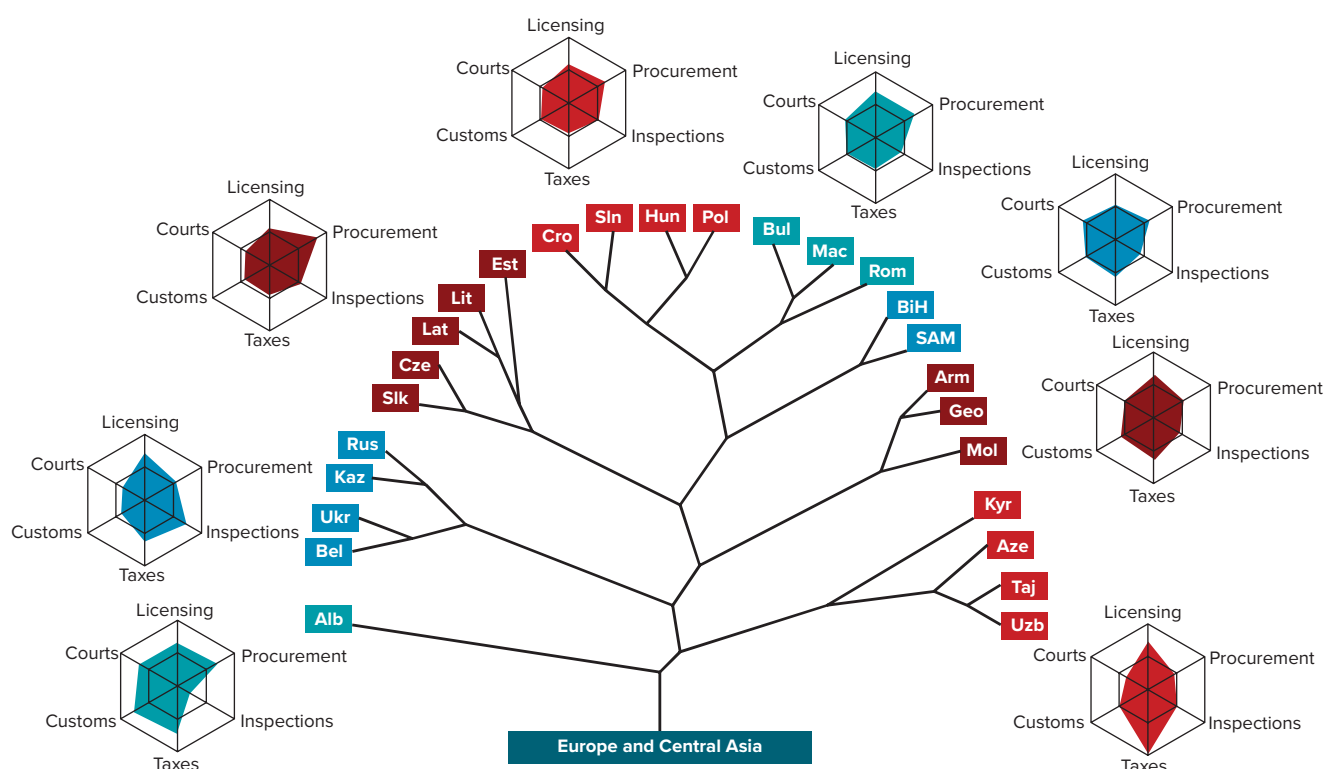
Corruption is most likely where the culture and norms within the public service enable it. The more prevalent corruption seems to public servants—the more it constitutes a norm—the more likely they may be to engage in corruption themselves (Köbis et al. 2015). Looking specifically at public servants, studies have found that certain motives, such as personal and social norms and opportunities not to comply with government rules, are significantly correlated with public servants' propensity to engage in corruption (Gorsira, Denkers, and Huisman 2018). These personal and social norms can also spur or hinder conflicts of interest, cases in which public servants' private interests unduly influence how they behave in public office, which have been a significant challenge in many countries. Initiatives led by various international organizations provide information on laws regulating conflicts of interest and their measurement.<sup>2</sup> However, business and ethical principles may vary across countries and within a single country. A report on Vietnam highlights different definitions of conflict of interest for public servants and how they can be culturally informed (World Bank 2016).

Identifying the attitudes and beliefs that motivate public servants to engage in corruption can help determine the root causes of corruption and inform strategies to curtail it at an international level as well. This task is crucial for practitioners examining corruption in the public service across countries, due to disparities in the understanding of corruption within different societies (Anderson and Gray 2006; World Bank 2016). Figure 8.2 shows clusters of countries based on the relative frequency of bribery in different sectors, enabling cross-country comparisons across different areas of corruption or bribes. These clusters map closely onto traditional groupings—for instance, northern and southern members of the Commonwealth of Independent States, Baltic states, and countries closest to accession to the European Union—suggesting that shared histories and similar paths of institutional development play a role in the types of corruption seen today.

The attitudes and motivations of individual public servants toward corrupt practices are primarily shaped by experiences and beliefs, making surveys a method well suited to analyzing them. However, self-reporting allows respondents to give inaccurate responses or not respond at all, distorting the resulting data. For example, if social or legal repercussions might arise from revealing a disposition toward corruption, public servants may try to mask their true attitudes or behavior. By applying methods in the *Handbook* to the design of surveys that aim to capture attitudes toward corruption, practitioners can mitigate distortions resulting from biases or nonresponse among public officials.

Chapter 22 presents findings about which questions can better solicit responses as well as a conceptual framework for understanding this phenomenon. One source of survey nonresponse is question sensitivity. A public servant asked about their views on violations of social norms or formally prohibited behavior may be hesitant to respond due to social-desirability bias or fear of legal sanctions. The chapter, however, suggests that

**FIGURE 8.2 Country Clusters Based on Relative Frequency of Bribes in Specific Areas**



Source: Adapted from Anderson and Gray 2006, figure 4.15.

Note: Alb = Albania; Arm = Armenia; Aze = Azerbaijan; Bel = Belarus; BiH = Bosnia and Herzegovina; Bul = Bulgaria; Cro = Croatia; Cze = Czechia; Est = Estonia; Geo = Georgia; Hun = Hungary; Kaz = Kazakhstan; Kyr = Kyrgyz Republic; Lat = Latvia; Lit = Lithuania; Mac = North Macedonia; Mol = Moldova; Pol = Poland; Rom = Romania; Rus = Russian Federation; SAM = Serbia and Montenegro; Slk = Slovak Republic; Sln = Slovenia; Taj = Tajikistan; Ukr = Ukraine; Uzb = Uzbekistan.

public servants *are* willing to answer sensitive questions on, say, their attitudes and behaviors toward corruption (though they may, of course, do so in a socially desirable way). Instead, the strongest predictor of nonresponse is complexity—specifically, a question’s “unfamiliarity” and “scope of information”—as when officials are asked general questions about the work environment rather than about their immediate experiences. To address this, survey questions should ask about public officials’ perceptions of corruption within their own organizations.

Merely eliciting a response, however, does not ensure that the data being collected through surveys are reflective of true norms surrounding corruption in public administration. For instance, to address whether face-to-face or online surveys better reduce response bias, chapter 19 examines the impact of survey mode on civil servant survey responses. Face-to-face surveys tend to offer several benefits, including significantly higher response rates and lower break-off rates. Online surveys, by contrast, limit the ability of an enumerator to probe public servants for responses, which risks distorting the true prevalence of attitudes and behaviors tolerant of corruption. Online formats tend to elicit more candid responses to potentially sensitive questions about topics such as ethics violations. Indeed, the chapter presents evidence that face-to-face surveys produce more “desirable” responses compared to online surveys—for instance, fewer public servants report that employees “observe unethical behavior among colleagues.” Survey designers must therefore consider the trade-offs of each survey mode, recognizing that the choice of survey mode can impact the accuracy of results. The pilot phase and focus group discussions can help validate survey results.

To draw comparisons regarding the norms that enable corruption, practitioners may compare survey results across different demographics, organizations, and countries. To do so, these different groups must understand survey measures in the same way. However, norms around corruption differ across countries and, at times, across organizations within a single country. Determining what public servants view as corruption or as an ethical violation is therefore necessary to understanding and contextualizing the results

of civil servant surveys.<sup>10</sup> Chapter 24 provides an approach to this task by measuring the comparability of a latent statistical concept among different groups. While this chapter looks specifically at the concept of transformational leadership, its approach can be applied to latent concepts relating to attitudes toward corruption. By using this framework, practitioners can better ensure that when they measure corruption norms against certain benchmarks, those benchmarks enable valid comparisons.

Finally, due to the limitations of survey data and the potential rigidity of attitudes and norms regarding corruption, qualitative analyses can be valuable tools for interpreting the data obtained through surveys. Principles for using qualitative analyses to analyze norms and the behavior of public servants are presented in chapter 30. For example, through participant observation, observers can document how public servants interact in an office environment, which may be difficult to capture using a survey instrument. Meetings and other forms of interaction between public servants may reveal power dynamics—in particular, how gender, race, and seniority stratify hierarchies in public administration. Incorporating qualitative analyses such as these into investigations of corruption norms can give practitioners more robust insights into the roots of corruption and tools to remedy it.

## DIRECT OUTPUTS OF PUBLIC ADMINISTRATION

Corruption may affect the outputs of governments in multiple ways. Politicians may exert pressure on public servants to relax procedures for electoral gain. Public servants may accelerate cases for firms in exchange for bribes. These forms of corruption can be measured through household and citizen surveys. This section draws on *Handbook* chapters to outline several illustrative indicators to measure corruption in case and task data, as well as citizens' perception of corruption. Since direct outputs are the products of public administration, these can be linked to multiple data sources, such as personnel, budget, and procurement. This enables practitioners to assess the efficiency of different personnel recruitment and management practices as well as norms and behaviors in producing important outcomes for citizens.

### Case and Task Data

To implement policy, governments generate large amounts of administrative data on the deliberations and actions of public servants. These case data are rich records of how work is carried out within governments, and they enable practitioners and researchers to better understand public servants' performance within public administration. For example, exploring data on social security cases in Italy, Fenizia (2022) estimates the effect of management changes on office productivity, finding that improvements in manager talent increase productivity. Dasgupta and Kapur (2020) collect data on time-usage diaries in India to analyze bureaucratic task overload in block development offices. In other cases, public servants may face pressure from superiors to expedite bidding processes and cases or to disregard due process in the development of new regulations (World Bank 2021).

It is possible to detect fraudulent activity by public servants by analyzing case data. Chapter 15 provides a simple measurement, the error rate, to identify cases of potential fraud risk in public administration. Calculating the fraction of claims and cases processed incorrectly allows governments to measure the extent of administrative irregularities and provide remedial measures. For example, in the case of social security claims, there are two types of mistake: a government agency may incorrectly give beneficiaries a social transfer or erroneously deny a transfer to the correct party. Keeping track of appeals by denied beneficiaries would only capture the latter case. To provide a comprehensive measure of the error rate, and better identify fraudulent behavior, governments should regularly audit a random subset of claims by government offices.

Public servant surveys can also provide insight into the extent to which cases are affected by corruption. Chapter 13 provides indicators to assess the quality and completeness of claims in government through enumerator reviews of extant cases. To measure the comprehensiveness of reporting on a claim across a

series of tasks, the following survey question was asked: “Where applicable, are minutes, memos, and other necessary records present and complete?” Another indicator is the overall commitment to effective process: “In general, to what extent does the file organization adhere to government procedure?”

## Household and Citizen Surveys

Household surveys are a valuable tool to measure corruption in public administration, generating evidence on citizens’ trust in government, the quality and accessibility of service delivery, and experiences with bribes (Anderson, Kaufmann, and Rekanatini 2003; UNODC and UNDP 2018). A 1998 report on Latvia shows, using household surveys, that over 40 percent of households and firms agreed with the statement “A system whereby people could anonymously report instances of corruption would not be successful because corruption is a natural part of our lives and helps solve many problems” (Anderson 1998). Additionally, citizens reported that bribes often occurred in reaction to difficulties in processing cases with public servants, such as intentional delays to resolve firms’ requests or vague explanations of legal requirements. Numerous studies have found negative correlations between citizens’ perception of corruption in government and their level of trust and satisfaction with the government, two foundational components of effective political systems (Park and Blenkinsopp 2011; Seligson 2002).

Citizen surveys can assess the extent to which citizens trust their government, focusing on issues of corruption, such as bribery. To provide standardized metrics for measuring this relationship, chapter 28 develops a framework concerning drivers of trust in public institutions. This framework is based on four components of institutional trust, including a competence indicator that measures whether citizens believe public institutions “minimize uncertainty in the economic, social, and political environment” and a values indicator that measures whether public institutions “make decisions and use public resources ethically, promoting the public interest over private interests, while combatting corruption.” Disaggregating these components of trust allows practitioners to gauge both the level of corruption citizens expect from their government and the extent to which citizens believe corruption affects the reliability of public services. By applying this framework through the Organisation for Economic Co-operation and Development (OECD) Survey on Drivers of Trust in Public Institutions, as described in chapter 28, practitioners can examine the multidimensional effects of corruption on trust in public institutions and compare their results with other countries.

While corruption may impact trust in different ways across countries, a few practices can strengthen government transparency and accountability to mitigate the effects of corruption. Chapter 25 outlines several of these practices, which include disseminating public servant survey results across all levels of government as well as to the public. By disseminating survey results, which may include public servants’ perceptions of corruption within their agencies, governments can increase awareness of areas for improvement and incentivize stakeholders to act on results. One example comes from a recent corruption report in which a survey of federal civil servants in Brazil revealed that a majority had witnessed unethical practices while in the public service (World Bank 2021). By gathering and revealing this information, the comptroller general demonstrated his public commitment to addressing corruption.



## FRONTLINE AGENCIES

Frontline agencies in government are responsible for the delivery of public services directly to citizens. These agencies include hospitals, schools, and police stations. Frontline agencies have been the subject of extensive research and policy work because their work has a direct impact on social welfare and their tasks, such as the delivery of health care or education, are often by nature amenable to measurement exercises. This section provides illustrative examples from the *Handbook* of how these agencies can be affected by corruption, such as bribery or budget leakages. Frontline agencies are amenable to different types of data integration:

personnel data help identify the extent to which service providers are qualified, management practice data provide information on how well services are being managed, and procurement data enable assessment of the extent to which important materials—for example, school textbooks or medical equipment—are sourced and distributed.

## Service Delivery

Corruption may occur in different public services and sectors in the economy due to public servants' power to allocate benefits or impose additional costs on citizens, such as bribes (Rose-Ackerman and Palifka 2016; World Bank 2003). An example of this type of corrupt behavior is when a service provider—such as a government clerk issuing licenses or a health official providing services—extracts an informal payment from a citizen or business to grant or expedite access to the service (World Bank 2020). The World Bank Enterprise Surveys find evidence that, across the world, around 12.3 percent of firms are expected to give gifts to get an operating license.<sup>11</sup> Data from the World Bank Business Environment and Enterprise Performance Survey (BEEPS) suggest a higher prevalence of corruption in taxation and government contracts in contrast to utilities or environmental inspections (Anderson and Gray 2006). Collecting survey data on different respondents—for example, public servants, business owners, and citizens—paints a more holistic picture of corruption in service delivery. For example, the Governance and Anti-corruption (GAC) diagnostic surveys, developed by the World Bank, identify where a public servant asked for a bribe, or if a citizen first offered it.

In the *Handbook*, we provide different methodologies to measure corruption in both service delivery and sectoral data. For example, chapter 28 provides an indicator for measuring the integrity of public servants while they perform their duties, as observed by citizens. Specifically, the survey question is “If a government employee is offered a bribe in return for better or faster access to a public service, how likely or unlikely is it that they would accept it?” Asking citizens this question directly may reduce concerns over social-desirability bias that arise when surveying public servants, but both responses provide insight into how corruption in service delivery occurs.

Another indicator that may assist in measuring the prevalence of corruption is the time it takes to process social security cases, as outlined in chapter 15. While a delay in processing cases may not directly imply corruption, combining this information with citizens' perception of the prevalence of corruption may help identify particular sectors—such as social protection or business licenses—where delays in case processing are used as leverage to extract bribes from citizens.

Chapter 29 provides measures of service delivery (MSD) that can be used to identify cases of corruption in public service delivery. For health care in particular, patients may be asked questions regarding their confidence in the health care system, such as their level of satisfaction and recommendation, as well as regarding care uptake and retention. Low indicators for satisfaction may signal issues regarding public service delivery and can be followed by more direct questions about corruption. Additionally, indicators on the availability of medical equipment and supplies may help evaluate whether these resources are being allocated to relevant facilities or leaked along the way.

Finally, procurement indicators can be also used to evaluate contracts between a government's frontline agencies and contractors, as outlined in chapter 12. The prevalence of single-bidder contracts and short time periods for bid preparation point to problems in the bidding process, while a low number of renegotiations can shed light on collusion between government officials and contractors.

## Regulation of Economic Sectors

Corruption is often linked to government officials' power to regulate what firms and individuals can do in particular sectors of the economy. Corruption in regulation affects environmental regulation, building inspections, labor and safety standards, and land-use management, among other areas. The rationale for government regulation emerged as a response to the rise of natural monopolies, such as in the telecommunications and energy industries, where the government should act in the public interest to reduce negative



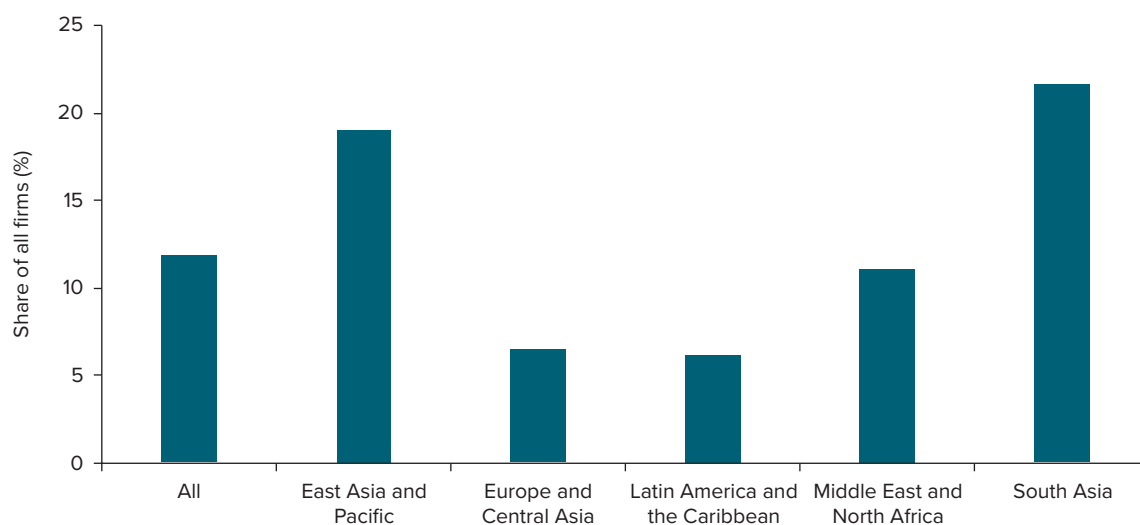
market externalities (Berg and Tschirhart 1988). However, regulation is subject to the pressure of interest groups and may be exposed to regulatory capture due to informational asymmetries and direct bribes (Laffont and Tirole 1991) or the promise of postgovernment employment in regulated industries (Dal Bó 2006). Because improving the informational environment can reduce the potential for regulatory capture, sectoral reforms have often focused on increasing transparency and access to data. A report on land management in Vietnam highlights how data can be used to track transparency and, in turn, how that transparency can reduce corruption in land-use regulation (World Bank 2014). The justice system plays an important role in enforcing regulation but can itself be exposed to corruption through bribery and other forms of capture.

The *Handbook* provides a few tools to understand corruption in sectoral cases, from both the supply and the demand side. On the demand side, chapter 28 highlights how measures of integrity and reliability can be applied to citizens' experiences with regulatory agencies. For example, if a citizen is to apply for a business license, an indicator of the fairness of the process is the question "If you or a member of your family would apply for a business license, how likely or unlikely do you think it is that your application would be treated fairly?" Evaluating how business owners have been treated depending on age, education, and gender may yield additional insights about fairness. Questions regarding the integrity of public servants when adjudicating business license applications may yield insights into the extent to which regulatory agencies have been captured. Chapter 13 provides indicators on the supply side of government regulation. The degree to which internal government procedures are followed when adjudicating licensing and regulatory cases, as well as the degree to which cases are complete, provides evidence on whether governments are being impartial and thorough when adjudicating the enforcement of regulations on businesses and citizens.<sup>12</sup>

## Revenue Collection

Corruption in revenue collection affects tax collection from citizens and revenue from customs, licensing fees, fines, and the sale of goods and services. Enterprise Surveys routinely point to tax as a source of corruption, often in the form of bribes to avoid paying the full tax amount, as illustrated in figure 8.3. This can weaken the tax base of the country and lead to government underfunding. Furthermore, corruption in tax administration can cause inefficiency by diverting resources from productive activity to bribery and

**FIGURE 8.3** Percentage of Firms Expected to Give Gifts in Meetings with Tax Officials, by Region



Source: World Bank Enterprise Surveys, <http://www.enterprisesurveys.org>.

undermining equity, as honest taxpayers bear the brunt of these costs (Purohit 2007). In customs, corruption tends to take two forms: customs officials may pocket a portion of import revenue or may extract bribes from importers in exchange for some benefit (Yang 2006). Both forms of corruption can harm citizens by siphoning off revenue from the government, of which customs duties often constitute a significant share. Beyond financial costs, bribery in customs risks harming citizens by allowing illegal and potentially dangerous goods into the country. Through targeted incentives and heightened monitoring, corruption in customs can be curtailed (World Bank 2020).

Methods for mitigating corruption in revenue collection and administration are examined in chapter 15 using a case study from Punjab, Pakistan. The tools used to ensure the quality of tax administration include standardizing the process of reporting collected taxes and cross-checking the tax department's administrative records against the national bank's tax receipts for discrepancies. The chapter also assesses performance-pay mechanisms to improve revenue collection. Although these programs may increase revenue, they may also lead to greater bribes as performance rewards increase tax collectors' bargaining power over citizens. Governments should therefore consider the costs of potential incentive schemes for tax administrators.

Chapter 14 provides tools for identifying bottlenecks in customs processes, which lessen incentives to bribe officials to expedite the processes. These include the Automated System for Customs Data (ASYCUDA), time-release studies (TRS), trader perception surveys, and GPS trackers. Furthermore, the chapter describes revenue collection indicators that can be used to detect fraud. Where fraud is suspected, authorities can look at the value of identical goods or similar goods to determine the revenue that could have been collected had goods been correctly declared. Monitoring customs can be reinforced with mirror analyses that compare the quantity of goods declared by the exporting country to the quantity of similar goods declared by the importing country.

## CONCLUSION

This chapter has provided a framework for both conceptualizing and measuring corruption in public administration. It recognizes corruption as a challenge with many faces and with particular characteristics and ways of operating in each sector of public administration. By assessing corruption along the public sector production function, we have highlighted the range of microdata approaches available in the *Handbook* to understand corruption in public administration. While our thematic focus has been on corruption, other topics in public administration are amenable to this approach as well. Of course, our review—drawing in particular on the chapters in this book—has only lightly engaged the vast literature on this topic. For example, hundreds of surveys on corruption in public administration have been conducted, with techniques ranging from list experiments to direct questions to randomized response techniques, to name a few.

Our goal in this chapter has been to highlight the different microdata sources and methodologies available to assess corruption and to show how measurement along the public administration production function can help analysts assess corruption holistically. This approach showcases the benefits of employing multiple sources of information in holistically assessing substantive topics in public administration, such as corruption. In so doing, we hope to have highlighted the benefits and challenges associated with the holistic use of government analytics in reducing corruption.

Beyond the comprehensiveness of analytics on corruption, a multipronged approach enables the integration of multiple data sources to reveal corruption in innovative ways. For example, integrating HRMIS with procurement data can reveal which procurement officials are more likely to manage corrupt bidding practices that benefit colluding companies. Additionally, information on wages, management practices, and other data sources can help diagnose what risk factors are associated with procurement officials' corrupt behavior. This integration of data provides exciting possibilities for understanding how corruption operates, generating novel insights that would not be possible if analytics were restricted to a single sector.

Finally, we have emphasized that measurement can provide direction for reducing corruption, but it does not reduce corruption itself. Challenging administrative reforms need to embed analytics of corruption into a broader public sector reform strategy, proactively using indicators to identify and enact policies against corruption. This requires not only technical knowledge but political negotiation based on compromise and consensus building. We hope that evidence-based reforms, informed by analytical insights, can guide practitioners in their efforts to understand and reduce corruption.

## NOTES

We are grateful to Alexandra Jasmine Chau for excellent research assistance in completing this chapter. We are grateful for helpful comments from Daniel Rogger.

1. For a report on the multiple faces of corruption, see Campos and Pradhan (2007). For a comprehensive discussion of definitions of corruption, see Fisman and Golden (2017). Anderson et al. (2019) include a review of other definitions.
2. For a sample of innovative research and development on the use of data analytics to understand corruption, see video of the sessions from the Symposium on Data Analytics and Anticorruption (World Bank and Korea Development Institute School of Public Policy and Management 2021).
3. For a discussion, see chapter 4.
4. Focus groups and interviews enable involved actors to share their experiences with corruption and reveal these sources of pressure (Benjamin et al. 2014).
5. The experience of the National Institute of Statistics and Geography (INEGI), the national statistical agency of Mexico, sets an example for other countries about how to integrate data collection on corruption with regular data efforts. INEGI centralizes information on audits, experiences with corruption by citizens, and sanctions against civil servants. See the topic “Transparency and Anti-corruption” on the INEGI website: <https://en.www.inegi.org.mx/temas/transparencia/>.
6. This problem is highlighted in the Brazilian newspaper article “Problema não é número de servidores, mas salários altos, diz Temer a Bolsonaro,” *Folha de S.Paulo*, November 16, 2018, <https://www1.folha.uol.com.br/mercado/2018/11/problema-nao-e-numero-de-servidores-mas-salarios-altos-diz-temer-a-bolsonaro.shtml>.
7. More information about the PEFA assessment is available at the PEFA website, <https://pefa.org/>. For more information about PETS, see Koziol and Tolmie (2010).
8. A sampling tool for survey designers to assess the scale of sample required to investigate different topics in the public service is described in more detail in chapter 20 and can be found here: [https://encuesta-col.shinyapps.io/sampling\\_tool/](https://encuesta-col.shinyapps.io/sampling_tool/).
9. Examples of the measurement of conflict of interest include the Public Accountability Mechanisms (PAM) Initiative, EuroPAM, and the work done under the Stolen Asset Recovery (StAR) Initiative. Data on PAM are available at <https://datacatalog.worldbank.org/search/dataset/0040224>. Information about EuroPAM can be found on its website, <https://europam.eu/>. For more information about StAR, see World Bank (2012).
10. Anderson (2002) and World Bank (2002) also include direct survey questions—in the Kyrgyz Republic and Kazakhstan, respectively—about whether public servants define certain behaviors as “corruption.”
11. Data on corruption from the World Bank Enterprise Surveys are accessible at <https://www.enterprisesurveys.org/en/data/exploretopics/corruption>.
12. There are also de jure measures that look at gaps in the existing legal or regulatory structure that allow government officials to exercise discretion in favor of or against regulated entities. Mahmood and Slimane (2018) look at the existence of these structural weaknesses that allow firms to exercise privileges.

## REFERENCES

- Anderson, James. 1998. “Corruption in Latvia: Survey Evidence.” Working Paper 33257, World Bank, Washington, DC.
- Anderson, James. 2002. *Governance and Service Delivery in the Kyrgyz Republic: Results of Diagnostic Surveys*. Washington, DC: World Bank. <http://documents.worldbank.org/curated/en/814871468046449300/Governance-and-service-delivery-in-the-Kyrgyz-Republic-results-of-diagnostic-surveys>.
- Anderson, James, David S. Bernstein, Jean Pierre Brun, Alexandra M. Habershon, Francesca Recanatini, Emile J. M. Van Der Does De Willebois, and Stephen S. Zimmermann. 2019. *Anticorruption Initiatives: Reaffirming Commitment to a Development Priority*. Washington, DC: World Bank. <http://documents.worldbank.org/curated/en/365421591933442799/Anticorruption-Initiatives-Reaffirming-Commitment-to-a-Development-Priority>.

- Anderson, James, and Cheryl W. Gray. 2006. *Anticorruption in Transition 3: Who Is Succeeding ... and Why?* Washington, DC: World Bank.
- Anderson, James, Daniel Kaufmann, and Francesca Recanatini. 2003. "Service Delivery, Poverty and Corruption—Common Threads from Diagnostic Surveys." Background Paper for the 2004 World Development Report Making Services Work for Poor People, World Bank, Washington, DC.
- Anderson, James, Gary Reid, and Randi Ryterman. 2003. "Understanding Public Sector Performance in Transition Countries: An Empirical Contribution." Working Paper 30357, World Bank, Washington, DC.
- Benjamin, Nancy, Kathleen Beegle, Francesca Recanatini, and Massimiliano Santini. 2014. "Informal Economy and the World Bank." Policy Research Working Paper 6888, World Bank, Washington, DC.
- Berg, Sanford V., and John Tschirhart. 1988. *Natural Monopoly Regulation: Principles and Practice*. New York: Cambridge University Press.
- Besley, Timothy J., Robin Burgess, Adnan Khan, and Guo Xu. 2022. "Bureaucracy and Development." *Annual Review of Economics* 14: 397–424. <https://doi.org/10.1146/annurev-economics-080521-011950>.
- Bosio, Erica, Simeon Djankov, Edward L. Glaeser, and Andrei Shleifer. 2020. "Public Procurement in Law and Practice." NBER Working Paper 27188, National Bureau of Economic Research, Cambridge, MA.
- Bragança, Arthur, Claudio Ferraz, and Juan Rios. 2015. *Political Dynasties and the Quality of Government*. Unpublished manuscript. <https://web.stanford.edu/~juanfrr/bragancaferrazrios2015.pdf>.
- Brollo, Fernanda, Pedro Forquesato, and Juan Carlos Gozzi. 2017. "To the Victor Belongs the Spoils? Party Membership and Public Sector Employment in Brazil." The Warwick Economics Research Paper Series (TWERPS) 1144, Department of Economics, University of Warwick, Coventry. <https://doi.org/10.2139/ssrn.3028937>.
- Cadot, Oliver. 1987. "Corruption as a Gamble." *Journal of Public Economics* 33 (2): 223–44. [https://doi.org/10.1016/0047-2727\(87\)90075-2](https://doi.org/10.1016/0047-2727(87)90075-2).
- Campos, J. Edgardo, and Sanjay Pradhan. 2007. *The Many Faces of Corruption: Tracking Vulnerabilities at the Sector Level*. Washington, DC: World Bank. <https://openknowledge.worldbank.org/handle/10986/6848>.
- Colonnelli, Emanuele, Mounu Prem, and Edoardo Teso. 2020. "Patronage and Selection in Public Sector Organizations." *American Economic Review* 110 (10): 3071–99. <https://doi.org/10.1257/aer.20181491>.
- d'Agostino, Giorgio, J. Paul Dunne, and Luca Pieroni. 2016. "Government Spending, Corruption and Economic Growth." *World Development* 84: 190–205. <https://doi.org/10.1016/j.worlddev.2016.03.011>.
- Dal Bó, Ernesto. 2006. "Regulatory Capture: A Review." *Oxford Review of Economic Policy* 22 (2): 203–25. <https://doi.org/10.1093/oxrep/grj013>.
- Das, Jishnu, and Jeffrey Hammer. 2014. "Quality of Primary Care in Low-Income Countries: Facts and Economics." *Annual Review of Economics* 6 (1): 525–53. <https://doi.org/10.1146/annurev-economics-080213-041350>.
- Dasgupta, Aditya, and Devesh Kapur. 2020. "The Political Economy of Bureaucratic Overload: Evidence from Rural Development Officials in India." *American Political Science Review* 114 (4): 1316–34. <https://doi.org/10.1017/S0003055420000477>.
- Evans, Peter. 1995. *Embedded Autonomy: States and Industrial Transformation*. Princeton, NJ: Princeton University Press.
- Evans, Peter, and James Rauch. 1999. "Bureaucracy and Growth: A Cross-National Analysis of the Effects of 'Weberian' State Structures on Economic Growth." *American Sociological Review* 64: 748–65. <https://doi.org/10.2307/2657374>.
- Fazekas, Mihály, and Gábor Kocsis. 2020. "Uncovering High-Level Corruption: Cross-National Corruption Proxies Using Public Procurement Data." *British Journal of Political Science* 50 (1): 155–64. <https://doi.org/10.1017/S0007123417000461>.
- Fenizia, Alessandra. 2022. "Managers and Productivity in the Public Sector." *Econometrica* 90 (3): 1063–84. <https://doi.org/10.3982/ECTA19244>.
- Ferraz, Claudio, Frederico Finan, and Diana B. Moreira. 2012. "Corrupting Learning: Evidence from Missing Federal Education Funds in Brazil." *Journal of Public Economics* 96 (9–10): 712–26. <https://doi.org/10.1016/j.jpubeco.2012.05.012>.
- Finan, Frederico, Benjamin A. Olken, and Rohini Pande. 2017. "The Personnel Economics of the Developing State." In *Handbook of Economic Field Experiments*, edited by Abhijit Vinayak Banerjee and Esther Duflo, vol. 2, 467–514. Amsterdam: North-Holland. <https://doi.org/10.1016/bs.hefe.2016.08.001>.
- Fisman, Ray, and Miriam A. Golden. 2017. *Corruption: What Everyone Needs to Know*. Oxford: Oxford University Press.
- Gorsira, Madelijne, Adriaan Denkers, and Wim Huisman. 2018. "Both Sides of the Coin: Motives for Corruption among Public Officials and Business Employees." *Journal of Business Ethics* 151 (1): 179–94. <https://doi.org/10.1007/s10551-016-3219-2>.
- Gurkan, Asli, Kai Kaiser, and Doris Voorbraak. 2009. "Implementing Public Expenditure Tracking Surveys for Results: Lessons from a Decade of Global Experience." PREM Notes 145, World Bank, Washington, DC.
- Hashim, Ali, Khuram Farooq, and Moritz Piatti-Fünfkirchen. 2020. "Ensuring Better PFM Outcomes with FMIS Investments: An Operational Guidance Note for FMIS Project Teams." Governance Discussion Paper/Guidance Note Series 6, World Bank, Washington, DC.
- Jain, Arvind K. 2001. "Corruption: A Review." *Journal of Economic Surveys* 15 (1): 71–121. <https://doi.org/10.1111/1467-6419.00133>.
- Kaufmann, Daniel, Sanjay Pradhan, and Randi Ryterman. 1998. "New Frontiers in Diagnosing and Combating Corruption." PREM Notes 7, World Bank, Washington, DC.

- Köbis, Nils C., Jan-Willem van Prooijen, Francesca Righetti, and Paul A. M. Van Lange. 2015. “‘Who Doesn’t?’—The Impact of Descriptive Norms on Corruption.” *PLoS One* 10 (6): e0131830. <https://doi.org/10.1371/journal.pone.0131830>.
- Kozioł, Margaret, and Courtney Tolmie. 2010. *Using Public Expenditure Tracking Surveys to Monitor Projects and Small-Scale Programs: A Guidebook*. Washington, DC: World Bank. <http://documents.worldbank.org/curated/en/388041468177874064/Using-public-expenditure-tracking-surveys-to-monitor-projects-and-small-scale-programs-a-guidebook>.
- La Cascia, Hunt, Izza Akram Malik, Eduardo Vicente Goncalves, Yasodara Maria Damo Cordova, and Michael Kramer. 2020. *Finding Fraud: GovTech and Fraud Detection in Public Administration*. Washington, DC: World Bank.
- Laffont, Jean-Jacques, and Jean Tirole. 1991. “The Politics of Government Decision-Making: A Theory of Regulatory Capture.” *The Quarterly Journal of Economics* 106 (4): 1089–127. <https://doi.org/10.2307/2937958>.
- Mahmood, Syed Akhtar, and Meriem Ait Ali Slimane. 2018. *Privilege-Resistant Policies in the Middle East and North Africa: Measurement and Operational Implications*. MENA Development Report. Washington, DC: World Bank. <https://openknowledge.worldbank.org/handle/10986/29353>.
- Meyer-Sahling, Jan-Hinrik, Christian Schuster, and Kim Sass Mikkelsen. 2018. *Civil Service Management in Developing Countries: What Works? Evidence from a Survey with 23,000 Civil Servants in Africa, Asia, Eastern Europe and Latin America*. Report for the UK Department for International Development (DFID).
- Park, Heungsik, and John Blenkinsopp. 2011. “The Roles of Transparency and Trust in the Relationship between Corruption and Citizen Satisfaction.” *International Review of Administrative Sciences* 77 (2): 254–74. <https://doi.org/10.1177/0020852311399230>.
- Purohit, Mahesh C. 2007. “Corruption in Tax Administration.” In *Performance Accountability and Combating Corruption*, edited by Anwar Shah, 285–302. Washington, DC: World Bank.
- Rauch, James E., and Peter B. Evans. 2000. “Bureaucratic Structure and Bureaucratic Performance in Less Developed Countries.” *Journal of Public Economics* 75: 49–71. [https://doi.org/10.1016/S0047-2727\(99\)00044-4](https://doi.org/10.1016/S0047-2727(99)00044-4).
- Recanatini, Francesca. 2011. “Assessing Corruption at the Country Level.” In *Handbook of Global Research and Practice in Corruption*, edited by Adam Graycar and Russell Smith, 34–64. Cheltenham: Edward Elgar.
- Riaño, Juan Felipe. 2021. “Bureaucratic Nepotism.” STEG Working Paper, Research Theme 5: Political Economy and Public Investment, Structural Transformation and Economic Growth, UK Foreign, Commonwealth & Development Office. <https://steg.cepr.org/publications/bureaucratic-nepotism>.
- Rose-Ackerman, Susan. 1978. *Corruption: A Study in Political Economy*. New York: Academic Press.
- Rose-Ackerman, Susan, and Bonnie J. Palifka. 2016. *Corruption and Government: Causes, Consequences, and Reform*. 2nd ed. Cambridge: Cambridge University Press.
- Seligson, Mitchell A. 2002. “The Impact of Corruption on Regime Legitimacy: A Comparative Study of Four Latin American Countries.” *Journal of Politics* 64 (2): 408–33. <https://doi.org/10.1111/1468-2508.00132>.
- UNODC and UNDP (United Nations Office on Drugs and Crime and United Nations Development Programme). 2018. *Manual on Corruption Surveys: Methodological Guidelines on the Measurement of Bribery and Other Forms of Corruption through Sample Surveys*. Vienna: UNODC. [https://www.unodc.org/documents/data-and-analysis/Crime-statistics/CorruptionManual\\_2018\\_web.pdf](https://www.unodc.org/documents/data-and-analysis/Crime-statistics/CorruptionManual_2018_web.pdf).
- World Bank. 2002. *Kazakhstan Governance and Service Delivery: A Diagnostic Report*. Washington, DC: World Bank. <https://openknowledge.worldbank.org/handle/10986/15098>.
- World Bank. 2003. *World Development Report 2004: Making Services Work for Poor People*. Washington, DC: World Bank.
- World Bank. 2012. *Public Office, Private Interests: Accountability through Income and Asset Disclosure*. Stolen Asset Recovery (StAR) Series. Washington, DC: World Bank.
- World Bank. 2014. *Land Transparency Study: Synthesis Report*. Washington, DC: World Bank. <http://documents.worldbank.org/curated/en/934431468121474463/Land-transparency-study-synthesis-report>.
- World Bank. 2016. *Managing Conflict of Interest in the Public Sector: Law and Practice in Vietnam*. Washington, DC: World Bank. <http://documents.worldbank.org/curated/en/580601478253123042/Managing-conflict-of-interest-in-the-public-sector-law-and-practice-in-Vietnam>.
- World Bank. 2019. *Anticorruption Initiatives: Reaffirming Commitment to a Development Priority*. Washington, DC: World Bank. <https://openknowledge.worldbank.org/handle/10986/34010>.
- World Bank. 2020. *Enhancing Government Effectiveness and Transparency: The Fight against Corruption*. Washington, DC: World Bank. <http://documents.worldbank.org/curated/en/235541600116631094/Enhancing-Government-Effectiveness-and-Transparency-The-Fight-Against-Corruption>.
- World Bank. 2021. *Ethics and Corruption in the Federal Public Service: Civil Servants’ Perspectives*. Washington, DC: World Bank. <https://openknowledge.worldbank.org/handle/10986/36759>.
- World Bank and Korea Development Institute School of Public Policy and Management. 2021. “Symposium on Data Analytics and Anticorruption.” Filmed October 25–28, 2021, at the World Bank, Washington, DC, and online. Video, 3:50:10. <https://www.worldbank.org/en/events/2021/10/25/symposium-on-data-analytics-and-anticorruption#1>.
- Yang, Dean. 2006. “The Economics of Anti-Corruption: Lessons from a Widespread Customs Reform.” In *International Handbook on the Economics of Corruption*, edited by Susan Rose-Ackerman, 512–46. Cheltenham: Edward Elgar.

The background features a series of wavy, horizontal lines in a light tan color, overlaid with a pattern of binary code (0s and 1s) in a slightly darker tan. The lines and code create a sense of digital flow and movement.

## **PART 3**

# Government Analytics Using Administrative Data





## CHAPTER 9

# Creating Data Infrastructures for Government Analytics

*Khuram Farooq and Galileu Kim*

### SUMMARY

Creating the conditions for effective data analytics in the public sector requires reforming management information systems (MIS). This chapter outlines a road map to guide practitioners in the development of analytically driven information systems, drawing on the experience of World Bank practitioners and government officials. The conceptual framework and cases presented focus on human resources management information systems (HRMIS), foundational data infrastructures that provide information on personnel and compensation. However, the chapter articulates broader lessons for designing and managing information systems for government analytics. The chapter first discusses the stages in reforming an HRMIS, outlining key decision points and trade-offs involved in building public sector data architectures. It then demonstrates how this framework can be applied in practice and the associated risks, drawing on case studies of analytical transformations of HRMIS in Luxembourg, Brazil, and the United States.

### ANALYTICS IN PRACTICE

- Many government information systems, such as human resources management information systems (HRMIS), can be thought of in terms of the interaction of distinct data modules. Thinking about data within these “islands” of like data allows the analyst to define more precisely the reforms such modules might need and identify interdependencies between them. For example, an HRMIS can be thought of as multiple modules, including payroll and personnel data, that interact to provide insights any single type of data could not. As a result, reforms must be precise regarding which modules will be modified and why. The implementation team needs to be aware of how modules depend on one another in order to understand how the flow of data within the information system might disrupt the use and interpretation of other modules in the system.

---

Khuram Farooq is a senior financial management specialist in the World Bank's Governance Global Practice. Galileu Kim is a research analyst in the World Bank's Development Impact Evaluation (DIME) Department.

- HRMIS data modules are not created equal: rather, some provide the foundation upon which all other systems rely. Reforms should prioritize the comprehensiveness and quality of foundational modules in human resource management and then transition to developing more complex analytical modules. In HRMIS reforms, foundational modules—including payroll compliance and basic personnel information—should take precedence over other layers, such as talent management and analytics. Without a quality foundation, other modules will produce imprecise or inaccurate analytics. Analytical modules, including dashboards and reports, require that foundational modules be set in place and that their data be accurate.
- However, almost any government data system that has prioritized accurate measurement can be useful. Thus, small reforms targeted at strengthening basic data quality can generate analytical insights even when other foundational modules need further reform. Most developing countries are in the process of building foundational HRMIS modules, such as payroll or basic information on HR headcount. Accurate measurement of these data can be useful. Even if these foundational modules are incomplete in terms of the wider vision of the implementation team, it is still possible to develop analytics reports from foundational modules, such as wage bill analysis and sectorwise employment. Analytical reports can be produced, even if manually, without implementing an analytics module. Though data analysis and visualization might be narrow in scope, this approach can provide quicker results and build even greater political will for further reform.
- HRMIS reform processes should be informed by the policy objectives of practitioners, such as improving the budgetary compliance of the wage bill. This facilitates political commitment to HRMIS reforms and ensures their policy relevance, although institutional coordination should be secured as well. HRMIS reforms should be anchored in problems the government considers policy priorities to secure commitment from political leadership. These problems typically include wage bill controls, identifying ghost workers, and providing analytics for decision-making about workforce composition and planning. Since HRMIS reforms are often cross-cutting, institutional coordination among the various government agencies involved in public administration is critical. An institutional mandate and the inclusion of key stakeholders may facilitate this effort.
- The reform process should sequentially strengthen the maturity of the wider system, with defined stages guiding the implementation process. A sequential approach to HRMIS reforms is illustrated in the figures throughout this chapter. They imply the preparation of a rollout strategy—however limited—that plans for the political obstacles ahead and considers the constraints of the existing legal framework. Implementation requires both repeated testing of the solution and creating accessible technical support for users of the HRMIS. Finally, monitoring the reform includes credibly shutting down legacy systems and tracking the use of new solutions.
- A gradual and flexible approach can enhance the sustainability and future development of the HRMIS, due to unexpected data coverage and quality issues. Because HRMIS and other public sector data systems are so complex, unexpected challenges may arise along the way. Data coverage may be incomplete, requiring that additional information be integrated from other modules. Data quality may also be compromised because incorrect human resources (HR) records may be widespread. Therefore, reform teams should build contingency plans from the start, make choices that provide them with multiple options, and be ready to adapt their plan even during the implementation phase.
- The design of the reform should carefully consider the trade-offs involved in choosing different specifications. Design choices have different implications for reform, regarding both the breadth of the reform and its sustainability. For instance, outsourcing the solution to private firms for the implementation of HRMIS reforms may reduce the need to build in-house capacity to develop the software and accelerate the reform timeline, but this choice may still require building capacity for maintenance in the long run. Building internal capacity and managing these operational trade-offs is at the heart of a public service that is most likely to capitalize on technological progress.

## INTRODUCTION

The *World Development Report 2021* (World Bank 2021b) outlines the potential for data to improve developmental outcomes. However, as the report highlights, the creative potential of data can only be tapped by embedding data in systems, particularly information systems, that produce value from them. In other words, data must be harnessed into public intent data, defined as “data collected with the intent of serving the public good by informing the design, execution, monitoring, and evaluation of public policy” (World Bank 2021b, 54).

For data to serve this purpose, robust data infrastructures are necessary. Governments across the world collect vast amounts of data, particularly through statistical offices—when gathering information on society—and through the internal use of management information systems (MIS) (Bozeman and Bretschneider 1986).<sup>1</sup> These forms of data infrastructure are used to generate measures in multiple policy domains described elsewhere in *The Government Analytics Handbook*: from budget planning (chapter 11) to customs (chapter 14) and public procurement (chapter 12). Government-owned data infrastructure can generate data analytics to support policy making through the application of statistical techniques (Runkler 2020).

However, data infrastructures that provide analytical insights are still at various levels of maturity in the public sector in general and developing countries in particular. A 2016 report highlights that governments only explore 10–20 percent of the potential value of analytics, in contrast to 40–50 percent in private sector retail (Henke et al. 2016). Multiple factors account for the relative underdevelopment of analytics in public administration. In contrast to the private sector, governments respond to multidimensional demands and diverse stakeholders (Newcomer and Caudle 1991). Siloed and legacy systems inhibit data integration and analytics pipelines (Caudle, Gorr, and Newcomer 1991).

Promoting the use of data analytics in the public sector requires a combination of both technological innovation and organizational change, the analog complements to data analytics (World Bank 2016). In particular, the development of data analytics within the public sector requires a coordinated effort to both transform how data are stored and analyzed and embed these analytical insights into the decision-making processes of public sector agencies. These reforms are often part of a larger digitalization strategy (World Bank 2021a). It is these reforms in data infrastructure that make possible the use of data analytics in the public sector, often led by a public sector reform team.

This chapter provides a road map to the implementation of analytically driven data infrastructures in the public sector, drawing on the experiences of World Bank practitioners and government officials across the world. The substantive focus is on human resources management information systems (HRMIS), a core function within public administration.<sup>2</sup> The conceptual framework outlined provides a foundational perspective on data analytics in the public sector, exploring the established domain of human resource management to illustrate key design decisions in the transformation of data infrastructure into analytics. However, the road map described in this chapter is generalizable to a variety of settings, and throughout the chapter, we emphasize its adaptability to other settings.

The conceptual framework is divided into two parts. The first section provides a typology of the modules that comprise an HRMIS, describing both their content and how they relate to one another. The emphasis is on the distinction between foundational and analytics modules—in particular, how the foundational modules feed into analytical products. Equipped with conceptual clarity about the structure of the information system, we move on to the operational framework for HRMIS reforms. This section describes in detail a framework for HRMIS reforms, outlining a sequential approach to reform (Diamond 2013). The operational framework describes the different stages in HRMIS implementation, their requirements, and best practices for each.

After laying out this conceptual framework, the chapter focuses on a set of case studies to illustrate how it can be applied in practice. Luxembourg showcases the development of a human resources business intelligence competency center (HR BICC), an intricate dashboard that has revolutionized how HR analytics are conducted. The case of Brazil describes how a machine-learning-empowered fraud detection system reduced

costs and improved the efficiency of an audit team responsible for overseeing the federal government payroll. Finally, the case of the United States highlights the experience of a team in the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) that developed a dashboard enabling the fast and intuitive use of employee engagement surveys for policy making.

We recognize that these cases are drawn primarily from developed countries (the United States and Luxembourg) and a developing country with a relatively mature HRMIS (Brazil). Practitioners should be aware that while the lessons are generalizable to contexts in which MIS may be less mature, the challenges faced may differ. For instance, in each of these cases, an HRMIS was already in place with foundational modules to use in analytical transformation. This may not be the case in countries where the foundational modules have not been set in place. As a result, while a similar operational framework may be deployed, the types of products (analytical or foundational) may differ substantially. Having said this, we believe these cases illustrate general principles that are useful for all practitioners interested in applying an operational framework for HRMIS reforms. Each case study is described in greater detail in case studies 9.1–9.3 of the *Handbook*.

A set of practical lessons emerge from the conceptual framework and case studies. First, a modular approach to HRMIS reforms enables a precise definition of the reform's scope and how the intervention will affect the broader data ecosystem. Reform teams should consider available resources when deciding which modules to target for reform, as well as how data flow across modules. Second, foundational modules, which focus on payroll, personnel management, and position management, should take precedence over analytical layers, in large part because analytics requires well-structured data records. Nevertheless, analytical layers can be designed for specific modules within an HRMIS if their scopes are sufficiently narrow and well defined. In general, a sequential and flexible approach to data infrastructure reform is recommended. An ex ante assessment of key issues in human resource and wage bill management (both in terms of their likelihood and the severity of their impact on the system as a whole) will enable governments to hedge risks to their initial design.

Finally, the implementation of data infrastructure reforms needs to navigate political-economic issues and ensure leadership commitment and institutional coordination among agencies. To do so, it helps to anchor measurement and analytical outputs in problems the government considers priorities to address. In an HRMIS, these are typically wage bill controls, analytics for personnel decision-making, and workforce planning. Coordination can be facilitated by including key stakeholders and clarifying institutional mandates over data collection and processing. On a more technical note, capacity issues should be considered before and during implementation. Governments often implement large-scale data infrastructure reforms for the first time and may require external assistance from experts who have engaged in similar reforms before.

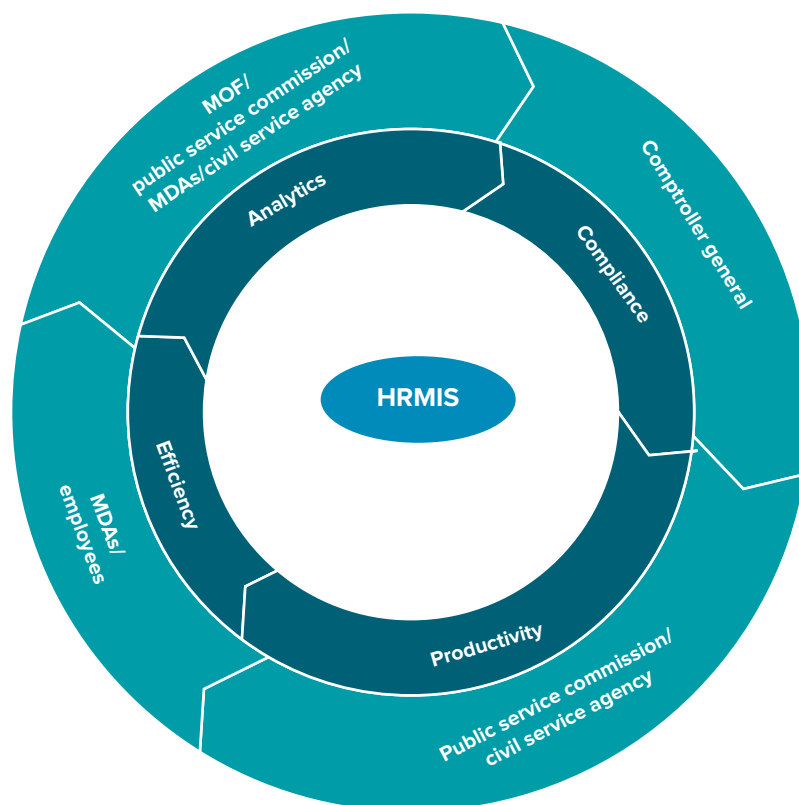
These lessons are generalizable to other data infrastructures. A modular approach to reform can be applied in non-HRMIS settings as well. For instance, reforms for public procurement information systems may focus on the bidding process or on contract implementation. Additionally, analytical insights can be derived from each one of these modules, but only once data quality and completeness are ensured.<sup>3</sup> A sequential and flexible approach to reform is beneficial in non-HRMIS settings as well. Shocks, whether political or technical in nature, can occur regardless of the content of the information system, and reform teams should be ready to address them in both the preparation and implementation phases.

This chapter is structured as follows. Section 2 presents policy objectives and associated issues for an HRMIS. Section 3 presents a typology of HRMIS core modules and how these modules relate to one another. Section 4 presents a framework for HRMIS reform implementation. Section 5 provides an overview of the case studies and applies the conceptual framework to the cases. Finally, section 6 concludes.

## HRMIS POLICY OBJECTIVES

In this section, we present four objectives common to HRMIS: compliance for fiscal sustainability, employee productivity, process efficiency, and analytics for decision-making. In so doing, we also highlight issues in their implementation. Each HRMIS module maps onto key policy objectives for an HRMIS: the compliance

**FIGURE 9.1** Policy Objectives for Human Resources Management Information Systems and Respective Stakeholders



Source: Original figure for this publication.

Note: HRMIS = human resources management information systems; MDAs = ministries, departments, and agencies; MOF = ministry of finance.

objective is associated with core modules, such as payroll, employee productivity is associated with talent management modules, and analytics for decision-making is associated with analytical modules. Process efficiency is cross-cutting and often directly relates to the way HR records are produced. For example, the appointment of civil servants may require long verification processes, often done manually. Automation of the process could increase HRMIS efficiency. Figure 9.1 highlights these policy objectives and their stakeholders.

### Compliance for Fiscal Sustainability

Wage bill compliance with the established budget is necessary for fiscal sustainability. The comptroller general, or another agency responsible for government payroll, manages payroll controls for budgetary compliance. At a more granular level, payroll controls include verifying the authenticity of employment and ensuring accurate payroll calculations and reconciliations. Verifying the authenticity of employment requires identifying and eliminating ghost workers.<sup>4</sup> Ghost workers adversely impact fiscal sustainability and often draw negative attention from the public and policy makers. Another important control is compliance with budgetary ceilings. In many jurisdictions, the approved budget is not directly linked to payroll, leading to overspending. HRMIS regulations should be interpreted correctly and consistently to calculate pay, allowances, and deductions. Employee records should be updated regularly, with bank reconciliations and payroll audits to ensure integrity and compliance.



## Employee Productivity

The public service commission and the civil service agency are entities focused on employee productivity and engagement. An HRMIS should give them an accurate overview of employees to help them improve recruitment, develop the performance of the workforce, and enhance their skills. Information on employee qualifications and skills can inform a strategic view of workforce training so that these entities can design training strategies around skills shortages and respond to emerging capacity needs. Recruitment patterns can be analyzed to improve talent pools and reduce potentially discriminatory practices. The performance of the workforce can be monitored using metrics on engagement and attrition rates. In the absence of these measurements, stakeholders are unable to approach employee productivity in an evidence-based and strategic manner.

## Process Efficiency

Government agencies, including the ministry of finance, the public service commission, and the civil service agency, are also interested in improving operational efficiency. In some settings, the HR department manually calculates the salaries of employees each month using spreadsheets. This process is not only extremely inefficient but also prone to error. Another example of operational efficiency lies in the hiring process. Hiring departments perform multiple verifications for the first-time appointment of civil servants. These may involve verifying hard copies of key information, such as prosecution history or educational qualifications, from multiple departments and ministries. Manual procedures and hard-copy files delay administrative actions, leading to lower morale, productivity, and efficiency. Process efficiency is therefore another key policy objective for an HRMIS.

## Analytics for Decision-Making

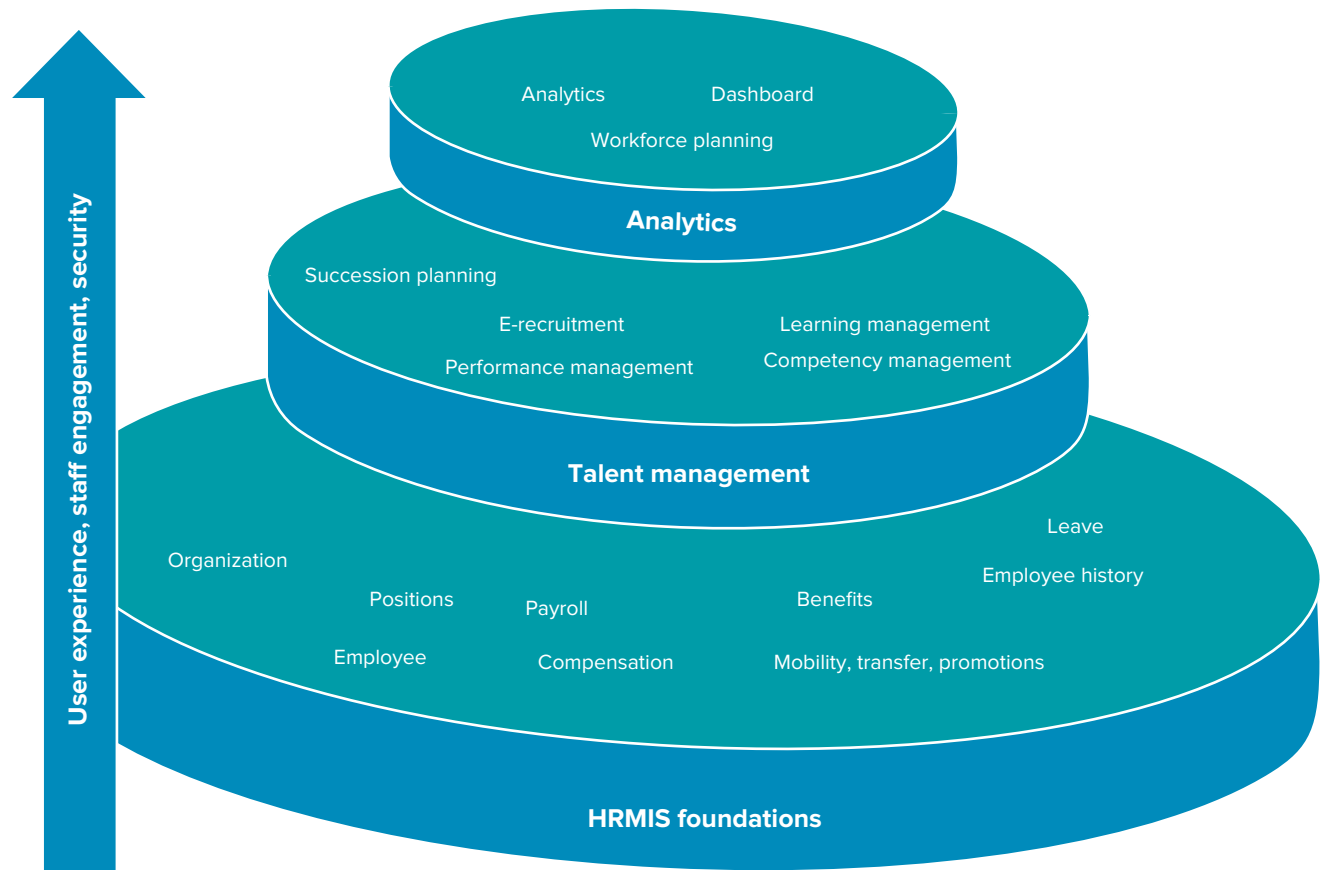
The ministry of finance—together with the civil service agency and the establishment or human resources (HR) department—needs data on HR for making evidence-based, strategic decisions. These decisions concern a range of issues, such as the overall size of the wage bill, salary and pension increases, cuts on allowances, and the financial impact of employee benefits, like medical insurance. Decision-makers need reliable HR reports in a timely manner. In most jurisdictions, these reports are collected manually, through ad hoc measures that take weeks or months, adversely impacting efficient decision-making. In advanced settings, analytics, dashboards, and business intelligence applications are used to enhance effective decision-making.

While these policy objectives are specific to an HRMIS, note that a more general exercise can be done for other information systems. For example, chapter 11 highlights how policy objectives such as budgetary compliance can inform the analytical use of public expenditure data. Ultimately, information systems are designed to assist policy makers in accomplishing their goals. It is only sensible that, depending on the policy area, these goals may differ, but this policy orientation remains the same.

## HRMIS CORE MODULES

An HRMIS is an information system designed to enable the management of human resources in the public sector. As such, these technological solutions offer a variety of functionalities, which correspond to modules such as payroll or talent management (figure 9.2). HRMIS reforms require careful consideration of the scope of an intervention—in particular, consideration of which module within the HRMIS will be targeted.

**FIGURE 9.2** Human Resources Management Information Systems Data Infrastructure Modules



Source: Original figure for this publication.

Note: HRMIS = human resources management information systems.

Identifying what modules comprise an extant HRMIS enables the consideration of interdependencies across modules, as well as the feasibility of reform.

HRMIS modules comprise four broad categories, ordered from foundational to analytical:

1. **Foundational modules** ensure compliance and control of both the wage bill and personnel. They include payroll (benefits and compensation) management, position and organizational management, career management, and employee profiles, among others.
2. **Talent management modules** include recruitment and performance management, competency, and learning, as well as succession planning. Talent management is used primarily to improve employee productivity.
3. **User experience and employee engagement modules** improve user experience and employee engagement. These modules encompass employee and manager self-service and staff survey systems.
4. **HR analytics** can be developed for workforce planning, as well as strategic and operational decision-making, once the data infrastructure layer has been developed.

These modules provide the basic infrastructure layer for HR data and associated analytical outputs. The foundational modules are responsible for the accurate and reliable storage of any form of HR record. Talent management modules monitor career life cycles, and user experience modules monitor the overall experience

of users who interface with the MIS or whose data comprise its case data. Finally, the analytics layer extracts value from the underlying data infrastructure to inform strategic decisions at an operational level.

An analogous structure can be found in other MIS. For instance, when considering customs data (see chapter 14), foundational modules include revenue collection and the release time of goods. These modules comprise their own records (“What was the monetary value of the customs declaration for a particular good?”) and potentially their own indicators (“What was the total revenue on exported goods for the month of June?”). Analytical modules provide these indicators to inform policy making. As an example, if customs authorities detected an atypical decrease in tax revenues for a given month, they might send a team of auditors to verify why this occurred. This example highlights how, while the data content of these systems differs, the logic by which they are organized remains largely the same.

Note that each of the modules outlined in figure 9.2 is connected to a set of records and measurements, described in further detail in table 9.1. The table highlights the variety of available HRMIS indicators and

**TABLE 9.1 Human Resources Modules and Associated Measures**

Module	HR measures
<i>Foundational</i>	
Payroll	Size of wage bill and budget/position compliance; deviation of wage bill expenditures from the budget; sector, administration, and geographical breakdown; percentage of employees in various categories—civil servants, public servants, part-time, wage bill arrears
Position management	Establishment control—employees paid against the approved positions in the budget; average tenure on a position; regional quotas; ratio of public servants, civil servants, political appointees, temporary workforce, and other employee categories
Organization management	Organization architecture reflecting government structure
Employee master data	Tracking of policy effectiveness on gender ratios, regional quotas, and minorities, disabled, and other minority groups; education profiles—degrees and certifications; experience; history of service; ghost workers as a percentage of total workers
Personnel development	Competency improvement measures; promotions; secondments
Benefits management	Benefits and their cost impact
Compensation management	Salaries and allowance structures; compensation equity/disparities in allowances/pay across sectors, administrations, and geographies
Time management	Absentee rate; overtime; staff on various types of leave
Pension	Pension as a percentage of the wage bill; future liabilities; impact of pension increases on budget
<i>Talent management</i>	
Performance management	Top-rated and lower-rated employees disaggregated by ministry, department, and agency; rate of performance reviews completed—ministrywide
E-recruitment	Time to hire; time to fill; applications per job posting; recruitment patterns; applicant profiles; recruitment method—through public service commission, direct contracting, contingent workforce, political appointments, or internal competition; ministry-level appointments
Learning management	Training and skills metrics
Succession planning	Percentage of identified positions that have an identified successor
Workforce planning	Ratios—gradewise and sectorwise; promotions; vacancies
Career development	Promotion rate; average time to promote in a grade per service category
Competency management	Percentage gaps in required competencies

(continues on next page)

**TABLE 9.1 Human Resources Modules and Associated Measures (continued)**

Module	HR measures
<i>User experience and employee engagement</i>	
Employee self-service	Time taken to complete an HR transaction; number of self-service systems available
Manager self-service	Time taken to decide or approve HR transactions; number of manager self-service systems available
Mobile apps	Number of mobile apps available for various HR functions—leave, profile
Employee engagement surveys	Number of employees responding to surveys; satisfaction rate with management and HR policies
<i>HR analytics and reporting</i>	
HR reports, analytics, or dashboards/workforce planning	Levels and distribution of employment; wage bill and its distribution across sectors, administration, and geographies; wage bill and its impact on fiscal sustainability; wage bill as a percentage of revenue; wage bill as a share of GDP; public vs. private employment; sector, administration, and geographic distribution of public employment

Source: Original table for this publication.

Note: HR = human resources.

their corresponding modules, which can be selected and adjusted to practitioners' needs. A payroll module may include different sets of measurements, from the size of the wage bill to a breakdown of contract types for civil servants. This diversity implies that practitioners may select measurements that are relevant to their use case, prioritizing some modules and indicators over others.

## OPERATIONAL FRAMEWORK FOR HRMIS REFORMS

HRMIS reforms are designed to address issues and bottlenecks that prevent stakeholders from accomplishing their policy objectives, such as improving compliance and employee productivity. The implementation of HRMIS reforms can be divided into three stages: preparation, implementation, and monitoring. Figure 9.3 outlines the different phases and their respective components, and the following subsections discuss each of the steps outlined.

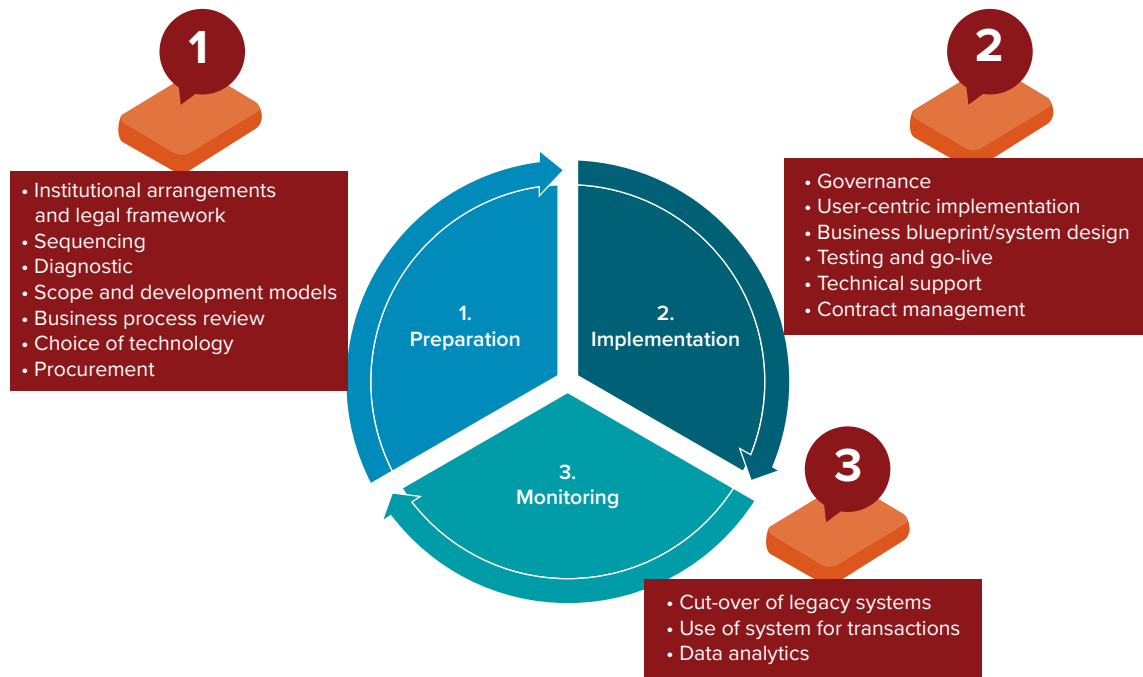
Note that the stages of HRMIS reforms described in the subsequent sections are agnostic with respect to the particular module targeted for reform. Different HRMIS reforms, whether in building an analytics dashboard or improving foundational modules, require a similar approach with regard to the sequence of implementation. Of the multiple elements contained in each of these phases, practitioners are encouraged to begin by assessing which elements are of the greatest importance to the implementation of a particular HRMIS reform project in their setting.

### Preparation

The preparation phase lays the groundwork for the implementation of HRMIS reforms. In this phase, key design choices occur, such as defining the scope and the technology to be deployed. Additionally, the preparation phase is an opportunity to engage in a comprehensive diagnostic of the current HRMIS, as well as define the scope of the reform and identify the modules that will be addressed in the intervention.

The preparation phase is an opportunity for the reform team to familiarize themselves with their institutional context as well as the extant data infrastructure, adjusting their strategy in the process. In settings

**FIGURE 9.3** Human Resources Management Information Systems Reform Sequence



Source: Original figure for this publication.

where a decentralized HRMIS is in place, the implementation team may require senior management support to promote reforms to different agencies. Much of the effort in the preparation phase should be spent on ironing out institutional coordination issues.

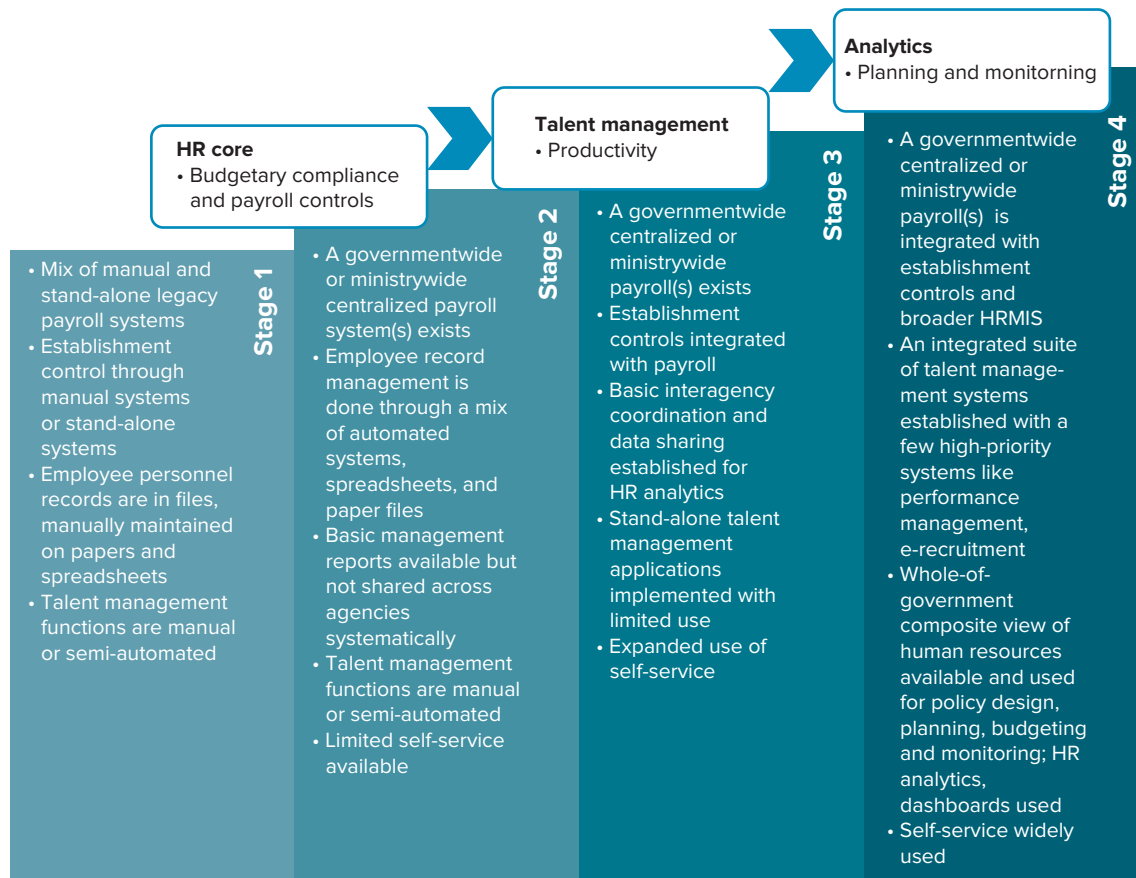
### **Rollout Strategy and Sequencing**

A system rollout strategy should be prepared to guide the implementation. The strategy should address key issues of the scope and sequencing of the rollout. The rollout strategy builds on the mapping of the different modules of the HRMIS. In contrast to the broader HRMIS implementation sequence, the rollout strategy defines how the intervention will achieve the development of the HRMIS. This sequencing accounts for the level of maturity of the current HRMIS, as outlined in figure 9.4.

In terms of sequencing, most countries spend considerable effort on first establishing the basic data infrastructure layer and ensuring compliance and controls (stage 1). Compliance generally centers on the financial compliance of payroll with budget but also on the accurate calculation of payroll to avoid paying ghost workers. Once the data infrastructure layer has been established, it would be prudent to work on data analytics, HR reports, and workforce planning through the implementation of these modules. However, it is worthwhile to note that raw analytical reports can be shared with decision-makers even during the foundational modules, without waiting for the full-blown implementation of analytics modules.

We can generalize stages of maturity in data infrastructure to other cases beyond HRMIS. Budgetary compliance and control in an HRMIS refer to more general principles of data quality and comprehensiveness, which allow for a transition from stage 1 to 2. This is a foundational step for data infrastructure in the public sector: it allows for the reliable collection and retention of data on a variety of HR procedures. The next step is the transition from stage 2 to 3, focusing on productivity: the use of HRMIS data to generate performance or productivity metrics on personnel. This requires access to reliable data, produced by

**FIGURE 9.4 Human Resources Management Information Systems Maturity**



Source: Original figure for this publication.

Note: HR = human resources; HRMIS = human resources management information system.

the careful implementation of step 1, as well as defining how to measure productivity and which indicators would be necessary to do so. The final step is planning and monitoring, where reliable data produce indicators that inform policy making. This characterizes an HRMIS that has transitioned to stage 4 of data analytics.

### ***Institutional Coordination and Legal Framework***

The implementation of a new HRMIS requires institutional coordination due to its complexity and its multiple stakeholders. One of the first steps in institutional coordination is identifying the principal stakeholders in an HRMIS. In most settings, the comptroller general of accounts, under the ministry of finance, is responsible for the payroll and leads the initial implementation to improve budget compliance, payroll calculations, and overall efficiencies. However, in other settings, it is the civil service agency that is responsible for an HRMIS and leads the initial implementation.

An HRMIS may also be decentralized. In this context, the ministry of finance is more focused on ensuring payroll compliance with existing regulations. It requires line ministries and respective HR departments to send payroll payment requests to the comptroller general to ensure budgetary control through an integrated financial management system. This decentralized arrangement could pose additional challenges to reform due to the multiplicity of stakeholders, siloed data, and the need for coordination. In such a case, reform may require additional coordination and buy-in from the implementation team.



Legal authorizations and associated reform strategies should be secured prior to reforms. For instance, the implementation team may require a set of legal documents authorizing the implementation of the reform, which may include terms of reference and procurement requirements to hire a vendor. These documents clearly articulate the scope of the reform to the implementation team and affected agencies. They provide assurance that necessary permissions have been secured for the project, reducing uncertainty and potential negative repercussions for not complying with the existing regulatory framework.

To avoid political resistance to reform, a softer approach can be adopted. Under this approach, line agencies can continue with their previous HRMIS but are asked to provide HR data to a central repository directly from the legacy system. However, this approach does not address inefficiencies associated with duplicative investments and siloed approaches, nor does it ensure compliance once the reform is implemented. Considerable delays and noncompliance can occur, though analytics information for decision-making may become initially available.

### **Defining the Scope of an HRMIS**

Changes to an HRMIS can target one or more of the HRMIS modules outlined in section 2. Note that the choice of scope entails important trade-offs between the breadth of the intervention and the feasibility of the reform project. Three possible scopes for HRMIS reforms are:

- **Increasing the number of indicators and modifying existing indicators for a particular HRMIS module.** For instance, reform stakeholders may be interested in obtaining additional information on the types of appointments of civil servants. This may include further disaggregation of appointment indicators: whether civil servants are tenured, political appointees, or on temporary contracts.
- **Expanding HRMIS coverage by increasing the number of modules.** A key decision may be what institutions will be covered under the HRMIS and how that coverage will be phased in. Certain institutions may differ in important ways from those in the wider service. How will these differences affect the types of indicators available?
- **Ensuring national coverage of the HRMIS, including subnational governments.** Civil servants may be spread across the country in various regions, provinces, or districts. Is the HRMIS meant to reach the entire country or some subset of the country? An analysis of employee coverage in these geographical areas would help define the scope and logistical effort in the implementation of the HRMIS.

By choosing the scope of the HRMIS along each of the above dimensions, the implementation team identifies the key issues about which choices must be made. Key areas to be covered could include legal, policy, and institutional frameworks; HR and payroll regulations (including pay scales and allowances and the extent of their consistent application across ministries, departments, and agencies [MDAs]); budgetary allocation; and key issues of compliance, productivity, efficiency, and analytics.

### **Choice of Technology**

Another key decision point during the preparation phase is the choice of technology (figure 9.3). Two major categories of software technology are available: custom-developed software and commercial off-the-shelf (COTS) software, though some open-source software choices are also available.

Under the custom-developed software approach, the likelihood that users accept the new software and procedures is higher because these implementations can be adapted to user requirements. For example, case study 9.2 outlines how the implementation team in Brazil tailored a solution to detect payroll irregularities using custom-developed software. The solution extracted HRMIS data and presented them in exactly the way payroll analysts required to execute their tasks. This level of customization, however, comes at a cost. Custom-developed systems require higher in-house capacity because all parts of the software have to be coded from the bottom up, rather than relying on a prepackaged solution.

Additionally, maintenance for custom-developed software tends to be higher in the long run because any changes to the underlying data infrastructure require changes to the software itself. If the original implementation team is no longer present—as is often the case—a new implementation team has to start from ground zero.

COTS software often contains prepackaged good practices and tighter integration between different parts of the information system. It also frequently comes with regular software updates, reducing the risk that the technology becomes obsolete. Major COTS packages include SAP, Oracle, and PeopleSoft, though financial management software like FreeBalance also provides public-sector-relevant HRMIS modules. As a result, COTS software applications are more robust and easier to maintain than their customized counterparts. However, this robustness comes at a cost. Adaptation to user needs—such as introducing novel indicators or modules—is, in general, difficult if not impossible to implement within the existing COTS software. Because the software is proprietary, modifications to the underlying software are not available to the implementation team.

Overall, custom-developed software is more suitable for nonfoundational modules, while a COTS solution is better suited to foundational modules because it ensures tighter linkages of these modules with each other. For modules like e-recruitment or performance management, governments can choose any technology platform from the market that meets their requirements and is cost efficient. Integration of these modules with the foundational HRMIS modules will ensure data integrity.

## Procurement

In most cases, HRMIS reforms are not fully delivered “in-house” by governments, for a variety of reasons. For instance, COTS solutions require that an external vendor build and deploy an HRMIS for use by a government agency. This includes the introduction of new modules and the training of government officials on how to properly use and manage the software. For customized solutions, the government may lack access to a team of software developers and data engineers to fully develop the solutions. As a result, it may have to rely on external vendors with the required expertise to do so.

As a result, an external vendor must be procured to support reform implementation. The culmination of the preparation phase is preparing a procurement package for the HRMIS implementation partners. The procurement document should cover multiple aspects of implementation: business process review, the deployment and rollout plan, quality assurance and testing of the solution, and the help desk and support strategy, among others. Client and vendor responsibilities, risks, and rights—such as intellectual property—should be protected equitably, in line with industry good practices.

## Implementation

The second stage of the implementation of HRMIS reforms, as outlined in figure 9.3, is the actual implementation of the reform plan. The implementation stage includes the management of HRMIS reforms, which first requires considering and defining the governance structure. Additionally, during the implementation phase, it is the responsibility of the implementation team to provide and adapt the business blueprint that guides the project. Iterative testing must take place to ensure that the project scope is being successfully developed. Technical support and a help desk ensure that users are supported throughout the implementation phase. Contract management ensures that expectations are aligned between government clients and external vendors.

The implementation stage of HRMIS reforms thus requires clear authority by the implementation team to make decisions and communicate them clearly to potential external vendors and end users. Flexibility is also required during the implementation stage, as well as proper documentation of any changes in the project design as a result of the implementation stage. Due to this flexibility, it is important to coordinate with external vendors during the rollout of implementation and to collaboratively decide whether changes are

indeed feasible under the existing contract or if additional resources—financial or time—may be necessary to successfully roll out the reform. We provide further detail below.

### **Governance Structure**

For effective implementation of HRMIS reforms, it is often necessary to form a steering committee to provide strategic guidance and ensure support from the project sponsor. This steering committee should ensure that key stakeholders are fully represented and consulted. The committee should have the authority to make strategic decisions and resolve strategic-level conflicts. To improve the efficiency of decision-making and the quality of implementation, the steering committee in some settings can also issue basic principles of implementation. These principles can be customized by context and include standardized business processes, user-centric system design, security, and privacy, among others.

The project director should be supported by a project management team, where possible, including procurement and financial management specialists, a project manager, a communications team, and change management teams, among others. A core team of subject matter experts from the ministries should be consulted to ensure they codesign and codevelop the system with the implementation partners. The core team should have a say in decisions carried out by the steering committee, ensuring co-ownership of the solution.

### **System Design Document**

The implementation team should prepare and revise a system design document throughout the implementation of the project. The system design document defines the needs and requirements for the new design of the HRMIS and should be approved by the steering committee before implementation. After launch, any modifications to it should be subject to steering committee approval as well. This living document becomes the final scope document with the technical details of the implemented solution. It also becomes the reference technical design document for future upgrades, and for any new implementation team if the existing vendor changes.

### **Iterative Testing**

Changes to the HRMIS should be developed iteratively. Iterative testing allows for controlled and reversible innovation within the HRMIS reform project, relying on feedback from senior management and staff who will ultimately use the new HRMIS. For instance, an implementation team may be interested in developing an interactive dashboard to measure employee engagement. However, an initial focus on indicators such as employee satisfaction may have to be replaced by employee exit surveys after an initial round of feedback from the steering committee, which is concerned about employee turnover. Iteration preserves flexibility and identifies features that, in the implementation stage, may not be considered relevant. Additionally, it enables adjustment to happen in reversible and controlled stages that do not jeopardize the wider integrity of the project. All changes made during iterative testing should be documented in the system design document.

### **Technical Support and Help Desk**

Technical support allows users to successfully navigate the transition to the reformed HRMIS. Clear documentation on how to use the remodeled HRMIS, as well as a help desk, should be implemented during the project rollout. This ensures users have sufficient information to use the HRMIS during and after the reform process. Failure to do this may result in increased user resistance because users may be confused and unable to operate the new system. Standardized help desk software tools, together with a telephone helpline, should be provided to ensure that user requests are appropriately logged, assigned, resolved, and monitored. Frequently asked questions should be compiled and shared, empowering users to find solutions to their own problems, minimizing help desk calls, and building a knowledge base of solutions.

## Contract Management

Contract management is another critical aspect of implementation. Implementation failures are often the result of inadequate contract management. Issues like the scope of the contract and any modifications require that both the steering committee and the vendor align expectations before and during the implementation of HRMIS reforms. Expectations should also be aligned regarding the payment schedule and the responsibilities of the contractor and vendor during the implementation process to avoid confusion and ensure smooth implementation of the project. A collaborative approach in contract management, which considers vendors as partners and not as contractors, is recommended. This collaborative approach creates a mindset of shared responsibility for successful HRMIS reforms.

## Monitoring

The third phase shown in figure 9.3 is monitoring the HRMIS once it has been implemented and is in place. The monitoring phase focuses on issues the implementation was meant to address and quantifies the benefits in terms of business results. Often, the implementation team monitors the project in terms of module development, user acceptance, trainings, and so on. While this approach could be useful for internal project management, it has limited utility at the strategic level if the modules have been developed but the business results are not delivered. Therefore, utilization of the system and its coverage should be the key focus of monitoring. If user departments continue to use legacy arrangements while the newly developed HRMIS is only used as a secondary system, the business benefits will be limited.

## Transition from Legacy Systems

Even after HRMIS implementation, it is often difficult to fully transition from the legacy system to the redesigned HRMIS. The use of the legacy system as the primary system of records and transaction processing poses a serious challenge. Continued use increases the workload by requiring the constant synchronization of old and new data systems. If the legacy system is still used as the primary system of records after the reform, this reduces the likelihood that the newly developed HRMIS will be used as the primary system. Therefore, during and after the implementation of HRMIS reforms, the legacy system should be gradually shut down to ensure there is a complete switchover to the new system. If required, governmentwide regulations and directives should be issued to ensure the use of the new HRMIS.

## Key Performance Indicators

Key performance indicators can help implementation teams gauge the relative success of the implementation process. These indicators should allow the implementation team to monitor how well the reform has performed. For instance, if the implementation team is intervening in a payroll module, it may develop an indicator on the proportion of the wage bill processed through the new HRMIS. Additionally, if one of the goals of the reform is to ensure payroll compliance, indicators can be developed to detect ghost workers. The proportion of employees with verified biometrics is an example of a key performance indicator that enables measurement of this goal.

## Monitoring Analytics

The use of monitoring analytics can provide stakeholders with immediate feedback on implementation. An HRMIS should be used to provide analytical information to key ministries involved in strategic decision-making. Initial monitoring should be provided even of foundational modules while the data analytics pipelines and dashboard applications are not fully developed. This will maximize the business value of the data gathered in the HRMIS. It will also provide a political support base for the system when the key

decision-making ministries harness the benefits of these investments. These ministries could include the ministry of finance, the public service commission, the civil service agency, and other large MDAs.

## CASE STUDIES: HRMIS REFORMS IN PRACTICE

To illustrate the implementation of HRMIS reforms in practice, we provide a set of HRMIS case studies that showcase how government officials and practitioners have employed the techniques outlined above in the reform process. In so doing, we highlight patterns in the development of data infrastructures, common challenges, and the design choices that guided these teams in their development efforts. These cases describe the HRMIS reform process as it was experienced by practitioners. We recognize that these cases represent two developed countries and one developing country with access to a mature HRMIS. As a result, practitioners should tailor lessons in this section to their own context. We highlight how the operational framework for HRMIS reforms is generalizable to other settings as well, from building foundational modules to implementing analytical modules. Subsequent case studies provide a fuller description of the cases, while this section provides a comparative analysis of all three.

### Luxembourg

In Luxembourg, the State Centre for Human Resources and Organisation Management (CGPO) is a central government administration, located in the Ministry of the Civil Service. Its mandate spans multiple responsibilities, from managing the life cycle of civil service personnel to strategic workforce planning. In 2016, the CGPO faced growing demands and follow-up needs from HR specialists and decision-makers in the federal government of Luxembourg. As the volume of these requests increased, it became clear to the CGPO that its HRMIS had to change.

In 2017, the CGPO developed and deployed a comprehensive HRMIS reform, which enabled the CGPO to build a comprehensive HR data infrastructure and framework to plan and monitor HR in the government of Luxembourg. The solution developed was large in scale, involving multiple data sources and HR specialists. This analytics center, the HR BICC, was developed over the course of a year and had important transformational consequences for the way HR was conducted. An illustration of the novel dashboard is presented in figure 9.5. It integrates both HRMIS data and strategic planning documents in a comprehensive dashboard portal (in orange).

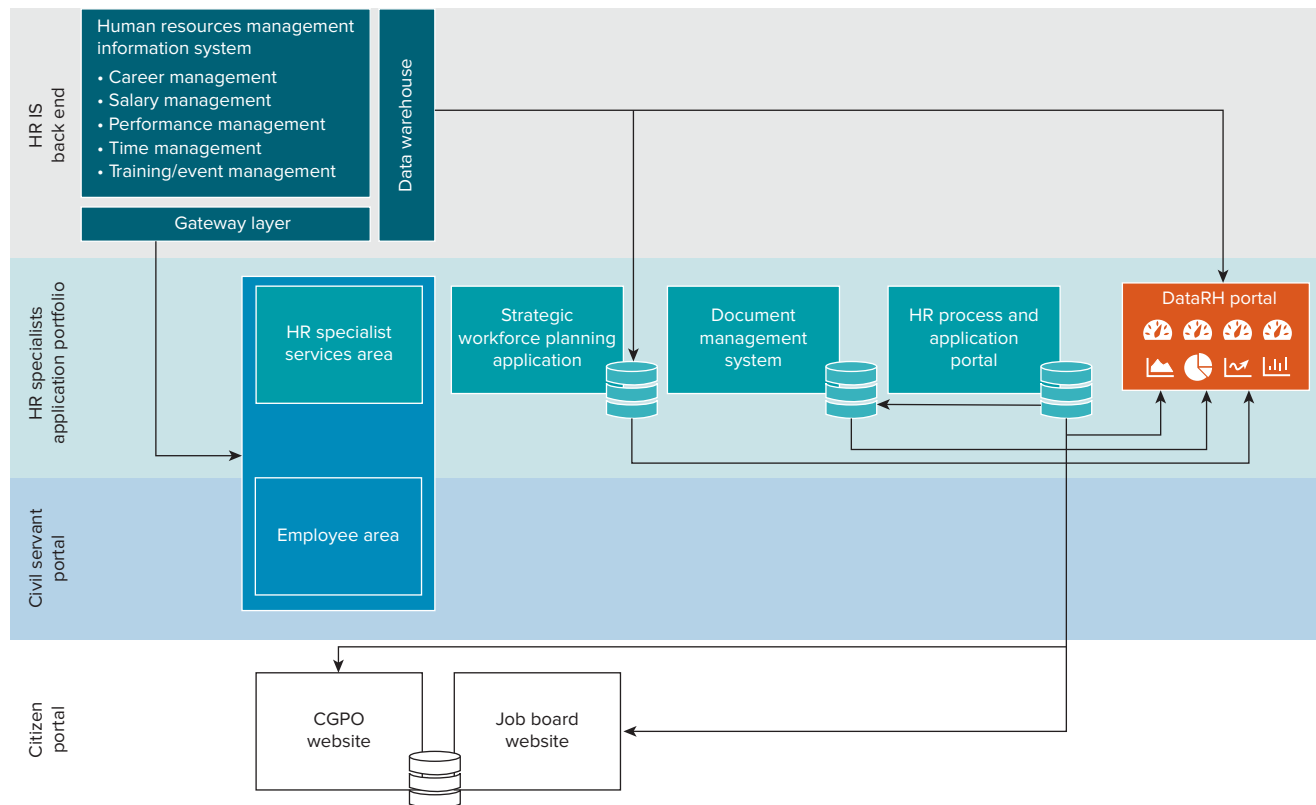
The Luxembourg case presents a fully integrated HRMIS pulling together all the major databases that are typically the focus of such exercises. As such, this case is the most comprehensive example of the implementation of a full HRMIS analytics module, as outlined in figure 9.2.

### Brazil

In Brazil's federal government, payroll quality control is the responsibility of the Department of Compensation and Benefits (DEREB). DEREb flags paycheck inconsistencies before disbursement, which are then forwarded to federal agencies' HR departments for correction. The task is challenging. The case volume is large, with tens of thousands of individual paychecks processed daily. Additionally, a complex set of regulations governs how payments should be disbursed. To enforce these rules and detect inconsistencies, a team of payroll analysts individually verifies each paycheck. The implementation team sought to improve this process.

In 2019, a partnership between DEREb and a private data science consulting firm (EloGroup) resulted in the development of a machine-learning-empowered fraud detection system. To generate the necessary data

**FIGURE 9.5 Human Resources Management Information System, Luxembourg**



Source: Adapted from CGPO.

Note: CGPO = State Centre for Human Resources and Organisation Management; HR = human resources; IS = information system.

to train this algorithm, a thorough restructuring and integration of the extant data infrastructure on payroll, compensation rules, and HR were developed. Through the development of new extraction, transformation, and loading (ETL) processes, this solution enabled auditors to better detect irregular payroll entries, increasing savings and improving efficiency.

The Brazil case illustrates that, although some HRMIS reforms may be relatively narrow and aimed at a particular outcome—in this context, fraud detection—many of the themes outlined in earlier sections are still of relevance to their implementation. Many of the steps taken in the development of the fraud detection system are foundation stones for wider HRMIS reforms, highlighting how the same methodology can be applied even in smaller contexts.

## United States

Every year, the Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS) is administered to over 1 million federal civil servants in the United States.<sup>5</sup> The FEVS measures employees' engagement through a variety of survey questions and provides valuable information to government agencies. In theory, it presents data on agency strengths and opportunities for improvement in employee engagement. However, extracting insights from the FEVS is challenging. Once given access to the survey, government agencies spend weeks analyzing the data to operationalize their findings. This effort is labor intensive and costly. An HRMIS reform team sought to accelerate this process.

In 2015, the NIDDK, within the National Institutes of Health (NIH), developed the Employee Viewpoint Survey Analysis and Results Tool (EVS ART) to extract rapid and actionable insights from



the FEVS. While not generating a new data infrastructure, EVS ART relied on the creative use of Microsoft Excel to extract and transform data to produce dashboards automatically from a single data file. Effectively, the Excel worksheet developed by the NIDDK team integrated a data infrastructure and a dashboard into a single platform.<sup>6</sup>

The US case illustrates how a grassroots initiative, undertaken within the public service rather than as a centralized effort, faced some of the key issues outlined above in its HRMIS reform process. While limited in scope, the implementation team found creative solutions to derive strategic value from the FEVS. It adapted the solution to its needs and was able to effectively improve how survey evidence could be operationalized into improvements to employee engagement.

### **What Modules Were Targeted for HRMIS Reform?**

All the cases we have presented directly relate to HRMIS and can be mapped directly onto the HRMIS core modules in table 9.1. In the case of Luxembourg, a new HR analytics module was developed, with dashboards and reports, to enable HR management by the CGPO. In Brazil, a machine-learning algorithm was deployed to ensure that payments followed established regulations, in a clear example of the compensation management module. Finally, EVS ART is an example of an employee engagement module: a federal engagement survey to guide strategic planning in the NHS.

Note that in the cases of Luxembourg and Brazil, although the end product targeted a single module, the solutions required the deployment of multiple modules. For instance, in Brazil, both the compensation module as well as the employee and organizational modules were combined to provide data for the machine-learning algorithm. In the case of Luxembourg, the analytics dashboards were supplied with data from various core modules in the HRMIS, such as compensation, organizational management, and performance management. For the United States, since EVS ART was based on a single employee engagement survey (the FEVS), no additional HRMIS modules were integrated.

### **Preparation: Laying the Groundwork for the Intervention**

This section outlines general principles involved in the two initial stages of the development of data analytics: preparation, where practitioners lay down the groundwork for the intervention, and implementation, where a decision-making process that is collaborative and adaptable plays a crucial role. We highlight what is generalizable and particular about each phase for the specific cases analyzed in this chapter. The accounts are not designed to be exhaustive: rather, they illustrate key concepts and sequential logics that may apply to the practitioner.

#### ***Institutional Coordination***

A key factor in the preparation phase is obtaining the necessary support from senior leadership. This support is what confers on the reform team the authority to make executive decisions and secure collaboration for the intervention. In general, a centralized authority with a mandate over a policy area makes reform easier. In Luxembourg, the implementation team was commissioned by the CGPO. The CGPO enjoyed a broad mandate that focused specifically on HR, from the management of the life cycle of personnel to strategic workforce planning. This broad mandate meant that once the decision to develop a new dashboard was made, no additional permissions were necessary.

In the United States and Brazil, leadership support was granted by senior management within the respective agencies. In the United States, the implementation team was based in the NIDDK, situated within the NIH. The NIDDK's senior leadership understood the importance of the effort and supported the team's effort—granting time, flexibility, and necessary resources. In Brazil, the senior leadership of the Department of Personnel Management and Performance (SGP), which oversees DEREb, gave full support to the project.

## **Respecting the Legal Framework**

The development of innovative technologies, such as data analytics, requires careful consideration of the existing legal framework, particularly in the public sector. It is necessary to assess whether there are rules and regulations in place that may limit the scope of the intervention and to ensure that the proper permissions are obtained. Depending on the mandate of the agency, as well as the regulatory environment, different legal permissions may be necessary. For instance, in Luxembourg, due to the CGPO's broad legal mandate to generate analytical insights on HR, it was not necessary to request additional permissions to implement the analytics pipeline. In the US, likewise, due to the limited scope of the intervention, no extensive legal framework was needed.

In Brazil, however, where regulations and norms govern how projects are implemented, extensive legal consultations were necessary. The agency partnered with the consulting firm, as well as with another agency familiar with technological innovation projects, to draft the project proposal and obtain the necessary permissions and legal documents. These cases highlight how interventions operate within the boundaries of existing legal frameworks and need to abide by laws and regulations to ensure their legality and feasibility.

## **Choice of Technology**

As outlined above, COTS solutions strengthen sustainability in the long run because they offer the technical assistance of a dedicated enterprise and tightly integrated tools. On the other hand, COTS solutions often lack the precision of custom-developed solutions, which are tailored to the specific needs of clients. COTS solutions may also cost more due to the high cost of licenses and upkeep. Custom-developed solutions, while more adaptable and flexible, require costly investment in a team of skilled developers to create as well as a long period of iterative maturation. Additionally, upkeep may be expensive if proper code documentation and dedicated maintenance staff are not set in place.

Our cases illustrate these trade-offs. Luxembourg opted for a COTS solution—in particular, a dashboard tool that had already been deployed by the implementation team in another, non-HRM context. The team opted to repurpose that tool for their needs, capitalizing on accumulated experience from a previous project, with a relatively short maturation period. The United States also opted for a COTS solution, Microsoft Excel, which was heavily customized for the requirements of EVS ART. The tool allowed the team to generate indicators and dashboards through the development of scripts that automatically converted data input from the FEVS into dashboard outputs.

Brazil opted for custom-developed, open-source software, developing its solution using Python and open-source machine-learning packages. The solution was deployed in a computing cluster on the cloud, where both a data pipeline and a fraud detection statistical model were hosted. The solution was tailored to the specific requirements of the auditing team, capturing both business process regulations and anomaly detection algorithms with the available HR and payroll data. Due to the technical nature of the project, its implementation was outsourced to a consulting firm.

## **Scope and Deployment Models**

There are clear trade-offs embedded in the choice of the scope of a project. Narrow scopes allow for quicker implementation and greater ease of use. However, they make it more difficult to scale across agencies due to their highly specialized nature. Broad solutions require intensive training and adaptation by users, as well as additional resources for the building of complex tools.

Luxembourg's CGPO opted for a broad scope, commensurate with its broad HRM mandate. The dashboard ecosystem was expansive and provided a wide array of insights, ultimately producing over 157 applications (HR dashboards) and over 2,600 sheets. This complexity required extensive data-quality assurance processes, as well as the training of HR specialists to learn how to use these different tools. A dedicated helpline provided additional assistance.

In contrast, Brazil and the United States had a narrower scope for their solutions. Brazil's solution focused specifically on fraud detection in the federal payroll for the subset of manually imputed payments only. This tailored approach was limited in use to a specific agency and was not amenable to scaling. The NIDDK in the United States focused exclusively on generating insights from the FEVS to guide the agency's decision. The focus was on employee engagement and methods to improve the agency's responsiveness. Due to the broad coverage of the survey itself, however, other agencies expressed interest in deploying the dashboard, proving that it was, in fact, generalizable.

## Implementation: An Adaptive Journey

### User-Centric Implementation

In user-centric implementation, the data infrastructure and solution requirements are defined by how users will use information. Data analytics pipelines are designed to answer user queries and provide answers to a well-defined set of problems, which then inform the required data infrastructure to provide these input data.

For Luxembourg, the mandate for the solution was broad, and the user base varied. The final design of the dashboard attended to multiple user bases, from citizens to HR specialists within the government. Mapping out each user to their use case and ensuring that the dashboards could attend to those needs separately but simultaneously was a key design choice by the implementation team. Multiple data pipelines and dashboards were designed, each for particular areas and users, and within each of these dashboards, multiple data visualizations were available. Figure 9.6 outlines the multiple modules contained in the dashboard solution, including information on pensions and recruitment processes.

For Brazil, extensive consultation occurred among frontline providers (auditors) who were going to use the solution. Feedback regarding the necessary data structure and how it would feed into their auditing decisions was crucial. The team opted for a simple risk score associated with each payment, along with flag indicators for the type of rule violated. In the United States, the users were primarily the management making strategic planning decisions for the agency. As such, the indicators were actionable, such as worker

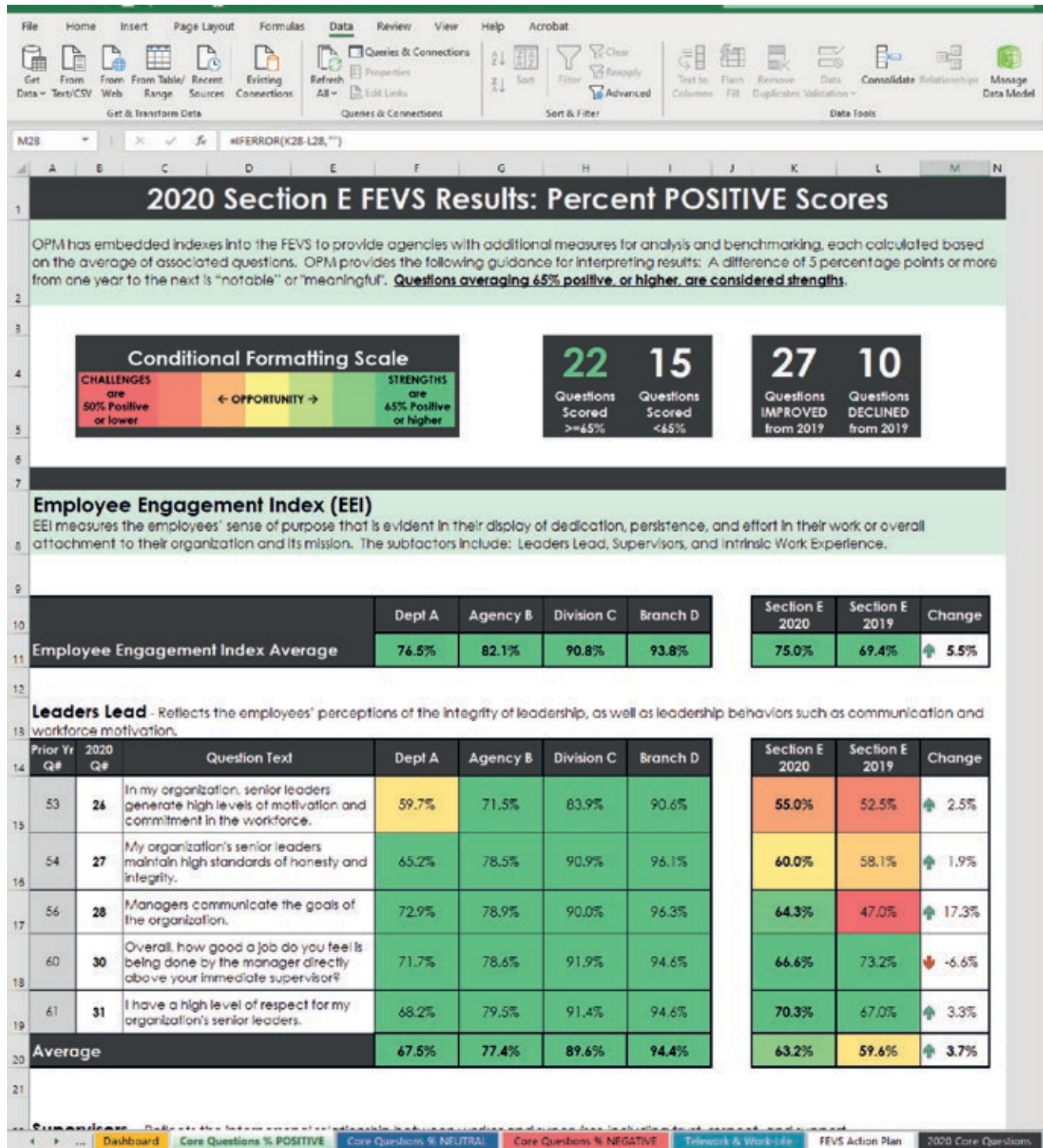
**FIGURE 9.6** Luxembourg's Dashboard Modules



Source: State Centre for Human Resources and Organisation Management (CGPO).

engagement and performance metrics. Organization-level indicators, with positive, neutral, and negative scores, provided ready-access insights into the relative performance of the agency compared to the previous year (figure 9.7). EVS ART also provided an action-planning tab to facilitate strategic planning.

**FIGURE 9.7** Percentage of Positive Employee Engagement Scores from the Federal Employee Viewpoint Survey



Source: Screenshot of EVS ART 2020, NIDDK.

Note: EVS ART = Employee Viewpoint Survey Analysis and Results Tool; FEVS = Federal Employee Viewpoint Survey; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases.



## Iterative Testing

The development of each data infrastructure and analytics pipeline was gradual and flexible. All implementation teams demonstrated a willingness to test, adapt, and redeploy their analytics solution at each stage of the implementation process. For instance, the Luxembourg CGPO first developed a set of dashboards on legacy career and payroll data. Once the initial dashboards were complete, the team realized that quality issues compromised the integrity of the data analysis. As a result, additional quality controls were set in place to ensure that the dashboards were generated as expected. User feedback and demands gradually expanded the scope of the dashboards, and the implementation team worked on iteratively expanding the scope of the HR BICC.

In the United States, the NIDDK had its own learning curve. The team had to work through a process of backward induction, starting from their conceptualization of the final product while researching and learning how to accomplish each step along the way. From the visual appearance of graphs to the data pipeline required to feed them, each task was iteratively resolved and incorporated into the final product. In Brazil, the implementation team tested alternative machine-learning algorithms to improve fraud detection. Repeated consultation with the auditor team generated new ideas, such as incorporating business-rules flags.

## Technical Support

Technical support is crucial to help users navigate the complexity of novel data analytics pipelines—and to be able to build on them. These support systems reduce confusion and facilitate the adoption and diffusion of the technology by new users. In Luxembourg, the CGPO created a helpline to assist HR specialists in the use of the newly developed tools, increasing uptake and facilitating the transition from the legacy system to the new one. The NIDDK team in the United States organized training workshops with different teams and agencies to explain how to use EVS ART as a planning tool. The implementation team created an instruction manual that allows users to navigate the dashboard easily, along with clear and accessible explanations for key indicators and metrics.

In contrast, the Brazil payroll team developed its solution through outsourcing and did not provide a robust system to assist users. The consulting firm, while communicating about the development of the tool, did not create formal channels to address questions and bugs. Rather, support was provided in an ad hoc fashion, depending on user feedback, to address bugs in the code or deployment. While the intense coordination between the consulting firm and the agency reduced user confusion, the lack of a dedicated support team, particularly after the completion of the project, raises concerns regarding the future correction of unexpected changes in the data infrastructure.

## LOOKING BEYOND: FEASIBILITY AND SUSTAINABILITY

The cases presented in this chapter illustrate key design choices and their feasibility constraints. Luxembourg's HR BICC has high entry barriers: its implementation relied on an in-house team that had previously deployed a similar business intelligence solution, as well as on hiring a team of in-house data scientists and IT staff to maintain and develop the solution. These investments are costly and rely on institutional changes that may be prohibitively difficult in other contexts. However, these investments facilitate the sustainability of the solution and its continued development.

Brazil's solution was agile and less costly but in many respects brittle. In less than a year, the implementation team was able to produce a machine-learning-based fraud detection system, but technical and political-economic issues raise concerns regarding its sustainability. Reliance on an external development team meant that in-house capacities were not developed. The sustained development and maintenance of the solution are at risk. Additionally, changes in the management of the agency mean that accumulated expertise during the implementation phase can be lost through succession cycles.

In EVS ART, a narrow scope and sustained implementation—over the course of two years—meant that the solution was widely disseminated and consolidated within the NIDDK. Additionally, it was designed proactively for planning and monitoring within the agency, resulting in tight integration with the analytics component. In many respects, the project is replicable with low feasibility constraints, given the ubiquity of Microsoft Excel in the public sector. At the same time, the highly specialized scope of the solution means that it is not easily modularizable and portable to other domains. Excel spreadsheets are, in general, not amenable to scaling. Furthermore, manual data imputation and modification make developing an automated analytics pipeline challenging.

Note, additionally, that monitoring of the solution was an important component in some but not all of these cases. For Luxembourg, the cutover of legacy systems was implemented in tandem with technical support and the training of HR specialists. The deployment of an extensive analytics dashboard gave administrators live feedback and an overview of the new HRMIS. In the United States, EVS ART replaced the manual approach to obtaining insights from the FEVS survey. In Brazil, a new data pipeline was built on top of an existing legacy system but did not seek to replace it.

These concerns are generalizable to non-HRMIS settings. If the implementation team lacks the financial resources and capacity to engage in a large overhaul of the information system, the scope of the project should be limited to a single module or two. At the same time, the team should consider whether a smaller intervention could have potential linkages to other data modules. Additionally, when engaging with external actors, implementation teams should consider how to ensure the sustained development of the solution after implementation concludes. To reiterate, these lessons are not restricted to HRMIS and can be applied in other administrative contexts, such as public procurement and customs data information systems.

## CONCLUSION

This chapter has outlined the practical issues and challenges in developing data infrastructure for improved government analytics. It has focused on MIS targeted at HR data, but the lessons presented in the chapter apply to public sector data systems more generally.

The chapter has presented the key stages through which HRMIS implementation or reform typically occurs, structured around an operational framework for HRMIS reforms. It has grounded this conceptual discussion by illustrating these stages using case studies from Brazil, Luxembourg, and the United States. The discussion is based on the World Bank's experience implementing data systems in government agencies across the world, as well as on the experiences outlined in the case studies.

There are trade-offs involved in each of the design choices presented. Without robust quality assurance processes in place, the validity of analytical insights is fragile. But expansive quality assurance may be prohibitively costly and is not feasible for all contexts. Deciding the optimal, feasible level of data quality for an analytical pipeline is a design choice, which highlights how the pipeline of data analytics is highly adaptable. Some countries opted for COTS solutions, while others opted for more customized approaches. Agile development, outsourced to external companies, may provide quick results, but it raises sustainability concerns.

The case studies presented in this chapter demonstrate the complexity and diversity of HRMIS implementation. While defying a one-size-fits-all approach, the cases illustrate how a set of different tools, when applied by a dedicated implementation team, can carve out the space for a more analytically driven HRMIS and data infrastructure more generally. Developing systems that both store and extract analytical insights from public data requires widely applicable methodologies. While the specific applications of data systems may vary, the methodology outlined in this chapter and illustrated here in practice provides a conceptual framework with which to approach this challenge. More detailed expositions of the chosen case studies now follow for those readers who want to better understand the individual HRMIS solutions described in summary here.



Beyond the examples presented in this chapter, we highlight the innovative uses of an HRMIS beyond payroll and HR. The underlying theme for this innovation is the use of disruptive technologies like data lakes and artificial intelligence (AI) to cross-reference HRMIS data with multiple other data sources in order to accomplish a policy objective. For example, HR data can be used to analyze procurement and economic activity data in order to identify corruption. In Brazil, HR data on civil servants were cross-referenced with public procurement contracts through the use of big data and AI. The AI tool identified more than 500 firms owned by public servants working at the same government agency that executed a public contract.<sup>7</sup> HR data can also be used to cross-reference budget data in order to improve performance by identifying which civil servants lead particular budgetary programs.

In sum, HRMIS—and MIS more generally—can play a crucial role in the innovative use of data to further policy objectives such as reducing corruption and improving the overall performance of the public sector. The conceptual framework presented here extends beyond HRMIS: the identification of data infrastructure modules and an operational framework for reforms can be applied in a variety of policy settings, as highlighted in other chapters of this book. Ultimately, extracting value from data—transforming them into public intent data—means anchoring them to clearly articulated policy objectives. Articulating what these policy objectives are, and what data are required to measure the achievement of these goals, is the first step toward creating data infrastructures for government analytics.

## **CASE STUDY 9.1 HRMIS CASE STUDY: HUMAN RESOURCES BUSINESS INTELLIGENCE COMPETENCY CENTER (LUXEMBOURG)**

*Ludwig Balmer, Marc Blau, and Danielle Bossaert*

### **SUMMARY**

In 2017, the State Centre for Human Resources and Organisation Management (CGPO) developed and deployed a human resources business intelligence competency center (HR BICC), which enabled it to build a comprehensive HR data infrastructure and framework to plan and monitor HR in the government of Luxembourg. The solution developed was large in scale, involving multiple data sources and HR specialists. This analytics center, developed over the course of a year, had important transformational consequences for the way HR was conducted.

### **INTRODUCTION**

A seemingly narrow question—how much does the government spend on personnel?—requires integrating human resources (HR) data from multiple modules. Which employees (position), types of payment (payroll), and government agencies (organization module) should be included in the wage bill? Policy makers require immediate answers to these questions to make informed personnel decisions. However, a human

---

Ludwig Balmer is the head of information technology for the Centre for Human Resources and Organisation Management (CGPO). Marc Blau is the director of the CGPO. Danielle Bossaert is the head of the Observatory of the Civil Service (Ministry of the Civil Service).

resources management information system (HRMIS) often reacts to ad hoc queries rather than proactively offering a system of answers. This project sought to change that.

In Luxembourg, the State Centre for Human Resources and Organisation Management (CGPO) developed a human resources business intelligence competency center (HR BICC) to provide an integrated overview of HRMIS data in accessible dashboards. This case study shows how this complex technology was developed. The comprehensive scope of the project meant that it integrated a variety of modules, from payroll to talent management. This contrasts with the more tailored approaches of Brazil (case study 9.2) and the United States (case study 9.3). It also provides the clearest example of what chapter 9 describes as an analytics module, the use of HRMIS for strategic and operational decision-making.

A few key lessons emerge from this project. First, quality assurance is paramount to the integrity of the analytics module. The team iteratively cleaned the data and established control protocols to protect its integrity. Second, it is important to reduce the burden of visualization on users. Ensuring visual coherence across dashboards and providing different choices of visualization reduces confusion and increases accessibility. Finally, it is important to provide users with additional support outside the dashboard itself. A helpline can guide users in the proper use of the dashboard as well as generate feedback on whether it is functioning as intended.

This case study is structured as follows. Section 1 provides institutional context on the HRMIS and its management. Section 2 describes the initial challenge and gives an overview of the solution itself. Section 3 explains the project's rollout strategy and reform sequence. Section 4 outlines the lessons learned in the project. Section 5 outlines the impact of the solution. Finally, we conclude.

## INSTITUTIONAL CONTEXT

The CGPO is a central government administration in Luxembourg, located in the Ministry of the Civil Service. Its mandate spans multiple responsibilities, including:

- Management of the entire life cycle of personnel, including candidate selection, onboarding, and professional development
- Calculation and management of remuneration and the careers of active state officials
- Management of retired state officials and pension beneficiaries
- Strategic workforce planning management, as well as HR data dashboard publication.

Alongside these responsibilities, the CGPO also provides consulting services. These include business process management and optimization, organizational development, digitalization, and project management. To manage HR data, the CGPO uses an integrated HRMIS, customized to suit its needs. Before the deployment of this solution, the system included information on the careers and salaries of civil servants in Luxembourg. HRMIS data were already centrally managed and stored. Regular and ad hoc extraction routines were executed to provide data insights to CGPO users as well as other public institutions.

## INITIAL CHALLENGE AND PROPOSED SOLUTION

In 2016, the CGPO faced growing demand and daily follow-up needs from internal HR specialists and decision-makers in the government. As the volume of demands increased, the CGPO decided to design and deploy an HR BICC. The purpose of the center was to facilitate a comprehensive overview of HRMIS data through the

development of dashboards. This would reduce the burden on the CGPO to respond reactively to demands and would empower consumers of HRMIS data to formulate questions and search for answers within each dashboard.

User orientation was an important principle in the project and was reflected in the development of interactive dashboards (an example is given in figure 9.8). The dashboard included two components: a more general data analysis perspective and an operational HR perspective including key indicators for HR specialists to track. In contrast to the previous reactive approach, the project generated a set of readily available visualizations to inform policy making by HR specialists and other agencies in Luxembourg's government.

Before the project's implementation, the legacy HRMIS only considered the management of careers and salary computation. The project expanded the set of modules in the HRMIS, including performance and training modules. The simplified diagram presented earlier in figure 9.5 shows the main applications and the workflow of the solution. The HR BICC integrates multiple databases and dashboard applications, each tailored for different use cases, including HR specialists, employees, and citizens.

Note that in figure 9.5, the DataRH portal (in orange) is fed by multiple databases beyond the HRMIS itself. Its data pipeline includes more strategically oriented databases, such as the strategic workforce planning application. This tight integration between databases designed for strategic workforce planning and the HRMIS data promotes a strategic orientation for the HR BICC.

## ROLLOUT STRATEGY AND REFORM SEQUENCE

In mid-2016, the initial decision was made to develop the HR BICC (table 9.2). In October of the same year, the project was formally launched. The first step was procurement and the launch of data warehouse deployment. The CGPO identified a business intelligence (BI) team that would be responsible for the implementation of the dashboard. After completing the selection process, the CGPO opted to hire an in-house team that had developed a similar solution in another, non-HR area within the government. It therefore opted against procuring the BI tool externally, in contrast to the Brazil HRMIS case study.

**FIGURE 9.8** Sample Dashboard from Luxembourg's HR BICC



Source: Screenshot of HR BICC dashboard, CGPO.

Note: CGPO = State Centre for Human Resources and Organisation Management; HR BICC = human resources business intelligence competency center.

**TABLE 9.2 Project Timeline for Luxembourg's HR BICC**

Period	Main steps
Mid-2016	Decision to put an HR BICC in place
October 2016	HR BICC project kickoff <ul style="list-style-type: none"> <li>• Launch of BI tools procurement process (call for proposals)</li> <li>• Launch of data warehouse deployment project</li> </ul>
February 2017	BI tool selection and deployment, start of governance process and documentation
March 2017	Setup of data warehouse architecture
March 2017	Start of dashboard production

Source: Original table for this publication.

Note: BI = business intelligence; HR BICC = human resources business intelligence competency center.

The main consideration was that the solution that had previously been deployed by the BI team would not only fit the CGPO's initial needs but would also be scalable in the future. The skills developed by the in-house team were transferrable: they had already developed data infrastructure and a previous version of the dashboard tool in another area. This procurement strategy allowed the CGPO to capitalize on previous experience and substantially accelerate the deployment of the solution. As a result of this decision, dashboard production was initiated shortly after the BI tool was selected, in March 2017. In the same month, the redesign of the data warehouse architecture for the HRMIS commenced.

The legal framework was an important consideration for the project. The General Data Protection Regulation (GDPR) impacted both the source side of the data export routines as well as user access management. Monitoring technologies were built into the BI tool to address security concerns. Plug-in tools tracked user activity, tracing how apps, sheets, and data were used or visited by users. This allowed the CGPO both to understand how the HR BICC was used and to ensure that user access was carefully monitored.

The implementation team faced several challenges during the rollout of the project. The first was ensuring quality control of HRMIS data. Because the HRMIS was initially built to perform specific operations, such as salary computation and career management, HRMIS data were not always complete or consistent. As a result, in the initial stages of statistical analysis and dashboard preparation, the team identified missing data series and inconsistent results. To overcome this issue, the team designed a data relevance and quality review process while, in parallel, training civil servants on how to respect it. This quality review process is now part of the CGPO's daily routines.

The second main challenge was providing technical support and a help desk for CGPO staff. The dashboard introduced a new way of working for HR internal specialists. Due to the novelty of the dashboard, internal teams had to adapt their business activities and processes to benefit from the new sources of information and ways of interacting with it. The implementation team also had to respond to new requests by users. Their responses ranged from converting legacy worksheets to operational dashboards to improving existing dashboards in response to user needs.

## LESSONS LEARNED

Valuable lessons were learned in the implementation of the project. The implementation team faced data infrastructure constraints as well as pressure to deliver quick results. To address this, the team opted for a pragmatic and flexible approach to exporting data from the HRMIS data warehouse. This meant simplifying extraction to a few data pipelines that would clean and load the HRMIS data to the HR BICC itself.

Another lesson was the importance of data quality and how to establish processes to protect it. The team defined a data glossary to establish a common understanding of expectations regarding data structure and shared this glossary with users. It also established data governance practices and quality checks to ensure the integrity of data fed into the HR BICC. The team implemented automated controls and routines for data entered and managed by HR departments and also conducted regular trainings and communication to increase awareness of data quality concerns.

The team also learned that standards and development guidelines improve user experience and accessibility. It designed uniform layouts, chart types, navigation practices, and colors, while documenting dashboard-development requirements. However, it also learned that end users should not be tasked with developing dashboards. Even with proper documentation, developing a dashboard is a complex task. Although BI tools can convey and promote a self-service approach, end users rarely master dashboard development without proper training. Different users may not follow the guidelines for building dashboards, resulting in heterogeneous dashboards.

A final lesson was that, while limiting the scope for end users, the dashboard development team has to remain flexible and respond to user needs. Responsibilities for the implementation team include developing new dashboards, modifying existing analyses, and generating reports. The team should consult with clients until dashboards meet end users' expectations. Finally, support systems for users are strongly recommended. A helpline proved particularly useful, with a service-desk phone number and an online form to receive and answer user questions and requests.

## IMPACT OF THE SOLUTION

As a result of the project, the HR BICC provides a comprehensive and detailed view of HRMIS data across the government (ministries and administrations/agencies) of Luxembourg. It includes multiple dashboards to visualize HRMIS modules, such as career management and pensions (see figure 9.6). This dashboard ecosystem keeps growing. As of today, the HR BICC maintains over 56 streams containing 157 HR dashboards with over 2,600 sheets.<sup>8</sup> In addition, it hosts 320 active users with more than 20,000 connections per year.

The HR BICC accommodates a variety of use cases. Active users are, on the one hand, internal HR specialists for whom dashboards provide a new tool to monitor and verify HRMIS data. Other users include HR managers and members of HR teams within ministries and agencies. For these users, the dashboards offer a better overview of their own HR, better control over the key dates in their HR processes, and better follow-up on their personnel.

The overall benefits of such an approach are, for all users, a gain in the quality of HRMIS data and a clear and guided HR data journey. This journey ranges from a broad overview of the HRMIS to deep dives into a particular topic, such as compensation. One example of a key daily benefit is the use of aggregated trend data to project new HR initiatives, orientations, decision-making, and negotiation arguments at the ministry level. Additionally, the HR BICC provides users with accurate and fast information, accelerating business processes and decision-making. Because some of the dashboards are shared with decision-makers at the ministry level, it helps build, improve, and adapt laws and regulations. Overall, this also increases the data literacy of government organizations.

## CONCLUSION

This case study has described how Luxembourg's CGPO developed an integrated dashboard system to inform policy making. The project included both the development of a dashboard ecosystem and the necessary data infrastructure to maintain it. The solution has grown considerably since its launch, hosting over

150 applications, each with its own set of dashboards. Together, these applications cover a variety of topics, from career management and pensions to recruitment.

The dashboard ecosystem has had a considerable impact on the way HRMIS data are consumed and analyzed. It provides immediate access to information on HR that allows policy makers to make better-informed decisions. It establishes quality controls and trains civil servants to better use the platform. Dashboards also increase data literacy within ministries and among HR specialists. However, it is important to note that the CGPO relies on an in-house team with experience in developing and deploying dashboards. This means that the project rollout and implementation were both fast and sustained over time. This experience contrasts with other cases, such as Brazil, more common in developing contexts, where solution maintenance and improvement were constrained by dependency on external actors.

The case study highlights the benefits of a systematic approach to HRMIS analytics, supported by a civil service with the capacity to implement and maintain it. Not all governments have access to these human capital resources. As a result, their dashboard ecosystems may require a more limited approach. Yet beyond the technical expertise, a valuable lesson can be learned from CGPO's methodical approach. The CGPO carefully developed a systematic array of protocols and documentation to protect the integrity of HRMIS data and dashboard visualizations. This requires not a group of IT experts but a careful consideration of the bureaucratic protocols necessary to both maintain and grow the solution. This approach could certainly be replicated in government agencies elsewhere.

## CASE STUDY 9.2 HRMIS CASE STUDY: FEDERAL PAYROLL CONTROL AND COMPLIANCE (BRAZIL)

*Luciana Andrade, Galileu Kim, and Matheus Soldi Hardt*

### SUMMARY

In 2019, a public-private partnership between a federal payroll auditing team and a consulting firm resulted in the development of a novel payroll irregularity detection system. The solution included an integrated data pipeline to train a statistical model to detect irregularities as well as automated identification of violations of payroll regulations. The fraud detection system was used to assist payroll auditors in their daily work. This complementary approach enabled auditors to better detect irregular payroll entries, increasing savings and improving efficiency.

### INTRODUCTION

Governments are responsible for the accurate and timely disbursement of payroll to civil servants. As the volume and complexity of payroll increase, manual approaches to quality control are not sustainable. In 2019, the Department of Compensation and Benefits (DEREB), a federal agency in Brazil, was responsible

---

Luciana Andrade is a senior regulatory agent with Anvisa. Galileu Kim is a research analyst in the World Bank's Development Impact Evaluation (DIME) Department. Matheus Soldi Hardt is a partner at EloGroup.



for overseeing over 80 million paychecks annually. To improve the process, DEREb introduced a new technology to support payroll analysts in their quality checks, which combined machine learning and automation. The Federal Payroll Digital Transformation project ultimately increased recovery rates on inconsistent paychecks and is used daily by payroll analysts in Brazil's federal government.

This case study describes how the project improved the workflow for control and compliance in payroll, a foundational module in a human resources management information system (HRMIS). Although the project had a narrow focus compared to the case of Luxembourg (case study 9.1), this limited scope enabled the development of a highly specialized solution to payroll management, analogous to the case of the United States (case study 9.3). This specialization allowed for the relatively quick and low-cost deployment of the solution. However, it also meant that the project was context specific and not necessarily scalable to other modules in the HRMIS.

Here are the key lessons from the case. First, the foundational steps of problem definition and scope were conducted through extensive dialogue with end users. Payroll analysts who would ultimately use the technology were consulted and offered input to the solution itself. Second, an iterative approach reduced risk aversion and secured buy-in from leadership in public administration. Because the payroll system was complex and the analysts themselves did not have complete knowledge of it, the team opted for gradual refinement of the solution. Finally, reliance on external actors allowed for rapid implementation, but due to this external reliance, the solution was not further developed once the intervention was finalized. In-house technical capacity was never built.

The case study is structured as follows. First, we provide institutional context about the federal payroll system. Section 2 outlines the solution. Section 3 highlights the rollout strategy for the solution. Section 4 describes risk aversion in bureaucratic organizations and how iterative disruption overcame it. Section 5 outlines the impact of the solution. Section 6 draws some lessons and cautionary observations about the external implementation of digital solutions. Finally, we conclude.

## INSTITUTIONAL CONTEXT OF THE FEDERAL PAYROLL SYSTEM

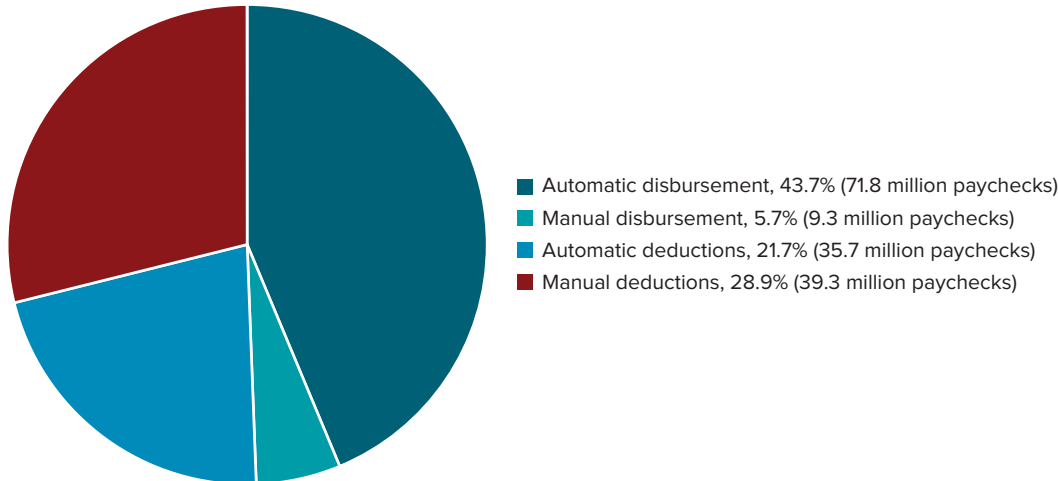
Brazil's federal government disburses over R\$150 billion (US\$30 billion) in the federal payroll every year, accounting for 1.4 percent of the national GDP in 2019. Of the total paychecks issued, over 43 percent are fully automated, meaning that payments are automatically disbursed according to pre-established rules and procedures (figure 9.9). However, 5.7 percent are still manually submitted entries, amounting to 9.3 million manual entries in 2018. While payroll data are centrally stored and managed by the Ministry of Finance, disbursement and deductions are submitted through claims by human resource (HR) departments in different federal agencies.

As noted in chapter 9, one of the foundational modules in an HRMIS is payroll compliance and control. In Brazil's federal government, payroll quality control is the responsibility of DEREb, which is overseen by the Department of Personnel Management and Performance (SGP). While it does not have the mandate to punish infractions, DEREb flags paycheck inconsistencies prior to disbursement, which must be addressed by HR departments in federal agencies.

The task is challenging. The case volume is large, with tens of thousands of individual disbursements transacted daily. Additionally, a complex set of regulations governs how payments should be disbursed. To enforce these rules and detect inconsistencies, a team of payroll analysts individually verify each paycheck. Over the course of a day, analysts check hundreds of entries to verify whether the values are in accordance with the existing rules, whether the amount issued is too high, and whether the public servant that would receive the value has the actual benefit, among other inconsistencies.

Before project implementation in 2019, payroll monitoring was done through a combination of selecting the highest-value paychecks and random sampling. At this stage, DEREb first determined the

**FIGURE 9.9** Brazil's Federal Payroll, 2018



Source: Original figure for this publication.

Note: Payroll excludes the municipal government of Brasília (GDF) and state-owned enterprises.

number of manual entries to be verified based on the productivity of each payroll analyst multiplied by the number of payroll analysts working that day. DEREb would then select payroll entries according to the following rules: 90 percent of the sample was selected from the highest-value entries and the remaining 10 percent was randomly selected. This approach was designed to reduce workload and maximize fund recovery since large entries were overrepresented in the sample.

Although this legacy approach represented an initial attempt to automate the sampling of entries for monitoring, it identified few inconsistencies. In total, only 2 percent of entries were notified for corrections, and of those, 40 percent were corrected. In total, inconsistencies that represented less than R\$10 million per year were corrected, less than 0.1 percent of the total amount disbursed by the federal payroll. Management at DEREb wanted to improve this process and opted for an HRMIS reform project in collaboration with a consulting firm.

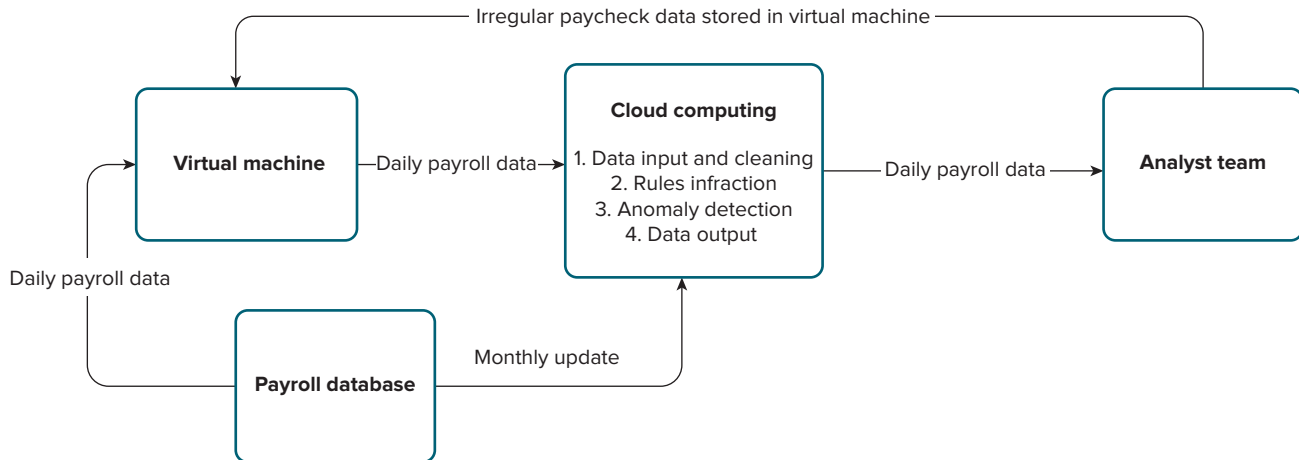
## THE SOLUTION: FEDERAL PAYROLL DIGITAL TRANSFORMATION

The Federal Payroll Digital Transformation project changed the workflow for payroll quality control through the implementation of new technologies. The project was a public-private partnership between DEREb and the consulting firm EloGroup. At its core, the solution generated flags and rankings for federal payroll analysts in their effort to detect and notify agencies of potential inconsistencies in their payrolls. The solution was open source and deployed through cloud technology. The development cycle took approximately eight months to complete.

The solution relies on two complementary approaches: qualitative flagging of regulations governing payroll and quantitative analysis through anomaly-detection statistics. The development of the business-rules module relied on translating regulations governing payroll into automated flags indicating whether an infraction has occurred. The quantitative approach adopts statistical techniques developed by credit card companies to detect anomalies in payments. Payroll values that are far off from a predicted value are assigned a greater risk score and prioritized for payroll analysts.

The solution is executed daily. The first step in the pipeline is the extraction of data on paychecks created in the previous working day, reduced to the subset of manually imputed disbursements (figure 9.10). The data

**FIGURE 9.10** Brazil's Solution Workflow



Source: Original figure for this publication.

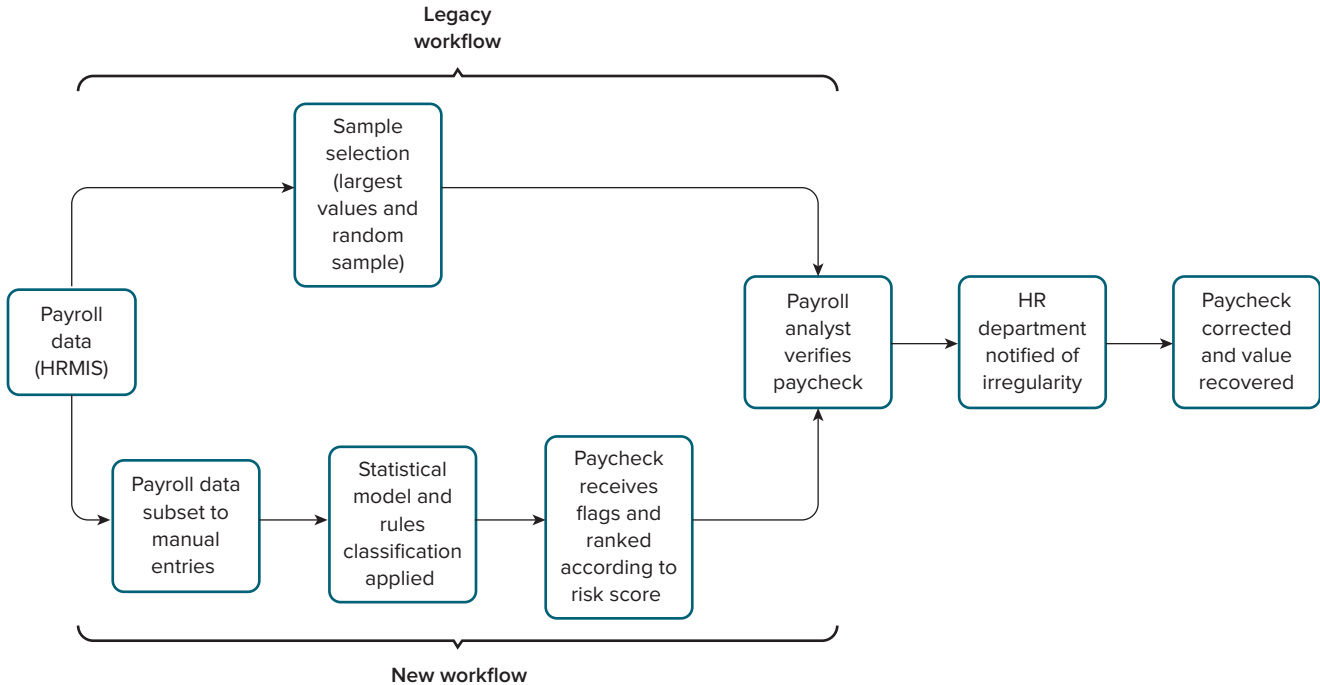
are fed directly from the payroll database into a virtual machine (VM), which receives and stores the daily payroll data. The data are then transferred to a computing cluster in the cloud, where a set of tasks is performed. The data are first cleaned and then go through a rules-infraction module, where they are flagged for potential violations. For example, one rule may be that civil servants are not allowed to claim over 1,000 reais in reimbursement for travel expenses. If the rules-infraction module detects claims that exceed that threshold, it would flag that paycheck and send it directly to the payroll analyst team, indicating that this rule has been violated. If no rule infractions are detected, the paycheck is fed into a machine-learning model that classifies paychecks as anomalous, attributing to them a risk score.

Once the business rules and the statistical model classification are applied, paychecks that are considered most likely to be inconsistent are ranked first and sent to the analyst team. The format in which the data are exported is a simple worksheet, with predetermined labels identifying the risk score and the rule-infraction flags, as well as usual paycheck fields, such as issuing agency and beneficiary. Payroll analysts have discretion over which paychecks to verify and can rank paychecks according to priority, regardless of the classification exercise. It is only at this stage that paychecks are verified and flagged for additional verification by the issuing agencies. Note that the decision to issue a flag remains under the jurisdiction of the analyst.

As a result, the workflow from the analyst's perspective has not changed significantly. The value added is curated information for the analyst, through automated rule-compliance flags and risk scores to facilitate the analyst's decision-making process. Each step in the solution workflow outlined in figure 9.10 is an additional layer of verification, which transparently encodes how the data are cleaned and classified before reaching the analyst's visual dashboard. This choice of design was agreed upon by the monitoring team and the data science team, who opted to make insights from the solution accessible and easy to use. Figure 9.11 compares the new approach with the legacy one.

The machine-learning model and the rules classification do not replace the monitoring team—rather, they enhance its workflow by automating procedures before the data even reach the individual analyst. This complementarity between analog and digital processes is what enabled the new workflow to be well received and adopted by analysts, in contrast to other experiences of technological innovation in which human decisions are eliminated. This hybrid solution provides a more gradual approach toward the goal of digital transformation, accommodating the need for preserving human autonomy while increasing humans' productivity through the use of technology.

**FIGURE 9.11** Comparison between Brazil's Legacy and New Payroll Workflows



Source: Original figure for this publication.

Note: HR = human resources; HRMIS = human resources management information system.

## ROLLOUT STRATEGY AND SEQUENCING

The director of DEREb decided to improve the existing monitoring system by leveraging the use of digital technologies. Given the agency's capacity constraints and lack of familiarity with technological innovation, the director outsourced the implementation and rollout strategy for the solution to an external consulting firm. The initial legal groundwork was crucial. The director of the consulting firm EloGroup leveraged its experience in the development of digital technologies for other government agencies and guided the drafting of the proposal. The General Coordinator for Special Projects of the Secretariat of Public Policies for Employment was familiar with the regulatory process and provided guidance on obtaining legal approval and initial funding for the solution.

The political environment was favorable for the project. Senior leadership was committed to fighting inefficiency and possible cases of corruption, and the federal payroll was under scrutiny due to its large size and perceived inefficiency. The SGP leadership team gave wide discretion to DEREb regarding the HRMIS reform to be enacted. This autonomy allowed the director of DEREb to make difficult decisions regarding personnel, who initially resisted modifying the existing monitoring process. To obtain funding for the project, the team submitted a project proposal to a technology company that provided seed funding for the project.

The monitoring system was developed by a small but agile team of technology consultants at the consulting firm EloGroup. The initial goal was to design a prototype of the workflow outlined in figure 9.10 to detect inconsistencies that would validate the approach. An intensive consultation process preceded the implementation of the technical solution. Workshops and open discussions with federal agencies highlighted what data would be available to develop the prototype, what unique identifiers there were for merging the data, and what kinds of variables would be available to the machine-learning algorithm. An initial workshop covered

problem definition and project scoping, defining how the solution would be embedded into the monitoring tasks performed by the auditors.

Once the project was launched, it faced resistance from staff. Personnel within the monitoring team at DEREb expressed concern regarding the proposed solution because they feared displacement and the disruption of existing procedures. Staff also worried that the digital update would lead to a technological dead end, as had occurred in previous collaborations with external consulting firms. Anecdotally, there was a perception among participating Brazilian public servants that private initiatives introduced off-the-shelf solutions without considering the needs or opinions of public servants who had worked for years in the area.

A collaborative design aimed to assuage these concerns. During the kickoff workshop with multiple federal agencies, staff from different areas within DEREb were able to express their views on the flaws and strengths of the payroll system. On more than one occasion, a public servant in one area identified that his challenge was shared across departments. These open conversations made even the most reluctant employees of the project express interest, or at least not boycott the initiative. In making these concerns transparent and sharing them in an open forum, the team included payroll analysts in the development of the project. Obtaining buy-in within and across departments proved crucial to the success and sustainability of the solution.

Buy-in was necessary not only for personnel but for upper management as well. Due to budget constraints, Brazil's federal bureaucracy had only limited access to cloud resources, for which agencies needed to petition. As a result, after the initial seed funding was spent, it was necessary to secure access to cloud computing through a formal project proposal. To do this, the team presented the results of the initial stage of the solution, highlighting the benefits of the approach and how it could assist the government in saving money. This effort was ultimately successful, securing additional funding to complete the solution.

## RISK AVERSION AND ITERATIVE DISRUPTION

Bureaucratic agencies are risk averse, and with good reason: they perform key roles in government and, while doing so, comply with rules and regulations. A task executed improperly or failure to abide by existing norms can have severe consequences, both for the general functioning of the state apparatus and for the individual careers of civil servants. The solution for this project was not to revamp the regulatory framework or standard operations. Instead, the reform team identified small opportunities to improve the workflow of the analyst team through multiple cycles of disruption.

Coordination was key to this approach. The consulting team was responsible for implementing the solution in terms of software and data engineering. Meanwhile, the payroll analysts and the management team at DEREb provided feedback and prototyped beta versions of the solution. To strengthen this partnership, communication channels between both teams were reinforced. The method deployed for the development of the solution was short but agile.

One of the main challenges in implementing the solution was a mutual lack of knowledge between DEREb and EloGroup regarding the other's area of expertise. For the consulting team, the payroll data and governance structures of Brazil's federal bureaucracy were so complex that most of their initial effort focused on learning how the payroll system operated. To address this, the consulting team had to communicate extensively with the monitoring team at DEREb to ensure that relevant data were extracted and that regulations were incorporated into the automated rules and statistical model.

On the other hand, the monitoring team at DEREb had limited exposure to statistics and software development and therefore needed to be introduced to novel techniques without prior knowledge. Conversations

revolved around how to formalize the substantive knowledge of analysts in software, but ultimately, analysts had to rely on the consulting team to implement the solution. Lack of familiarity with software development and the platform meant that when bugs in the operations were identified, the consulting team had to address them, and workflow was interrupted.

With the initial data pipeline designed, the business rules and the statistical model were put into production. Anomalous paychecks were sent directly to the monitoring team for validation. The initial results were positive, with the algorithm-empowered monitoring consistently outperforming the previous approach, based on the size of paychecks. As additional resources were necessary to expand the project, the director of DEREb presented the results to government leadership as promising evidence that the approach was correct. This initial buy-in proved key: having an actual solution in production and demonstrating results reduced uncertainty in higher levels of management.

The deployed solution combines two key insights: first, it formalizes existing laws and regulations governing payments in an automated pipeline. This means that the analyst no longer has to verify whether a paycheck complies with regulations; the business-rules module does this automatically. Second, the anomaly-detection algorithm relies on statistical modeling to leverage information about public servants, their departments, and their payment histories. This process fully leverages the information methodically collected by the Brazilian government on its payroll and public servants without imposing additional burdens on the analyst team.

Additionally, the current algorithm is designed to reduce workload and help analysts prioritize paychecks with higher risk. This complementary approach to improving payroll analysts' workflow is key: after initial resistance regarding these changes, the monitoring team realized the benefits of the new digital approach over previous approaches. This hybrid model, incorporating both analog and digital processes, can provide a template for public sector technological innovations.

## IMPACT OF THE SOLUTION

The clearest gains from the solution were in efficiency: despite the reduction in personnel, performance increased. Due to staff attrition unrelated to the project, the team of payroll analysts had been reduced in size. Despite this reduction, the reduced analyst team could flag the same amount of resources as inconsistent compared to a larger team, while dedicating less time to each task. This reduction in the cost and maintenance of performance was an important selling point to other departments within the federal bureaucracy, highlighting the gains in efficiency from technological innovation.

An unintended consequence of the project was an increase in data literacy and a change in mindset. Users of the dashboard displayed greater interest in learning how the solution was implemented, with analysts expressing willingness to learn how to code to better understand the data. This growth in data literacy resulted from initial exposure to a set of techniques that had not been available before. Additionally, because of data integration, new linkages were formed between DEREb and other departments in the bureaucracy. Because the solution relied on data generated in other departments, there was a need for communication and transparency to make it work.

Finally, there was a shift in mindset regarding how to monitor payrolls. While previously, analysts had relied on their accumulated experience and intuition, the solution complemented this approach by emphasizing the use of data and regulatory infractions. The analytical framework of the solution provided a new template that analysts could use to assess whether a paycheck was indeed inconsistent. In a sense, the new technology changed the way payroll analysts approached their task.



## SUSTAINABILITY OF EXTERNAL IMPLEMENTATION

External solutions are brittle. They introduce dependency on the technical know-how of external actors, and once the engagement is finalized, the beneficiary is no longer able to maintain or improve on the external solution. In this case, technical know-how—including software and data engineering—for the implementation of the project remained with the consulting team once it left. The analyst team at DEREb did not acquire the necessary skills or capacity to develop the solution further, even though it was open source. Although data literacy in the monitoring team increased, the analyst team was not formally trained to modify or further develop the software.

Additionally, changes in the management structure of DEREb after the implementation of the technical solution put the sustainability and continued development of the project at risk. While the previous director locked in the current version of the solution, it has not evolved since. Turnover in management and a contract-based approach meant that desirable additions to the solution—such as the extension of automation to all HR departments across federal agencies—were never implemented. The loss of institutional leadership and the lack of in-house capacity meant that while the product survived, it did not continue evolving.

## CONCLUSION

Technological innovation is disruptive, but the costs and uncertainty associated with it can be reduced by adopting a gradual approach. Risk aversion—an important feature of bureaucracies—can be overcome through communication and small modifications to existing workflows. The Federal Payroll Digital Transformation project outlined in this case study showcases this approach. Instead of a complete transformation of the payroll monitoring process, the technology focused on complementing existing workflows by payroll analysts.

A collaborative approach helped build trust in the relevance of the solution and its applicability to daily operations by end users. Iterative cycles of feedback and adaptation ensured that the algorithm proposed was appropriate to the use case and understood by payroll analysts. In addition, this reduced resistance to the final adoption of the solution. Technological disruption can thus be managed and incorporated into existing procedures, giving rise to hybrid solutions that provide a stepping stone for more extensive and intensive solutions.

While the current version of the solution has been finalized, its future development is uncertain. Due to the project's outsourcing, the necessary expertise to implement and develop the solution was not developed in-house. Technological innovation through a public-private partnership therefore comes with associated costs and benefits. There is a trade-off between the agility and rapid gains from outsourcing to external agents and the lack of development of in-house expertise to continue growing solutions. External solutions therefore generate dependency on external actors for developing solutions, lowering the likelihood of maintenance and expansion in the long run.

Finally, the implementation team has emphasized the need for spaces within public administration to incubate technological innovation. These spaces would allow for calculated risks—and mistakes—within the public sector. While the team identified and opened spaces within which the solution could grow, it is important to ensure that those spaces are already set in place. This would incentivize not only managers willing to lead innovations but also staff members, who would prove more willing to engage in changes without fear of reprisal. It would also create incentives for agencies to develop the in-house capacity for technological innovation and reduce dependence on external actors.

## CASE STUDY 9.3 HRMIS CASE STUDY: EMPLOYEE VIEWPOINT SURVEY ANALYSIS AND RESULTS TOOL (UNITED STATES)

*Camille Hoover and Robin Klevins*

### SUMMARY

In 2015, the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), within the National Institutes of Health (NIH), developed the Employee Viewpoint Survey Analysis and Results Tool (EVS ART) to extract insights from the Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS). The solution relied on the creative use of worksheet software to extract and transform data to produce dashboards automatically from a single data file. Effectively, the worksheet developed by the NIDDK team integrated a data infrastructure and a dashboard into a single platform, reducing implementation costs. The tool provides valuable information for senior leadership to promote employee engagement and guide policy making.

### INTRODUCTION

It is a leader's responsibility to care for the people in an organization and to create and sustain a culture where employees can flourish—one in which performance is central and employee engagement is maintained. To be successful, these values must be integrated into the function and mission of the organization, not treated as distinct or separate. To create this type of culture, leadership must secure buy-in from staff at all levels. Staff must embrace the organization's vision and emulate its core values.

It is important that the core values not just be lofty or aspirational goals but translate into action on the frontlines, where the people of the organization are doing the work. Values can and should be measured through employee engagement surveys. This measurement allows leaders to keep a finger on the organization's pulse. It is important to combine data analytics with the voices of employees to inform strategies and resource allocation and to verify whether actions are paying off. Employee feedback must inform and orient action, whether in the form of focus groups, town halls, stay or exit interviews, or crowdsourcing.

This case study describes how the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) developed an analytics dashboard to measure and promote employee engagement. The project was named the Employee Viewpoint Survey Analysis and Results Tool (EVS ART). EVS ART provided NIDDK leadership with immediate and informative data analytics on their employees' perceptions of whether, and to what extent, conditions characterizing a successful organization were present in their agencies. Using EVS ART, the NIDDK was able to transform the enormous amount of data provided by the United States Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS) into a user-friendly format in mere minutes. The survey topics, in response to which employees candidly shared their perceptions about their work experience, organization, and leaders, covered employee engagement, employee satisfaction,

---

Camille Hoover is an executive officer at the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Robin Klevins is a senior management analyst at the NIDDK.

and several submeasures, including policies and practices, rewards and recognition, opportunities for professional development, and diversity and inclusion—all of which were used to inform decision-making.

EVS ART is an example of human resources management information system (HRMIS) analytics, similar in purpose to the case study of Luxembourg (case study 9.1). However, in contrast to Luxembourg, which generated analytical insights on the entirety of its HRMIS, this case focuses on the employee engagement module within an HRMIS. This module is diverse and expansive as a result of the rich data provided by the FEVS. The FEVS measures employees' perceptions of whether, and to what extent, conditions characteristic of successful organizations are present in their agencies. It is a survey in which employees can candidly share their perceptions about their work experiences, organizations, and leaders. EVS ART therefore includes indicators on employee satisfaction, global satisfaction, compensation, and organization, as well as more customized questions about remote work and COVID-19. The focus on improving a particular module of an HRMIS makes this case similar to the approach in Brazil (case study 9.2), which reformed how the payroll module operated.

The project provided a set of lessons that may be helpful for practitioners. First, keep the solution simple. While the inner workings of a tool itself may venture over to the complex side, do not make the act of engaging with the analysis complex for the user. Second, make the solution accessible to all types of users. This means two things. One, ensure that the tool is accessible to those with disabilities, and two, make the tool available to the broadest audience possible. If people do not know about the tool, they will continue to spend unnecessary time recreating analyses and will not obtain insights from the data. Finally, remember that transparency ensures that data analytics can be trusted by those it benefits. When working with data, leadership should not shy away from difficult conversations, because survey takers already know whether something is working well or not. It is incumbent on leadership to be honest, dig deeper, and let staff know that their input will drive organizational change.

This case study is structured as follows. We first describe the institutional context, with particular attention to the FEVS, the largest civil servant engagement survey in the United States. Section 2 explains the initial rollout of the solution. Section 3 provides a detailed overview of the solution. Section 4 outlines the lessons learned during the implementation of the project. Section 5 describes the impact of the solution. Section 6 reflects critically on the importance of looking beyond analytics and effectively promoting change. Section 7 reviews challenges faced and future improvements to EVS ART. Finally, we conclude.

## INSTITUTIONAL CONTEXT: THE FEVS

Each year, the OPM administers the FEVS to over 1.4 million full- and part-time permanent, nonseasonal employees governmentwide.<sup>9</sup> The FEVS measures employee engagement, including employees' perceptions of whether, and to what extent, conditions characterizing successful organizations are present in their agencies. It therefore provides valuable insight into agencies' strengths and opportunities for improvement. In 2020, 44.3 percent (624,800) of those receiving the FEVS completed it—each spending, on average, 25 minutes to do so (OPM 2021). This translates to over 260,000 federal hours and equates to over US\$10 million worth of staff time taking the survey.<sup>10</sup>

The FEVS provides valuable information because the OPM proactively designed the FEVS to include multiple index measures and key categories, such as employee engagement and satisfaction, to help agencies identify important patterns and themes.<sup>11</sup> Each index is valuable, aggregating multiple answers.<sup>12</sup> While much can be learned from the index measures and key categories, on average, there is a three- to four-month period during which the OPM processes the raw data before distributing it to agencies.

The FEVS allows agencies to obtain valuable feedback from all levels of the organization. Subgroups within an agency that have 10 or more survey participants can receive their own area-specific results, and those with fewer than 10 participants roll up to the next level of report to ensure at least 10 responses. This protects the confidentiality of the survey respondent, which is crucial when the goal is to obtain honest

feedback (NIH 2018). In 2018, over 28,000 organizations within the federal government had 10 or more survey participants, for a total of over 280,000 survey respondents—and the number continues to grow (Kamensky 2019).

The FEVS's granular and large-scale data allow organizational leaders within the federal government to tap into the perspective of those on the frontlines and learn from the voices of employees. In turn, the same information can be used to design employee-informed programs and initiatives. It is important for staff to be made aware of changes informed by their feedback. Informed change is noticed, creates ownership, and leads to increased engagement—and engagement is the foundation on which successful missions are built.

Despite this valuable information, extracting insights from the FEVS and putting them into action is challenging. Once given access to the survey, government agencies spend weeks culling large amounts of data to operationalize the survey's feedback. This effort is extremely labor intensive, time-consuming, and costly. Some agencies spend thousands of dollars on manpower or on procuring outside support to analyze the data. In addition, by the time the results are received and the analysis completed, agencies are often on the heels of the next survey—with little time to act on the feedback provided. It is difficult to launch meaningful initiatives with old data, and the lack of timely action, or perceived inaction, often leaves employees wondering whether taking the survey is of value.

## INITIAL ROLLOUT

A small team at the NIDDK, within the National Institutes of Health (NIH), took it upon themselves to work with the data and create a framework to deliver results quickly, accurately, and intuitively. The NIDDK's senior leaders appreciated the importance of these data and made it the highest priority to construct a way to translate them. They fully supported the NIDDK team's efforts—giving them time, flexibility, and necessary resources.

The NIDDK team set out to design a tool that brought to life the voice of the people, one that was unlike other tools. As analysts, they wanted to ensure that users could arrive at actionable data quickly. However, they approached it differently from a traditional report. It was important that the tool was easy to look at, that the flow of information made sense, and that it told a story. They also wanted to ensure that actionable target areas—and themes—jumped out at the user. It was of great importance that the tool be both easy to use and accessible to all federal employees.

The team worked for two years to create a tool that would enable leaders to drill down and compare data, have a better pulse on engagement levels, and view FEVS scores in an actionable and targeted way. They began by utilizing a resource that they already had at their fingertips, a common program used across the federal government: Microsoft Excel. The team worked to design an easy-to-use template that provided a report with an easy-to-understand flow, and they ensured that the templates were password protected so that links could not be broken and results would not be compromised. The team also worked to ensure that the tools and associated resources followed the guidelines of Section 508 of the Rehabilitation Act.<sup>13</sup>

## OVERVIEW OF THE SOLUTION

The team created the EVS ART—an Excel-based tool that allows users simply to copy data provided by the OPM and paste them into a similarly formatted template. Upon clicking “Refresh,” users can review conditionally formatted results, thoroughly compare prior years' data, and conduct a deeper-dive analysis of their outcomes.

EVS ART is different from other tools available to analyze FEVS data because users can arrive at actionable data quickly: the tool and output are easy to look at, the flow is intuitive, and the tool tells a story in a way that allows actionable target areas—and themes—to jump out. It is designed to be easy to use: it requires only basic Excel knowledge, it generates a user-friendly dashboard, and it captures and displays all OPM index measures and key categories.

The tool's utility lies in its simplicity of use but power in transforming massive amounts of information, allowing leaders to home in on important themes and compare prior years' data. EVS ART was designed so this can all be done in a few steps and as little as five minutes. EVS ART pulls data points from each of the main themes in the FEVS, such as employee engagement and global satisfaction. The tool organizes the survey results based on those themes by agency, subcomponent, and office, and it shows the change in responses for a specific item from year to year. This allows NIDDK senior leaders to monitor progress and evaluate the impact of strategies and interventions.

### Instructions Tab

The first tab in EVS ART is the instructions tab (figure 9.12). Users enter the organization acronyms for the areas they wish to analyze and the year(s) of the results they wish to use. This information will automatically populate the headers and table titles on tabs throughout the Excel workbook.

Using FEVS data provided by the OPM, users copy and paste the information from their original FEVS data report into the corresponding EVS ART tab. No reformatting is required. This is done for each organization being compared. If prior year data are available, this step is repeated by pasting the data into the appropriate prior year tab(s). When this is completed, the user refreshes the data and EVS ART automatically populates the dashboard itself.

### Dashboard Design

Upon feeding the data to EVS ART, users gain access to a dashboard that provides an overarching view of the organization's results. The dashboard delivers top-scoring questions for "positive," "neutral," and "negative" results, as well as the largest positive and negative shifts from one year to the next (figure 9.13). Below the charts, users are provided with a heat map that shows the average scores for each of the index measures and key categories, as well as their subcategories. This is helpful because it provides a clear visual at a high level and allows users to easily compare one organization to another.

The dashboard also provides a side-by-side visual comparison of FEVS results (figure 9.14). This helps users to determine areas of focus across the organization and identify areas that need more targeted intervention. The conditionally formatted heat-map feature uses color to show managers their highest and lowest scores and identifies areas that might be strengths or challenges for the agency or a specific office. While the dashboard shows where to start looking, the information behind it—in the remainder of the report—provides a path that intuitively narrows the broader topics down to specific focus areas.


### Analysis Tabs

While the dashboard is a great place to start, the deeper-dive portion of the report takes the user from a general overview to more specific focus areas, where the organization's scores begin to tell a story. Figure 9.15 shows an example of an organization's percent-positive employee engagement index scores. At the top of the tab is the OPM's guidance for interpreting the results. In the case of the FEVS,

- Questions averaging 65 percent positive or higher are considered "strengths,"
- Questions averaging 50 percent neutral or higher may indicate "opportunities" for improved communication, and
- Questions averaging lower than 50 percent are considered "challenges."



FIGURE 9.12 Instructions Tab in the EVS ART



## EVS... at the heART of a healthy organization!

For use with OPM 2020 "All Items All Levels" Report

The Employee Viewpoint Survey Analysis & Results Tool (EVS ART) allows for quick, easy, and accurate analysis of Federal Employee Viewpoint Survey (FEVS) results! Something that once took hours can now be done in minutes - and in **5 simple steps** - providing valuable and actionable data for decision-making in a timely manner.

**IMPORTANT:** This template accommodates the revised 2020 FEVS and can be used to analyze survey's Core, Telework & Work Life questions. It is designed for use with data provided in the OPM "All Levels" report format, allows for the comparison of up to 5 organizations, and offers users the ability to conduct a year-to-year comparison for the "Primary" Org.

STEP 1: Enter Required Information:	
Organization 1 Acronym (for Governmentwide or other comparison Org)	Dept A
Organization 2 Acronym (for Department or other comparison Org)	Agency B
Organization 3 Acronym (for Agency or other comparison Org)	Division C
Organization 4 Acronym (for Office or other comparison Org)	Branch D
Organization 5 Acronym (for Primary Org being analyzed)	Section E
Current Year (Do not change from 2020)	2020
Prior Year (When comparing prior year data for Org 5/Primary Org)	2019

STEP 2: Enter Participation Rates and/or Custom Message

Enter Participation Rates and/or Custom Message

STEP 3: Copy & Paste 2020 Data

**ACTION:** Open the OPM 2020 FEVS "All Items All Levels" report for the organization(s) you wish to analyze.

Follow the instruction at the top of the "2020 Core Questions" and "2020 Telework & Work-Life" tabs below to copy and paste information from the OPM 2020 FEVS "All Items All Levels" report to this EVS ART template.

**RESULT:** The Dashboard, Positive, Neutral, Negative, and Telework & Work-Life worksheets will autofill with 2020 survey data.

STEP 4: Copy & Paste Org 5/Primary Org PRIOR Year Data (if available)

**ACTION:** Open the PRIOR year FEVS "All Levels" report that you wish to use to do a year to year comparison for Org 5.

Follow the instruction at the top of the "Org 5 PRIOR Year Data" tab below to copy and paste information from the prior year OPM FEVS "All Items All Levels" reports to this EVS ART template.

**RESULT:** The Dashboard, Positive, Neutral, and Negative worksheets will autofill with Org 5's PRIOR year survey data.

STEP 5: Activate the Dashboard!

**ACTION:** Click the "Data" tab in the toolbar above and select "Refresh All". Then go to the "Dashboard" tab to begin reviewing your 2020 results.

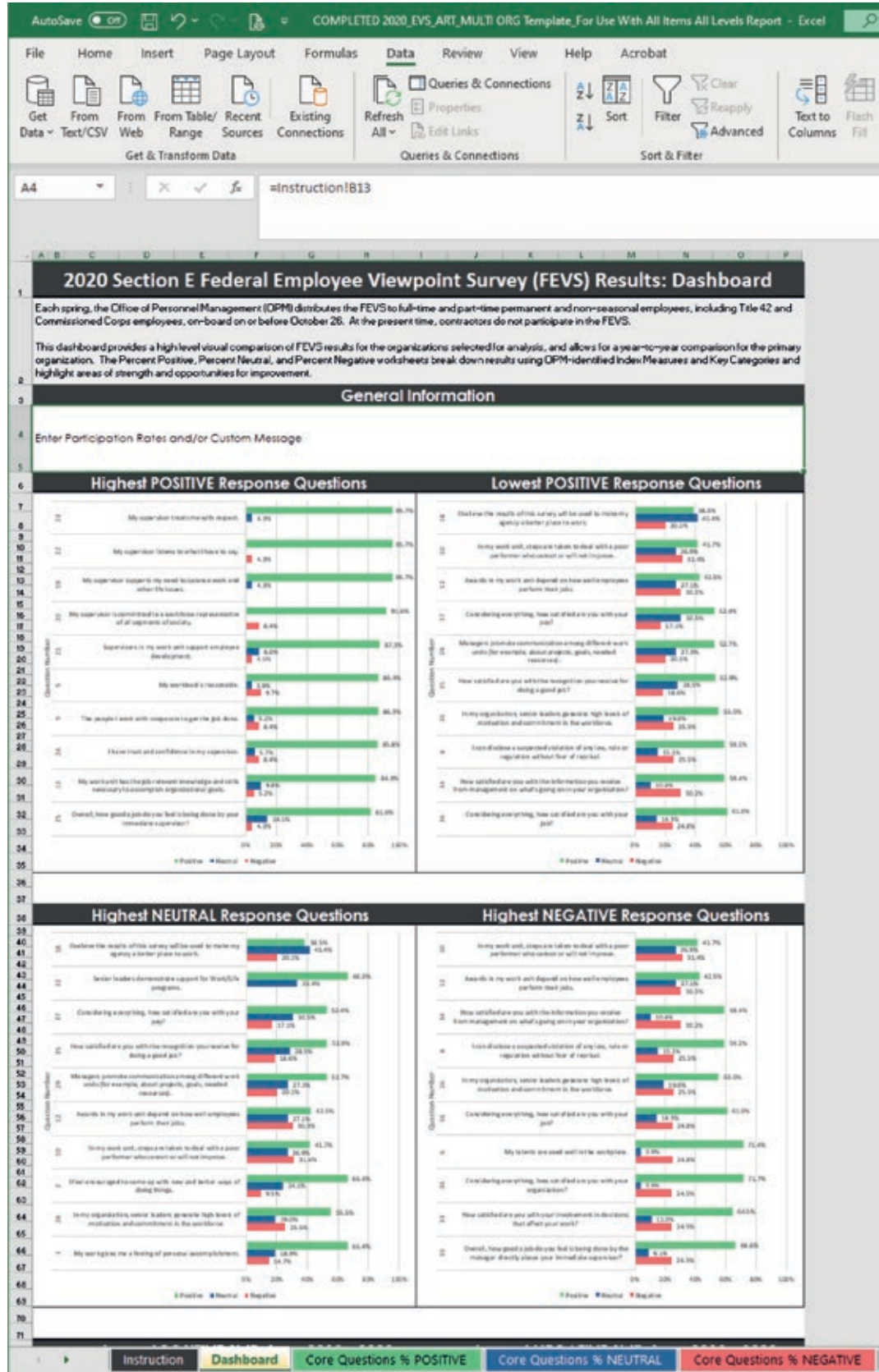
**RESULT:** The Dashboard graphs and heat map will autofill with the 2020 results. When prior year data is present, the largest positive and negative shifts will be shown and the difference between 2020 and prior year will be reflected in the heat map.

Source: Screenshot of EVS ART 2020, NIDDK.

Note: EVS ART = Employee Viewpoint Survey Analysis and Results Tool; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases; OPM = Office of Personnel Management.



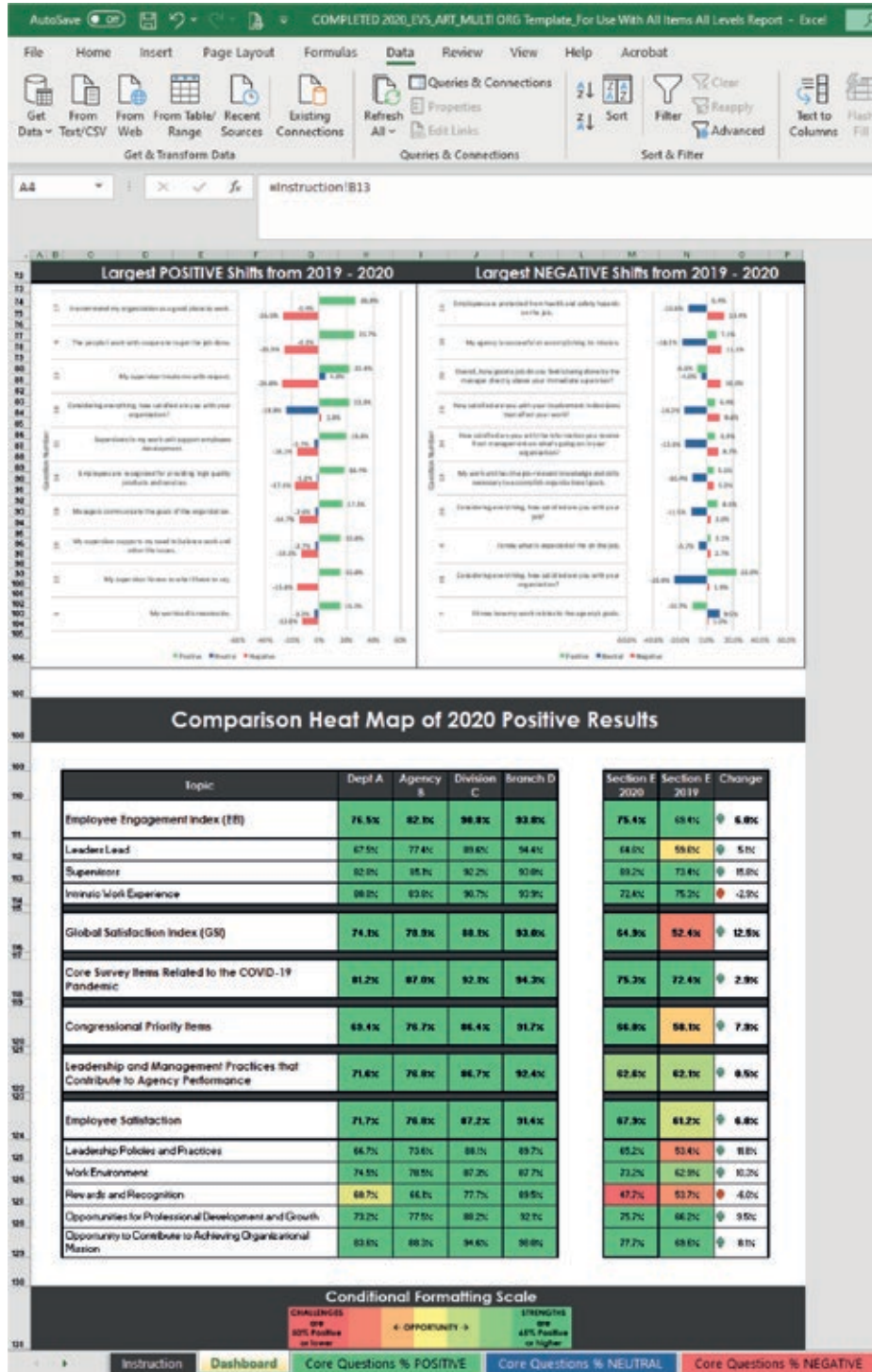
**FIGURE 9.13** Landing Page of the EVS ART Dashboard



Source: Screenshot of EVS ART 2020, NIDDK.

Note: EVS ART = Employee Viewpoint Survey Analysis and Results Tool; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases.

FIGURE 9.14 Results Comparison in the EVS ART Dashboard

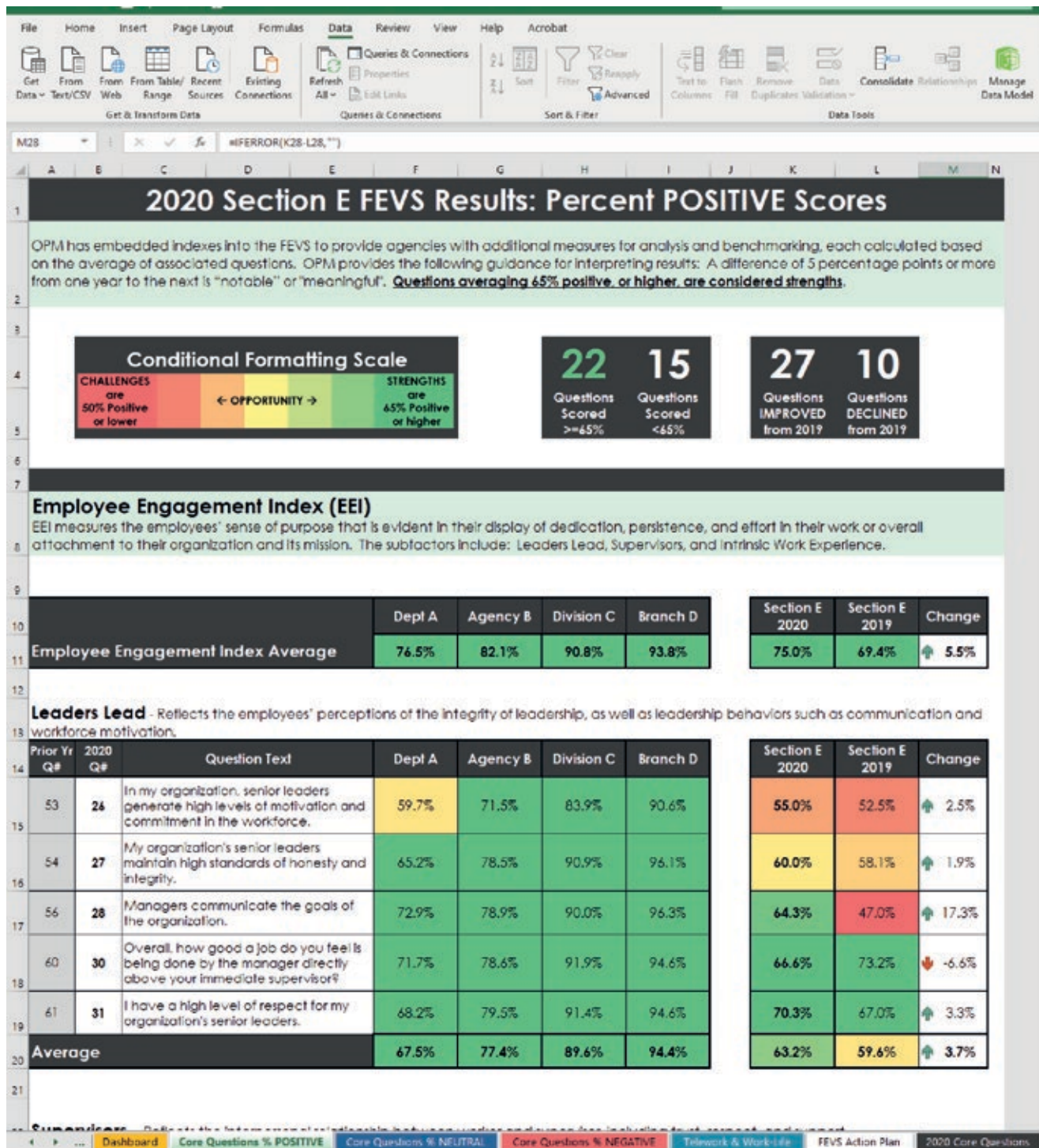


Source: Screenshot of EVS ART 2020, NIDDK.

Note: EVS ART = Employee Viewpoint Survey Analysis and Results Tool; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases.



**FIGURE 9.15** Percentage of Positive Employee Engagement Scores from the Federal Employee Viewpoint Survey

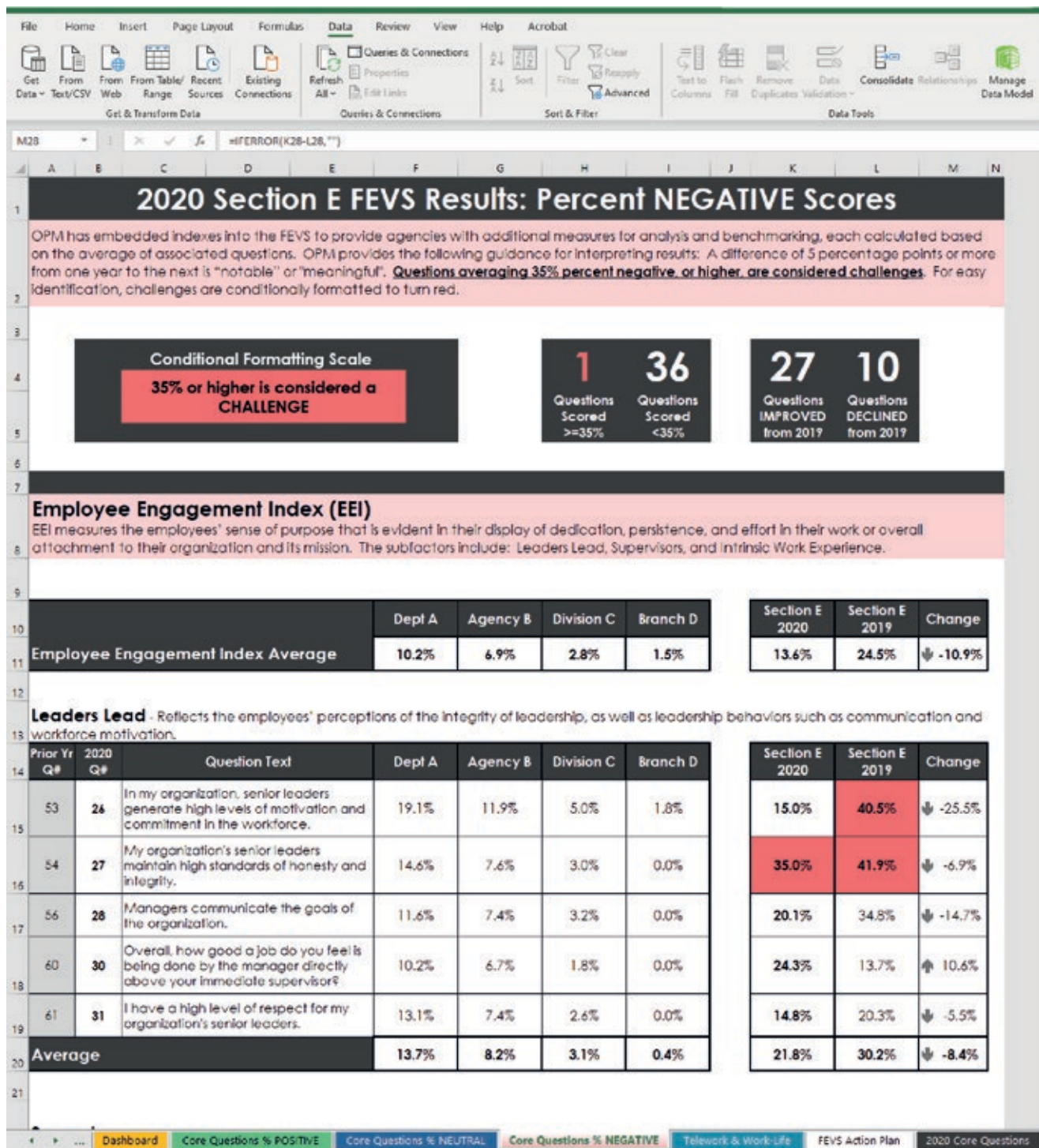


Source: Screenshot of EVS ART 2020, NIDDK.

Note: EVS ART = Employee Viewpoint Survey Analysis and Results Tool; FEVS = Federal Employee Viewpoint Survey; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases; OPM = Office of Personnel Management.

EVS ART is conditionally formatted so that themes are easily identified. Users do not have to know how the tool works to be able to interpret the story or determine where they need to focus, where they have strengths, and where there are opportunities for improvement.

**FIGURE 9.16** Percentage of Negative Employee Engagement Scores from the Federal Employee Viewpoint Survey



Source: Screenshot of EVS ART 2020, NIDDK.

Note: EVS ART = Employee Viewpoint Survey Analysis and Results Tool; FEVS = Federal Employee Viewpoint Survey; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases; OPM = Office of Personnel Management.

It became clear that one should look beyond whether scores are positive or not. Often, federal leaders focus solely on questions that average 65 percent positive or lower. While this is important, going beyond to review both neutral and negative scores can provide clarity (figure 9.16). For instance, there is a big difference between a *low positive score with a high neutral score* and a *low positive score with a*

*high negative score.* While a low positive score is not preferable, if it is paired with a high neutral score, it could indicate an opportunity for communication and clarification, whereas a low positive score paired with a high negative score clearly indicates a problem area.

## Action-Planning Tab

Effective action planning can transform data into meaningful change. The EVS ART action-planning tab is designed to help initiate the process and determine next steps (see figure 9.17). After reviewing the results, users can

- Identify focus areas (these areas can align with OPM index measures and key categories or can be customized to reflect areas of interest),
- Enter related FEVS question numbers (data will automatically populate based on the question number selected),
- Brainstorm initiatives and interventions geared toward improving focus areas, considering both the potential impact and available resources,
- Designate a lead person or office to address each focus area, and
- Assign target completion dates.

## Implementation and Reform Sequence

When initiating the development of the tool, the team first identified the questions that made up each of the FEVS index measures. This was a bigger challenge than anticipated because no one document contained all the information needed, so they created their own. The team scoured the OPM's FEVS technical guides going back to 2012 to identify each measure, its definition, and the associated survey questions. They compiled a master document with this information that is still in use today.

The team also faced their own learning curve. They had a creative vision of what they wanted to accomplish, what they wanted the tool to look like, and what they wanted it to do, but they did not necessarily have the expertise to accomplish it—or so they thought. So the team began to work backward, peeling back the layers of what they anticipated the final product would look like, then researching and teaching themselves how to accomplish each step along the way.

Whether it was the visual appearance and flow or the inner workings of many hidden pivot tables and charts, each task was new, each was important, and each was tackled and then painstakingly built out, tested, adjusted, and then tested again. With each success came a small victory that fueled the next challenge. The analyst team looked for gaps, identified opportunities for improvement, and created efficiencies—and this project provided all of that and more. They knew that what they were creating could feasibly make a difference in the way the FEVS was used and valued governmentwide.

Upon completion, the NIDDK team recognized that the dashboard could be useful in other contexts and decided to share it broadly. Little did they know that getting the word out and giving the tool to other departments and agencies would prove to be more of a challenge than building the tool itself. First and foremost, the creation of EVS ART began as a grassroots effort, far removed from those who managed and administered the FEVS. The NIDDK team began sharing their tool across their agency, but the department had little influence in sharing it broadly.

When the team gained the attention of the US Office of Management and Budget (OMB) and the OPM, all of that changed. The NIDDK team was invited to present to the OMB and the OPM. The OMB was impressed with EVS ART and praised the work done by the NIDDK.<sup>14</sup> The OMB and the OPM organized a venue during which the NIDDK shared the tool with federal chief human capital officers (CHCOs) governmentwide. With the amplification of this extraordinary tool, the team received numerous requests for



**FIGURE 9.17** Sample Action-Planning Tab in the EVS ART

Focus Area	Q#	Question Text	2020 Positive %	2020 Neutral %	2020 Negative %	Initiatives & Interventions	Lead/Office	Target Completion Date
	18	I believe the results of this survey will be used to make my agency a better place to work.	38.5%	41.4%	20.1%			

Source: Screenshot of EVS ART 2020, NIDDK.

Note: EVS ART = Employee Viewpoint Survey Analysis and Results Tool; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases; OPM = Office of Personnel Management.

demonstration of EVS ART from agencies and departments outside of their own. This was a challenge in itself for the team of three because many of the organizations expressing interest in the tool were within the same agency or department but siloed from one another, resulting in multiple requests from each. Additionally, due to turnover in political leadership, there were requests to return to organizations to share with new leaders the capabilities of EVS ART and the progress recognized by the NIDDK when their FEVS results were used to inform change.

The enormity of the US federal government made it more and more difficult to manage such requests. The NIDDK team established an online presence, which allowed federal employees to access the tool and



its training resources. The OPM also offered EVS ART as a resource to agencies and departments as part of work being done under the President's Management Agenda. The collaboration between the NIDDK, the OPM, and the OMB blossomed, and since 2017, the NIDDK team has conducted hundreds of presentations, trainings, and customized workshops that have reached personnel in each of the 15 executive departments. These meetings and training sessions continue because the team has found that organizations at all levels within departments are interested in learning about successful practices.

It was important to the team that other federal employees knew of EVS ART's availability, and equally important that they benefited from it, but it was also important that no outside entity unduly profit from its use. EVS ART was created by federal employees, for federal employees, using resources readily available and with no contractor support. Realizing its benefit and potential, the NIDDK elected to share its tool with any and all federal entities expressing interest, free of charge. Its view as a steward of federal funds was that it had done the work and, by sharing its tool, others could avoid duplicating its efforts and could create extraordinary efficiencies within their own organizations. Many organizations had spent weeks, if not months, and sometimes thousands of federal dollars on outside consultants to do what they could now do for themselves in minutes using EVS ART. The NIDDK has received numerous requests from outside vendors and consultants related to the use of its tool in support of work they are doing for other federal organizations—and even requests for unlocked versions that they can modify for their own use with federal clientele. This goes against the grain of the NIDDK's vision for sharing the tool at no cost. The team does not want federal funds to be used, perhaps unknowingly, to pay for a tool available for free.

### Feedback, Flexibility, and Continuous Improvement

End users of EVS ART have expressed gratitude for the tool.<sup>15</sup> Having this tool helps leadership to see data in one place, or by field office if they like. This tool gives the flexibility to do that, and quickly, economizing time. The way the analysis and reports are organized makes the data clearer, which makes for faster analysis of employee feedback and allows leadership to address the question “what now?” so that agencies can develop a plan of action based on employee responses.

EVS ART was designed to provide users with many ways to view data. It offers a dashboard, heat maps, breakouts by index measure, and bar charts. However, there is always the desire to display data in different ways. Early on when the team received requests from users to modify the tool, they provided unlocked versions to those requesting to make modifications. After seeing the inner workings of EVS ART, and the thought that went into the creation of the tool, a user remarked that “it look[ed] easier than it really is,” and this is true.

The team learned, through trial and error, that it was not wise to share unlocked versions of the tool. There are numerous pivot tables and charts and thousands of links and formulas in each of the templates. Breaking any one of them could compromise the analysis. Because of this, they decided to no longer provide unlocked versions and instead to collect the feedback received and use that information to improve the templates each year.

## LESSONS LEARNED

The project taught the implementation team a set of lessons:

- **Cost does not equal worth.** A tool does not have to be expensive to provide extraordinary value.
- **Keep the solution simple.** While the inner workings of a tool may venture over to the complex side, do not make the act of engaging with the analysis complex for users, or they will not use it.
- **Make the solution accessible to all.** This means two things. One, ensure that the tool is accessible to those with disabilities, and two, make it available to the broadest audience possible. If people do not

know about the tool, they will continue to spend unnecessary time re-creating analyses and unnecessary money on contracts to conduct analyses, or they may simply do nothing with the valuable information they have at their fingertips.

- **Ensure that the output of the tool is intuitive and useful.** Do not make users reanalyze the analysis—the tool should do the work for them the first time. Provide results in a format that can be utilized for presentation.
- **Tell the story.** Do not overwhelm end users. Offer a high-level visual overview and then direct them down an intuitive path to more specific details.
- **Be transparent.** When working with results, whether positive or negative, do not shy away from difficult conversations. Survey takers already know whether something is working well or not. Be up front, dig deeper, and let them know that their input will drive change.
- **Tie actions back to survey feedback.** When creating initiatives based on feedback obtained through a survey, it is important to tie the organization's actions back to the voice of the people. This will increase engagement, add validity to the survey, and in most cases, increase future participation.

What was the most basic lesson learned? Great things can come from grassroots efforts.

## IMPACT OF THE SOLUTION

The introduction of EVS ART created immediate efficiencies in both the time and cost of completing the FEVS analysis. Colleagues at the Centers for Disease Control and Prevention (CDC) experienced a significant reduction in the time spent conducting FEVS analysis. Prior to EVS ART, they produced 24 reports in 72 workdays at a cost of approximately US\$30,861. Reporting can now be done in one workday at a cost of approximately US\$1,129—a savings of US\$29,732 and a 96 percent reduction in both time and cost. These efficiencies have allowed the CDC to increase its reporting sixfold to 150 individual analyses—meaning that 126 additional managers now receive their own customized FEVS results.

An NIH analyst who once spent 30 hours creating one report at an average cost of US\$1,350 can now complete an analysis in less than 5 minutes at a cost of US\$3.75. Simplifying the analysis process means that frontline managers can access meaningful data to better inform policies, programs, and initiatives much sooner. They can also have confidence that the information they are using to create or bolster initiatives is coming directly from those whom their actions impact most.

## BEYOND ANALYTICS: CREATING MEASURABLE AND SUSTAINABLE CHANGE

While the efficiencies created by EVS ART have helped save both time and money, the most important aspect, by far, has been the increased ability to identify themes and measure organizational change (see figure 9.18).

One example of a success story concerns the transformation of an underperforming organization. This organization was forward facing and interfaced with all 1,300 institute employees. To remedy its underperformance, the NIDDK Executive Officer stepped in with a multipronged approach and, over the course of a year,

- Put in place new standards and forms of accountability, including metrics to measure productivity (industry standards),

**FIGURE 9.18** Identifying Challenges through the EVS ART



Source: Original figure for this publication, NIDDK.

Note: EO = executive office; EVS ART = Employee Viewpoint Survey Analysis and Results Tool; FEVS = Federal Employee Viewpoint Survey; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases.

- Worked closely with leaders to create a new vision for the group,
- Changed out leaders who did not embrace the new vision,
- Taught necessary competencies to supervisors,
- Created opportunities for high performers,
- Ensured that mediocrity was not acceptable and that there were consequences for poor performance, not only for employees but also for leaders, and
- Worked closely with the employees of the organization so they knew in real time what changes were happening and why, ensuring that each employee within the organization had a voice.

Over the course of a year, the organization was transformed. Employees knew it because service improved, complaints were greatly reduced, and partnerships began to form. By using EVS ART, the NIDDK was able to prove that its targeted interventions were working. Figure 9.19 illustrates the transformation from one year to the next. The employee engagement index went up by 22.8 percentage points, the global satisfaction index went up 36.6 percentage points, and the new inclusion quotient increased from 51.6 percent to 76.6 percent positive.<sup>16</sup>

NIDDK staff recognized the transformation, and confidence in the organization returned. The work continues to pay off, and five years later, the success of the interventions is still clearly demonstrated (see figure 9.20).

**FIGURE 9.19** Changes in Federal Employee Viewpoint Survey Index Measures, 2015–16



Source: Original figure for this publication, NIDDK.

Note: EO = executive office; FEVS = Federal Employee Viewpoint Survey; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases.

## Performance Management in Practice

The same success has played out across the institute. In addition to targeted interventions, to be a truly performance-centric organization, performance management must be incorporated into an organization's culture continuously. NIDDK leadership routinely finds opportunities across the institute to highlight the importance of performance standards and conversations.

At the NIDDK, people throughout the organization shared via the FEVS that they wanted discussions with their supervisors about their performance to be more worthwhile: they wanted their supervisors to provide them with more constructive suggestions about how to improve their job performance and to give them meaningful recognition when they had done a good job. To address this, the NIDDK Executive Officer initiated the following practices:

- Reviewing performance ratings across the entire organization to make sure that all supervisors were interpreting “outstanding” rating requirements, versus “excellent” and “satisfactory” ones, in the same way and that, where appropriate, they were giving lower ratings when deserved rather than ignoring underperformance

**FIGURE 9.20** Changes in Federal Employee Viewpoint Survey Index Measures, 2015–19



Source: Original figure for this publication, NIDDK.

Note: EO = executive office; FEVS = Federal Employee Viewpoint Survey; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases.

- Reviewing all awards and retention incentives to make sure there was equity and fairness in who received awards and in what amounts
- Sending mid-year and end-of-year communications to the NIDDK's supervisors, reiterating what employees had said and emphasizing that staff played an active role in their performance conversations
- Sending communications to staff reminding them that what they said was important and encouraging them to play an active role in their performance conversations
- Sharing the document "Performance Management Tips and Templates" with both supervisors and staff to equip them with the tools they needed to have more robust performance conversations.

Over time, the people of the organization saw a noticeable change in performance management, which the NIDDK has validated using the FEVS and EVS ART. Traditionally, one of the lowest-scoring questions across government has been "In my work unit, steps are taken to deal with a poor performer who cannot or will not improve." This is one of the most difficult questions to tackle across the government. Many



**FIGURE 9.21** Improving Measures of Accountability at the National Institute of Diabetes and Digestive and Kidney Diseases

## Accountability...

Organization	2015	2016	2017	2018	2019	Change from 2015 to 2019
Governmentwide	28%	29%	31%	32%	34%	6%
HHS	34%	35%	38%	39%	40%	6%
NIH	39%	41%	43%	46%	46%	7%
NIDDK/EO	57%	56%	70%	73%	86%	29%

### FEVS question:

In my work unit, steps are taken to deal with a poor performer who cannot or will not improve.



Source: Original figure for this publication, NIDDK.

Note: EO = executive office; FEVS = Federal Employee Viewpoint Survey; HHS = Health and Human Services; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases; NIH = National Institutes of Health.

federal leaders have said that it should be removed from the FEVS because, due to the confidential nature of employee relations, it is nearly impossible to demonstrate that actions are being taken.<sup>17</sup>

However, the NIDDK proved that it is possible. Leaders throughout the institute devoted resources to assist supervisors and employees early on when there were problems with performance or conduct. The key was creating a culture where early intervention occurs and clear standards and accountabilities are established and transparent. When this was done, staff began to notice underperforming organizations improve (see figure 9.21).

## CHALLENGES FACED AND THE PATH FORWARD

The biggest challenge for the NIDDK team has been balancing their traditional responsibilities with the demands of creating, modifying, and supporting a tool that has gained in popularity. Since its inception, EVS ART has been enhanced to expand its capabilities several times due to the NIDDK's desire to strive for continuous improvement based on feedback received from users. The FEVS itself has undergone changes over the last three years, and EVS ART has required substantial modification to adapt to those changes as well. The team has learned that, while there is great satisfaction in being able to provide their federal colleagues with a tool that evolves with their needs, this also means that their work is never really done.



One function not yet incorporated by the tool's creator is the ability of the tool to determine statistical significance in changes from one year to the next, or between organizations of similar or different sizes. The addition of this capability could help to "win over" survey cynics. Last, with the topics of diversity, equity, inclusion, and accessibility at the forefront of conversations within the United States, it would be helpful to have the ability to compare data using different demographics, such as race and ethnicity. This is something that the NIDDK team is actively working on. While not all users have access to this level of data, the team would like to provide a similar user-friendly reporting format to those who do have access.

## CONCLUSION

All leaders aspire to create and sustain high-functioning organizations. How can organizations achieve high-functioning performance, much less provide measurable evidence that they have reached this goal?

The synergy between technology for data analytics and the voice of the people can be powerful. It can inform a leader's strategy and resource allocation and provide evidence that an organization's performance and engagement activities are paying off. The NIDDK now uses the FEVS as a launchpad for moving forward, not as a report card looking back. It is with this in mind that it put into effect the year-round campaign "You Speak ... We Listen ... Things Happen!" to reiterate to employees that it is constantly listening to their voices and taking their feedback into account in the planning of programs and initiatives. Leadership incorporates this campaign into email communications, posters, and all-hands meetings to remind employees that their voices make a difference.

The NIDDK Executive Officer also conducts workshops to build communities and connect with staff. Early on, these workshops were used as part of targeted interventions. Now, as a very high-functioning organization, the NIDDK has transitioned to more strategic initiatives. It does this by harnessing the talents of staff who have relevant interests and technical expertise that extend beyond their functional areas to deliver workshops that continue to strengthen employee development—in lieu of bringing in outside facilitators. It focuses on career development, offering world cafés that allow staff one-on-one interaction with senior leaders from across the NIH who volunteer to share experiences, as well as specialized workshops on resilience, problem solving, conducting difficult conversations, and managing up.

Another part of the NIDDK's success has been in creating many strategic initiatives. Some of the more novel programs it has put in place, which have resulted in an increase in FEVS scores and in employee engagement across the institute, include using crowdsourcing to initiate conversations and capture ideas, incorporating pulse surveys into town halls, and conducting stay and exit interviews with staff. In addition, it has created a novel awards program to recognize "rising stars," innovative problem solving, and personification of the organization's core values. It has also focused on the professional and career development of staff through the launch of a formal mentoring program, a shadowing program, a new supervisors program, and the novel Career Climbers Cohort, which was specifically designed for staff who were junior in experience or brand new to the federal workforce. Each of these initiatives, programs, and activities has been informed by the institute's employees. This largely explains the institute's success in the FEVS's "Belief in Action" question: "I believe the results of this survey will be used to make my agency a better place to work." Across government, this question has traditionally scored incredibly low—but at the NIDDK, that has changed.

In 2015, the NIDDK was able to do its first deeper-dive analysis using an early version of EVS ART. Armed with this information, they set out to create employee-informed change, and this did not go unnoticed. Between 2015 and 2016, the NIDDK Executive Office's positive responses to the "Belief in

Action” question jumped by 14 percentage points, from 52 percent to 66 percent. In 2020, this same office recognized a “Belief in Action” score that was 90 percent positive—a jump of 38 percentage points from 2014 (see figure 9.22).

With the increase in “Belief in Action” scores, survey response rates increased as well. The NIDDK’s overall employee participation increased from 36.8 percent to 74.5 percent (see figure 9.23).

Very basic things are needed to help ensure an organization’s success. An organization requires reliable and real-time data, the ability to keep a finger on its own pulse, and the ability to tie organizational interventions and strategic initiatives back to employees’ voices. Data are only meaningful when they are accounted for, acted upon, and shared with staff. They must be incorporated into the organization’s culture and practices on a daily basis. The result is an amazing ripple effect (figure 9.24).

In closing, EVS ART is an incredible resource, but it is important to remember that the tool itself cannot create change—it can only inform it. The magic lies in what is done with the information it provides. The importance of leadership buy-in and action, at all levels, is critical, and a leader’s level of buy-in can either help or hinder an organization’s success. When leaders effectively use employee feedback to create timely, well-informed, and meaningful initiatives, the rest will begin to fall into place—and that is a wonderful cycle to be in.

**FIGURE 9.22 “Belief in Action” Scores from the Federal Employee Viewpoint Survey, 2014–20**

## Belief in Action...

Organization	2014	2015	2016	2017	2018	2019	2020	Change from 2014 to 2020
Governmentwide	38%	39%	41%	42%	41%	41%	43%	5%
HHS	47%	49%	52%	54%	55%	56%	57%	10%
NIH	46%	48%	53%	56%	59%	59%	62%	16%
NIDDK	50%	55%	69%	71%	74%	73%	76%	25%
NIDDK/EO	52%	66%	76%	83%	77%	88%	90%	38%

### FEVS question:

I believe the results of this survey will be used to make my agency a better place to work.



Source: Original figure for this publication, NIDDK.

Note: EO = executive office; FEVS = Federal Employee Viewpoint Survey; HHS = Health and Human Services; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases; NIH = National Institutes of Health.

**FIGURE 9.23** Federal Employee Viewpoint Survey Participation Rates, 2014–20

## FEVS Participation...

Organization	2014	2015	2016	2017	2018	2019	2020	Change from 2014 to 2020
NIDDK	36.8%	45.1%	54.3%	60.7%	64.0%	71.9%	74.5%	38%

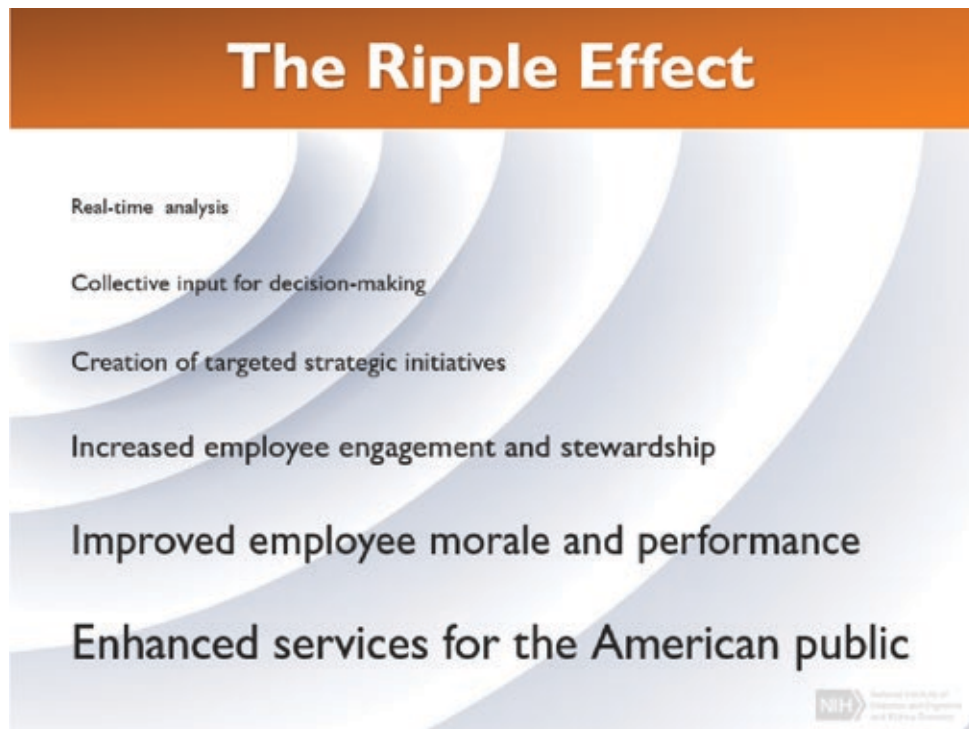
...often reflects an employee's "Belief in Action"



Source: Original figure for this publication, NIDDK.

Note: FEVS = Federal Employee Viewpoint Survey; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases.

**FIGURE 9.24** The Ripple Effect



Source: Original figure for this publication, NIDDK.

Note: NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases.

## NOTES

The authors are grateful for contributions by government counterparts in Brazil—Luciana Andrade (Anvisa) and Matheus Soldi Hardt (EloGroup)—Luxembourg—Ludwig Balmer (CGPO), Marc Blau (CGPO), and Danielle Bossaert (Observatory of the Civil Service)—and the United States—Camille Hoover (NIDDK) and Robin Klevins (NIDDK).

1. For an overview of different statistical capacities produced by the World Bank, see the Statistical Performance Indicators (SPI) page on the World Bank website, <https://www.worldbank.org/en/programs/statistical-performance-indicators>.
2. There is an important distinction between data infrastructure and information systems. Data infrastructure often entails both the digital and physical infrastructure required to store data, while the information system refers specifically to the software architecture in which data are stored.
3. For additional details, see chapter 12.
4. Ghost workers can appear in many forms: employees without a legal appointment who are nevertheless being paid; dead employees who continue to appear on the payroll and whose salaries are drawn by local administrative staff or shared with descendants; employees who draw both a pension and regular salaries; employees who draw multiple salaries from different ministries or agencies; employees who have been dismissed or have retired or resigned from service but continue to draw salaries; employees who are not working or showing up but continue to be paid; and employees who use false or multiple identities to draw multiple salaries.
5. More information about the FEVS can be found on the website of the OPM, <https://www.opm.gov/fevs/>.
6. See case study 9.3.
7. The project is described in more detail on the World Bank website, <https://projects.worldbank.org/en/projects-operations/project-detail/P176877>.
8. A sheet is a page that contains the charts, key performance indicators, and tables that compose a dashboard. An application contains multiple sheets.
9. More information about the FEVS can be found on the website of the OPM, <https://www.opm.gov/fevs/>. For summary statistics on full-time permanent, nonseasonal federal employees, see OPM (2017).
10. This staff time estimate does not include the time spent administering the survey or analyzing its results.
11. Index measures are aggregates of positive questions regarding perceptions of employee engagement. Key categories are generally described as survey modules—for instance, work experience and relationship to supervisors.
12. Some indexes contain four, and others, as many as 39 answers.
13. This act requires all electronic and information technology that is created by the federal government to be accessible to people with disabilities. Compliance allows users with assistive technology, such as screen readers, to use the tool.
14. One participant noted that the dashboard was “proof of concept that with strategic initiatives and targeted interventions, a federal leader can affect positive change and realize significant measurable improvement from 1 year to the next.”
15. A colleague from the Department of Homeland Security “shed a tear when we learned of [the] tool, watched the video, used it, and saw how fast the analysis was performed.” A senior user from the Centers for Disease Control and Prevention (CDC) shared that “EVS ART is a great tool to convey more data in less time and in a more meaningful way.” And a senior colleague from the Office of the Secretary of the Department of Health and Human Services stated that “EVS ART will allow us more time to analyze the data and focus more time on strategic planning.”
16. These are not jumps in specific questions but changes to the average of index measures. Many of the specific questions within the index measures went up by 40–50 percentage points in just one year.
17. Note that the specific question has been removed for 2022 and exchanged for a multiple-choice question. The wording of the question has been changed as well to address these concerns.

## REFERENCES

- Bozeman, Barry, and Stuart Bretschneider. 1986. “Public Management Information Systems: Theory and Prescription.” *Public Administration Review* 46: 475–87. <https://doi.org/10.2307/975569>.
- Caudle, Sharon L., Wilpen L. Gorr, and Kathryn E. Newcomer. 1991. “Key Information Systems Management Issues for the Public Sector.” *MIS Quarterly* 15 (2): 171–88. <https://doi.org/10.2307/249378>.
- Diamond, Jack. 2013. *Good Practice Note on Sequencing PFM Reforms*. Washington, DC: Public Expenditure and Financial Accountability (PEFA). <https://www.pefa.org/resources/good-practice-note-sequencing-pfm-reforms>.
- Henke, Nicolaus, Jacques Bughin, Michael Chui, James Manyika, Tamim Saleh, Bill Wiseman, and Guru Sethupathy. 2016. *The Age of Analytics: Competing in a Data-Driven World*. McKinsey Global Institute. <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-age-of-analytics-competing-in-a-data-driven-world>.

- Kamensky, John. 2019. "Why Engagement Matters and How to Improve It." *Government Executive*, April 19, 2019. <https://www.govexec.com/management/2019/04/why-engagement-matters-and-how-improve-it/156424/>.
- Newcomer, Kathryn E., and Sharon L. Caudle. 1991. "Evaluating Public Sector Information Systems: More Than Meets the Eye." *Public Administration Review* 41 (5): 377–84. <https://doi.org/10.2307/976406>.
- NIH (National Institutes of Health). 2018. "FEVS Privacy." NIH Videos, May 4, 2018. Video, 2:14. <https://www.youtube.com/watch?v=k2umYftXKCI>.
- OPM (Office of Personnel Management). 2017. "Profile of Federal Civilian Non-Seasonal Full-Time Employees." US Office of Personnel Management, US Government, September 30, 2017. <https://www.opm.gov/policy-data-oversight/data-analysis-documentation/federal-employment-reports/reports-publications/profile-of-federal-civilian-non-postal-employees/>.
- OPM (Office of Personnel Management). 2021. *Governmentwide Management Report: Results from the 2020 OPM Federal Employee Viewpoint Survey*. Washington, DC: US Office of Personnel Management, US Government. <https://www.opm.gov/fevs/reports/governmentwide-reports/governmentwide-management-report/governmentwide-report/2020/2020-governmentwide-management-report.pdf>.
- Runkler, Thomas A. 2020. *Data Analytics: Models and Algorithms for Intelligent Data Analysis*. Cham, Switzerland: Springer Vieweg.
- World Bank. 2016. *World Development Report 2016: Digital Dividends*. Washington, DC: World Bank. <https://www.worldbank.org/en/publication/wdr2016>.
- World Bank. 2021a. *Europe and Central Asia Economic Update, Spring 2021: Data, Digitalization, and Governance*. Washington, DC: World Bank. <https://openknowledge.worldbank.org/handle/10986/35273>.
- World Bank. 2021b. *World Development Report 2021: Data for Better Lives*. Washington, DC: World Bank. <https://www.worldbank.org/en/publication/wdr2021>.

## CHAPTER 10

# Government Analytics Using Human Resources and Payroll Data

*Rafael Alves de Albuquerque Tavares, Daniel Ortega Nieto, and Eleanor Florence Woodhouse*

### SUMMARY

This chapter presents a microdata-based approach for governments to improve their strategic human resource management and fiscal planning. Using a series of examples from Latin American countries, the authors demonstrate how basic statistics created using payroll and human resource management data can help policy makers gain insight into the current and future state of their government's wage bill. The authors argue that this constitutes an important first step toward tapping the potential of existing bodies of payroll and human resources microdata that are currently underused. This approach can help policy makers make difficult decisions by breaking down the causes of problems and putting numbers to the ways in which certain policy choices translate into longer-term consequences.

### ANALYTICS IN PRACTICE

- *Data collection practices.* It is recommended that where possible, governments centralize their human resources (HR) data collection systems and render these data accessible to insights teams. If such data do not exist, even in a disparate fashion, we strongly advise governments to begin collecting, in a centralized manner, payroll and human resources management information system (HRMIS) microdata. We advise governments to make these data public where possible (anonymizing the data, naturally) to improve transparency.
- *Fiscal planning.* We advocate for better integration of HR data analysis with fiscal planning. To be able to leverage payroll and HRMIS microdata, governments must encourage civil servants from the treasury and HR department(s) to collaborate more closely. This could be achieved by allocating dedicated

---

The authors' names are listed alphabetically. Rafael Alves de Albuquerque Tavares is a consultant at the World Bank. Daniel Ortega Nieto is a senior public sector management specialist at the World Bank. Eleanor Florence Woodhouse is an assistant professor in the Department of Political Science and School of Public Policy, University College London.



portions of civil servant workload to the task of sharing and analyzing data or by creating dedicated interdepartmental roles to push forward and undertake the collection and analysis of payroll microdata for strategic human resource management (SHRM).

- *Service delivery.* By better integrating HR data and wage bill planning, policy makers can improve service delivery to citizens. For example, projections of which categories of public servants will retire or transfer across posts allow managers to identify where additional resources will be required to ensure the continuity of service provision. This logic can be extended to the integration of personnel data with the wider dataverse available to policy makers. For example, data on demographic changes among citizens allow policy makers to predict changes in service demands. The interaction of these analytics on the demand and supply sides of service delivery allows policy makers to use their resources intelligently.
- *Insulation of SHRM and fiscal planning.* Political considerations can impede the implementation of successful SHRM and fiscal planning. We recommend that governments insulate certain aspects of planning offices' work from the ebb and flow of politics. This could go hand-in-hand with our second lesson, to carve out explicit portfolios or roles dedicated to collecting and analyzing HR microdata, by ensuring that this work is undertaken by public servants reporting to an independent agency rather than to a minister.

## INTRODUCTION

This chapter offers policy makers ways to use HR microdata to improve SHRM and fiscal planning. More specifically, practical examples are presented of how to use payroll and HRMIS data to strengthen wage bill projections, gain better insights into the dynamics of the public sector labor market, and strengthen evidence-based personnel policy. The approach offers ways to tap the potential of HRMIS data that are widely available but underused. The types of analysis that we propose can support public sector managers in their decision-making by rendering explicit some of the consequences of key human resource management (HRM) choices and simulating distinct scenarios.

The approach described uses administrative data related to individual employment and compensation to model the dynamics of the public sector workforce and its associated costs. By applying an analytical lens to the administrative data the public sector holds on its employees, these data become a means of better understanding the characteristics of public administration. This includes determining simple statistics, such as the ratio of pay and allowances across distinct groups of employees, the different job placements and training opportunities secured by officials across time and institutional environments, and extrapolations of core variables, such as the wage bill under current laws and regulations.

With these generally straightforward statistics—to which any government with an HRMIS should have access—significant improvements can be made to addressing potential HRM shortcomings and related fiscal issues, including strategic workforce planning, replacement rates, salary inequalities within and across government agencies, the distribution of pay-for-performance benefits, the retirement of personnel, and projections of payroll costs, among others.<sup>1</sup>

Data analytics based on personnel data have proliferated in recent years and enable organizations to understand analytics across the HRM cycle (Davenport 2019)—from the attractiveness of distinct positions advertised by an organization (measured by the number of applicants) to diversity and inclusion (measured by, for instance, the ethnic diversity in different ranks of an organization's hierarchy), to name just two examples. Yet, as outlined in chapter 9 of the *Handbook*, many government organizations lack an HRMIS with which to register these data. For this reason, we limit the HRMIS analysis in this chapter to personnel data that are often more readily available and registered by governments—such as age (by registering date of birth) and gender—while acknowledging that this only presents a small fraction of the HRMIS data analytics possible with more widely available data.

## Common Sources of Microdata

Two key terms used throughout this chapter are the *government wage bill* and *payroll and HRMIS microdata*. Harborne, Bisca, and Dorotinsky (2017, 267n48) define the government wage bill as

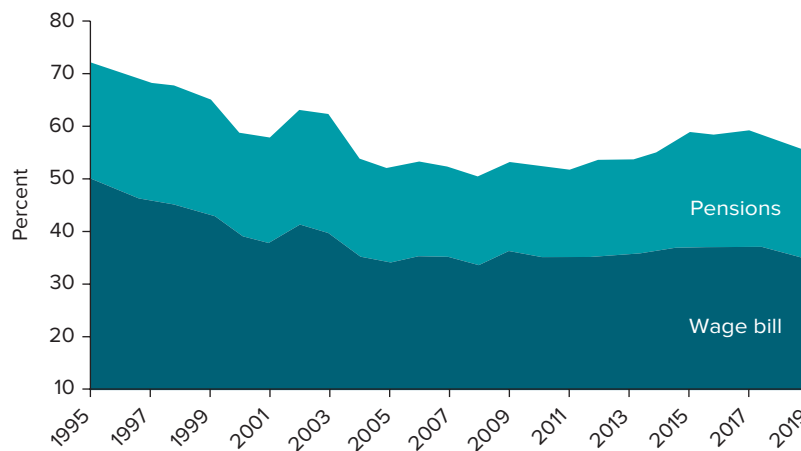
the sum of wages and salaries paid to civilian central government and the armed forces. Wages and salaries consist of all payments in cash (no other forms of payment, such as in-kind, are considered) to employees in return for services rendered, before deduction of withholding taxes and employee pension contributions. Monetary allowances (e.g., for housing or transportation) are also included in the wage bill.

Pensions, by contrast, are generally not included in the wage bill. Pensions remain, however, an important share of a government's payroll. Indeed, a recent report has found that subnational governments in Brazil have been under growing pressure from pension expenses, at times spending the same amount of money on pensions as on the wage bill (World Bank 2022). Figure 10.1 presents the percentage of the state budgetary ceiling allocated to pensions and the wage bill in Brazil. In 2019, over 20 percent of the budgetary ceiling was spent on pensions, almost the same proportion as the wage bill itself. The same type of analysis that we perform for the wage bill in this chapter could be replicated for pensions. For example, projections of pension expenses could aid governments in making strategic decisions to prepare for and potentially reduce their burden.

Payroll and HRMIS microdata are two separate data sources that we leverage in our analyses. Both capture individual-level information about employees and should be easily available to most governments. Payroll data include information pertaining to an employee's contract (job position, contract type, etc.), personal details (date of birth, national insurance number, address, etc.), salary and tax details (amount and date of payments, tax codes, etc.), and leave, holidays, and benefits. These data are generally collected by the HR or finance department that administers the salaries of public employees. For this reason, these data are automatically updated because they must reflect promotions or changes in role or leave allocations.

HRMIS data, on the other hand, can be used to enrich payroll data because they also capture information such as an employee's gender and educational qualifications or prior professional experience. However, they tend not to be updated in the same way as payroll data because they are usually taken as a

**FIGURE 10.1** Wage Bill and Pensions as a Percentage of Subnational States' Budget, Brazil, 1995–2019



Source: World Bank 2022.

snapshot at the recruitment stage and thus capture an employee at only a single point in time (for example, if the employee earns a degree after starting a position, this will not necessarily be reflected in the HRMIS data).

When we refer to HR data more broadly, we refer to the combination of payroll and HRMIS data for active (or currently employed) civil servants.<sup>2</sup> While many governments have access to data collected via their HRMIS, they sometimes struggle to extract and use them to their full potential. This can be due to a number of issues, including outdated HRMIS or analytical capacity, the decentralization of HR departments (which means that central government administrators only have access to partial data), or a lack of long-term strategic HR planning. In the section “Payroll and HRMIS Microdata and Related Challenges,” we offer insights into how to get these data into shape for analysis, and in the section “Descriptive Statistics,” we discuss how to undertake simple and effective analyses using said data to improve SHRM.

## Capitalizing on Government Microdata

We focus on SHRM and the government wage bill for a number of reasons. First, the wage bill has considerable fiscal impact because it represents a significant portion of government spending: around one-fifth of total spending, according to the International Monetary Fund (IMF) (Gupta et al. 2016, 2), or around 9–10 percent of gross domestic product (GDP) and roughly a quarter of general government expenditures, according to the World Bank (Hasnain et al. 2019, 8). Second, it is likely that across the globe, pressures on wage spending will increase in the coming years and decades because “advanced economies are facing fiscal challenges associated with aging populations while also needing to reduce high public debt levels” and because “emerging markets and low-income countries have pressures to expand public service coverage in the context of revenue and financing constraints and the need for higher public investment” (Gupta et al. 2016, 1). Thus, in order to ensure that they are able to continue to deliver essential public services when facing increasing financial constraints, governments must invest in fiscal planning and SHRM.

The approach we propose allows government organizations to better leverage their HR data and make use of evidence for decision-making. Such strategic use of HR data can also have a significant fiscal impact, helping to avoid short-termism and here-and-now pressures that may cast a long shadow over government organizations’ ability to undertake their work and offer the best services to citizens under government budget constraints. A case in the Brazilian state of Alagoas offers a good example, illustrating the potential of using payroll microdata to provide empirical evidence for the pros and cons of policy decisions. Here, estimates of a decreasing pupil-per-teacher ratio helped inform the government’s decision to recruit fewer teachers while maintaining the quality of the public education delivered to its citizens by opening up fiscal space to better provide other, more needed public services.

One of the major advantages of applying an analytical lens to SHRM and wage bill data is that it supports governments to improve workforce and fiscal planning jointly and in a coordinated way. These two aspects of government work should occur in tandem, but in practice, this is rarely the case. *Workforce planning* “is a core HRM process that helps to identify, develop and sustain the necessary workforce skills” and that “ensures that the organisation has the right number of people with the right skills in the right place at the right time to deliver short and long-term organisational objectives” (Huerta Melchor 2013, 7). The ultimate goal of public sector workforce planning is to optimize the number and type of staff employed and the budget of the department or government in question. By the *type* of staff, we mean their professional skill set: are they able to contribute to completing the mission of the organization they serve? Identifying the needs of an organization and the HR required to achieve its goals is the heart of strategic workforce planning (Jacobson 2009; Kiyonaga 2004; Selden 2009): “a goal of workforce planning is to identify the gap between those needs and the available labor supply for government to continue providing quality services and fulfill its mission” (Goodman, French, and Battaglio 2015, 137). *Fiscal planning*, by contrast, refers to the way in which governments use their spending and taxation to influence the economy. As such, it can be improved by developing a better understanding of when certain groups of employees are going to be hired or retire, for example, allowing for more accurate revenue forecasting, which influences the budget approved by the

government. One area that the IMF has identified as important for improving fiscal planning is precisely strengthening links between wage bill management—specifically, wage determination processes—and fiscal frameworks (Gupta et al. 2016, 2).

### Strengthening Traditional Approaches

One additional application of our microdata-driven approach is to help bridge the gap between *macroanalysis* and traditional *functional reviews*, two common approaches to the analysis of the government wage bill and the distribution of work functions across the civil service, respectively. The former relies on macro-level analysis that leverages indicators such as the wage bill as a share of GDP and government employment per capita to gauge the appropriate size and cost of the civil service. By relying on macro indicators, these analyses have often led to simplistic policy prescriptions in the context of fiscal crises.

The latter strain of analysis relies on functional reviews. Using mostly legal documents, regulations, and interviews, these reviews scrutinize the goals, tasks, and resources of units inside the government to improve efficiency and effectiveness. Functional reviews thus have multiple goals but generally aim to assess how work is distributed across the civil service and to identify potential duplication of work through the functions performed by different departments. The analysis may produce results that are not integrated with an overarching strategy of reforming the civil service based on fiscal constraints.

By undertaking microdata analyses, one can complement functional reviews by not only looking at government functions but also gaining greater insight into other relevant dimensions of government organization, such as staffing and competencies. For instance, if one undertook a functional review and discovered that two departments perform similar functions, a parallel microdata-powered analysis could identify the distribution of competencies across the two departments. Perhaps one department has a natural advantage in taking full responsibility for the function because of the greater strength of its staff. Or perhaps there needs to be a redistribution of staff to more effectively distinguish the roles and activities of the two departments.

Micro-level analysis can be used to help reconcile and complement the fiscally oriented nature of macro-analysis and the flexible and detailed nature of functional reviews. This can be done through the use of simple descriptive statistics, such as the drivers of payroll growth (variation in total payroll, wages, and number of employees), the distribution of the workforce according to levels in the career ladder, and progressions and promotions over time and how much they cost, among others, and via a model-based simulation of the wage bill with the fiscal impacts of policies that improve and consolidate wage bill spending. One contribution that our chapter makes is to demonstrate some of the potential uses of and synergies between payroll and HRMIS data. By breaking down data silos, governments can start to better leverage data that are already at their disposal to gain insights into how to manage certain processes, such as adjusting the wage bill and improving fiscal planning.

In short, our chapter aims to lay out a practical, practitioner-friendly approach to the government wage bill that can improve SHRM and fiscal planning with relatively little technical expertise and data that should be accessible (with a relatively low cost of extraction) to any government with an HRMIS. This approach offers significant advantages in helping governments to use the untapped potential of lakes of payroll and HRMIS microdata and, more broadly, to use evidence in order to navigate difficult policy decisions.

## STRATEGIC HUMAN RESOURCE MANAGEMENT AND FISCAL PLANNING

The public administration literature on SHRM focuses on identifying how it is used across different levels of government (Choudhury 2007; Goodman, French, and Battaglio 2015; Jacobson 2010), evaluating the effectiveness of different types of SHRM (Selden 2009; Selden and Jacobson 2007), and determining which factors influence the successful implementation of SHRM strategies (Goodman, French, and

Battaglio 2015; Pynes 2004). However, it is widely recognized that there is a paucity of empirical research on public sector SHRM (Choudhury 2007; Goodman, French, and Battaglio 2015; Reitano 2019), with much of the existing literature being normative in nature, relying on small samples, or being significantly dated. Moreover, the extant literature has a strong focus on the United States, with little to no evidence from the rest of the world.<sup>3</sup> Broadly, SHRM and wage bill data are underused as a source of analytics data for better understanding the characteristics and nature of public administration and public service.

One central finding of the existing literature is that many local governments do not have workforce plans in action (Jacobson 2010). In their survey of the largest US municipal governments, Goodman, French, and Battaglio (2015, 147) find that “very few local governments make use of comprehensive, formal workforce plans.”<sup>4</sup> This is confirmed by other studies focusing on specific geographical regions, such as Jacobson (2010) and Frank and Zhao (2009). Local governments have been shown to lack the technical know-how and resources required to undertake SHRM (Choudhury 2007; Huerta Melchor 2013; Jacobson 2010). Small local governments, in particular, often lack the fiscal, professional, and technical expertise to innovate successfully (French and Folz 2004). For this reason, local governments may shy away from more complex econometric approaches to processes such as budget forecasting because they lack the know-how (Frank and Zhao 2009; Kavanagh and Williams 2016). This is precisely where our approach comes into its own. With very few, simple statistics that any public organization with an HRMIS should have access to, local and national HR departments can make a marked improvement in the use of their SHRM data.

Although the lack of capacity for SHRM seems to be most acute at the local level, it has also been documented in national governments. The Organisation for Economic Co-operation and Development (OECD) describes how its member states have “experienced problems with developing the necessary institutional capacity to engage in workforce planning both at the level of the central HRM body and the budget authority, and at the level of HR departments, professionals and front line managers” (Huerta Melchor 2013, 15). Strategic human capital management was identified by the US General Accounting Office (GAO) in 2001 as a governmentwide high-risk area because many agencies were experiencing “serious human capital challenges” and the combined effect of these challenges placed “at risk the ability of agencies to efficiently, economically, and effectively accomplish their missions, manage critical programs, and adequately serve the American people both now and in the future” (GAO 2001b). Strategic human capital management remains “high risk” to this day and is proving difficult to improve upon, with “skills gaps . . . identified in government-wide occupations in fields such as science, technology, engineering, mathematics, cybersecurity, and acquisitions” and “emerging workforce needs in the wake of the COVID-19 pandemic” (GAO 2021). For this reason, simple, timely ways to improve SHRM—such as the approach that we propose—are urgently needed.

Another important obstacle to successful SHRM and fiscal planning highlighted by the existing literature is political considerations. Successful SHRM requires support and planning from top management because data have to be systematically collected and analyzed over long periods of time. If elected figures are more interested in satisfying concerns “here and now” and are unwilling to invest in longer-term HRM and fiscal strategies, this can pose a significant challenge. This is especially true in smaller local governments, where leadership tends to be more centralized and informal and where, frequently, no separate personnel departments exist (Choudhury 2007, 265). Thus, local governments appear more susceptible to a lack of long-term planning because they are more likely to lack technical know-how or to face direct political pressures (Kong 2007; Wong 1995). It seems especially important, then, to take into consideration the nature and size of a government when examining SHRM (Reitano 2019). As Choudhury (2007, 265) notes, “the conditions of effective human resource management at the federal, state, or large urban levels often are not a good fit for smaller jurisdictions.” That said, we believe that our approach can cut across different levels and sizes of government because it relies on data that should be widely available to small and large governments alike.

The extant literature has also paid significant attention to what Goodman, French, and Battaglio (2015, 147) refer to as the “perfect storm” of “human capital crisis that looms for local governments due to the number of employees who will be eligible for retirement or early retirement in the near future,” which “offers significant opportunity for the use of workforce planning to help with forecasting the labor pool and fine tuning recruitment efforts.” Such a storm is still brewing in many countries around the world, both at the local and



national levels. A significant number of studies explore the issue, which was becoming evident already in the early 2000s, with predictions that over 50 percent of US government senior management would retire as the baby boomer generation came to retirement age (Dychtwald, Erickson, and Morison 2004; GAO 2001a; Jacobson 2010; Kiyonaga 2004; Pynes 2009; Wilkerson 2007). Today, the issue of retirement, and the subsequent talent shortage due to a smaller pool of younger public officials available to replace retiring officials, is aggravated by significant budget constraints in the public sector. Agencies are “freezing recruitment and not replacing employees who retire. The problem is that countries continue cutting budgets without scaling back agencies’ and ministries’ missions, compromising the ability to serve the public” (Huerta Melchor 2013, 15). This makes SHRM all the more important because governments need to use their available resources as wisely as possible to continue to deliver essential services to the public.

Another obstacle to successful SHRM that has been identified by the existing literature is a lack of adequate data (Anderson 2004). For example, in the empirical context of Queensland, Australia, Colley and Price (2010, 203) argue that there were “inadequate workforce data to support workforce planning and thereby identify and mitigate workforce risks.” Several other studies echo the finding that public organizations in many countries find it difficult to obtain an accurate picture of their workforce composition (OECD 2007; Pynes 2004; Rogers and Naeve 1989). Colley and Price (2010, 204) note that “there is general agreement in the public service HR literature that the ideal is a centralised whole-of-service database to meet the common workforce planning needs of agencies. However, establishing such databases is time-consuming and costly, which limits its appeal to an incumbent government focused on short term budget and election cycles.” Again, then, we see that political short-termism can obstruct successful SHRM before one even considers the lack of technical expertise or time and capacity that HR professionals may suffer (as we saw earlier in this section). Our proposed approach speaks to this obstacle to SHRM because it requires only a few basic statistics to better leverage HR data.

In addition to the direct challenges of enacting SHRM, SHRM and fiscal planning also interact in important ways. In order to enact more effective and sustainable fiscal planning, there are numerous ways in which the management of government wages can be improved and better take fiscal concerns into consideration. For example, the IMF notes that wage bill increases have been shown to be associated with worsening fiscal balances: “rather than crowding out other items in the budget, increases in the wage bill have on average been associated with increases in other government spending and with a deterioration of the overall balance” (Gupta et al. 2016, 14). For this reason, policy makers should be especially wary of increasing the wage bill when the budget is tight. Furthermore, if SHRM is not undertaken so as to employ the right type and amount of workers, this can have a negative fiscal impact. If there is a wage premium in the public sector, this can “increase private production costs, including wage costs, as well as result in additional ‘deadweight losses’ associated with distortionary taxation” (15). In fact, wage penalties can also have detrimental fiscal effects because difficulty recruiting and retaining qualified workers adversely affects the quality of publicly provided goods and services and can also contribute to corruption (Hasnain et al. 2019, 8). For this reason, public sector salaries should be calibrated to those of the private sector for comparable jobs and adjusted according to broader changes in the population, society, and the economy at large (Somani 2021). Indeed, advanced economies have been found to struggle to adjust employment levels in response to demographic changes—such as the decline in school-aged children, which led to an oversupply of teachers (Gupta et al. 2016, 20)—which can lead to significant fiscal concerns that could be avoided with a more forward-thinking HRM strategy.

## **PAYROLL AND HRMIS MICRODATA AND RELATED CHALLENGES**

Before delving into what analysis can be done with payroll and HRMIS microdata, it is important to further discuss the kind of data we are talking about and the type of variables one can extract from such data sources. We describe payroll microdata first, before turning to HRMIS microdata. Payroll microdata are



drawn from the administrative data sets that governments use to follow and register the monthly compensation of civil servants and their underlying items. They usually cover most of the government's contracts with its employees and sometimes contain demographic characteristics of civil servants and their occupational information (for example, the department or unit where the civil servant is located, the type of contract, the date of their entry in the civil service, etc.). In some contexts, sets of information are collected independently by different teams. HRMIS microdata, on the other hand, as anticipated in the introduction, are additional data, often collected by recruitment units, that can enrich payroll data with information about employees' gender, education level, and professional sector, for example. To undertake our analyses, we combine these two types of microdata.

In table 10.1, we present an example of a hypothetical combined payroll-HRMIS microdata set with the main variables (columns) and observations (lines) needed for the type of analysis we propose in this chapter. This table represents the minimum data required to undertake the analyses we propose. Each line represents an individual and that individual's respective contract with the government, and each column points to some relevant variable for analysis, such as the unit where the civil servant is located, age, gender, date of entry in the civil service, type of contract, and so on. An individual might have more than one contract with the government: for example, a teacher with two part-time job positions. Ideally, the database should have information about the government's employees for the last 10 years so that one can retrieve variables of interest based on historical data (for example, the average number of years of service before retirement).

Ideally, governments should have the aforementioned information for all their public employees readily available, but based on our experience working with several governments from Latin America and the Caribbean (LAC) countries, we know that governments face challenges when it comes to their wage bill microdata. These challenges can be organized along two dimensions. First, governments may not be able to collect information about all their employees, potentially leading aggregate figures to be wrong or biased. This can happen if wage bill microdata collection is not centralized and the information of some units or departments is missing in the data. In table 10.1, this would be reflected in fewer observations (lines) in the data than in the actual government bureaucracy. A second dimension relates to the number of different aspects that are collected to describe the bureaucracy. In table 10.1, these are captured in the number of columns in the data set. For example, in a recent analysis undertaken in the context of a project with a LAC country, the wage bill data did not have information about when public employees started their careers in the civil service, making it difficult to determine how experience in a position, measured by years of service, was related to wage levels and, as a consequence, the total cost of hiring a new civil servant for that position. With these issues in mind, practitioners should be cautious about what the available wage bill microdata can tell them about the current situation of bureaucracy in aggregate terms and about which aspects can be explored to provide insights for governments to better manage their SHRM and fiscal planning.

In figure 10.2, we propose a simple wage bill microdata “quality ladder” to help practitioners separate good data from bad data. We organize the ladder into five levels, with the first level representing the lowest-quality microdata and the fifth level the highest-quality microdata. At level 0, there is a missed opportunity for HRMIS data analysis because the *minimum required data* are not available (see table 10.1 for reference). This is because the information on public employees is scarce, inaccurate, inconsistent, and scattered across government units or career lines, such that almost any indicator or statistic based on such data would be wrong or biased. Statistically, it is impossible to draw inferences from incomplete data, especially where there are worries that the missingness is correlated with relevant features of the underlying values of the variables in the data. To see this, you need only think of some reasons why a government agency would not report HR microdata: because they lack the capacity or manpower to do so (in this case, only agencies with greater capacity would present their data, offering a skewed vision of the performance of the government at large) or because they are not mandated to do so and thus will not spend precious resources reporting HR data (again, in this case, drawing inferences from such data would give a misleading impression of the government at large because only the agencies with reporting mandates would provide their microdata for analysis).

**TABLE 10.1 Example of Payroll + Human Resources Microdata Set Showing Minimum Data Required**

March 2020																					
Year	Month	Individual ID	Job ID	Date of birth	Gender	Education	Date of entry	Type of contract	Area	Job position	Weekly working hours	Career level	Date of last progression	Base salary	Allowance 1	Allowance 2	Allowance 3	Vacation	Pension contribution	Gross wage	Net wage
2020	March	100,001	1	1987-03-05	Female	Secondary	2015-01-01	Statutory	Education		20	III	2016-03-01	3,500	0	0	0	0	440	3,500	3,060
2020	March	100,001	2	1987-03-05	Female	Secondary	2010-11-10	Statutory	Health		20	IV	2013-03-01	1,000	0	100	0	0	110	1,100	990
2020	March	100,004	1	1980-06-04	Female	Superior	2008-03-02	Temporary	Safety		30	VI	2020-03-05	4,000	0	0	0	0	440	4,000	3,560
2020	March	100,005	1	1985-02-03	Female	No schooling	2009-05-03	Political appointee	Other		40	III	2020-03-31	2,500	200	0	0	0	275	2,700	2,425
March 2021																					
Year	Month	Individual ID	Job ID	Date of birth	Gender	Education	Date of entry	Type of contract	Area	Job position	Weekly working hours	Career level	Date of last progression	Base salary	Allowance 1	Allowance 2	Allowance 3	Vacation	Pension contribution	Gross wage	Net wage
2021	March	100,001	1	1987-03-05	Female	Secondary	2015-01-01	Statutory	Education		30	III	2016-03-01	3,500	0	0	0	0	440	3,500	3,060
2021	March	100,002	1	1980-06-05	Male	Primary	2010-11-10	Statutory	Health		40	IV	2013-03-01	1,000	0	100	0	0	110	1,100	990
2021	March	100,004	1	1980-06-04	Female	Superior	2008-03-02	Temporary	Safety		30	VI	2020-03-05	4,000	0	0	0	0	440	4,000	3,560
2021	March	100,005	1	1985-02-03	Female	No schooling	2009-05-03	Political appointee	Other		40	III	2020-03-31	2,500	200	0	0	0	275	2,700	2,425

Source: Original table for this publication.

At level 1, some analysis can be performed for the units or careers for which there are data available. However, for the reasons outlined above, such analyses must be applied only to the units or career lines for which data are available, and careful consideration must be given to why and how the missingness in the data is occurring. A good example of this is a situation where the wage bill data gathering is decentralized and some government units collect data while others do not. For instance, if only the education and health departments could fill table 10.1 with information about their employees, the analysis should be restricted to these units, and the government should start collecting data from other units of the government.

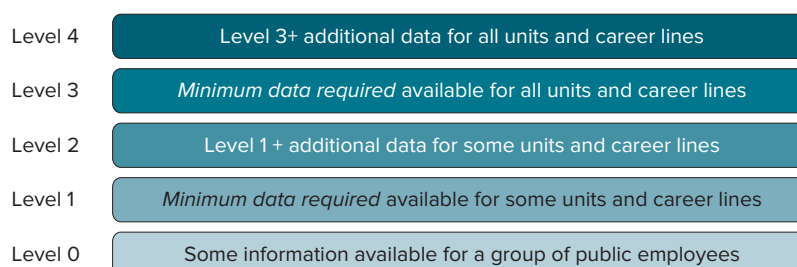
At level 2, not only is the basic information shown in table 10.1 readily available, but one is also able to connect these data with additional data sources and explore specific features of job contracts. Using the above example, this would be the case if the wage bill data for teachers could be connected to students' performance in standardized tests, allowing for the analysis of teachers' productivity in the public sector.

Level 3 illustrates a situation in which the information outlined in table 10.1 is collected for a large part of the bureaucracy in such a way that one can undertake an aggregate analysis of wage bill expenditures based on the microdata. In the section "Wage Bill Projections," we present an example of such an aggregate analysis, with a projection of the wage bill for future years based on data from the Brazilian federal government. We would like to note that levels 2 and 3 of the quality ladder can be ranked differently depending on the objectives of the analyses to be performed. For example, when analyzing the impact or value added of teachers on students' performance, having a productivity measure in the wage bill data for teachers can be especially useful. Given the fiscal nature of the analyses undertaken in this chapter, having a wage bill data set that allows the analyst to create aggregate figures is particularly important. Because of this, we have decided to rank a comprehensive data set for all civil servants without productivity measures above a data set with partial productivity measures in our quality ranking.

In level 4, one can not only undertake the analysis described in level 3 but can also merge other available data sources and connect them with the overall fiscal landscape of the government. Building on the example in level 2, one could assess both the fiscal impacts and the productivity impacts of adding a pay-for-performance scheme to teachers' compensation based on the performance of students on standardized tests.

Building an HRMIS that climbs the ladder described in figure 10.2 can be politically costly and requires sustained investment in the technical skills that underlie data management. The benefit is the improved understanding of the public sector that such an effort provides. The next section outlines the basic analytics for which such databases provide the foundation. Without the qualities outlined in figure 10.2, these analytics are undermined and can be distortionary. But with a sound foundation of quality and comprehensive data collection, these descriptives can support substantial fiscal efficiencies and improved service delivery. In the country cases described in the following section, these investments have paid off many times over.

**FIGURE 10.2 Human Resources Microdata Quality Ladder**



Source: Original figure for this publication.

## DESCRIPTIVE STATISTICS

In this section, we present descriptive statistics that can help policy makers gain insight into the current and future state of their government's wage bill. Along with each insight, we present examples from wage bill analyses that we undertook in different LAC countries. As mentioned before, the data required for these analyses should be available to any government that has an HRMIS. That said, we recognize that there are sometimes significant challenges to obtaining these data—especially in contexts where these data sets are not held centrally—and organizing them in order to undertake these analyses. We posit that there is great untapped potential in the payroll and HRMIS data that governments collect and propose a way to start using these data lakes, where they exist. Where they do not exist, we recommend starting to centralize HR microdata to undertake these types of analyses.

We present our proposed descriptive statistics in three groups. The first provides a general overview of the wage bill and HRM systems to give the reader a sense of how HR practices can impact the wage bill. The second addresses how these HR microdata can be used to identify inequalities in terms of representation within the public sector. Finally, the third proposes a way to address some of these inequalities by adopting a forward-looking perspective that applies changes to fiscal policy to avoid such inequalities or inefficiencies in the future.

### General Overview of the Wage Bill and HRM

We first address how HRM practices can impact the wage bill and offer some examples of the insights that can be gained by better exploiting payroll and HRMIS microdata.

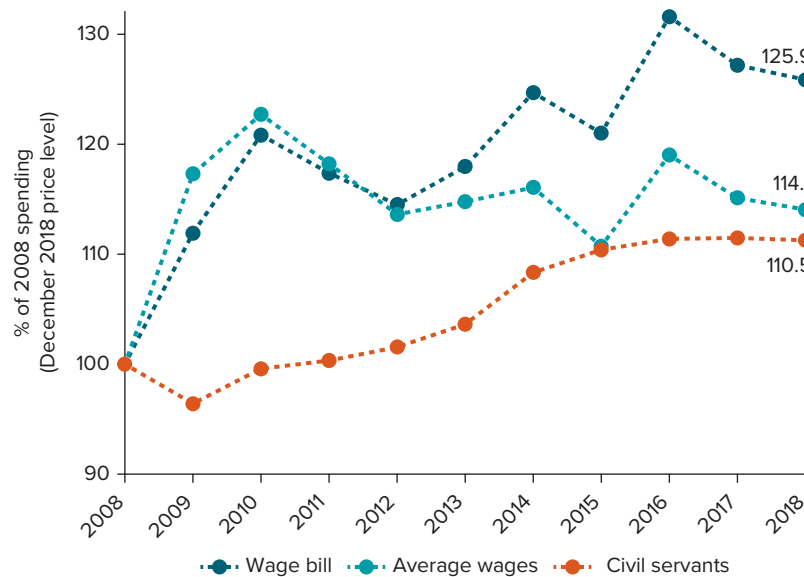
#### *Drivers of Payroll Growth*

Changes in wage bill expenditures can be attributed to changes in employment levels and changes in the average wages of civil servants. A wage bill increase resulting from increased employee hiring is usually accompanied by an expansion in the coverage of public services. Wage increases do not have an immediate impact on the provision of public services, but they may have a medium- and long-term impact on the attraction, retention, and motivation of civil servants that could enhance the productivity of the public service and lead to better service provision. Figure 10.3 presents a simple way of analyzing what is driving wage bill variation. By setting the starting year as a baseline, we can see in this example from the Brazilian federal government's wage bill that most of the increase in wage bill expenditures came from increases in civil servants' compensation. In fact, between 2008 and 2017, spending on Brazilian federal executive personnel rose by 2.9 percent per year in real terms. This growth was made up of a 1.8 percent increase in average salaries and a 1.2 percent increase in the number of public servants. This kind of figure can also be applied to analyze specific sectors and career lines in the government, such as the education sector and, within that, teachers. Undertaking a sector- or career-specific analysis is also a way of providing insights with partial data, since one should be cautious when making aggregate claims from microdata if not all wage bill data are available.

#### *Breakdown of the Wage Bill by Sector*

Breaking down the change in overall wage bill expenditures into changes in the number of civil servants and in average wages can also lend itself to understanding how civil servants and wage bill expenditures are distributed among priority areas. Extending the analysis to the sector level can shed light on the needs and targets of the government in areas such as education, health, and security. For example, in the case of the Brazilian state of Rio Grande do Norte (see figure 10.4), 86 percent of civil servants are distributed in

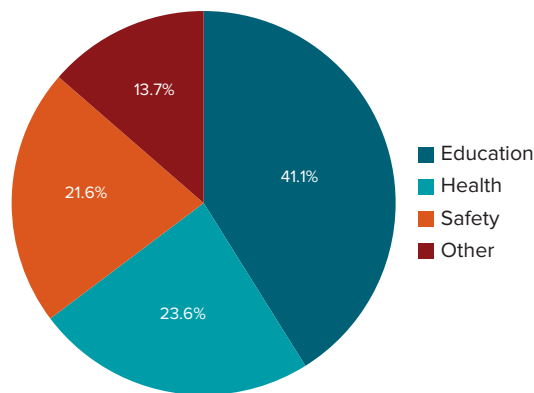
**FIGURE 10.3 Drivers of Wage Bill Variation, Brazilian Federal Government, 2008–18**



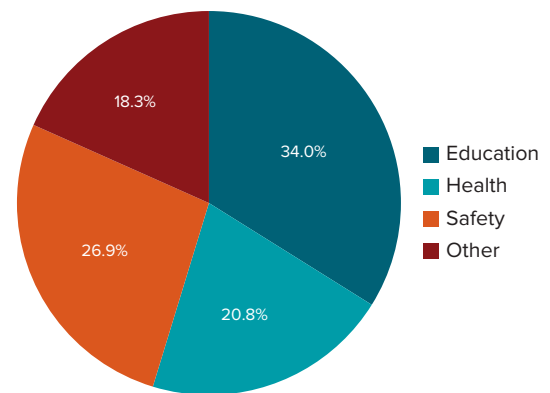
Source: World Bank 2019.

**FIGURE 10.4 Wage Bill Breakdown, by Sector, Brazilian State of Rio Grande do Norte, 2018**

**a. Share of public employees, by sector**



**b. Share of wage bill expenditures, by sector**



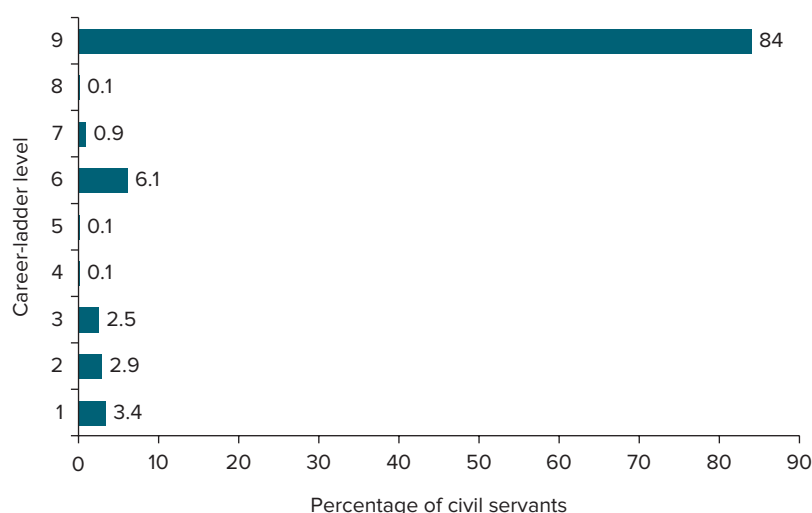
Source: Original figure for this publication.

priority areas, while the wage bill expenditures for these same sectors amount to 82 percent of the total wage bill spending. In particular, the education sector employs 41 percent of the public servants and accounts for 34 percent of the total wage bill.

### ***Distribution of Civil Servants by Career-Ladder Level***

Progressions and promotions along career ladders are a common path for governments to connect higher labor productivity to wage increases. Based on this link between productivity and wages, we can expect that a longer tenure in the civil service reflects a knowledge gain that should equip employees with better tools

**FIGURE 10.5** Distribution of Civil Servants, by Career-Ladder Level, Brazilian Federal Government, 2018



Source: Original figure for this publication.

Note: The figure shows the career-ladder levels for the Tax Auditor career track in the Brazilian federal government.

with which to deliver public services. By analyzing how civil servants are distributed along career-ladder levels, policy makers can assess whether the ladder structure of civil service careers reflects increases in productivity. Ideally, we should expect to see a smooth distribution of civil servants across the different levels. In figure 10.5, we use as an example the career of tax auditors in the Brazilian federal government. We can see that more than 80 percent of public employees are in the final step of their careers, which suggests that there may be margin for improving the design of the career structure and the requirements for progression or promotion to better reflect labor productivity gains.

### Strict Progression Rules and High Turnover of Civil Servants

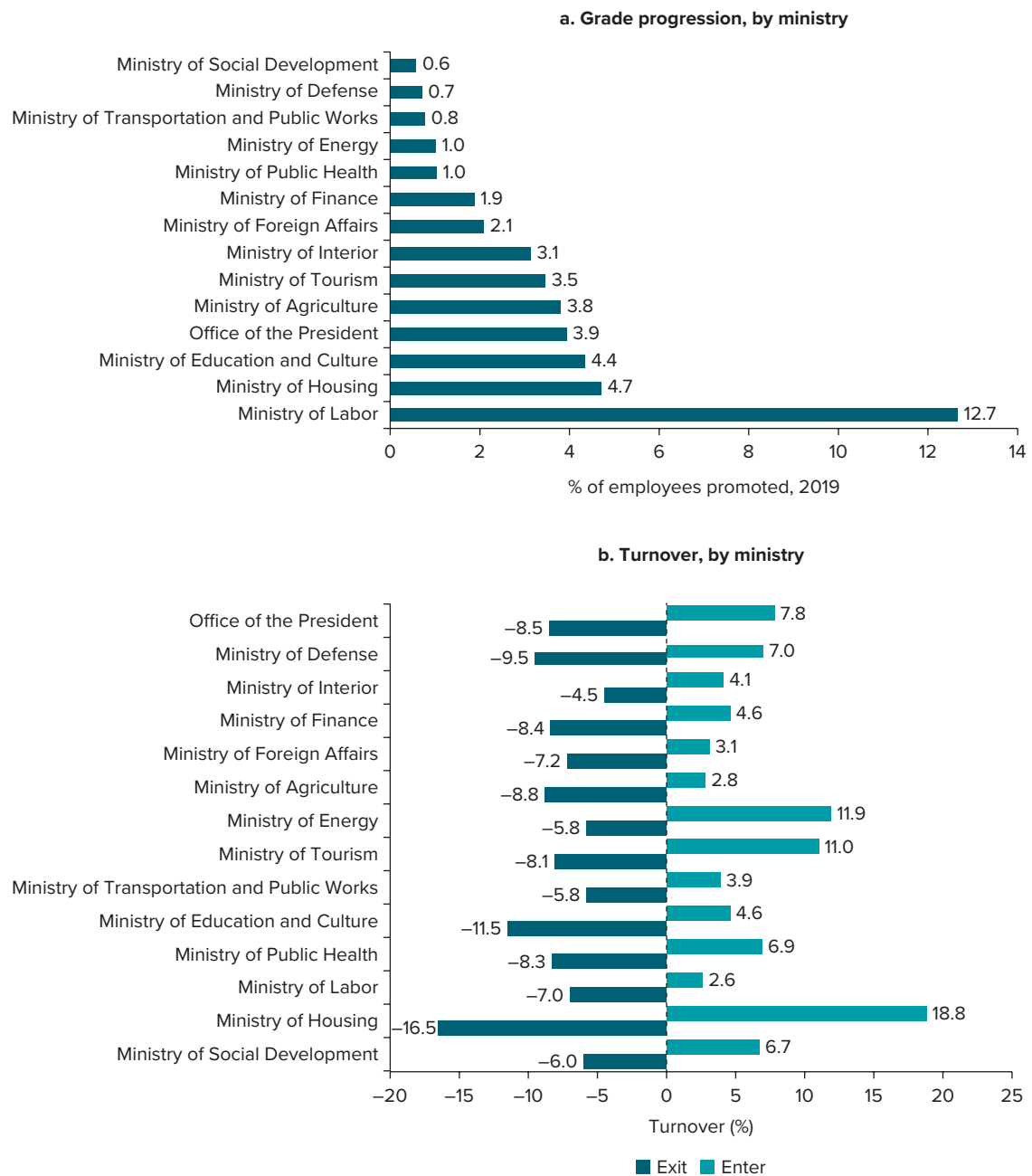
On the other hand, situations where the rules for career progression and promotion are too strict may lead to difficulty retaining public employees, along with their acquired knowledge and expertise. To illustrate such a situation, we can examine the case of Uruguay's central administration, where civil servants are assigned to one scale (*escalafón*) and ministry. As movement across ministries and scales is rare and can only take place with special authorization, grade progression is the only career path available for civil servants. As a result, this limited room for vertical promotions may end up hindering productivity and motivation, as well as increasing turnover. In figure 10.6, we can see the share of employees who were promoted in 2019 (figure 10.6, panel a), and the turnover of employees (figure 10.6, panel b) by ministry in Uruguay's central administration. Less than 5 percent of employees were promoted to a higher grade in almost all ministries, while 7 percent of employees entered the central administration in 2019, and 6 percent exited that same year. In some ministries, the exit rate was even higher than the entry rate. This high turnover can be interpreted as a sign of the challenges in retaining civil servants. It also represents a hidden cost for the government due to the loss of expertise and the cost of training new staff.

### Distribution of Pay-for-Performance Allowances

Pay-for-performance is a useful tool to stimulate productivity in the civil service. In theory, it rewards high-performing public employees and inspires low performers to perform better. However, there is much debate regarding the extent to which performance pay succeeds in improving civil service performance



**FIGURE 10.6 Grade Progressions and Turnover, Uruguay Central Government, 2019**

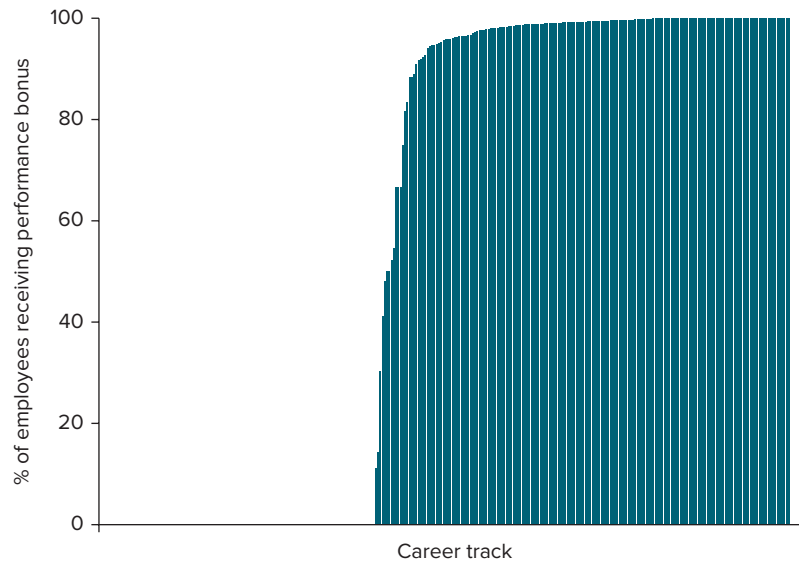


Source: World Bank 2021, based on data from the Contaduría General de la Nación, Government of Uruguay, 2019.

Note: In panel b, the year of entry of an employee is the first year when her/his ID appears on the list of employees related to a specific job category.

(cf. Hasnain, Manning, and Pierskalla 2014). We posit that our approach can help policy makers understand whether pay-for-performance is working in the context in which they work. For example, one problem arises when *all* employees receive performance payment. In figure 10.7, using data from the Brazilian federal government, we display on the *x* axis all careers (each vertical line represents a specific career track) and on the *y* axis the percentage of each profession that received a performance bonus. We show that in 2017, at least 90 percent of employees received performance-related payment in 164 of the 187 careers that offered such schemes. This could indicate that the pay-for-performance scheme in question is not successful in differentiating between good and bad performers.

**FIGURE 10.7** Distribution of Pay-for-Performance Allowances, Brazilian Federal Government, 2017



Source: Original figure for this publication.

Note: The x axis shows all career tracks in the Brazilian federal civil service, ranked by the y-axis variable.

### Inequality in the Public Sector Wage Bill

Having given a general overview of key features of the public service, we turn to the use of HRMIS data to understanding inequalities in the public service. Such inequalities may come in different forms and have correspondingly different impacts on the efficiency or other qualities of the state.

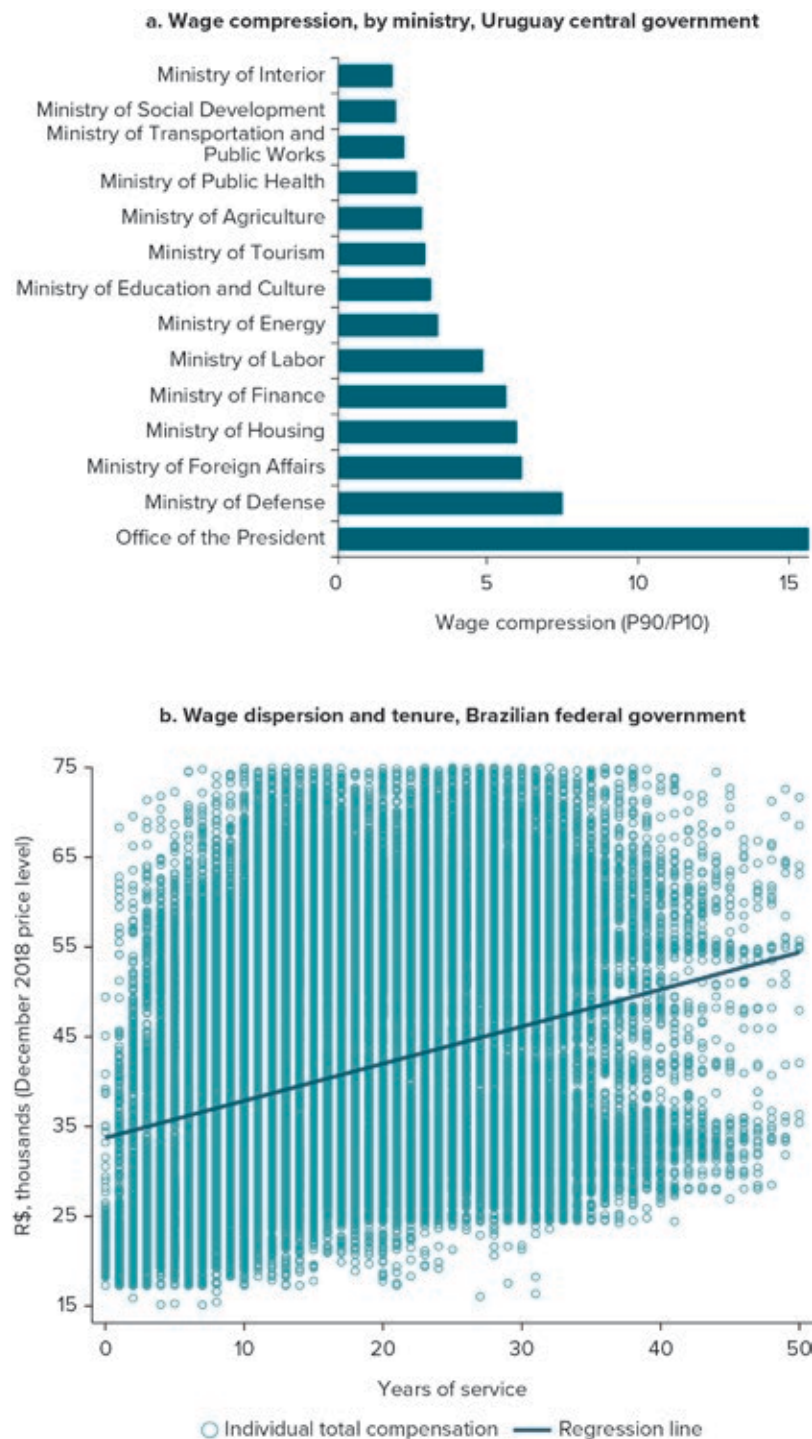
#### Representativeness

Many governments strive to recruit officials in a way that ensures the administration as a whole is broadly representative of the population it serves: for example, by having personnel from across the country's regions in rough proportion with the distribution of the population across those regions. Normatively, such considerations are important, given that in a democratic setting, bureaucracies should represent the populations they serve. Moreover, it has been empirically demonstrated that more representative bureaucracies—dependent on the policy domain—can affect important phenomena such as citizens' trust in the government and willingness to cooperate with the state (see, for example, Riccucci, Van Ryzin, and Lavena 2014; Theobald and Haider-Markel 2009; Van Ryzin, Riccucci, and Li 2017). Though there may be good reason for this principle not to hold strictly, HRMIS data allow the degree of representativeness of the administration to be accurately articulated and to act as the foundation of an evidence-based debate on the matter.

#### Pay Inequity

Inequality in payments in the public sector can reflect underlying differences in responsibilities or can be a sign that inconsistent compensation rules are being applied. For example, we expect the government to reward managers and high-performing employees with better compensation than entry-level civil servants, but we do not expect it to award significantly different levels of compensation to employees with the same attributes, jobs, and tenure, following the generally observed principle of equal pay for equal jobs. In the case of the Brazilian federal government tax auditors (see figure 10.8b), we can see that there is huge wage dispersion for similar workers. Gross pay can vary fivefold for workers with similar levels of experience, which is largely a result of nonperformance-related payments and is not related to base salary.

**FIGURE 10.8 Measuring Pay Inequity in the Uruguayan and Brazilian Governments**



Source: Panel a: World Bank 2021, based on data from the Contaduría General de la Nación, Government of Uruguay, February 2020. Panel b: Original figure for this publication.

Related to this is the need for governments to devise pay schedules that incentivize officials to keep exerting effort to rise up the career ladder while also being aware that equity in pay is a key issue for some officials' motivation. To measure inequality due to differences in responsibilities and career level, we can analyze the pay scale compression of the government's units.<sup>5</sup> Higher wage compression (a smaller wage gap between management level and entry level) is associated with greater difficulty in motivating personnel

to progress through the public service because increased responsibility is not adequately compensated. For example, in the case of Uruguay (see figure 10.8a), wage compression in the central administration is low by international standards but varies greatly across ministries. Having low wage compression by international standards is good for equity, but the implications for civil servants' productivity and motivation are unclear. Low pay compression can generate positive attitudes across civil servants if responsibilities are also spread accordingly across the civil service, but it might also indicate that the salary structure is not sufficiently able to incentivize and reward workers' efforts or reward workers who have additional responsibilities.

### **Pay Inequity Based on Increasing Wage Components**

A good compensation system should allow the government to select high-quality candidates and offer incentives to align each public servant's interests with those of society. Desirable characteristics of a payment system include the ability to link wage gains with skills and performance and the transparency of the wage components. Having a large number of salary components can hinder transparency and generate inequalities. For example, in the case of Uruguay's central administration, there are 297 different salary components, of which 53 are "basic" and 244 are "personal."<sup>6</sup> Each entity has some discretion to define the compensation its employees receive, thereby reducing transparency and potentially creating payment inequalities. From figure 10.9, we can see that this discretion is reflected in the distribution of personal payments (figure 10.9, panel b), which, unlike the distribution of basic payments (figure 10.9, panel a), follows a nonstandard distribution. The nonstandard distribution of personal payments suggests both a lack of transparency and an unequal pay structure, based on the increase of payment line items.

### **Wage Inequality by Gender**

Gender equality is a key indicator of progress toward making the public sector workforce more diverse, representative, and innovative, and better able to provide public services that reflect citizens' needs. According to the OECD (2019), women are overrepresented in the public sector workforce of OECD countries.

However, this is not true across the globe; in fact, the Worldwide Bureaucracy Indicators show that public sector gender equity is correlated with country income (Mukhtarova, Baig, and Hasnain 2021). Part of the issue lies in providing similar levels of compensation for women and men where some systems discriminate against women. In some cases, the wage gap can discourage women from entering the civil service or applying for higher positions in an organization. In this sense, identifying potential gender wage gaps in the public sector is important to fostering the diversity of public employees. In figure 10.10, we analyze the gender wage gap in Uruguay's public sector workforce. The results suggest that overall, after controlling for working hours, age, type of contract, grade, tenure, and occupation, there is not a statistically significant gender wage gap, but this varies across ministries.

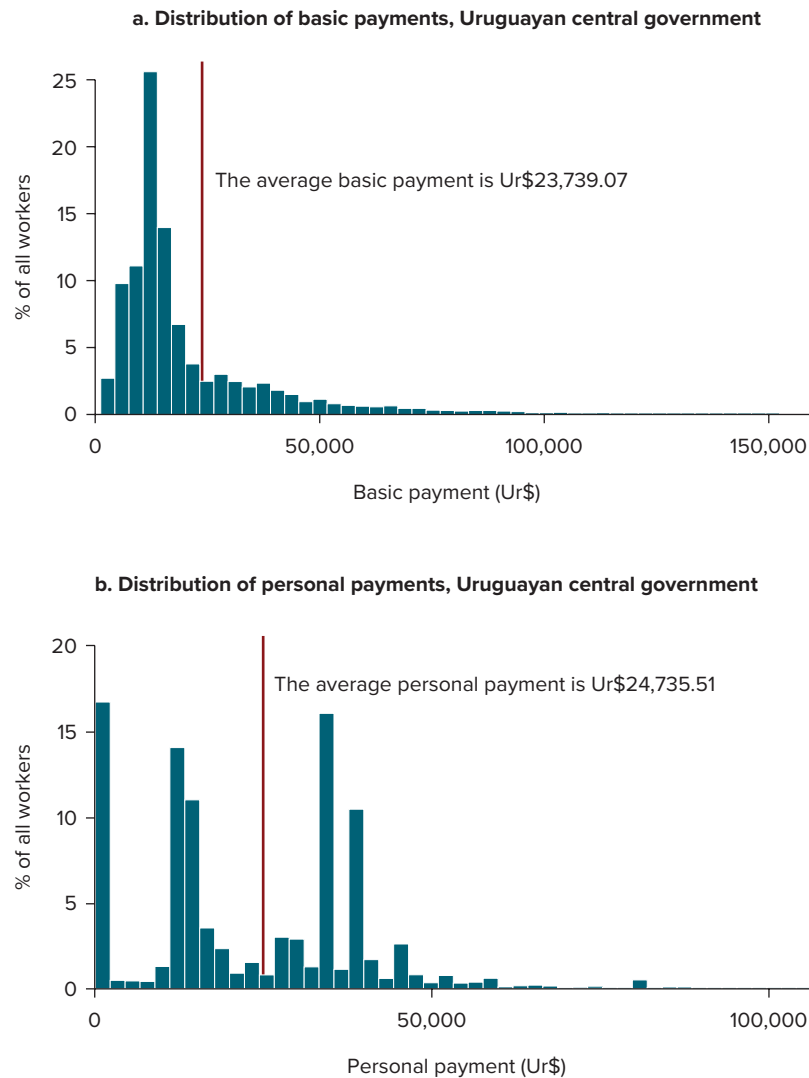
While there are many other margins of potential inequality in the service, and between the public service and the rest of society, these examples showcase the power of government microdata in identifying the extent and distribution of inequities across public administrations.

### **Fiscal Analysis**

Having considered what the wage bill is, how HRM can affect it, and how HR practices can affect the character and equity of the bureaucracy, we now turn our attention to how such practices can affect the fiscal health of a polity.

Setting compensation schemes, including initial wages and wage increases related to progressions and promotions, is a key tool to attract, retain, and motivate civil servants. But it can also be a cause of long-term fiscal imbalance because public sector employees usually work for more than 15 years.<sup>7</sup> For example, careers with high starting salaries may attract qualified candidates, but when combined with slow or small wage increases related to progressions, this can lead to demotivated public employees. In such a situation, a reform

**FIGURE 10.9 Inequity of Pay in Wage Components, Uruguay, 2019**

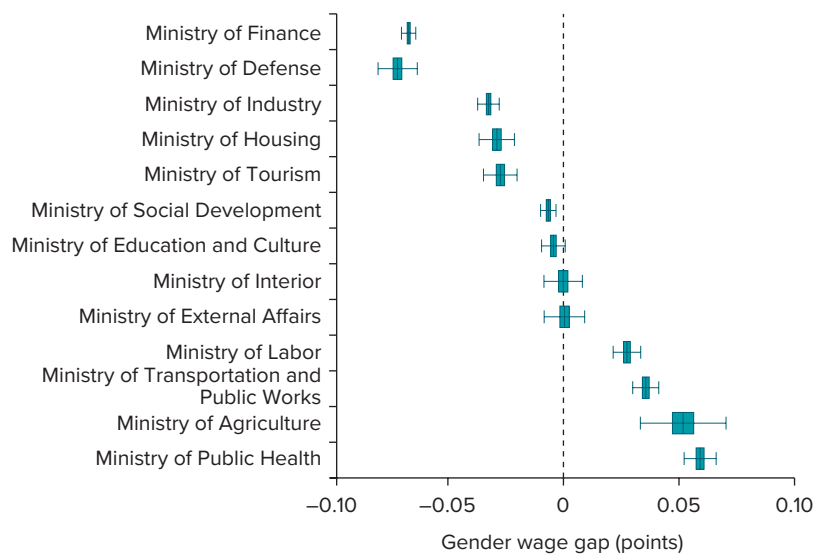


Source: World Bank 2021, based on data from the Contaduría General de la Nación, Government of Uruguay, February 2020.

that kept starting salary levels high and increased the additional pay related to progressions and promotions might cause the wage bill to be fiscally unsustainable. By understanding the fiscal impact of current career compensation schemes and potential reforms, policy makers can better manage the public sector's HR in the long term. In figure 10.11, we present examples of how these compensation features can be visualized. In the case of the Brazilian state of Mato Grosso (figure 10.11, panel b), we find that for some of the careers, the first three progressions more than double public employees' salaries.

Besides starting salaries and wage increases, another important piece of information for policy makers implementing strategic workforce planning is when public officials retire. Getting a clearer picture of when public employees retire is of critical importance for strategic workforce planning and fiscal planning. One needs to understand who will retire and when in order to plan successfully for incoming cohorts of civil servants, both in terms of their numbers and the competencies they will need. When large numbers of public servants are all due to retire at the same time, this can offer a window of opportunity for policy reform. For example, in the case of the Brazilian federal administration, the World Bank projected, using 2017 data, that 22 percent of public servants would have retired by 2022 and that 40 percent would have retired by 2030 (see figure 10.12). This situation presented an opportunity for administrative reform to restructure career systems

**FIGURE 10.10 Gender Gap, by Ministry, Government of Uruguay, 2010–20**



Source: World Bank 2021, based on data from the Contaduría General de la Nación, Government of Uruguay, 2010–20.

Note: The graph shows regression coefficients and 95 percent confidence intervals for the interaction between the female dummy and for the ministry fixed effect. Each point represents the average salary difference with respect to the Ministry of Public Health, after controlling for workers' characteristics.

and rationalize the number of existing civil servants in order to better plan, both in terms of the workforce and in fiscal terms. The use of HR microdata to undertake this analysis helped to inform the debate about civil service reform.<sup>8</sup>

## WAGE BILL PROJECTIONS

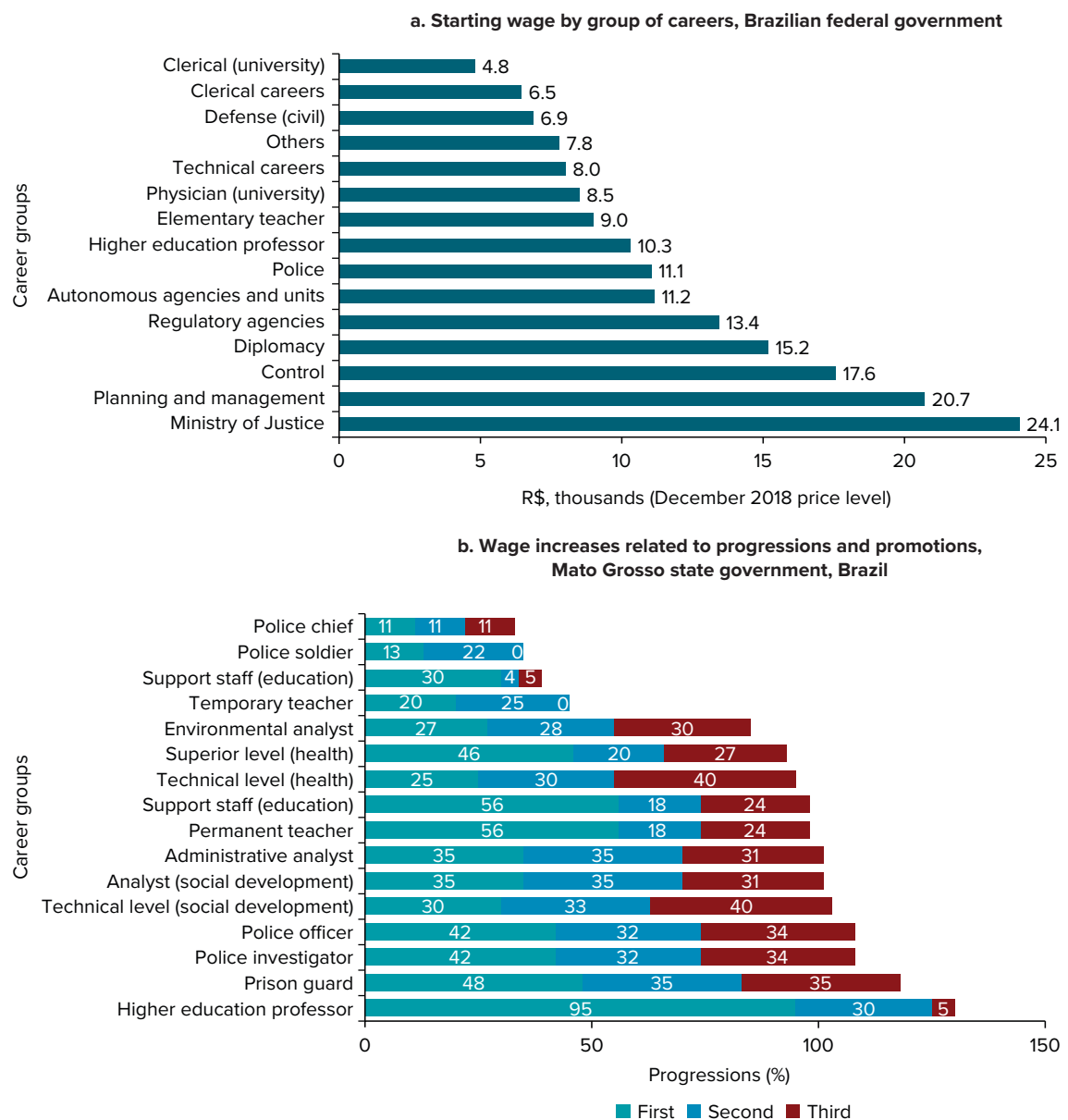
In this section, we present an HR-microdata-based model on the basis of the building blocks presented so far. With information about initial wages, wage increases related to career progressions, and expected dates of retirement, policy makers can project the expected fiscal impact of civil service reforms, the design of new careers, and fiscal consolidation policies. Using counterfactual scenarios can also help governments promote diversity and reduce inequalities in the civil service, fostering policies and services that better reflect citizens' needs.

Payroll and HRMIS microdata represent an important tool for the analysis of HR and fiscal policies. They can help policy makers lay out the trade-offs among competing policy objectives. For example, in the Brazilian state of Maranhão, the government sought to understand the fiscal impacts of wage increases for teachers along with increased recruitment of police personnel. By representing graphically the relevant statistics and comparing, first, the decreasing trend of the pupil-per-teacher ratio and its effect on the demand for new teachers and, second, levels of violence in the state when compared with its peers and the ratio of policemen per inhabitant, decision-makers obtained a more realistic picture of the available employment policies. In this section, we use some of the figures from the previous section to lay out the building blocks of a policy-oriented model for projecting wage bill expenditures. This model can help policy makers make difficult choices more transparent by showing the real costs and benefits of potential civil service reforms.

In practice, this is how we make the projections. First, we set up the HR microdata in a structure similar to the one described in the section “Payroll and HRMIS Microdata and Related Challenges”



**FIGURE 10.11 Career Types and Wages, by Career Group, Brazil**

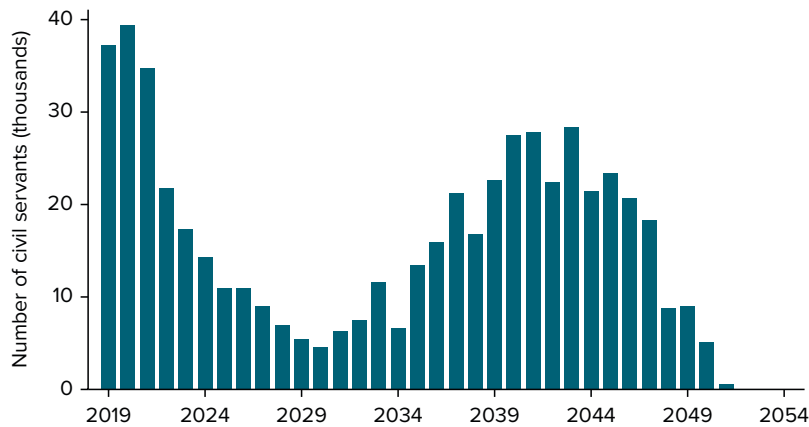


Source: Original figure for this publication.

and reported in table 10.1. Ideally, the database should contain payroll and HR information for the last 10 years. If monthly data are not available, it is possible to use a representative month of the year.<sup>2</sup> The wage bill data from previous years are then used to estimate some of the parameters of the model, and the most recent month or year data are used as a starting point for the projections.

Second, with the microdata set up, we group civil servants according to similarities in job position or common legal framework. The inputs of the government's HR managers are critical to this first part of the model because the number of groups set should both reflect the bulk of civil service careers and allow for more fine-grained policy options. In this sense, there is no "magic number" of groups; the number is based on context. In practice, we tend to cluster civil servants in a range of 5–20 groups.

**FIGURE 10.12 Retirement Projections, Brazilian Federal Government, 2019–54**



Source: Original figure for this publication based on Brazilian government data from 2008–18.

For example, in the case of some Brazilian states, we defined seven main groups: teachers, military police, investigative police, physicians, education support staff, health support staff, and others. These groups were defined to reflect the main public services Brazilian states are responsible for: public security, secondary education, and mid- to high-complexity health care. In another example, for the Brazilian federal government, we defined 15 career groups, which included university professors because Brazilian public universities are mostly federal.

Third, after setting the clusters of careers, we estimate some basic parameters for these groups using the microdata from previous years: the number of retirees by year for the following years, average tenure when retiring, initial wages, years between progressions or promotions, real increases in salaries related to progression or promotion legislation, real increases in salaries not related to progression or promotion legislation, and the attrition rate, which is the ratio of new hires to leavers. Some of these parameters were shown in the previous section. For example, figure 10.12 shows estimates for the number of retirees by year for the Brazilian federal government with data from 2008 to 2018.

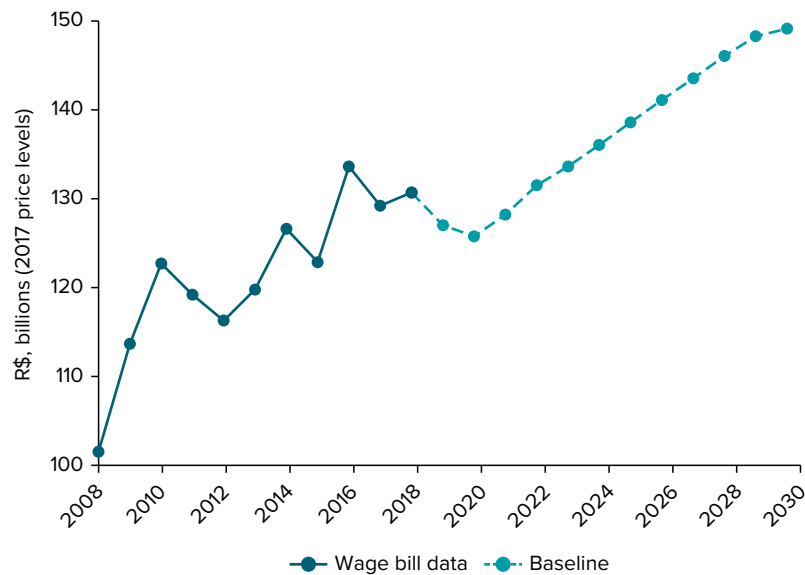
Fourth, we use the most recent month of the wage bill database and our estimated parameters to track the career path of current employees until their retirement and of the new civil servants who will replace retiring civil servants. Because of the fiscal nature of the model, the wage bill estimates tend to be less accurate for long-term projections. Based on experiences with LAC governments, we recommend using at most a 10-year span for projections. Using the estimated parameters, we come up with a baseline projection: the trajectory of wage bill expenditures assuming “business as usual,” as extrapolated from the data on past years. In other words, we project the expected wage bill spending if we assume the same wage increases as in past years, the same expected tenure before retirement, and the same replacement rate of new civil servants per retiring employee.

Finally, after making a baseline projection of the wage bill, we are able to simulate reforms that implement changes to the estimated parameters. For example, if the government wants to analyze the fiscal impacts of a reform that increases the recruitment of teachers, we simply change the rate of replacement of the career group of teachers. In another example, if the government wants to consolidate wage bill expenditures by freezing wages for the next two years, we change the parameter for salary increases that are not related to progressions or promotions. The list of potential policy scenarios includes hiring freezes or targeted pay increases for specific classes of employees. The model is meant to be flexible to adapt to the government’s needs so policy makers can test different reform options and hypotheses.

## Example from the Brazilian Federal Government

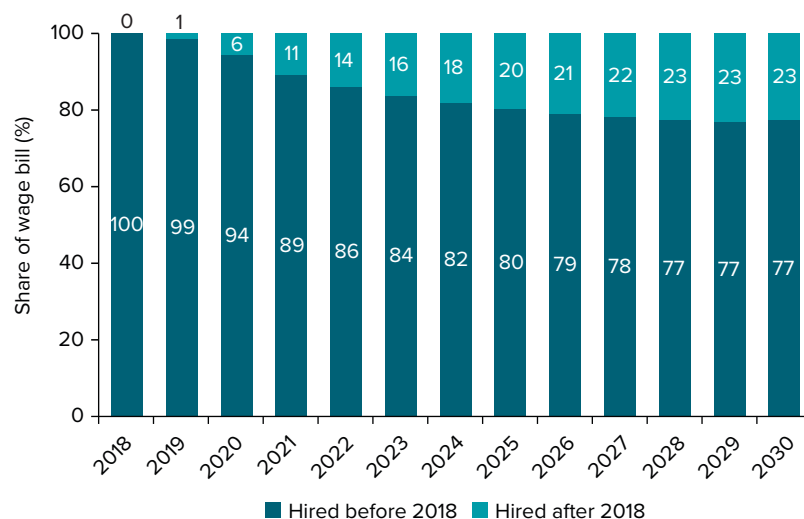
To exemplify the use of the model, in this section, we present wage bill projections for the Brazilian federal government for the period 2019–30, which were undertaken using HR microdata from 2008 to 2018. For example, figures 10.11, panel a and 10.12 in the previous section are graphical representations of the starting wages and the number of retirees by year, respectively. Figure 10.13 presents the baseline projection of the wage bill, and figure 10.14 provides a decomposition of the wage bill projection across current and new employees. Brazil is something of an outlier among LAC countries in that it has very high-quality administrative data, making it a good example of the more advanced types of analyses one can undertake with HR microdata once a comprehensive, centralized data collection system has been put in place.

**FIGURE 10.13** Baseline Wage Bill Projection, Brazilian Federal Government, 2008–30



Source: Original figure for this publication based on Brazilian government data from 2008–18.

**FIGURE 10.14** Decomposition of Wage Bill Projection between Current and New Employees, Brazilian Federal Government, 2018–30



Source: Original figure for this publication based on Brazilian government data from 2008–18.

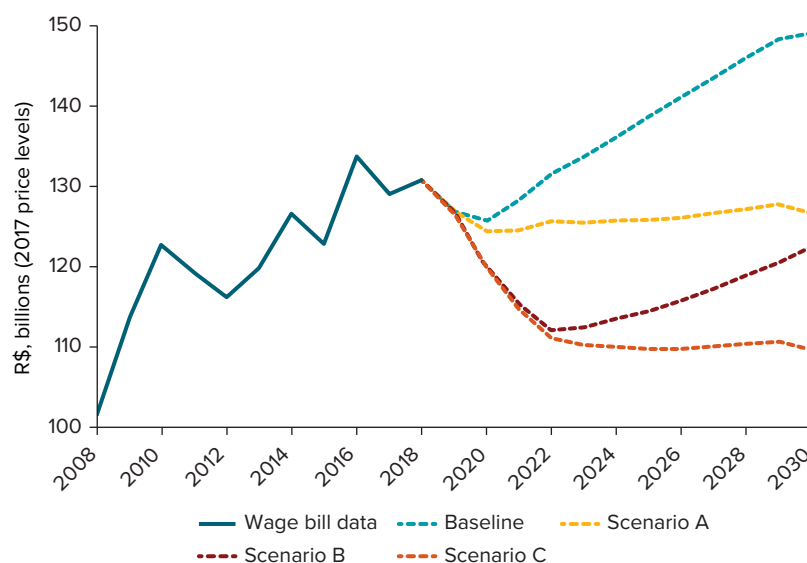
After projecting a baseline scenario for wage bill expenditures in the coming decade, we are able to compare it to different policy scenarios. To better organize reform options, we can separate them into pay-related and employment-related reforms. In the context of Brazil, the federal government's main objective was to simulate reforms that could lead to fiscal savings. We presented nine policy options, two of them related to employment reforms and the other seven related to pay policies. Based on these specific policies, we projected the following scenarios, each with a set of pay-related and employment-related policies:

- Scenario A: Replacement of 100 percent of retiring employees and no real salary increases for 10 years.
- Scenario B: Replacement of 90 percent of retiring employees and no nominal salary increases for the first three years.
- Scenario C: Replacement of 80 percent of retiring employees and no nominal salary increases for the first three years, and after that, no real salary increases for the next seven years.

Figure 10.15 provides a graphical presentation of the baseline projection along with the three outlined reform scenarios. In scenario A, a policy of no real wage increases is implemented starting in 2019. Since the y axis measures wage bill expenditures in real prices for 2017, the policy of correcting salaries only for inflation leads to an almost steady line in the chart. Scenarios B and C implement tighter policies, with a nominal freeze in salaries for the first three years starting in 2019, along with fewer hires of new employees to replace retiring civil servants. The bulk of the difference in savings between scenarios B and C comes from the years after 2022, in which scenario B returns to the baseline wage bill expenditures, while in scenario C, salaries are corrected for inflation.

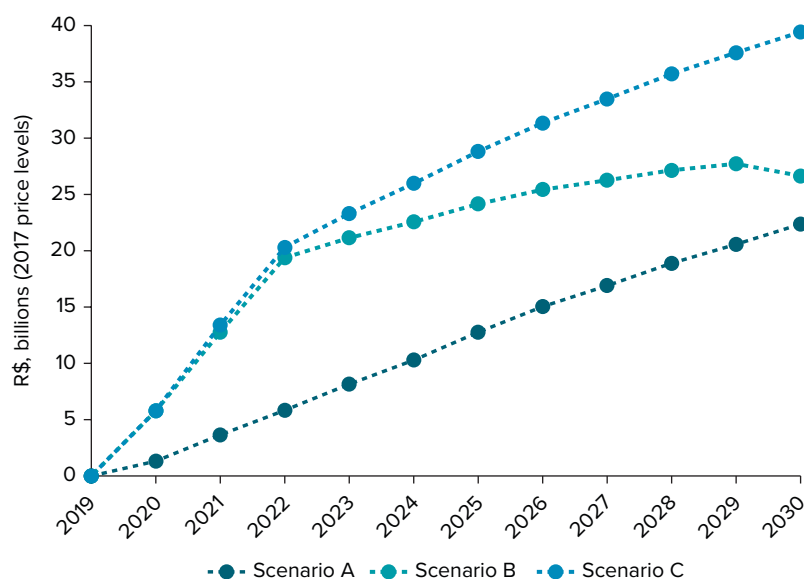
To put these different scenarios in perspective and compare their effectiveness in providing fiscal savings, we show in figure 10.16 the fiscal savings accumulated throughout the years in each reform scenario. In 2018, wage bill expenditures in the Brazilian federal civil service amounted to a total of R\$131 billion. The projections of the model used in this analysis indicate that in 2026, scenario A saves approximately 12 percent of the 2018 wage bill expenditures, scenario B saves 19 percent, and scenario C saves 24 percent. Besides these differences in total savings, in scenarios B and C, the government achieves larger savings in the short term while compensating with smaller savings after a few years, whereas, in scenario A, the total savings are spread out over the years.

**FIGURE 10.15 Wage Bill Projection and Policy Scenarios, Brazil, 2008–30**



Source: Original figure for this publication.

**FIGURE 10.16** Cumulative Fiscal Savings from Policy Scenarios, Brazil, 2019–30



Source: Original figure for this publication based on Brazilian government data from 2019.

Experimenting with combinations of policies before implementation to understand their fiscal impact has the potential to save a significant proportion of a government's wage bill. Similarly, such extrapolations can be extended to the descriptive analysis outlined in previous sections so governments can better understand how personnel policy reform will impact the character of the public service. With enough good-quality data, governments can leverage their SHRM and wage bill data for evidence-based planning of their fiscal expenditures and personnel dynamics into the future.

## CONCLUSION

We have presented a microdata-based approach for governments to improve their SHRM and develop realistic civil service compensation and employment strategies. We have also demonstrated how such strategies can allow policy makers to make better fiscal choices. We have used a series of examples from LAC countries to demonstrate how the use of relatively basic payroll and HRMIS statistics can help policy makers gain insight into the current and future state of their government's wage bill. We posit that this constitutes an important first step toward tapping the potential of existing bodies of payroll and HRMIS microdata that are currently underused. We believe that our approach can help policy makers make difficult decisions by breaking down the causes of problems and putting numbers to the ways in which certain policy choices will translate into longer-term consequences. On the basis of our experience using HR microdata for such analyses, we have a series of practical recommendations to make.

The first recommendation pertains to the collection of the data required to undertake the analyses we propose. Although, in theory, any government with an HRMIS should have access to these data, we know from our experience working with governments that extracting and cleaning these data can be a difficult task. As such, we recommend that where possible, governments centralize their HR data collection systems and render these data accessible to insights teams. If such data do not exist, even in a disparate fashion, we strongly advise governments to begin collecting, in a centralized manner, payroll and HRMIS microdata. If governments are able

to break down existing inter- and intradepartmental data silos and embed data analytics into their institutional culture, they stand to gain a much clearer idea of—among many other phenomena—the composition of their workforce, how to use their workforce more effectively, and how to plan, budget, and staff for future challenges. This is a central recommendation from our experience working with these microdata. As we laid out above, the quality and coverage of the data at one's disposal affect the usefulness of the analyses one can undertake and, consequently, the power of the insights one can gain.

The second recommendation is that the analysis of HR data be better integrated with fiscal planning. Our approach can both complement and help to bridge functional reviews and macroanalyses and, for this reason, can reconcile the fiscally oriented nature of macroanalyses with the detail of functional reviews. For this to be effective, however, governments must encourage civil servants from the treasury and HR department(s) to collaborate more closely. This could be achieved by allocating dedicated portions of civil servant workload (from both the treasury and the HR department) to the task of sharing and analyzing data in collaboration, or by creating dedicated interdepartmental roles to push forward and undertake the collection and analysis of HR microdata for SHRM. By better integrating HR data and wage bill planning, policy makers can also improve the services that are delivered to citizens. In the example we mentioned in the introduction, policy makers in Alagoas incorporated demographic changes into their projections of how many teachers to hire (given the falling pupil-per-teacher ratio caused by lower fertility rates) and were thereby able to identify an area in which they could achieve substantial savings and better target their HR strategy to hire different categories of civil servants that were not oversupplied. In this way, the state was able to provide better-quality services to its citizens by hiring civil servants in areas where greater personnel were needed, rather than in the education sector, where there was an excess of teachers.

The third recommendation relates to how political considerations can impede the implementation of successful SHRM and fiscal planning. We recommend that governments, in addition to centralizing HR data collection systems, seek to insulate certain aspects of planning offices' work from the ebb and flow of politics. This could go hand-in-hand with our second recommendation, to carve out explicit portfolios or roles dedicated to collecting and analyzing HR microdata, by ensuring that this work is undertaken by public servants reporting to an independent agency rather than to a minister.

All three recommendations pertain to how governments can better institutionalize SHRM and improve their analytical capabilities with data that should be relatively easy to collect and use. By developing a culture of centralizing and sharing such data—always anonymized, stored, and shared with full respect for employees' privacy and rights—governments can improve their ability to identify and resolve issues pertaining to the workforce and fiscal planning alike, as we have laid out. Moreover, such analyses are simple to undertake, meaning that governments can leverage these data through existing staff with even minimal data literacy, without hiring a significant number of data specialists. We hope we have illustrated the benefits of combining HRMIS and payroll data to inform SHRM and fiscal planning and that we have inspired practitioners to exploit these data's potential for more and better evidence-based policy making.

## NOTES

This chapter is based on technical support provided to several governments across Latin America. The team was led by Daniel Ortega Nieto. Our thanks go to Vivian Amorim, Paulo Antonacci, Francisco Lima Filho, Sara Brolhato de Oliveira, Alison Farias, and Raphael Bruce for their part in the work presented here. The findings, interpretations, and conclusions expressed in this chapter are entirely those of the authors.

1. The IMF, in fact, estimates that over 130 countries report comprehensive government finance statistics and that, on average, countries have about 25 years of data at their disposal (Gupta et al. 2016, 11).
2. Some governments also gather and record data on pensioners and survivors. Having this additional data can be useful, especially to improve the government's understanding of retirees' profiles and the overall fiscal impact of pensions. Given that this subject opens a whole set of new analyses, however, we do not comprehensively discuss the use of pension data in this chapter.



3. The studies that exist are limited analyses looking at very specific issues, often from the health care sector, with the notable exception of Colley and Price (2010), who examine the case of the Queensland public service.
4. Of the HRM professionals they surveyed, 47 percent reported engaging in little or no work-force planning for their municipalities, and only 11 percent reported that their municipalities had a centralized, formal workforce plan (Goodman, French, and Battaglio 2015, 148).
5. Wage compression is generally defined as the ratio between high earners and low earners in a specific organization. In this chapter, we define wage compression as the ratio between the 90th percentile and the 10th percentile of the wage distribution of the organization.
6. The salary structure in the public administration consists of multiple salary components, grouped into “basic” and “personal” components. Basic payments are determined based on the specific position (*plaza*), which represents the set of tasks, responsibilities, and working conditions associated with each civil servant, including *sueldos al grado* and *compensaciones al cargo*. All civil servants also receive personal payments, which are specific to each individual employee.
7. For example, a Brazilian federal government employee works for an average of 30 years before retiring.
8. See “Banco Mundial aponta urgência de uma reforma administrativa,” *Valor Econômico*, October 10, 2019, <https://valor.globo.com/brasil/noticia/2019/10/10/banco-mundial-aponta-urgencia-de-uma-reforma-administrativa.ghml>.
9. The “representative month” should allow for the extrapolation of monthly wage bill expenditures and the number of civil servants for the whole year.

## REFERENCES

- Anderson, Martin W. 2004. “The Metrics of Workforce Planning.” *Public Personnel Management* 33 (4): 363–78. <https://doi.org/10.1177/009102600403300402>.
- Choudhury, Enamul H. 2007. “Workforce Planning in Small Local Governments.” *Review of Public Personnel Administration* 27 (3): 264–80. <https://doi.org/10.1177/0734371X06297464>.
- Colley, Linda, and Robin Price. 2010. “Where Have All the Workers Gone? Exploring Public Sector Workforce Planning.” *Australian Journal of Public Administration* 69 (2): 202–13. <https://doi.org/10.1111/j.1467-8500.2010.00676.x>.
- Davenport, Thomas. 2019. “Is HR the Most Analytics-Driven Function?” *Harvard Business Review*, April 18, 2019. <https://hbr.org/2019/04/is-hr-the-most-analytics-driven-function>.
- Dychtwald, Ken, Tamara Erickson, and Bob Morison. 2004. “It’s Time to Retire Retirement.” *Public Policy and Aging Report* 14 (3): 1–28. <https://doi.org/10.1093/ppar/14.3.1>.
- Frank, Howard A., and Yongfeng Zhao. 2009. “Determinants of Local Government Revenue Forecasting Practice: Empirical Evidence from Florida.” *Journal of Public Budgeting, Accounting & Financial Management* 21 (1): 17–35. <https://doi.org/10.1108/JPBAFM-21-01-2009-B002>.
- French, P. Edward, and David H. Folz. 2004. “Executive Behavior and Decision Making in Small U.S. Cities.” *The American Review of Public Administration* 34 (1): 52–66. <https://doi.org/10.1177/0275074003259186>.
- GAO (US General Accounting Office/Government Accountability Office). 2001a. *Federal Employee Retirements: Expected Increase over the Next 5 Years Illustrates Need for Workforce Planning*. Report to the Chairman, Subcommittee on Civil Service and Agency Organization, Committee on Government Reform, House of Representatives, GAO-01-509. Washington, DC: US General Accounting Office. <https://www.gao.gov/assets/gao-01-509.pdf>.
- GAO (US General Accounting Office/Government Accountability Office). 2001b. *High-Risk Series: An Update*. GAO-01-263. Washington, DC: US General Accounting Office. <https://www.gao.gov/assets/gao-01-263.pdf>.
- GAO (US General Accounting Office/Government Accountability Office). 2021. *High-Risk Series: Dedicated Leadership Needed to Address Limited Progress in Most High-Risk Areas*. GAO-21-119SP. Washington, DC: US Government Accountability Office. <https://www.gao.gov/products/gao-21-119sp>.
- Goodman, Doug, P. Edward French, and R. Paul Battaglio Jr. 2015. “Determinants of Local Government Workforce Planning.” *The American Review of Public Administration* 45 (2): 135–52. <https://doi.org/10.1177/0275074013486179>.
- Gupta, Sanjeev, David Coady, Manal Fouad, Richard Hughes, Mercedes Garcia-Escribano, Teresa Currstine, Chadi Abdallah, et al. 2016. “Managing Government Compensation and Employment-Institutions, Policies, and Reform Challenges.” IMF Policy Paper, April 8, 2016, International Monetary Fund, Washington, DC.
- Harborne, Bernard, Paul M. Bisca, and William Dorotinsky, eds. 2017. *Securing Development: Public Finance and the Security Sector*. Washington, DC: World Bank. <http://hdl.handle.net/10986/25138>.
- Hasnain, Zahid, Nick Manning, and Jan Henryk Pierskalla. 2014. “The Promise of Performance Pay? Reasons for Caution in Policy Prescriptions in the Core Civil Service.” *World Bank Research Observer* 29 (2): 235–64. <https://doi.org/10.1093/wbro/lku001>.

- Hasnain, Zahid, Daniel Oliver Rogger, Daniel John Walker, Kerenssa Mayo Kay, and Rong Shi. 2019. *Innovating Bureaucracy for a More Capable Government*. Washington, DC: World Bank. <http://documents.worldbank.org/curated/en/24989154999073918/Innovating-Bureaucracy-for-a-More-Capable-Government>.
- Huerta Melchor, Oscar. 2013. "The Government Workforce of the Future: Innovation in Strategic Workforce Planning in OECD Countries." OECD Working Papers on Public Governance 21, OECD Publishing, Paris. <https://doi.org/10.1787/5k487727gwb-en>.
- Jacobson, Willow S. 2009. "Planning for Today and Tomorrow: Workforce Planning." In *Public Human Resource Management: Problems and Prospects*, edited by Steven W. Hays, Richard C. Kearney, and Jerrell D. Cogburn, 5th ed., 179–202. New York: Longman.
- Jacobson, Willow S. 2010. "Preparing for Tomorrow: A Case Study of Workforce Planning in North Carolina Municipal Governments." *Public Personnel Management* 39 (4): 353–77. <https://doi.org/10.1177/009102601003900404>.
- Kavanagh, Shayne C., and Daniel W. Williams. 2016. *Informed Decision-Making through Forecasting: A Practitioner's Guide to Government Revenue Analysis*. Chicago: Government Finance Officers Association.
- Kiyonaga, Nancy B. 2004. "Today Is the Tomorrow You Worried about Yesterday: Meeting the Challenges of a Changing Workforce." *Public Personnel Management* 33 (4): 357–61. <https://doi.org/10.1177/009102600403300401>.
- Kong, Dongsung. 2007. "Local Government Revenue Forecasting: The California County Experience." *Journal of Public Budgeting, Accounting & Financial Management* 19 (2): 178–99. <https://doi.org/10.1108/JPBAFM-19-02-2007-B003>.
- Mukhtarova, Turkan, Faisal A. Baig, and Zahid Hasnain. 2021. "Five Facts on Gender Equity in the Public Sector." *Governance for Development* (blog). *World Bank Blogs*, September 27, 2021. <https://blogs.worldbank.org/governance/five-facts-gender-equity-public-sector>.
- OECD (Organisation for Economic Co-operation and Development). 2007. *Ageing and the Public Service: Human Resource Challenges*. Paris: OECD Publishing. <https://doi.org/10.1787/9789264029712-en>.
- OECD (Organisation for Economic Co-operation and Development). 2019. "Gender Equality in Public Sector Employment." In *Government at a Glance 2019*, 88–89. Paris: OECD Publishing. <https://doi.org/10.1787/9735a9f2-en>.
- Pynes, Joan E. 2004. "The Implementation of Workforce and Succession Planning in the Public Sector." *Public Personnel Management* 33 (4): 389–404. <https://doi.org/10.1177/009102600403300404>.
- Pynes, Joan E. 2009. "Strategic Human Resources Management." In *Public Human Resource Management: Problems and Prospects*, edited by Steven W. Hays, Richard C. Kearney, and Jerrell D. Cogburn, 5th ed., 95–106. New York: Longman.
- Reitano, Vincent. 2019. "Government and Nonprofit Personnel Forecasting." In *The Palgrave Handbook of Government Budget Forecasting*, edited by Daniel Williams and Thad Calabrese, 361–76. Cham, Switzerland: Springer. <https://doi.org/10.1007/978-3-030-18195-618>.
- Riccucci, Norma M., Gregg G. Van Ryzin, and Cecilia F. Lavena. 2014. "Representative Bureaucracy in Policing: Does It Increase Perceived Legitimacy?" *Journal of Public Administration Research and Theory* 24 (3): 537–51. <https://doi.org/10.1093/jopart/muu006>.
- Rogers, James, and C. M. Naeve. 1989. "One Small Step towards Better Government." *Public Utilities Fortnightly* 3: 9–12.
- Selden, Sally Coleman. 2009. *Human Capital: Tools and Strategies for the Public Sector*. Washington, DC: CQ Press.
- Selden, Sally Coleman, and Willow Jacobson. 2007. "Government's Largest Investment: Human Resource Management in States, Counties, and Cities." In *In Pursuit of Performance: Management Systems in State and Local Government*, edited by Patricia W. Ingraham, 82–116. Baltimore, MD: The Johns Hopkins University Press.
- Somani, Ravi. 2021. "The Returns to Higher Education and Public Employment." *World Development* 144: 105471. <https://doi.org/10.1016/j.worlddev.2021.105471>.
- Theobald, Nick A., and Donald P. Haider-Markel. 2009. "Race, Bureaucracy, and Symbolic Representation: Interactions between Citizens and Police." *Journal of Public Administration Research and Theory* 19 (2): 409–26. <https://doi.org/10.1093/jopart/mun006>.
- Van Ryzin, Gregg G., Norma M. Riccucci, and Huafang Li. 2017. "Representative Bureaucracy and Its Symbolic Effect on Citizens: A Conceptual Replication." *Public Management Review* 19 (9): 1365–79. <https://doi.org/10.1080/14719037.2016.1195009>.
- Wilkerson, Brian. 2007. *Effective Succession Planning in the Public Sector*. Arlington, VA: Watson Wyatt Worldwide.
- Wong, John D. 1995. "Local Government Revenue Forecasting: Using Regression and Econometric Revenue Forecasting in a Medium-Sized City." *Journal of Public Budgeting, Accounting & Financial Management* 7 (3): 315–35. <https://doi.org/10.1108/JPBAFM-07-03-1995-B001>.
- World Bank. 2019. *Gestão de Pessoas e Folha de Pagamentos no Setor Público Brasileiro: O Que os Dados Dizem*. Washington, DC: World Bank.
- World Bank. 2021. *Uruguay Public Expenditure Review: Civil Service Diagnosis*. Washington, DC: World Bank.
- World Bank. 2022. *Subnational Civil Servant Pension Schemes in Brazil: Context, History, and Lessons of Reform*. Washington, DC: World Bank. <https://openknowledge.worldbank.org/handle/10986/37267>.



## CHAPTER 11

# Government Analytics Using Expenditure Data

*Moritz Piatti-Fünfkirchen, James Brumby, and Ali Hashim*

### SUMMARY

Government expenditure data hold enormous potential to inform policy around the functioning of public administration. Their appropriate use for government analytics can help strengthen the accountability, effectiveness, efficiency, and quality of public spending. This chapter reviews where expenditure data come from, how they should be defined, and what attributes make for good-quality expenditure data. The chapter offers policy makers an approach to reviewing the adequacy and use of government expenditure data as a means to strengthen the effectiveness of government analytics that builds on these data. Case studies are used to illustrate how this can be done.

### ANALYTICS IN PRACTICE

- Be clear about how *expenditure* is defined. *Expenditure* is a term that is often interpreted and used loosely. This can lead to confusion and misunderstandings in analytical assessments. The literature around public expenditure data is clear on a series of standard definitions and offers guidance as to their application. It is recommended to take advantage of this, where feasible, and to minimize the ambiguous use of terms, to the extent possible.
- Understand and document the origins of government expenditure data. Government expenditure data have enormous potential to inform the accountability, efficiency, impact, and equity of operations. It is important to understand and document how transactions across spending items in government are created, what control protocols they are subject to, how this information is stored, and how microdata are aggregated for analysis.
- Do not take data at face value. The usefulness of analysis from government expenditure data hinges upon the quality of the underlying microdata. It is recommended that the origins of government expenditure

---

Moritz Piatti-Fünfkirchen is a senior economist and James Brumby is a senior adviser at the World Bank. Ali Hashim is an independent consultant.

microdata be periodically reviewed for data provenance and integrity, comprehensiveness, usefulness, consistency, and stability. It is recommended that such work be publicly disclosed, to the extent possible. This can be used as a baseline upon which a reform program can be built to address deficiencies.

- Take a microdata-driven approach to expenditure analysis. An analysis of microlevel expenditure data can offer data-driven and objective insights into expenditure management practices and the impacts of expenditure policy. From this analysis, a government expenditure profile can be derived, which shows where large transactions with high fiduciary risks are taking place, how these compare to low-value transactions at points of service delivery, and where expenditure policy intentions are not being converted into the desired impact. Such analysis can offer operational insights for better expenditure management that serves expenditure control and service delivery objectives.

## INTRODUCTION

Public resources are scarce. Increasingly, competing demands on the public purse make prudent and evidence-based expenditure decisions ever more important. This requires, among other things, accurate and timely government expenditure data. Government expenditure data are central to the social contract between society and elected officials. They provide an important basis for accountability, insights into whether resources are being used for budgeted priorities, and assessments of whether spending is sustainable and equitable.

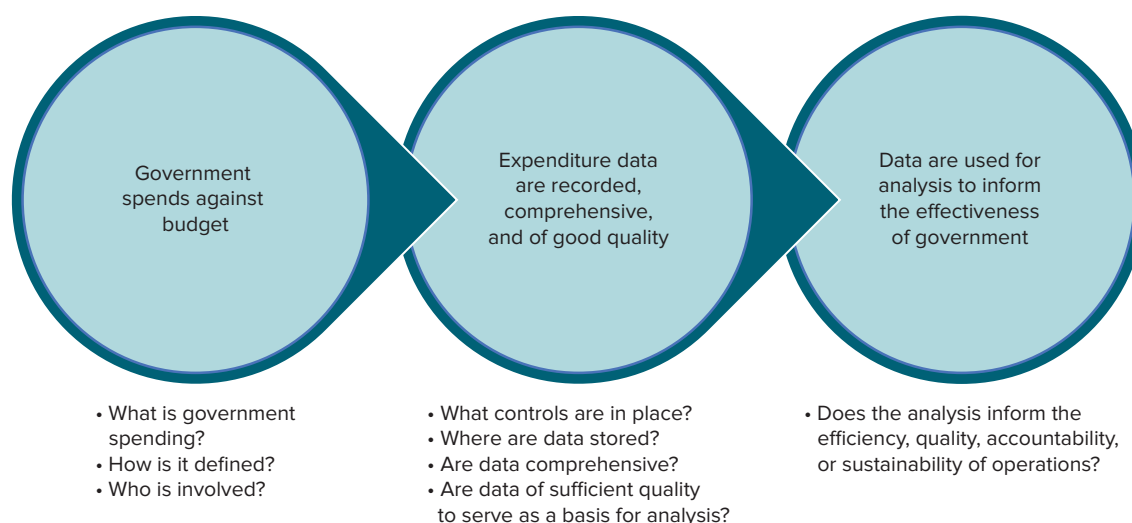
Expenditure data are central to assessing the fiscal health of a country (Burnside 2004, 2005) and necessary for debt sustainability analyses (Baldacci and Fletcher 2004; Di Bella 2008). These are core elements of government operations and often accompany World Bank Public Expenditure Reviews (PERs) or International Monetary Fund (IMF) Article IV reports. A government's commitment to deficit targets can, for example, be measured by identifying whether large expenditure items in the budget are subject to necessary internal controls (Piatti, Hashim, and Wescott 2017).<sup>1</sup>

Expenditure data can be used to assess whether spending is effective and efficient. They thus provide information about the functioning of public administration. Such assessments can be made at the very granular level of a department, unit, or even project. For example, budget data might indicate that a construction project is not disbursing as quickly as projected, limiting its progress. This is indicative of potential problems in expenditure management. Such analysis of government expenditure data is used by the executive branch, audit institutions, the legislative branch, and civil society to offer insights into the quality of the administration.

To get to the stage where expenditure data can effectively inform the functioning of government, a simplified, three-step logframe indicates two preparatory stages (see the figure 11.1; a detailed exposition of the stages in the creation of expenditure data is provided in appendix C). First, there needs to be clarity about what government spending means, who it involves, and what it covers. Second, there is the question of how spending is executed (for example, what controls it is subject to), where data are stored, how comprehensive they are, and what the quality of the data is. Third, high-quality expenditure data with high coverage can lend themselves to analyses that inform the effectiveness of government.

This logframe illustrates that the value of any analysis is a function of the quality of the underlying expenditure data. It is therefore important for practitioners to reflect carefully on how government expenditures are defined and where they come from and to critically assess the quality of government expenditure microdata (or transactions data). While there is a lot of guidance on how to analyze government expenditure data (step 3 in figure 11.1), the literature is relatively silent on how to assess the quality of expenditure data, as well as on how poor data may affect the validity of the conclusions drawn from such analyses (step 2). Despite clear guidance on definitions and coverage (step 1), the term *government expenditure* continues to be used to imply a multitude of different concepts that are frequently not well communicated, leading to confusion among analysts.

**FIGURE 11.1** Use of Government Expenditure Data for Government Analytics



Source: Original figure for this publication.

This chapter walks through each of these steps as follows. It starts by discussing issues related to defining government expenditure data (step 1). It then reviews the attributes of good government expenditure data and makes observations about how data can be strengthened and made more reliable and useful for analysis. The chapter highlights the importance of data provenance and integrity, comprehensiveness, usefulness, consistency, and stability as critical attributes (step 2). Examples of how to pursue these characteristics are provided and illustrated through case studies. These case studies indicate that deficiencies in any of these characteristics constitute a risk to the ability of analysts to use these data to inform an understanding of government functioning (step 3).

## WHAT ARE GOVERNMENT EXPENDITURE DATA?

Despite the centrality of government expenditure, definitional issues remain. The term *government expenditure* is often used with liberty among practitioners and analysts. For example, *budget*, *commitment*, and *expenditure data* are sometimes used interchangeably. Further, there is often insufficient differentiation between cash and accrual concepts. Suffice it to say, it is important to be clear and precise when using the term *expenditure* to allow for an effective dialogue and comparability over time and across countries.

*Expenditure* is defined by the Organisation for Economic Co-operation and Development (OECD) as “the cost of goods and services acquired, regardless of the timing of related payments.” Expenditures are, therefore, different from cash payments. Instead, “expenditures on goods and services occur at the times when buyers incur liabilities to sellers, i.e. when either (a) the ownership of the goods and services concerned is transferred from the seller to the new owner; or (b) when delivery of the goods and services is completed to the satisfaction of the consumer.” Conversely, the term *expense* “defines the set of transaction flows that reduce net worth over the accounting period” (Allen and Tommasi 2001, 452). This distinction reveals that while an *expenditure* may result in the acquisition of a capital item, an *expense* will apply to the use (depreciation) or care (maintenance) of the item.



Governments spend money as a result of a series of economic relationships. The main ones are as follows:

- To pay wages, salaries, and other emoluments for labor
- To purchase goods and services that are then used in the production of government outputs
- To purchase assets
- To transfer resources (unrequited) to other levels of government, households, or firms
- To meet the cost of servicing debts
- For various other purposes, such as meeting legal claims.

Expenses can be incurred for events that do not involve a same-time transaction—for instance, changes in the estimate of unfunded liabilities associated with government pensions or the impairment of an asset through its use (depreciation). The distinction considers an *expenditure* to acquire goods, with the *expense* occurring when the goods are used.

All expenditure transactions that are routed through a *financial management information system* (FMIS) are reflected in the government's accounts, or *general ledger*, without exception, providing a comprehensive data source for analysis. Each transaction originates from a spending unit within the government, ensuring that each transaction can be mapped to a particular office. Because these transactions must be executed against the index of allowed payments agreed upon in the budget, or *chart of accounts* (COA), and must specify the amount, the details of the payee (including the recipient's account number and the time of the transaction) are a natural component of expenditure data. Depending on the level of detail of the COA, the transaction may capture the source of funds, the organizational code, the purpose of the expenditure (economic classification or line item), the jurisdiction in which the transaction happened, and the program or subprogram it related to. The format that the data structure of financial transactions in an FMIS typically takes is given in table 11.1.

Transactions may also be processed manually, outside the FMIS, and then posted manually to the general ledger. These transactions are thus not automatically subject to the same set of FMIS internal controls, and the same level of transaction detail may not be available. Furthermore, these transactions may be aggregated and posted in bulk, making the desired analysis of microdata difficult.

## ATTRIBUTES OF GOOD-QUALITY EXPENDITURE DATA

Understanding definitional nuances and assessing the quality and credibility of the underlying microdata both benefit from an understanding of the government information system's architecture. There are multiple functions, processes, agencies, and associated systems at play. These include processes and systems for

**TABLE 11.1** Example of Expenditure Data, by Transactions

Transaction ID	Time stamp (date)	Chart of accounts segment					Amount	Payee (and account number)
		Source of funds	Organization code	Purpose code (line item)	Location code	Program/ subprogram code		
Transaction 1								
Transaction 2								
...								
Transaction <i>n</i>								

Source: Hashim et al. 2019.

macroeconomic forecasting; budget preparation systems; treasury systems; establishment control, payroll, and pension systems; tax and customs systems; debt management systems; and auditing systems. Together, these systems represent the information architecture for government fiscal management, underpinning government expenditure management, and are the basis for government expenditure data. A detailed account of these systems is provided by Allen and Tommasi (2001), Hashim (2014), and Schiavo-Campo (2017). Carefully designed, functional processes supported by adequate systems, and the good utilization of those systems, will yield good-quality government expenditure data that can be analyzed to inform policy. Weaknesses in any one of these processes, by contrast, will undermine the quality of expenditure data.

Spending, and the production of expenditure data, follows a process. Once the budget is authorized and apportioned to spending units, commitments can be made. The receipt of goods or services then needs to be verified before a payment order can be initiated. Bills are paid upon the receipt of the payment order. This is then accounted for against the full COA and provides the basis for expenditure data (figure 11.2). A full account of these processes, including differentiation by colonial history, is offered by Potter and Diamond (1999) and Shah (2007). Further details are provided in appendix C.

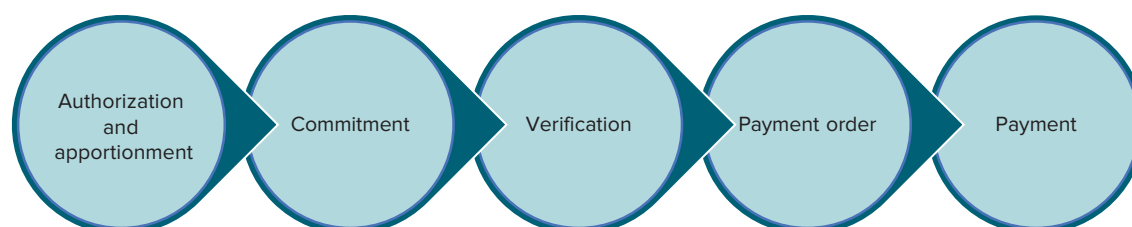
There are numerous agencies and processes involved in the production of government expenditure data. The quality and credibility of these data depend on how well the data production process is implemented across these agencies and processes. This chapter identifies five principles in the data production process that can help assess the adequacy of the data for further analysis. Each adds to the likelihood that expenditure is reliable. Unlike personnel data, discussed in chapter 10 of this *Handbook*, it is more difficult to present a “ladder” of quality for expenditure data. These five principles interact to determine the utility of the resulting data. For example, an incomplete data set that focuses on the 50 percent largest expenditure items is likely to cover a substantial portion of total expenditure. These principles should be seen as underpinning a high-quality expenditure-data-generating system, and what will yield the greatest improvement in overall quality will be specific to contexts and almost to data points.

## Data Provenance and Integrity

Expenditure data are useful for analysis if there is confidence in their integrity. *Data provenance*—the documentation of where data come from and the processes by which they were produced—is necessary to have this confidence. There should be a clear sense of what systems data have come from, who was involved in the production of the data, and where the data are stored. Internal controls for systems should ensure data provenance and integrity. If systems are used, controls are applied, and data are immutable (or there is a clear trail in any changes), there can be confidence in data integrity. The use of an FMIS, for example, should guarantee data provenance and integrity—if transactions were executed through the system and, therefore, were subject to FMIS internal controls.

If expenditures are not routed through the dedicated government system, data provenance and integrity are more difficult to guarantee (Chami, Espinoza, and Montiel 2021; Milante and Woolcock 2021).

**FIGURE 11.2** Stages in the Execution Process That Create Government Expenditure Data



Sources: Adapted from Potter and Diamond 1999; Shah 2007.

As evidenced in the literature, in many lower- and middle-income countries, such as Ghana, Pakistan, and Zambia, FMIS coverage remains limited (European Commission and IEG 2017; Hashim, Farooq, and Piatti-Fünfkirchen 2020; Hashim and Piatti-Fünfkirchen 2018; Piatti-Fünfkirchen 2016). In some instances, it may be for good reason that systems are not used. There may, for example, be information and communication technology limitations, a lack of access to banking services, or human capacity constraints in remote areas. In other instances, not using systems may be a purposeful choice to avoid said controls and, thus, clear data provenance. In either case, transactions posted manually to the general ledger are more susceptible to manipulation. In these cases, confidence that the reported expenditure reflects actual spending is likely to require a costly ex post audit. An example of how FMIS utilization helped resolve a major corruption episode in Malawi is illustrated in box 11.1.

Mixing good-quality data with questionable data calls into question the credibility of the entire data set because the provenance is not accurately tracked for each and every transaction. It is therefore important to understand which transactions were processed by the FMIS and where transactions that were *not* processed by the FMIS come from, as well as whether their integrity can be assured (Hashim and Piatti-Fünfkirchen 2018).

## Comprehensiveness

*Comprehensiveness* is defined with respect to the reporting entity. If the desire is to review expenditure performance across the entire government sector, then this requires data to be comprehensive across levels of government, sources, and, preferably, time. Data comprehensiveness is complicated by mismatches between the institutional setup of a government (consolidated fund or public account) and the definition of *government*, which may include bodies that are outside the consolidated fund or a jurisdictional structure that clearly separates the levels of government, as is the case with federations. What is important for the

### BOX 11.1 How Utilization of a Financial Management Information System in Malawi Supported Data Provenance and Helped Resolve a Major Corruption Episode

Adequate utilization of the financial management information system (FMIS) in Malawi helped ensure that most spending was transacted through the system and that expenditure data were recorded and stored on the general ledger. During a major corruption episode, data provenance—ensured through the FMIS—enabled the tracing of the transactions and events (Baker Tilly Business Services Limited 2014; Bridges and Woolcock 2017; Hashim and Piatti-Fünfkirchen 2018; World Bank 2016a).<sup>a</sup> This, consequently, allowed authorities to follow up, identify collusion, and prosecute. In an environment where transactions are posted manually, it is easier to tamper with records, which undermines the integrity of the data and, thereby, the ability of authorities to ensure accountability. The increasing penetration of banking innovations, such as mobile money or smart cards, offers governments the ability to make electronic transfers or payments even in remote areas (where access to conventional banking services is unavailable), which leave a digital footprint. Even if the FMIS does not execute the transaction, posting these onto the ledger would strengthen data provenance, transparency, and accountability (Piatti-Fünfkirchen, Hashim, and Farooq 2019). This practice has been widely applied for cash transfers in some countries, such as Kenya, Rwanda, Uganda, and Zambia.

a. There was a misconception that the FMIS was at fault for not preventing the misappropriation of funds. Collusion among main stakeholders was a human, not a system, error. The FMIS should be credited with ensuring data provenance and supporting prosecution in due course (World Bank 2016a).

integrity of the analysis is that the comprehensiveness of the reporting entity can be established. If transaction coverage for a reporting entity is not comprehensive, the findings will reflect this and may fall short of their intended purpose.

Guidance on how the *public sector* is defined is available in the IMF's *Government Finance Statistics Manual* (IMF 2014) and in the *Handbook of International Public Sector Accounting Pronouncements* (IFAC 2022). However, the application of this guidance can vary across countries and institutions, making meaningful cross-country comparisons for reporting purposes difficult (Barton 2011; Challen and Jeffery 2003; Chan 2006). In some cases, the public sector may be narrowly defined, with asymmetrical representation of the general government and public corporations. Reporting on the public sector may be partial—even at the aggregate level—in sensitive sectors, such as defense, the police force, or space programs, or it may be partial due to the funding source. The budget sector (and within it, the various forms of annual and standing appropriations), the public account, the general government sector, and the broader public sector may all justifiably be the entity of interest for some analyses; what is important is to understand *why* a given entity is the entity in question and what activity it excludes relative to a more relevant entity. For example, when seeking to communicate restraint, a central government may highlight its total spending, including transfers to lower levels of government, whereas a subnational government may wish to distinguish the central government's spending for its own purposes and programs from the funds it transfers to lower levels of government. This distinction may reveal that a large portion of the central government's "restraint" comes from cuts to others' programs rather than restraint in the delivery of the central government's own work.

The comprehensiveness of spending can suffer from a lack of transparency on debt. As countries are increasingly indebted, this factor becomes increasingly important. Drawing from new databases and surveys, Rivetti (2021) finds that nearly 40 percent of low-income developing countries (LIDCs) have never published debt data on their websites or have not updated their data in the past two years. When debt data are available, they tend to be limited to central government loans and securities, excluding other public sector components and debt instruments. For some LIDCs, debt data disclosed across various sources show variations equivalent to as much as 30 percent of a country's gross domestic product—often because of differing definitions and standards and recording errors. Data in the debt management system should comprehensively reflect all loans and liabilities and actual debt servicing requirements. Actual spending should be comprehensively reflected in the FMIS. Even here, it is important that expenditure controls apply in order to avoid expensive short-term borrowing that has not been budgeted for (Hashim and Piatti-Fünfkirchen 2018).

Comprehensiveness is equally important for sector expenditure analysis. Health spending, for example, is frequently benchmarked against the Abuja Declaration target of 15 percent of the government budget (African Union 2001). However, how well one can proxy this indicator depends on a country's ability to credibly populate the numerator (health spending) and the denominator (general government spending), and the literature has shown this to be difficult (Piatti-Fünfkirchen, Lindelow, and Yoo 2018). Estimating health spending typically goes beyond just one reporting entity. Thus, reporting comprehensively on all health spending, including relevant off-budgetary funds, the use of internally generated funds (for example, user fees), development partners (Piatti-Fünfkirchen, Hashim, et al. 2021), and the use of tax expenditures (Lowry 2016) becomes important in a consideration of the resources dedicated to the sector.<sup>2</sup> Estimating the comprehensiveness of the denominator, then, is complicated by all the factors outlined above.

Comprehensiveness also requires comprehensive reporting over time. A timing mismatch between receiving a good or service and the payment of cash can lead to the creation of *payment arrears*—a liability that is past due. Accurate reporting on such arrears is important for comprehensiveness. The 2020 Public Expenditure and Financial Accountability (PEFA) global report notes that countries' stock of expenditure arrears was around 10 percent of total expenditure, well above the 2 percent considered good practice (PEFA 2020).<sup>3</sup> If these are not adequately reported, any expenditure analysis will be inaccurate. The PEFA indicator for expenditure arrears (PI-22) is, however, one of the poorest-rated indicators in the framework (PEFA 2022). This is despite the fact that adequate expenditure controls tend to be in place, suggesting that these are frequently bypassed, leading to the aforementioned data provenance and integrity concerns.

Finally, many aspects of government expenditure are driven by trends that extend beyond the annual time cycle that generally applies to budgets. For example, changing demographics mean that societal needs for services such as education and health care change over time. Similarly, differences in timing between the creation of an obligation (such as a pension) and the payment of that obligation mean that it is important to consider the multiannual nature of spending to get a more complete picture. Spending (or not spending) today may create important obligations over time. Consumption- and investment-related spending are fundamentally different and need to be recognized as such. Yet annual expenditure reporting requirements tend to take a short-term perspective regardless of the nature of spending. Further, what is captured as expenditure may be influenced by what is not captured, which may nevertheless impact what is left to be performed by functions requiring expenditure—for example, regulation and its associated compliance and tax expenditures. If wider resource use is a concern, rather than narrow expenditure, then the analytical net should also be cast much wider (see, for example, the methods in Stokey and Zeckhauser [1978]).

## Usefulness

In order to analyze and interpret findings in a way that meaningfully informs government administration, government budget data also need to be structured in a meaningful way. Budget and expenditure data are generally presented by administrative, economic, programmatic, and functional segments (see table 11.1). The purpose of the administrative segment is clear: it allows the government to allocate, monitor, and hold to account spending within its administrative structures. The purpose of the economic classification is also clear. It classifies the expenditure according to what inputs it has been spent on, which is necessary for accountability. Countries with a program structure require program classification in the COA because appropriations happen accordingly. Functional classification is appealing because it allows decision-makers to readily identify how much has been allocated and spent according to specific functions, such as primary education and basic health care. If expenditure items can be clearly mapped to functions, this type of classification offers substantial analytical possibilities. An example of a classification of the functions of government (COFOG) pertaining to the health sector is offered in table 11.2. Together, these set of classifiers should let analysts cross-tabulate expenditure data in many meaningful ways.

Business intelligence strategies and technologies can then be used for the analysis of the information stored in the data warehouse. Appropriate tagging and data structure allow for automated reporting and analytical processing following business needs. Dashboards can be developed to provide information to management in government agencies on issues such as budget execution, cash position, and audit, allowing for real-time, evidence-based decision-making (Negash and Gray 2008).

**TABLE 11.2** Example of Classification of Functions of Government from the Health Sector

First level	Second level
Health	Medical products, appliances, and equipment
	Outpatient services
	Hospital services
	Public health services
	R&D health
	Health n.e.c.

Source: Eurostat 2019, 37.

Note: n.e.c. = not elsewhere classified; R&D = research and development.

However, classifying these functions may not be trivial. With reference to the health sector, the following issues may arise:

- **Classifying by some functions may not always be possible.** In the health sector, a hospital generally offers both inpatient and outpatient services. Unless it has dedicated departments drawing on distinct cost centers, it may not be possible to differentiate between these services. It may be possible to understand total hospital spending but not necessarily the functions to which spending was dedicated within the hospital. Furthermore, health staff may provide both inpatient and outpatient services, and it would be difficult to apportion wages without a robust time-recording system. Similarly, in countries where the region- or district-level administration is also the lowest spending unit, it can be difficult to apportion specific functions because district authorities (in health) generally need to offer primary and secondary care as well as public health services. Therefore, if the spending unit does not have a clear mandate that maps directly to the COFOG functions, it is necessary to make assumptions, and these may not always be helpful or appropriate. In case there is no clear fit, it may be more accurate to simply report by administrative segment than to fit a square peg into a round hole. A COA reform process can be pursued over time to make spending more meaningful from a functional perspective.
- **Reporting by function requires a discrete choice.** Spending can be classified as *either* health or education spending—but not *both*. However, there are teaching hospitals that could be classified as either. There may also be medical facilities managed by the defense sector where allocation could be contested. Further, it is unclear whether subsidies for enrolling the poor in health insurance should be considered a health or a social protection function.
- **Not all functions, per COFOG categories, can be clearly understood as government functions.** For example, in the health sector, the COFOG category of *medical products, appliances, and equipment* may more appropriately be classified as *inputs* in the economic classification rather than *functions*. This also raises the question of inconsistencies within classifications because these medical products also serve other functions in the COFOG, such as hospital care or outpatient services.
- **There may be an overlap between functional and program classifications** because programs are output oriented and should serve specific government functions. There can still be added value for having both, but this needs to be clarified.

Reporting by functional classification is useful as long as it can be done credibly. Whether and how this exercise is done should reflect local context, demand, and capacity. Recommendations to shift spending toward some functions will remain unhelpful if these cannot clearly be traced back to the government's administrative structures. For example, it may be appealing to recommend more spending on outpatient services (which tend to be more efficient than hospital services), but as long as the government cannot clearly differentiate between spending on inpatient and outpatient services at the hospital level, such recommendations will remain unhelpful. Furthermore, as long as functional classification remains subjective, based on the assumptions of the analyst, any recommendations to adjust spending will lack credibility. This problem was recognized by the *Rwanda Nutrition Expenditure and Institutional Review 2020*, which cautions that extensive allocative efficiency analysis will remain futile as long as it cannot be clearly mapped back to the budget (Piatti-Fünfkirchen et al. 2020). Instead, a COA reform process may be more meaningful to improve the functional classification toward what is needed for the easier interpretation of expenditure data (see box 11.2).

Data presented in a useful format with integrity will likely foster demand for analysis. To make data more useful to analysts, there are ongoing initiatives by development partners that systematically clean, process, and categorize FMIS data (see, for example, the World Bank's BOOST initiative).<sup>4</sup> There has been a lot of demand for these initiatives because they support the various other analytical products that require government expenditure data in an interpretable format. However, as long as this work is not produced domestically through domestic systems, it is unlikely to be sustainable and will not undergo the required domestic checks and balances. This is an essential task of government data management—data storage in an adequate format in the business warehouse, from which a business intelligence system can pull meaningful reports—possibly requiring investments in institutional, systems, and human capacity.



## BOX 11.2 How the Government of Rwanda Uses Budget Tagging to Implement a High-Priority, Cross-Cutting Agenda

The budget of the government of Rwanda, like in many countries, is organized vertically by ministry, department, and agency. This lends itself well to oversight and accountability. For some issues, such as climate change, gender, or nutrition, where implementation cuts across sectors and agencies, it can be difficult to identify what relevant activities were budgeted for and implemented. The government therefore introduced an upstream tagging process, in which ministries identify what interventions in the budget are related to these issues. This has provided a crucial management function because the financial management information system can now produce functional budget execution reports that reflect spending on these issues. At any point in time, it provides insight into what activities have been implemented, which activities remain to be implemented, and what the remaining cash flow requirements are. It thereby uses the budget as a tool for oversight, coordination, and mutual accountability for implementing a high-priority, cross-cutting agenda.

### Consistency

Consistency in data management enables the production of data that can interface across systems and over space and time to allow for meaningful analysis. The COA's definition and use in government systems are influenced by different public financial management (PFM) traditions. PFM traditions can leave countries with the application of different COAs across levels of decentralization (Cooper and Pattanayak 2011). As long as this is the case, it is difficult to have a unified data set that allows for the analysis of expenditure information across the country, which complicates management and decision-making (PEMPAL 2014). One example of such a case is Indonesia, where a long-standing reform process has aimed to unify the COA across the country.

Consistency is also required across the system landscape in a country. This means that the same COA should be used throughout the FMIS and that taxes, debt management, and payroll should be classified according to the codes in the COA. Without unified use of the COA, adequate integration across systems to conduct government analytics will not be possible. For example, understanding the full fiscal health of an organization requires an integrated data set on the budget, debt, and payroll. If development partners are an important source of revenue, they should be encouraged to use the same classification structure so that comprehensive expenditure reports can be produced (Piatti-Funfkirchen, Hashim, et al. 2021).

It is equally important that the COA is used as intended. If activities are posted as line items, or vice versa, this creates problems for the quality of expenditure data and, subsequently, for analysis (Farooq and Schaeffer 2017). Similarly, it is important not to confuse programs with projects. A *program* is a set of activities that contribute to the same set of specific objectives, an *activity* is a subdivision of a program into homogenous categories, and a *project* is a single, indivisible activity with a fixed time schedule and a dedicated budget or activities. In some instances, development partners are responsible for the introduction of line-item codes into the COA in order to give a certain engagement more visibility or allow for the allocation of resources to one specific engagement area. This can come at the cost of coherence and consistency. For example, in Zimbabwe's health sector, there is a line item called *results-based financing*, one called *hygiene and sanitation*, and one called *malaria control*. All of these are important engagement areas but not inputs. They should be reflected as such in the COA. Similarly, in Rwanda, there is a line item called *maternal and child health*, which is also not a reflection of inputs but rather a target group.

Finally, it is important to be clear about nomenclature. Mixing budget data, release data, commitment data, and actual expenditure data within the same data set will lead to inconsistencies and problems.

## Stability

The comparability of data over time is assisted by having a stable process to produce them and a stable classification system to parse and present them. But perfect stability does not occur: some degree of variation is natural and to be expected as conditions change, knowledge advances, and governments address evolving needs. Changes in reporting may be consequential, through the introduction of new spending agencies or the shift from input budgets to program structures. Stability does not require a static reporting structure, which would be unrealistic and unhelpful. It does, however, require the government to be able to connect current expenditure information to the past to be able to make use of trend data. This can be done by designing a coding scheme that can accommodate older and newer codes; by taking a sequenced, incremental approach to reforms; or by at least maintaining tables that form a bridge between data series to allow for reasonable consistency between the past, the present, and the future.

If such mitigation measures are not taken, change can be disruptive. For example, in Zimbabwe, the program structure in the health sector was substantially revised at both the program and the subprogram levels to accommodate an additional focus on research by the Ministry of Health and Child Care. A program and four subprograms were added, and four subprograms were removed. This meant that 35 percent of the approved 2020 budget had been allocated to programs that no longer existed in the 2021 budget. Instability in the classification of the program structure over time without adequate mitigation measures (for example, bridge tables) raises the question of what kind of actual reallocations accompanied these shifts. The possibility of multiyear analysis for costing or value for money remains severely limited in such scenarios. Similarly, the changes mean that performance targeting may be disrupted, as it was in the Zimbabwe health case, in which none of the 17 program outcome indicators in the 2020 budget remained available in the 2021 budget (World Bank 2022).

## EXPLORING MICROLEVEL GOVERNMENT EXPENDITURE DATA

Developing comprehensive, appropriately structured, consistent, and stable data with a clear provenance provides a foundation for effective analytics. Though a large literature on the analysis of expenditure data exists (see, for example, Robinson [2000], Tanzi and Schuknecht [2000], and some discussion in appendix C), there is less discussion of how these data might be used to understand the functioning of government itself.

There are many examples of how government expenditure data can be used to inform the efficiency of government spending and better understand how a government is functioning. Expenditure information is necessary for an administration to explore opportunities for reducing the cost of the resources used for an activity or for increasing the output for a given input while maintaining quality (McAfee and McMillan 1989). The National Audit Office in the United Kingdom assesses how well the administration makes use of resources to achieve intended outcomes (NAO 2020). In Kenya, “data envelope analysis” is used to compare the efficient utilization of resources across counties (Kirigia, Emrouznejad, and Sambo 2002; Moses et al. 2021).

Information on differences in the amounts paid for goods between the public and private sectors is also frequently used to measure inefficiencies and can point to deep-rooted problems in the quality of an administration (see chapter 12). In Zambia, for example, such an analysis found that the rapid accumulation of payment arrears led to suppliers’ building in a risk premium and, consequently, to the government’s paying higher prices and suffering an unnecessary efficiency loss (World Bank 2016b). Generally, efficiency analyses are a central component of many analytical products of governments and development partners, such as Public Expenditure Reviews (PERs), and guidance on how to conduct these is widely available (Coelli et al. 2005; Greene 2008; Pradhan 1996; Shah 2005).

Government expenditure data can also be used to inform allocative choices, determining which groups, geographic regions, or sectors receive the most resources. Equity analysis allows for reorienting spending

to better follow needs if resources are not flowing to the areas identified as requiring the most resources. In the health sector, benefit and expenditure incidence analyses are commonplace (Binyaruka et al. 2021; Mills et al. 2012; Mtei et al. 2012; Wagstaff 2012) and often accompany PERs. They provide insight into who pays for services and, separately, who utilizes services. They can thus offer concrete recommendations about how to restructure spending to be more equitable. More broadly, Commitment to Equity Assessments offer a methodology to estimate the impact of fiscal policy on inequality and poverty (Lustig 2011, 2018).

Government expenditure data are used as a foundation for accountability (Ball, Grubnic, and Birchall 2014; Griffin et al. 2010; Morozumi and Veiga 2016). If government expenditure data can be made publicly accessible for analytical purposes, this extends the benefits further. Groups across society can use published data to undertake their own assessments of government functioning and the distribution of public resources. A growing body of research tests the notion that transparency facilitates accountability and leads to a host of developmental outcomes. Using the frequency of the publication of economic indicators, including those related to government expenditure, Islam (2003) finds that countries with better information flows have better-quality governance. Hameed (2005) analyzes indexes of fiscal transparency based on IMF fiscal Reports on the Observance of Standards and Codes (ROSCs) and shows, after controlling for other socioeconomic variables, that more-transparent countries tend to have better credit ratings, better fiscal discipline, and less corruption. Similarly, an analysis of the Open Budget index shows that more-transparent countries tend to have higher credit ratings (Hameed 2011). Looking at each of the six PFM pillars covered by the current PEFA framework, de Renzio and Cho (2020) find that the “transparency of public finances” and “accounting and reporting” have the most direct effect on budget credibility. The authors stipulate that this may be because more information and timely reporting allow for more direct, real-time control of how public resources are being used.<sup>5</sup>

To provide practical details of this type of data analysis, this chapter now focuses on some of the most basic but useful analyses of expenditure data that can assist in understanding the functioning of government administration, with particular reference to a case study in Indonesia.

### Basic Descriptives from Government Expenditure Microdata

First, to gain a sense of the completeness of the data being used for analysis, analysts may wish to estimate the budget coverage, which requires the summation of the value of all expenditure transactions routed through the data source (usually an FMIS) in a given fiscal year and, subsequently, the division of this value by the total approved budget reported by the government. This is presented in equation 11.1, where  $t$  represents the fiscal year and  $i$  the individual transaction:

$$\frac{\sum_{t,i} (trans_{1,1} + trans_{1,2} + trans_{2,2} + \dots + trans_{t,i})}{Total\ approved\ budget} \quad (11.1)$$

Equation 11.1, in turn, provides inputs to a table of the form of table 11.3.

The FMIS budget coverage statistics can be calculated for the general government, subagencies, and provinces or other subnational levels of government separately. These calculations then give an idea of the

**TABLE 11.3 Sample Output of FMIS Budget Coverage Estimation**

	2019	2020	2021
Total approved budget			
Total volume processed through the FMIS			
Percentage processed through the FMIS			

Source: Original table for this publication.

Note: FMIS = financial management information system.

agencywide and geographic spread in the coverage of the FMIS, allowing analysts to assess what percentage of the approved budget is processed by the FMIS.

Second, budget expenditure data can be used to identify trends and patterns in *budget execution rates*: the proportion of intended expenditures that have been undertaken within a specific time period. Budget execution data are a basic but important representation of how an organization is using resources and, when coupled with other information, how well it is working. If it is spending well but producing no outputs or not spending despite important upcoming commitments, these are signals of problems within the administration. Execution analysis also serves as a foundation for accountability because it can shed light on whether funds have been used for their intended purpose.

The analysis of budget execution rates can be conducted for the government as a whole or for specific sectors, spending units, line items, or programs. The type of analysis done will depend on how analysts want to assess the effectiveness of the administration. The aggregate budget execution rate alone—say, at the agency level—only informs analysts of whether resources are being used in line with authorized amounts and spending within the budget. Such aggregate analysis can hide important details, such as overspending on some items and underspending on others.<sup>6</sup> Disaggregation in the analysis frequently leads to insights. For example, overspending on the wage bill in the health sector is often associated with expenditure cuts on goods and supplies or capital expenditures (Piatti-Fünfkirchen, Barroy, et al. 2021). This undermines the quality of the health services provided.

Third, a *transactions profile* can be developed as a useful way to map out expenditure patterns and management (Hashim et al. 2019). The transactions profile is a measure that gauges how government expenditure transactions are distributed by size. The actual pattern of financial transactions can have significant implications for how activities are actually being executed and, hence, can be useful for understanding what is driving effective government functioning. To do this, analysts can calculate the number of transactions, the percentage of transactions, the cumulative share of the number of transactions, and the cumulative share of the amount processed through the FMIS for specific sets of transaction types. Table 11.4 provides a sample template.

**TABLE 11.4** Template for a Government Expenditure Transactions Profile

Range (US\$ equivalent)	Number of transactions	Share of transactions (%)	Cumulative share (%)	Total amount of transactions (US\$)	Share of amount processed through FMIS (%)	Cumulative share of amount processed through FMIS (%)
<100						
100–200						
200–500						
500–1k						
1k–5k						
5k–10k						
10k–25k						
25k–100k						
100k–500k						
500k–1,000k						
1,000k–50,000k						
>50,000k						
<b>Total</b>						

Source: Hashim et al. 2019.

Note: FMIS = financial management information system.

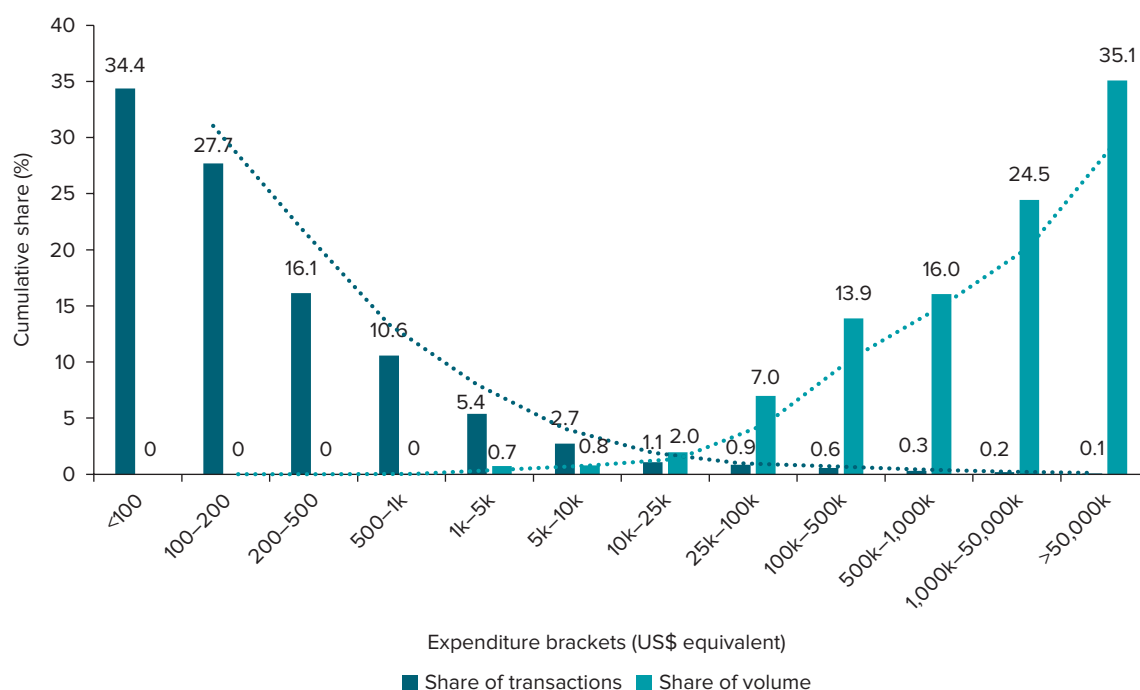
The transactions profile can then be displayed graphically (figure 11.3 provides an example from Bangladesh), where expenditure brackets are plotted against the cumulative share of the number of transactions and value of transactions. Typically, a larger percentage of transactions are small value transactions and even in sum cover only a small share of total spending. At the same time, high value transactions tend to be few in number but make up a large share of the total volume of spending.

### Assessing the Attributes of Expenditure Data

As well as providing useful descriptions of basic patterns in the expenditure data, budget execution data, FMIS coverage data, and the transactions profile offer useful information for analysts on the expenditure data's attributes (see the section above on Attributes of Good-Quality Expenditure Data). Specifically, analysts may further probe the data in the following ways.

To assess *integrity and data provenance*, analysts may first wish to get clarity on how various transactions are processed and what kinds of controls they are subject to. For example, how are debt payments, subsidies, wage payments, or payments for goods and services handled, and is this equal across all expenditure items? A useful starting point may be to document which transactions are processed through the FMIS and which ones are not. Follow-up questions may then relate to whether the current process for how various transactions are treated is adequate from an integrity and provenance perspective. Does it suffice to merely post some transactions, such as wage payments or debt payments, to the general ledger? This also opens an important political-economic dimension because it may show the revealed preferences of governments that wish to control spending on certain line items (for example, not using the FMIS would make it easier to adjust spending by the executive without legislative approval). Therefore, discussing this openly and bringing transparency into the process would be a useful first step. Second, analysts may wish to identify technical challenges in routing certain transactions through the FMIS and then explore how advancements in

**FIGURE 11.3** Expenditure Transactions Profile, Bangladesh



Source: Hashim et al. 2019.

maturing technologies (for example, financial technology innovations or the use of blockchain technology) could help strengthen the process.

As part of assessing the *comprehensiveness* of government expenditure data, analysts may wish to critically review how the government and the broader public sector are defined within the data. This should be followed by an assessment of whether these are appropriately reported across agencies. Identifying potential shortcomings in comprehensiveness, such as a lack of reporting on sensitive sectors or expenditure arrears, is another red flag for the FMIS data, as may be reporting against various select appropriation types. Such checks will minimize the risk of misinterpreting the findings and establishing poor indicators and targets that are poor representations of true spending patterns. These red flags are an opportunity for improvements in the comprehensiveness of expenditure reporting.

To assess the *usefulness* of government expenditure data, analysts may wish to explore what elements are captured and how they relate to government priorities. Do the data allow analysts to identify who spent funds, what they spent them on, and whether this usefully informs progress against the priorities set out in the agenda? On the question of who spends, it would be useful for the data to have sufficient detail in the administrative classification. Is it possible, for example, to know which hospital, health clinic, or school received what budget? What they spent it on should then be clear from the line item or activity segment. What purpose they spent it on (for example, malaria, primary education, and so on) can potentially be derived from the functional classification, but it can be difficult to establish this well. If the government has a functional classification, it may be useful to review how the mapping is generated and how well it serves its purpose. Given all of the above, the overarching questions for analysts will then be how well the established classification of expenditure data can be used to inform government priorities and what can be done to improve it.

To assess the *consistency* of the data, analysts can check whether there is consistency in the application of the COA across levels of decentralization and information systems across the government to allow for adequate integration. Analysts may also check for quality in the application of data entry to ensure the COA has been used as intended. Inconsistencies in the actual application can lead to problems in analysis and interpretation. Finally, in environments where development partners are an important source of revenue, analysts can review whether they have followed the same basis for accounting as the government to allow for the integration of expenditure data and comprehensive reporting.

Finally, to assess the *stability* of the data, analysts can review major changes in the expenditure data structure over time. If these are evident, analysts may explore whether tools to compare spending over time have been developed to give policy makers a multiyear perspective on important expenditure areas. With a solid understanding of the strengths and weaknesses of the underlying data, analysts can then use this expenditure and budget execution data to pursue efficiency, equity, or sustainability analyses to inform the effectiveness of government.

## Case Study: Investigating Ineffective Capital Expenditure in Indonesia

At the request of Indonesia's Ministry of Finance, the World Bank conducted an institutional diagnostic to understand the causes of "low and slow" capital budget expenditure execution (World Bank 2020). The study is an example of the value of drilling down deep on expenditure data, with information from 11,589 spending units and survey responses from nearly 2,000 spending units. By matching spending data and survey responses, the study identified that over 80 percent of capital budget allocations were directed to only 3 percent of spending units, and 78 percent were directed to four ministries, all of which had lower execution rates than others.<sup>7</sup>

The survey indicated that line ministries found planning difficult because they were not provided with predictable indicative budget ceilings for the next three years. They therefore prepared capital projects to align with annual budgets. Only 6 percent of spending units used multiyear contracts. The rest split their projects across annual contracts, leading to inefficiencies in contract implementation that contributed to low budget execution. For example, in 2019, disbursements were bunched at the end of the year, with 44 percent being made in the fourth quarter.



Compounding this, annual budgets tended to be very rigid, with expenditure authority lapsing at the end of the year. This led to a stop-start approach to projects due to the annual cessation of appropriation approval, limiting the administrative ability of agencies to implement the capital works program, given the multiyear nature of many projects.

The analysis also allowed World Bank staff to assess whether preexisting reforms to confront these problems were working. They did not seem to be. The spending units of only one ministry—the Ministry of Public Works and Housing—made use of early procurement, which was supported by a ministerial decree. While there was a government regulation that enabled spending units to begin the procurement process in the preceding year, 60 percent of spending units prepared their procurement plans after the start of the new fiscal year, thereby introducing bunching and delays in the execution of the program.

At least part of the root cause came from the supplier side. Half of all spending units faced difficulties in ensuring that vendors submitted invoices within five days of finishing work. Further, 73 percent reported that incomplete proof in vendors' invoices was the main cause for delays in preparing payment requests. The analysis also identified other areas of concern. Some 42 percent of spending units reported that difficulties in obtaining land approvals delayed contract implementation. A particular blockage occurred in cases where the land value, determined in a quasi-judicial proceeding for land acquisition, was greater than the budget. There was also a concern that fiduciary (audit) control discouraged spending units' performance in project implementation. Some 14 percent of spending units said that auditors created a delay in implementation, and 32 percent of respondents preferred splitting activities into multiple contracts to avoid the audit of large contracts.

Overall, this detailed diagnostic enabled specific, practical recommendations for improved government management. It was only made possible by triangulating microlevel expenditure data at the spending unit with survey data.

## CONCLUSION

Government expenditure data can assist our understanding of the functioning of government agencies, acting as a basis for conducting broader efficiency, equity, or productivity analyses.<sup>8</sup> Such analyses can be valuable and informative for policy and for improving the quality of the administration. However, expenditure data are only useful for these ends if they also have the right attributes.

All technical solutions require an enabling environment of government commitment, actionable political economy, and resilience to shocks. It is important that strong systems for government expenditure data are in place and protected during times of adversity. Governments are encouraged to put in place processes that identify deficiencies in routines to allow for strengthening over time. The root causes of distortions may take considerable effort to uncover. Political leadership and a willingness to embrace transparency in the identification process are key.

This chapter has provided health warnings that should be considered when using expenditure data and has identified the following five attributes of good-quality expenditure data:

- Data provenance and integrity
- Comprehensive across space and over time
- Usefulness
- Consistency
- Stability.

How well government expenditure data meet the above attributes is rarely emphasized in analytical work or considered directly in its underlying methodologies. Instead, expenditure data are often taken at

face value, with the implicit assumption that the above conditions are met. If they are not, it can render the analysis incorrect and misleading.

This chapter suggests a periodic and data-driven review of these issues in all budgeting systems. For example, expenditure data can be used to estimate FMIS budget coverage. Such statistics provide insight into whether budget managers have incentives to avoid FMIS internal controls. This chapter advocates for estimating budget coverage periodically and making it publicly available in an effort to deepen the understanding of the incentives and the underlying political economy of budget controls. A step beyond this is to assess how variation in expenditure management relates to government effectiveness.

Budget coverage statistics could accompany analytical products that draw on these data to offer cautions in the interpretation of the data. Audit institutions can report on why FMIS coverage may be low and what can be done to strengthen it in their management letters and reports to the legislature.<sup>2</sup> Alongside this indicator, a transactions profile can be mapped to identify where risks in current expenditure management may lie and what types of reform may be warranted to improve expenditure control and service delivery objectives.

High-quality government expenditure microdata can be used by analysts to provide insight into expenditure management practices, functional effectiveness, and the related pursuit of public policy. A basic analysis simply assesses how capable expenditure units are at absorbing and spending funds.

The analysis of expenditure data benefits from triangulation with performance information on spending units to guide a dialogue on public sector effectiveness. Just as reviewing the calories one takes in without considering the activities undertaken may shed little light on the fitness and workings of one's metabolism, so, too, is the consideration of expenditure data limited if not aligned with the impacts of the activities being funded.

The strongest analysis frames the discussion of expenditure in terms of a logframe of expenditure (figure 11.1): where do expenditure data come from and how is expenditure defined, what are their quality and comprehensiveness, and how do they impact government effectiveness? Framing the discussion within government in terms of these steps is important because it facilitates noticing and learning (Hanna, Mullainathan, and Schwartzstein 2012). The “failure to notice” systemic problems may be a key binding constraint in reaching the production frontier if practitioners only excel at one aspect of the logframe—in this case, the analysis of data without sufficient regard to their origins and quality.<sup>10</sup>

It almost goes without saying that expenditure data may not be everything in the pursuit of government effectiveness. Some organizations spend very little but have very important public mandates, such as a policy, coordination, or regulatory function. However, for some of the most important government functions—such as the building of large capital projects—expenditure data can be a critical lens for understanding government functioning.

## NOTES

1. Expenditure data can also capture governments' responses to shocks through reallocation and adjustments to their revealed preferences (Brumby and Verhoeven 2010). After the global financial crisis, expenditure analysis showed that countries were temporarily expanding safety nets, protecting social sector spending through loans, redirecting funding to retain social spending, and harnessing the crisis to achieve major reforms to improve efficiency and quality.
2. Lowry (2016) estimates that health-related tax expenditures in the United States involved almost US\$300 billion in 2019.
3. The PEFA program provides a framework for assessing and reporting on the strengths and weaknesses of public financial management (PFM), using quantitative indicators to measure performance. PEFA is designed to provide a snapshot of PFM performance at specific points in time using a methodology that can be replicated in successive assessments, giving a summary of changes over time.
4. More information about the BOOST initiative is available on the World Bank's website at <https://www.worldbank.org/en/programs/boost-portal>.

5. More broadly, Kaufmann and Bellver (2005) find that transparency is associated with better socioeconomic and human development indicators, higher competitiveness, and reduced corruption. They show that for countries with the same level of income, a country with a more transparent environment tends to have more-effective government agencies. Glennerster and Shin (2008) find that countries experience statistically significant declines in borrowing costs as they become more transparent.
6. PEFA assessments can offer valuable information on budget execution rates. Not spending as intended and spending more than intended are considered equally problematic. A 15 percentage point deviation from the original appropriation is considered poor practice by the PEFA because, at that point, it likely renders the budget not credible or effective.
7. Over 64 percent of the capital budget was allocated to spending units in Jawa.
8. Beyond governments, these data are also used by international organizations for Public Expenditure Reviews (PERs), Public Expenditure Tracking Surveys, Commitment to Equity Assessments, and Article IV agreements.
9. As blockchain technology matures, it may also offer a pathway to the immutability of records, making them less susceptible to manipulation.
10. The “learning through noticing” approach alters the standard intuition that experience guarantees effective technology use (see, for example, Foster and Rosenzweig 2010; Nelson and Phelps 1966; Schultz 1975).

## REFERENCES

- African Union. 2001. *Abuja Declaration on HIV/AIDS, Tuberculosis and Other Related Infectious Diseases*. African Summit on HIV/AIDS, Tuberculosis, and Other Related Infectious Diseases, Abuja, Nigeria, April 24–27, 2001. OAU/SPS/ABUJA/3. <https://au.int/sites/default/files/pages/32894-file-2001-abuja-declaration.pdf>.
- Allen, Richard, and Daniel Tommasi, eds. 2001. *Managing Public Expenditure: A Reference Book for Transition Countries*. Paris: OECD Publishing.
- Baker Tilly Business Services Limited. 2014. *National Audit Office Malawi: Report on Fraud and Mismanagement of Malawi Government Finances*. Report to the Auditor General of the Government of Malawi. London: Baker Tilly Business Services Limited.
- Baldacci, Emanuele, and Kevin Fletcher. 2004. “A Framework for Fiscal Debt Sustainability Analysis in Low-Income Countries.” In *Helping Countries Develop: The Role of Fiscal Policy*, edited by Sanjeev Gupta, Benedict Clements, and Gabriele Inchauste, 130–61. Washington, DC: International Monetary Fund.
- Ball, Amanda, Suzana Grubnic, and Jeff Birchall. 2014. “Sustainability Accounting and Accountability in the Public Sector.” In *Sustainability Accounting and Accountability*, 2nd ed., edited by Jan Bebbington, Jeffrey Unerman, and Brendan O’Dwyer, 176–96. London: Routledge.
- Barton, Allan. 2011. “Why Governments Should Use the Government Finance Statistics Accounting System.” *Abacus* 47 (4): 411–45.
- Binyaruka, Peter, August Kuwawenaruwa, Mariam Ally, Moritz Piatti, and Gemini Mtei. 2021. “Assessment of Equity in Healthcare Financing and Benefits Distribution in Tanzania: A Cross-Sectional Study Protocol.” *BMJ Open* 11 (9): e045807. <http://doi.org/10.1136/bmjopen-2020-045807>.
- Bridges, Kate, and Michael Woolcock. 2017. “How (Not) to Fix Problems That Matter: Assessing and Responding to Malawi’s History of Institutional Reform.” Policy Research Working Paper 8289, World Bank, Washington, DC.
- Brumby, Jim, and Marijn Verhoeven. 2010. “Public Expenditure after the Global Financial Crisis.” In *The Day after Tomorrow: A Handbook on the Future of Economic Policy in the Developing World*, edited by Otaviano Canuto and Marcelo Giugale, 193–206. Washington, DC: World Bank.
- Burnside, Craig. 2004. “Assessing New Approaches to Fiscal Sustainability Analysis.” Working paper, Economic Policy and Debt Department, World Bank, Washington, DC.
- Burnside, Craig, ed. 2005. *Fiscal Sustainability in Theory and Practice: A Handbook*. Washington, DC: World Bank.
- Challen, Don, and Craig Jeffery. 2003. “Harmonisation of Government Finance Statistics and Generally Accepted Accounting Principles.” *Australian Accounting Review* 13 (30): 48–53.
- Chami, Ralph, Raphael Espinoza, and Peter Montiel, eds. 2021. *Macroeconomic Policy in Fragile States*. Oxford, UK: Oxford University Press.
- Chan, James L. 2006. “IPSAS and Government Accounting Reform in Developing Countries.” In *Accounting Reform in the Public Sector: Mimicry, Fad or Necessity*, edited by Evelyn Lande and Jean-Claude Scheid, 31–42. Paris: Expert Comptable Média.
- Coelli, Timothy J., D. S. Prasada Rao, Christopher J. O’Donnell, and George Edward Battese. 2005. *An Introduction to Efficiency and Productivity Analysis*. New York: Springer.

- Cooper, Julie, and Sailendra Pattanayak. 2011. *Chart of Accounts: A Critical Element of the Public Financial Management Framework*. Washington, DC: International Monetary Fund.
- de Renzio, Paolo, and Chloe Cho. 2020. "Exploring the Determinants of Budget Credibility." Working paper, International Budget Partnership, Washington, DC.
- Di Bella, Gabriel. 2008. "A Stochastic Framework for Public Debt Sustainability Analysis." IMF Working Paper WP/08/58, International Monetary Fund, Washington, DC.
- European Commission and IEG (Independent Evaluation Group, World Bank). 2017. *Joint Evaluation of Budget Support to Ghana (2005–2015): Final Report*. Brussels, Belgium: European Commission.
- Eurostat. 2019. *Manual on Sources and Methods for the Compilation of COFOG Statistics: Classification of the Functions of Government (COFOG)*. Luxembourg: Publications Office of the European Union.
- Farooq, Khuram, and Michael Schaeffer. 2017. "Simplify Program Budgeting: Is There a Place for 'Activities' in a Program Classification?" *IMF Public Financial Management Blog*, October 30, 2017. International Monetary Fund. <https://blog-pfm.imf.org/en/pfmblog/2017/10/simplify-program-budgeting-is-there-a-place-for-activities-in-a-program-classifi>.
- Foster, Andrew D., and Mark R. Rosenzweig. 2010. "Microeconomics of Technology Adoption." *Annual Review of Economics* 2: 395–424.
- Glennerster, Rachel, and Yongseok Shin. 2008. "Does Transparency Pay?" *IMF Staff Papers* 55 (1): 183–209.
- Greene, William H. 2008. "The Econometric Approach to Efficiency Analysis." In *The Measurement of Productive Efficiency and Productivity Growth*, edited by Harold O. Fried, C. A. Knox Lovell, and Shelton S. Schmidt, 92–250. New York: Oxford University Press.
- Griffin, Charles C., David de Ferranti, Courtney Tolmie, Justin Jacinto, Graeme Ramshaw, and Chinyere Bun. 2010. *Lives in the Balance: Improving Accountability for Public Spending in Developing Countries*. Washington, DC: Brookings Institution Press.
- Hameed, Farhan. 2005. "Fiscal Transparency and Economic Outcomes." IMF Working Paper WP/05/225, International Monetary Fund, Washington, DC.
- Hameed, Farhan. 2011. "Budget Transparency and Financial Markets." Working paper, International Budget Partnership, Washington, DC.
- Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein. 2012. "Learning through Noticing: Theory and Experimental Evidence in Farming." NBER Working Paper 18401, National Bureau of Economic Research, Cambridge, MA.
- Hashim, Ali. 2014. *A Handbook on Financial Management Information Systems for Government: A Practitioners Guide for Setting Reform Priorities, Systems Design and Implementation*. Africa Operations Services Series. Washington, DC: World Bank.
- Hashim, Ali, Khuram Farooq, and Moritz Piatti-Fünfkirchen. 2020. *Ensuring Better PFM Outcomes with FMIS Investments: An Operational Guidance Note for FMIS Project Teams Designing and Implementing FMIS Solutions*. Guidance Note, Governance Global Practice. Washington, DC: World Bank.
- Hashim, Ali, and Moritz Piatti-Fünfkirchen. 2018. "Lessons from Reforming Financial Management Information Systems: A Review of the Evidence." Policy Research Working Paper 8312, World Bank, Washington, DC.
- Hashim, Ali, Moritz Piatti-Fünfkirchen, Winston Cole, Ammar Naqvi, Akmal Minallah, Maun Prathna, and Sokbunthoeun So. 2019. "The Use of Data Analytics Techniques to Assess the Functioning of a Government's Financial Management Information System: An Application to Pakistan and Cambodia." Policy Research Working Paper 8689, World Bank, Washington, DC.
- IFAC (International Federation of Accountants). 2022. *Handbook of International Public Sector Accounting Pronouncements*. New York: IFAC.
- IMF (International Monetary Fund). 2014. *Government Finance Statistics Manual 2014*. Washington, DC: IMF.
- Islam, Roumeen. 2003. "Do More Transparent Governments Govern Better?" Policy Research Working Paper 3077, World Bank, Washington, DC.
- Kaufmann, Daniel, and Ana Bellver. 2005. "Transparenting Transparency: Initial Empirics and Policy Applications." Draft discussion paper presented at the International Monetary Fund conference on transparency and integrity, July 6–7, 2005, World Bank, Washington, DC.
- Kirigia, Joses M., Ali Emrouznejad, and Luis G. Sambo. 2002. "Measurement of Technical Efficiency of Public Hospitals in Kenya: Using Data Envelopment Analysis." *Journal of Medical Systems* 26 (1): 39–45.
- Lowry, Sean. 2016. *Health-Related Tax Expenditures: Overview and Analysis*. CRS Report Prepared for Members and Committees of Congress. Washington, DC: Congressional Research Service.
- Lustig, Nora. 2011. "Commitment to Equity Assessment (CEQ): A Diagnostic Framework to Assess Governments' Fiscal Policies Handbook." Tulane Economics Working Paper 1122, Department of Economics, Tulane University, New Orleans, LA.
- Lustig, Nora, ed. 2018. *Commitment to Equity Handbook: Estimating the Impact of Fiscal Policy on Inequality and Poverty*. Washington, DC: Brookings Institution Press.

- McAfee, R. Preston, and John McMillan. 1989. "Government Procurement and International Trade." *Journal of International Economics* 26 (3–4): 291–308.
- Milante, Gary, and Michael Woolcock. 2021. "Fiscal Policy in Fragile Situations: Flying in Fog with Limited Instrumentation." In *Macroeconomic Policy in Fragile States*, edited by Ralph Chami, Raphael Espinoza, and Peter Montiel, 271–96. Oxford, UK: Oxford University Press.
- Mills, Anne, John E. Ataguba, James Akazili, Jo Borghi, Bertha Garshong, Suzan Makawia, Gemini Mtei, et al. 2012. "Equity in Financing and Use of Health Care in Ghana, South Africa, and Tanzania: Implications for Paths to Universal Coverage." *The Lancet* 380 (9837): 126–33.
- Morozumi, Atsuyoshi, and Francisco José Veiga. 2016. "Public Spending and Growth: The Role of Government Accountability." *European Economic Review* 89: 148–71.
- Moses, Mark W., Julius Korir, Wu Zeng, Anita Musiega, Joyce Oyasi, Ruoyan Lu, Jane Chuma, and Laura Di Giorgio. 2021. "Performance Assessment of the County Healthcare Systems in Kenya: A Mixed-Methods Analysis." *BMJ Global Health* 6 (6): e004707.
- Mtei, Gemini, Suzan Makawia, Mariam Ally, August Kuwawenaruwa, Filip Meheus, and Josephine Borghi. 2012. "Who Pays and Who Benefits from Health Care? An Assessment of Equity in Health Care Financing and Benefit Distribution in Tanzania." *Health Policy and Planning* 27 (suppl. 1): i23–i34.
- NAO (National Audit Office, UK). 2020. "Assessing Value for Money." *Successful Commissioning Toolkit*. United Kingdom Government. <https://www.nao.org.uk/successful-commissioning/general-principles/value-for-money/assessing-value-for-money/>.
- Negash, Solomon, and Paul Gray. 2008. "Business Intelligence." In *Handbook on Decision Support Systems*, edited by Frada Burstein and Clyde W. Holsapple, 2: 175–93. Berlin: Springer.
- Nelson, Richard R., and Edmund S. Phelps. 1966. "Investment in Humans, Technological Diffusion, and Economic Growth." *American Economic Review* 56 (1/2): 69–75.
- PEFA (Public Expenditure and Financial Accountability). 2020. *Global Report on Public Financial Management*. Washington, DC: PEFA Secretariat.
- PEFA (Public Expenditure and Financial Accountability). 2022. *Global Report on Public Financial Management*. Washington, DC: PEFA Secretariat.
- PEMPAL (Public Expenditure Management Peer Assisted Learning). 2014. "Integration of the Budget Classification and Chart of Accounts: Good Practice among Treasury Community of Practice Member Countries." PEMPAL.
- Piatti, Moritz, Ali Hashim, and Clay G. Wescott. 2017. "Using Financial Management Information Systems (FMIS) for Fiscal Control: Applying a Risk-Based Approach for Early Results in the Reform Process." Paper presented at the IPMN Conference on Reform, Innovation, and Governance: Improving Performance and Accountability in the Changing Times, School of International and Public Affairs, the China Institute of Urban Governance, and the Center for Reform, Innovation, and Governance of Shanghai Jiao Tong University, Shanghai, China, August 17–18, 2017. <http://dx.doi.org/10.2139/ssrn.3090673>.
- Piatti-Fünfkirchen, Moritz. 2016. "What Can We Learn from a Decade of Public Financial Management and Civil Service Reform in Malawi?" IEG Project Lessons, World Bank, Washington, DC, April 22, 2016. <https://ieg.worldbankgroup.org/news/what-can-we-learn-decade-public-financial-management-and-civil-service-reform-malawi>.
- Piatti-Fünfkirchen, Moritz, Helene Barroy, Fedja Pivodic, and Federica Margini. 2021. *Budget Execution in Health: Concepts, Trends and Policy Issues*. Washington, DC: World Bank.
- Piatti-Fünfkirchen, Moritz, Ali Hashim, Sarah Alkenbrack, and Srinivas Gurazada. 2021. *Following the Government Playbook? Channeling Development Assistance for Health through Country Systems*. Washington, DC: World Bank.
- Piatti-Fünfkirchen, Moritz, Ali Hashim, and Khuram Farooq. 2019. "Balancing Control and Flexibility in Public Expenditure Management: Using Banking Sector Innovations for Improved Expenditure Control and Effective Service Delivery." Policy Research Working Paper 9029, World Bank, Washington, DC.
- Piatti-Fünfkirchen, Moritz, Liying Liang, Jonathan Kweku Akuoku, and Patrice Mwitende. 2020. *Rwanda Nutrition Expenditure and Institutional Review 2020*. Washington, DC: World Bank.
- Piatti-Fünfkirchen, Moritz, Magnus Lindelow, and Katelyn Yoo. 2018. "What Are Governments Spending on Health in East and Southern Africa?" *Health Systems and Reform* 4 (4): 284–99.
- Potter, Barry H., and Jack Diamond. 1999. *Guidelines for Public Expenditure Management*. Washington, DC: International Monetary Fund.
- Pradhan, Sanjay. 1996. "Evaluating Public Spending: A Framework for Public Expenditure Reviews." World Bank Discussion Paper 323, World Bank, Washington, DC.
- Rivetti, Diego. 2021. *Debt Transparency in Developing Economies*. Washington, DC: World Bank.
- Robinson, Marc. 2000. "Contract Budgeting." *Public Administration* 78 (1): 75–90.

- Schiavo-Campo, Salvatore. 2017. *Government Budgeting and Expenditure Management: Principles and International Practice*. New York: Taylor & Francis.
- Schultz, Theodore W. 1975. "The Value of the Ability to Deal with Disequilibria." *Journal of Economic Literature* 13 (3): 827–46.
- Shah, Anwar, ed. 2005. *Public Expenditure Analysis*. Public Sector Governance and Accountability Series. Washington, DC: World Bank.
- Shah, Anwar, ed. 2007. *Budgeting and Budgetary Institutions*. Public Sector Governance and Accountability Series. Washington, DC: World Bank.
- Stokey, Edith, and Richard Zeckhauser. 1978. *A Primer for Policy Analysis*. New York: Norton.
- Tanzi, Vito, and Ludger Schuknecht. 2000. *Public Spending in the 20th Century: A Global Perspective*. Cambridge, UK: Cambridge University Press.
- Wagstaff, Adam. 2012. "Benefit-Incidence Analysis: Are Government Health Expenditures More Pro-Rich Than We Think?" *Health Economics* 21 (4): 351–66.
- World Bank. 2016a. *Evaluation of the Malawi Financial Management, Transparency, and Accountability Project*. Project Performance Assessment Report, Report 103060. Washington, DC: World Bank.
- World Bank. 2016b. *Zambia Public Sector Management Program Support Project*. Project Performance Assessment Report, Report 106280. Washington, DC: World Bank.
- World Bank. 2020. *Indonesia Revenue and Budget Management: Institutional Diagnostic of Low and Slow Central Government Capital Budget Execution*. Report AUS0001636. Washington, DC: World Bank.
- World Bank. 2022. *Zimbabwe Health Public Expenditure Review*. Washington, DC: World Bank.





## CHAPTER 12

# Government Analytics Using Procurement Data

*Serena Cocciolo, Sushmita Samaddar, and Mihaly Fazekas*

### SUMMARY

The digitalization of national public procurement systems across the world has opened enormous opportunities to measure and analyze procurement data. The use of data analytics on public procurement data allows governments to strategically monitor procurement markets and trends, to improve the procurement and contracting process through data-driven policy making, and to assess the potential trade-offs of distinct procurement strategies or reforms. This chapter provides insights into conducting research and data analysis on public procurement using administrative data. It provides an overview of indicators and data sources typically available on public procurement and how they can be used for data-driven decision-making, the necessary data infrastructure and capacity for optimizing the benefits from procurement data analytics, and the added value of combining public procurement data with other data sources. Governments can take various steps to create the conditions for effectively using data for decision-making in the area of public procurement, such as centralizing public procurement data, periodically assessing their quality and completeness, and building statistical capacity and data analytics skills in procurement authorities and contracting entities.

### ANALYTICS IN PRACTICE

- The increasing availability of public procurement administrative microdata should be exploited for evidence-based decision-making. The digitalization of national public procurement systems across the world has opened enormous opportunities to measure procurement outcomes through the analysis of administrative data now available in machine-readable formats on electronic government procurement (e-GP) systems. The full potential of e-GP reforms can be realized when data analytical tools are systematically applied at scale for the monitoring and evaluation of public procurement.

---

Serena Cocciolo is an economist at the World Bank. Sushmita Samaddar is a researcher at the University of Kansas. Mihaly Fazekas is an assistant professor at the Central European University and scientific director at the Government Transparency Institute.

- Procurement data analytics can be used for monitoring and characterizing public procurement. Public procurement data can be used to characterize national public procurement spending; describe time trends; compare procurement performance across procuring entities, regions, and types of contract, as well as across types of procedure, sector, or supplier; and identify performance and compliance gaps in the national public procurement system. Interactive dashboards are increasingly widespread tools for monitoring public procurement through descriptive analysis because they enable procurement authorities to track, analyze, and display key performance indicators through customizable and user-friendly visualizations.
- Procurement data analytics can be used for data-driven policy making. The analysis of public procurement data can enable procurement agencies to develop key procurement policies or refine and assess existing regulations. First, data analytics allows agencies to assess existing efficiency gaps and understand the drivers of performance; these empirical insights are useful to identify and prioritize potential areas for interventions and reform efforts. Second, data analytics allows agencies to monitor the consequences of new policies, assess whether they are delivering the expected outcomes, and understand potential trade-offs. Especially in cases where an e-GP system already exists at the time of piloting and implementing new strategies, public procurement can also be a rich space for research and impact evaluations because the necessary data for tracking key outcome indicators are readily available from the existing e-GP system.
- Appropriate data infrastructure and capacity are necessary for effectively using public procurement data for decision-making. First, procurement data should be homogeneously collected and maintained across procuring entities and connected to a centralized platform. Second, data generated from different stages of the procurement cycle (for example, tendering process, bidding process, bid evaluation, contract award, and contract signing) should be consistently organized and connected through key identifiers. Third, the availability of data should be expanded to cover the full public procurement and contract management cycle, including parts of the process that are not typically included in procurement data, such as data on public procurement planning and budgeting, tender preparation data, contract execution data, and complaints data. Fourth, data quality and completeness should be improved through relatively simple and practical steps by the government, such as automated data quality checks in the e-GP system and periodic data audits. Finally, the necessary capacity for statistical analysis should be built in the public procurement authority, potentially including the creation of a dedicated statistical unit.
- A “whole-of-government” approach should be adopted in procurement data analytics. Public procurement is multidimensional and critically interconnected with other functions of the public sector and public administration. Yet the integration of e-procurement systems into other e-government systems is not yet a common practice. Data innovations should enable the integration of public procurement data with administrative microdata from other parts of the public sector, such as justice, firm registries, and tax administration. This would provide a comprehensive picture of the procurement function, holistically explore the environment within which procurement is conducted, and enable the government to develop innovative and impactful procurement strategies.
- Procurement data analytics should move beyond traditional public procurement indicators and data sources. While there is widespread consensus about the measurement framework for some dimensions of public procurement, including costs, price efficiency, integrity risks, transparency, and competition, other relevant aspects of public procurement, such as the inclusiveness and sustainability of public procurement and the quality of contract implementation, currently lack well-defined and commonly used indicators. Using nontraditional public procurement data can contribute to the development of new measures and expand the scope of public procurement data analytics, such as survey data with firms or procurement officers.

## INTRODUCTION

While it is difficult to measure the size of public procurement transactions in each country, a global exercise by Bosio et al. (2022) estimates that around 12 percent of the global gross domestic product is spent on public procurement—the process by which governments purchase goods, services, and works from the private sector. Given this massive scale, public procurement has the potential to become a strategic policy tool in three crucial ways.

First, improved public procurement can generate sizeable savings and create additional fiscal space by reducing the price of purchases and increasing the efficiency of the procurement process (Bandiera, Prat, and Valletti 2009; Best, Hjort, and Szakonyi 2019; Singer et al. 2009).<sup>1</sup> Second, public procurement can support national socioeconomic and environmental aspirations by encouraging the participation of local small firms in the public contract market, promoting green and sustainable procurement, and creating jobs through large public works (Ferraz, Finan, and Szerman 2015; Krasnokutskaya and Seim 2011). Finally, efficient public procurement can improve the quality of public services through several channels, such as the selection of higher-quality goods, more timely delivery of goods and completion of public infrastructure, and better planning of purchases and stock management. Given these strategic functions, efficient and effective public procurement can contribute to the achievement of the development goals of ending poverty and promoting shared prosperity.<sup>2</sup>

Data and evidence are necessary to monitor public procurement spending and identify the optimal policies and strategies for efficient, inclusive, and sustainable procurement. The use of data can contribute to a problem-driven, iterative approach to strengthening and modernizing national public procurement systems through the identification of efficiency and integrity gaps, analysis of the trade-offs associated with alternative procurement strategies, the development of data tools for monitoring the public procurement function, and the generation of knowledge and evidence on the impact of certain policies.

The digitalization of national public procurement systems across the world has opened enormous opportunities to measure procurement outcomes through the analysis of administrative data now available in machine-readable formats on electronic government procurement (e-GP) systems. E-procurement refers to the integration of digital technologies to replace or redesign paper-based procedures throughout the procurement cycle (OECD 2021). While countries are increasingly digitalizing public procurement processes, the functionalities covered by e-GP systems vary widely across countries (box 12.1), and this has implications for the accessibility and quality of procurement and contract data for analysis and research. Map 12.1 shows advancements in e-GP adoption globally and highlights the different degrees of sophistication of national e-GP systems, depending on the extent to which various procurement stages—advertisement, bid submission, bid opening, evaluation, contract signing, contract management, and payment—can be implemented electronically.<sup>3</sup>

Governments can take various steps to create the conditions for effectively using data for decision-making in the area of public procurement, such as centralizing public procurement data, periodically assessing their quality and completeness, creating the data infrastructure for integrating data from various stages of the procurement cycle and from other e-government systems, measuring the socioeconomic and environmental dimensions of government purchases, integrating procurement data and systems into other e-government data and systems, and building statistical capacity and data analytics skills in procurement authorities and contracting entities.

This chapter provides insights and lessons on how to leverage administrative microdata for efficient and strategic public procurement. The chapter provides an overview of indicators and data sources typically available on public procurement and how they can be used for data-driven decision-making (section 2), the necessary data infrastructure and capacity for optimizing the benefits from procurement data analytics (section 3), and the added value of combining public procurement data with other data sources (section 4).

## BOX 12.1 Types of Digitalization of Public Procurement Systems

The degree to which the procurement process is digitalized and integrated with other functions of government plays an important role in determining the accessibility and quality of administrative procurement microdata and how they can be used for conducting data analysis and research on public procurement.

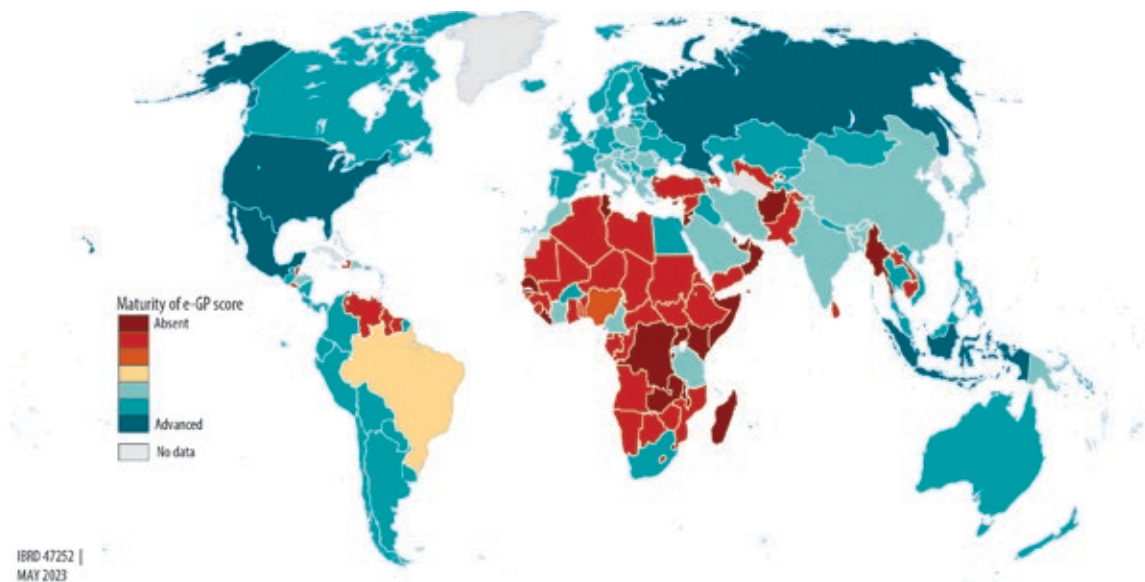
The digitalization of public procurement systems has been implemented in different ways across the world, with implications for data availability and quality. Most commonly, electronic government procurements (e-GP) systems are used to publish and store public procurement information. For example, in Pakistan and Tanzania, the online procurement system allows for the upload of tender and contract documents as scanned copies or PDFs.<sup>a</sup> In these cases, data would first need to be scraped from PDF documents and organized in a structured manner before any kind of data analysis could be performed. In fewer countries, the e-GP system includes functionalities related to the transactional aspects of public procurement, such as e-tendering, electronic submission of bids, e-evaluation, e-awarding, and, in the most advanced cases, electronic submission of invoices, e-catalogs, and contract management. In these cases, the e-GP system generates data in machine-readable formats, readily available for analysis. For example, in Colombia, a data management system has been implemented following the Open Contracting Data Standard guidelines on data transparency, so the data from the e-GP system can be downloaded in the form of Excel files and readily used for analysis.<sup>b</sup>

There are variations in the quality and completeness of data generated from e-GP systems, as well as in how well the data from different parts of the procurement process can be integrated or merged for a holistic view of government purchases. The integration of e-procurement systems into other e-government systems is not yet a common practice, and further work is needed to promote this “whole-of-government” approach from a data perspective.

a. For Pakistan, see World Bank (2017). For Tanzania, see the example of an invitation for bids from the Tanzania National Roads Agency (Tender AE/001/2020-21/HQ/G/79) available at [https://www.afdb.org/sites/default/files/documents/project-related-procurement/invitation\\_for\\_tenders\\_-\\_edms.pdf](https://www.afdb.org/sites/default/files/documents/project-related-procurement/invitation_for_tenders_-_edms.pdf).

b. For more information about the Open Contracting Data Standard, see the project website at <https://standard.open-contracting.org/latest/en/>.

MAP 12.1 Use of Electronic Government Procurements across the World, 2020



Source: World Bank, based on Doing Business 2020 Contracting with the Government database, <https://archive.doingbusiness.org/en/data/exploretopics/contracting-with-the-government#data>.

Note: The Maturity of e-GP score was calculated based on the number of features existing in the electronic government procurement (e-GP) system portal, as reported in the World Bank's Contracting with the Government database.

## HOW DO WE USE PUBLIC PROCUREMENT DATA FOR DECISION-MAKING?

### Procurement Indicators Used for Data Analysis

Based on the perspective that public procurement is a strategic function contributing to efficient public spending, as well as to the achievement of national socioeconomic and environmental objectives, this chapter provides a holistic view of public procurement. While the application of data analytical tools is often associated with the use of corruption flags to uncover malpractice, this focus risks discouraging governments from using and opening public procurement data. Data analytical tools' main purpose is strengthening the efficiency of public procurement and government spending in achieving national objectives, and a stronger focus on these more comprehensive goals could help reduce resistance from governments to adopting them.<sup>4</sup> Following this view, in this section, we present a broad set of procurement indicators and uses of procurement data analytics corresponding to a wide range of objectives, including (but not only) anticorruption goals. Table 12.1 provides an example of public procurement indicators that can be used to measure the performance of the public procurement system along the dimensions described in the following paragraphs: economy and efficiency, transparency and integrity, competition, inclusiveness, and sustainability.

The procurement and contracting cycle refers to a sequence of related activities, from needs assessment through competition and award to payment and contract management, as well as any subsequent monitoring or auditing (OECD 2021). It is typically divided into the following stages: (1) budget planning and tender preparation; (2) tendering process, bidding process, and bid evaluation; (3) contract award and contract signing; and (4) contract execution and monitoring. Traditional public procurement data often cover only stages (2) and (3) because the other stages are typically managed by other units (budget and financial management) and therefore recorded in separate systems. These data can be organized at the tender, lot, item (product), bid, and contract levels. Figure 12.1 provides a visual representation of how the different levels of public procurement data connect. Specifically, tenders can be divided into lots, and each lot can specify different product items. Firms submit bids to specific tenders or lots and can submit for specific tenders; tenders result in one or more contracts, which are then linked to contract amendments and payments. Understanding the structure of public procurement data and the links between different stages is the first step for effectively using and analyzing it. For example, the e-GP systems for Brazil, Romania, Croatia, and Honduras organize open procurement data at the tender, lot, contract, and bid levels, allowing users to connect these different stages of the process through unique identifiers for each data set.

**TABLE 12.1** Examples of Public Procurement Indicators

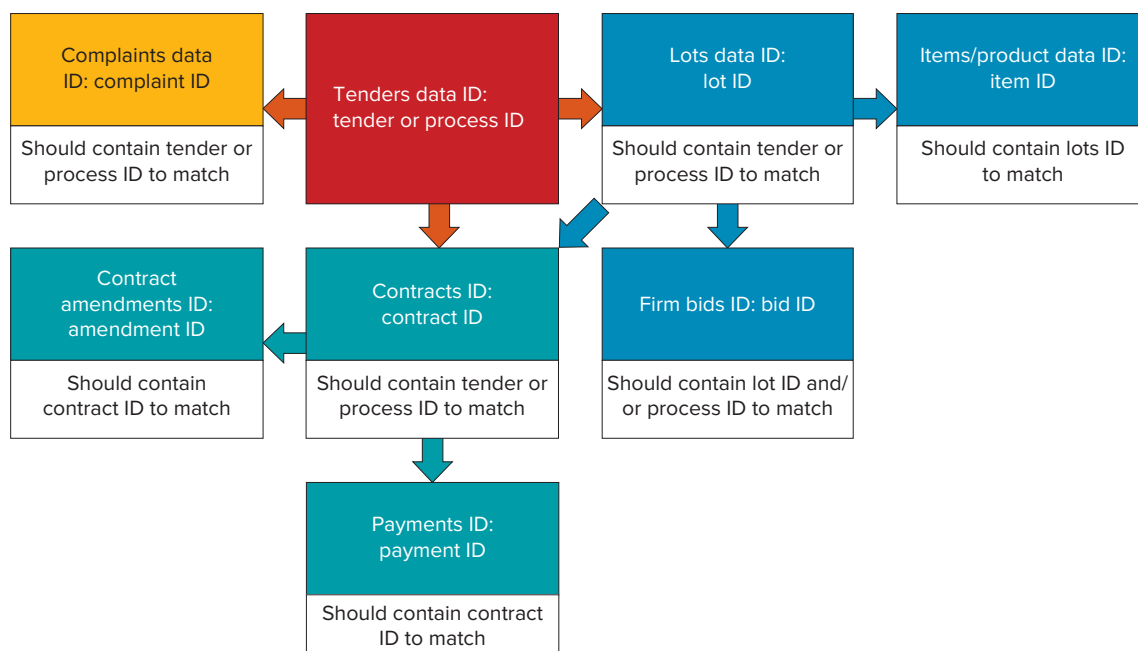
Economy and efficiency	Transparency and integrity	Competition	Inclusiveness and sustainability
<i>Tender and bidding process</i>			
<ul style="list-style-type: none"> <li>• Total processing time</li> <li>• Evaluation time</li> <li>• Contracting time</li> </ul>	<ul style="list-style-type: none"> <li>• Time for bid preparation</li> <li>• Single-bidder tender</li> </ul>	<ul style="list-style-type: none"> <li>• Open procedure</li> <li>• Number of bidders</li> <li>• Share of new bidders</li> </ul>	<ul style="list-style-type: none"> <li>• Share of SME bidders</li> <li>• Share of WOE bidders</li> </ul>
<i>Assessment and contracting</i>			
<ul style="list-style-type: none"> <li>• Awarded unit price</li> <li>• Final unit price after renegotiation</li> </ul>	<ul style="list-style-type: none"> <li>• Share of excluded bids</li> </ul>	<ul style="list-style-type: none"> <li>• Number of bidders</li> <li>• New bidders</li> </ul>	<ul style="list-style-type: none"> <li>• Share of SME bidders</li> <li>• Share of WOE bidders</li> </ul>
<i>Contract implementation</i>			
<ul style="list-style-type: none"> <li>• Final unit price after renegotiation</li> <li>• Time overrun</li> </ul>	<ul style="list-style-type: none"> <li>• Variation orders</li> <li>• Renegotiations</li> </ul>		

Source: Original table for this publication.

Note: SME = small and medium enterprise; WOE = women-owned enterprise.



**FIGURE 12.1 Data Map of Traditional Public Procurement Data**



Source: Original figure for this publication.

The academic literature and practitioners in the field have identified a common set of indicators that are typically used to measure the efficiency, effectiveness, and integrity of the public procurement function. These indicators cover dimensions of public procurement related to methods and procedures (for example, use of open methods), transparency and integrity (for example, time for bid submission), competition (for example, number of bidders), processing time (for example, time for bid evaluation), price (for example, unit prices), and contract implementation (for example, time overrun). (A full list of indicators is provided in appendix D.) In addition to performance indicators, public procurement microdata can also be used for the construction of red flags for corruption or collusion risk. The richness of the data available on public tenders has allowed economists, anticorruption authorities, and competition agencies to develop different screening techniques and has offered the opportunity to test them empirically. Red flags can be useful to identify unusual patterns in certain markets, but these patterns are not sufficient evidence of misbehavior. Rather, red flags can be used as the starting point for further investigation and as sufficient evidence for courts to authorize inspections of dawn raids (OECD 2013). One reason why red flags cannot provide sufficient proof of corruption or collusion is that by design, these data-driven methods can produce false positives (by flagging cases that do not merit further scrutiny) and false negatives (by failing to identify cases that do merit further scrutiny). Given that corruption risk indicators and cartel screens do not directly point to illegal activities, establishing their validity is of central importance.<sup>5</sup> Boxes 12.2 and 12.3 present the existing literature on corruption risk indicators and cartel screens and some recent advances in these techniques thanks to novel machine-learning applications.

Beyond a transactional view of public procurement, there is increasing interest in measuring dimensions of public procurement related to the strategic role it can play to promote inclusive and sustainable growth and the achievement of socioeconomic and environmental objectives. Recent studies and research on these topics have focused both on the development of new procurement indicators (for example, on green procurement and socially responsible procurement) and on linking public procurement data with other data sources to promote a holistic approach to data analytics (for example, firm registry microdata). These topics are discussed in more detail in section 5.

An area that would require further development and research is the measurement of contract implementation quality. Various approaches have been experimented with in the literature, but there

is no agreed-upon strategy yet, and this is a dimension where data constraints are particularly binding. One option would be to use audits data, but the limitations are that audits often focus on compliance with procurement regulations rather than on actual project implementation and that audits data are not typically integrated with public procurement data. Contract supervision data and project management reports could also be used to generate information on contract implementation. The potential for integrating data from various stages of the public procurement cycle and from other functions of the state is discussed further in section 4. Ad hoc data collection could also be considered for specific sectors—for example, through engineering assessments of the material used for the construction of infrastructure projects (Olken 2007) or through visits to hospitals to verify the availability of medicines and their quality. With respect to the construction sector, recent advances in technology (for example, drones and satellite images) can monitor the progress—but not necessarily the quality—of construction work, while information on quality can be obtained from citizen monitoring. More work is needed to assess the pros and cons of different measurement strategies, particularly in terms of the objectivity of different measurement approaches and their scalability.

## **BOX 12.2 What We Know about Corruption Risk Indicators**

The starting point for measuring any corrupt phenomenon is to define the particular behaviors of interest (Mungiu-Pippidi and Fazekas 2020). In public procurement, one definition widely used in both academia and policy considers corruption to be the violation of impartial access to public contracts—that is, a deliberate restriction of open competition to the benefit of a connected firm or firms (Fazekas and Kocsis 2020).

Corruption risk indicators identify the factors and traces of corrupt transactions defined by deliberate competition restrictions favoring connected bidders. Widely used corruption risk indicators in public procurement include single bidding in competitive markets, restricted and closed procedure types, or the lack of publication of the call for tenders (Fazekas, Cingolani, and Tóth 2018). These risk indicators have been shown to correlate with already established indexes of corruption, such as the Control of Corruption scores in the Worldwide Governance Indicators (Fazekas and Kocsis 2020), as well as with other markers of corruption, such as the price of auctions (Fazekas and Tóth 2018), the political connections of bidding firms (Titl and Geys 2019), and proven cases of corruption (Decarolis et al. 2020).

Novel machine-learning applications have been used to advance the measurement of corruption risks. For example, machine-learning approaches have been used on carefully curated data sets of proven corrupt and noncorrupt cases to train algorithms predicting corruption risks (Decarolis and Giorgiantonio 2022; Fazekas, Sberna, and Vannucci 2021). Advanced network science methods have also been increasingly used to detect high-corruption-risk groups of organizations (Wachs, Fazekas, and Kertész 2021).

Corruption risk indicators have been used in numerous civil society and journalistic applications, as well as by multilateral banks and national authorities for policy design and implementation. For example, the European Commission and Organisation for Economic Co-operation and Development's (OECD) Support for Improvement in Governance and Management (SIGMA) initiative (OECD and SIGMA 2019) has regularly monitored some risk indicators, such as single bidding and the publication of calls for tenders. The International Monetary Fund (IMF) has endorsed corruption risk indicators and models predicting the price impacts of such risks as valuable inputs to addressing macrocritical risks. The European Investment Bank uses public procurement risk indicators, combined with internal financial risk assessments, to select projects for prior integrity reviews (Fazekas, Ugale, and Zhao 2019), an approach highlighted as good practice by the European Court of Auditors (Adam and Fazekas 2019).

### BOX 12.3 What We Know about Collusion and Cartel Screens

The characteristics of collusive behavior in public procurement markets are similar to those in conventional markets: companies coordinate their behavior regarding price, quantity, quality, or geographic presence to increase market prices.

Cartel screens are defined according to two key competition and economy principles. First, it is expected that in competitive tendering processes, bids will be submitted independently; therefore, signs of coordination between bidders can be interpreted as signs of collusion. Second, bids submitted by independent competitors should appropriately reflect the costs of each bidder in a competitive market. Based on these two criteria, various elementary collusion risk indicators have been developed for the early detection of collusive bidding, such as the submission of identical bids, high correlation between bids, lack of correlation between the costs and the bid submitted by each bidder, the relative difference between the lowest and the second lowest bid price per tender, the relative standard deviation of bid prices per tender, and the range of submitted bid prices per tender.

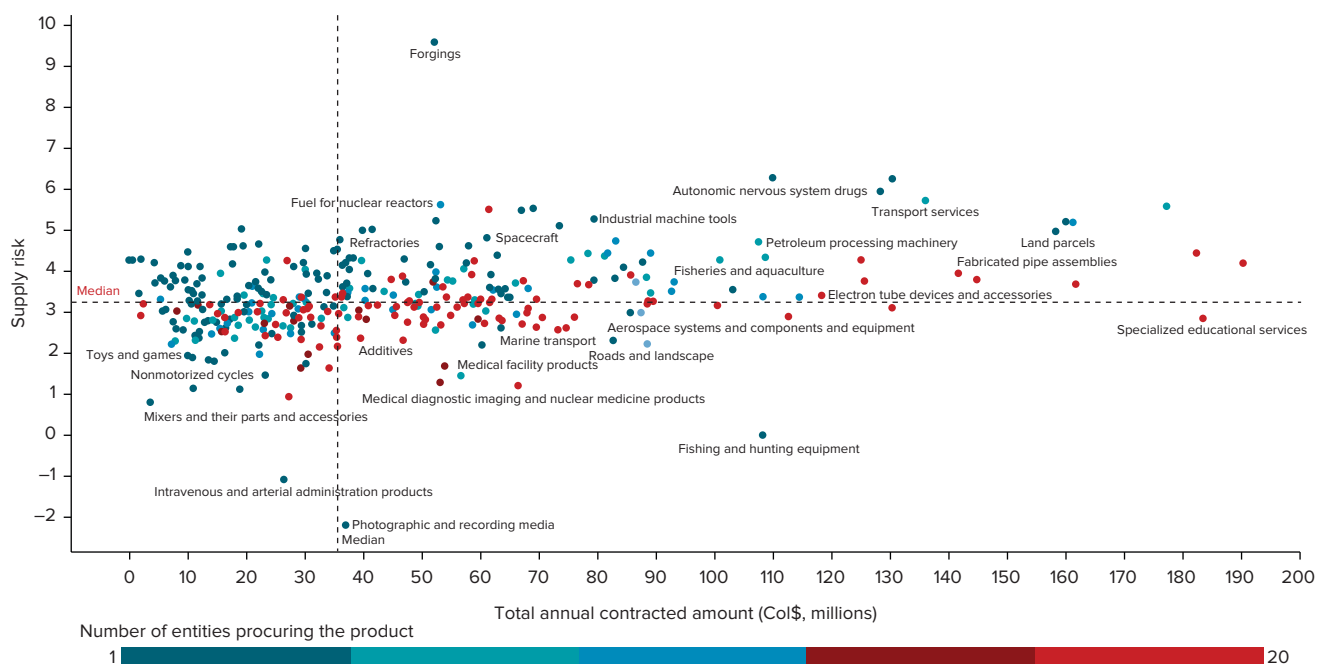
Increasingly, more advanced statistical techniques have been used to define cartel screens as well as develop algorithms that minimize the probability of both false positives and false negatives. For example, Conley and Decarolis (2016) have developed statistical tests of coordination based on randomization inference methods. These tests identify unexpected firm behaviors conditional on their characteristics—for example, the unexpected joint entry of firms within a group of bidders given observed firm and process characteristics. Huber and Imhof (2019) study the performance of different screening algorithms, specifically a lasso logit regression and a weighted average of predictions based on bagged regression trees, random forests, and neural networks. Most interestingly, these recent examples use machine-learning techniques to identify optimal algorithms thanks to the combination of public procurement data and judicial and auctions data for validation.

### Government Monitoring of Public Procurement

With the increasing use of e-GP systems and access to structured procurement data, public procurement authorities are often using the common procurement indicators discussed in appendix D to monitor the performance of their own public procurement systems. These public procurement authorities use the available procurement data to characterize national public procurement spending and identify performance and compliance gaps in the national public procurement system. This descriptive analysis can include time trends or comparisons of performance indicators across procuring entities, regions, and types of contract, as well as types of procedure, sector, or supplier. In some cases, this exercise may be mandated by international treaties or organizations, or as a prerequisite to access financing from multilateral development banks.<sup>6</sup> The results of this monitoring are often reported in the form of annual reports on the functioning of the procurement system, and they can be used for informing and guiding reform efforts and the development of new strategies and policies in public procurement. For example, in Poland, the Public Procurement Office (PPO) prepares the annual report on the functioning of the procurement system, which is posted on the PPO website following approval by the Council of Ministers.<sup>7</sup>

Public procurement agencies may use certain tools or mechanisms to describe their procurement data and trends. For example, spend analysis is a widespread approach for monitoring and assessing public procurement, consisting of various tools (for example, the Kraljic matrix; see figure 12.2) that provide a first overview of the procurement market and, specifically, what is being purchased, by whom, and from which suppliers. This analysis is used to identify the areas (products, entities, and suppliers) for which the improvement of efficiencies is expected to have the largest budget implications, to define procurement strategies, and to adapt relationship management for different types of suppliers.

**FIGURE 12.2 Kraljic Matrix for Colombia**



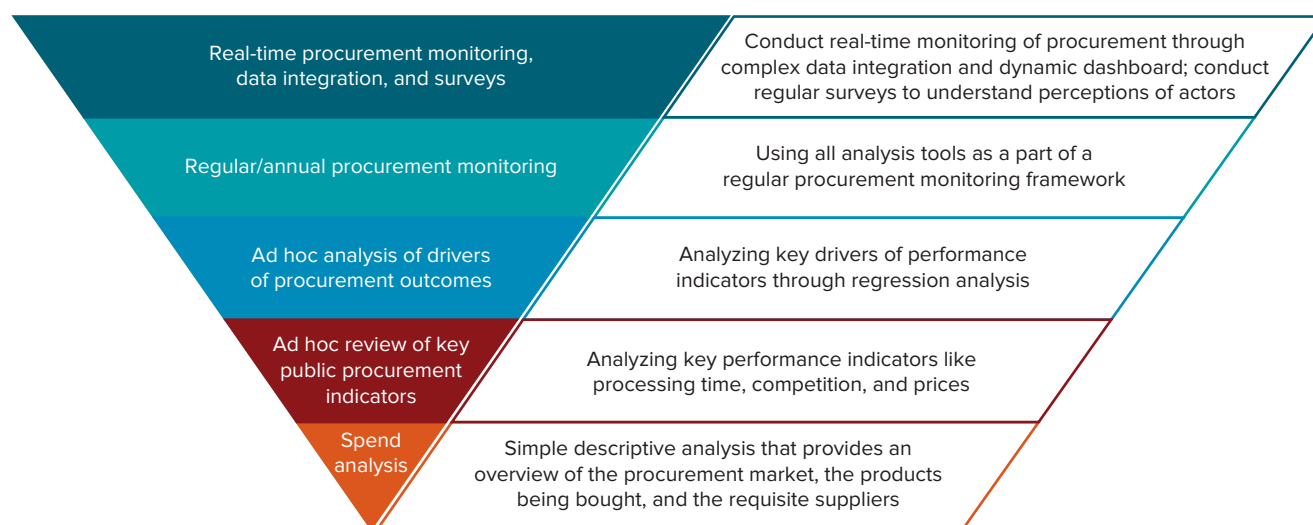
Source: Original figure for this publication based on Colombia's open public procurement data.

Note: Col\$ = Colombian peso.

The analysis of performance and compliance offers another set of tools typically used by public procurement authorities and audit authorities to monitor the national public procurement system. This monitoring may include the compliance of procurement regulations as reported by procurement agencies.<sup>8</sup> This type of descriptive analysis explores efficiency indicators, like competition, price, and processing time, as well as the extent to which procurement regulations (for example, regulations on contract thresholds, the use of open participation methods, or the use of centralized procurement methods) are met. This type of analysis is useful to describe the efficiency gaps that exist in the public procurement system and to help prioritize audit activities. For example, Best, Hjort, and Szakonyi (2019) show that in Russia, individuals and organizations of the bureaucracy together account for more than 40 percent of the variation in prices paid and that moving the worst-performing quartile of procurers to 75th percentile effectiveness would reduce procurement expenditures by around 11 percent, or US\$13 billion each year.

As illustrated in figure 12.3, these descriptive analysis tools are the least complex uses of public procurement administrative data. Figure 12.3 shows a ladder for analysis tools in procurement monitoring, in which each step of the ladder represents analytical tools conducted on procurement at different levels of complexity. Beyond descriptive analytics, diagnostic analysis (for example, regression analysis) can be used to identify the drivers of performance and therefore inform the government of potential strategies to improve efficiency and integrity. The following section discusses in detail diagnostic analysis tools for data-driven policy making. However, descriptive analysis tools can still be among the most advanced uses of public procurement when they are systematically embedded in the public procurement monitoring and reporting function—for example, for the preparation of annual reports or through interactive dashboards, which typically require institutional reorganization and the acquisition of necessary skills in the public procurement authority.

**FIGURE 12.3 Complexity Ladder for Analysis Tools in Procurement Monitoring and Evaluation**



Source: Original figure for this publication.

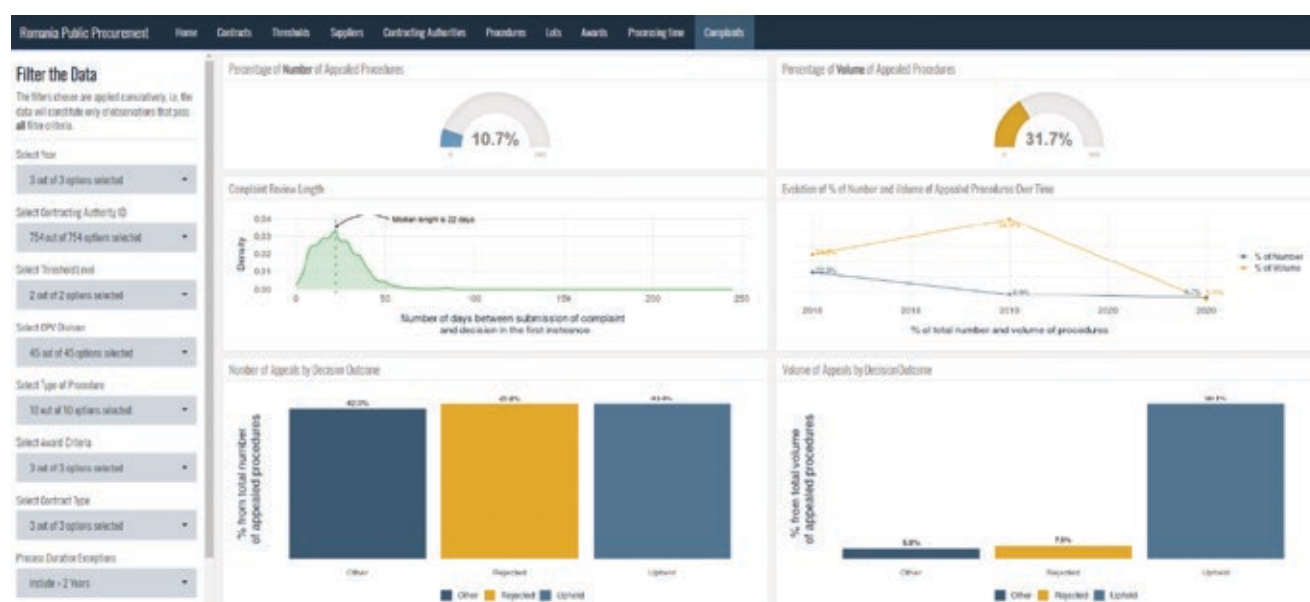
Interactive dashboards are increasingly widespread tools for monitoring public procurement through descriptive analysis because they enable procurement authorities to track, analyze, and display key performance indicators through customizable and user-friendly visualizations. One of the great advantages of these dashboards is that they allow users to focus their analysis at different levels of government or in specific markets. Depending on how the public procurement system is set up, these interactive dashboards can be connected directly with the underlying e-GP system or can be regularly updated. These dashboards can be built for the use of the national public procurement authorities and individual procuring entities for monitoring their procurement activity, or they can be made public for greater accountability of the public procurement system.

For example, between 2020 and 2021, the World Bank worked with the National Agency for Public Procurement (ANAP) in Romania to develop a monitoring mechanism, in the form of a dashboard that would enable the public procurement agency to track its own key performance indicators and would enable easy reporting to the EU (World Bank 2019). The dashboard (figure 12.4) was developed in close collaboration with the ANAP to ensure that the most relevant indicators were captured. Regular data analysis workshops conducted by the World Bank ensured that staff in the ANAP had the capacity and training to replicate and add to the dashboard to ensure its sustainability in the long run.

### Data-Driven Policy Making

The analysis of public procurement data can enable procurement agencies to develop key procurement policies or refine and assess existing regulations. Data analytics allows agencies to assess existing efficiency gaps and understand the drivers of performance, and these empirical insights are useful to identify and prioritize potential areas for interventions and reform efforts. For example, in 2019, the World Bank conducted a complete and comprehensive analysis of Uruguay's public procurement data that generated discussion and space for policy recommendations to improve the performance of the procurement system. This analysis identified demand consolidation as the most significant potential source of savings, with framework agreements being the most effective instrument to implement the strategy. Based on these empirical insights, in 2021, the World Bank worked with the Regulatory Agency for Public Procurement and the Central Procurement Unit within the Ministry of Economy and Finance to implement these recommendations, specifically building capacity in the generation and management of framework agreements and supporting the development of pilot framework agreements for goods and services with the greatest savings potential.

**FIGURE 12.4 National Agency for Public Procurement (ANAP) Dashboard, Romania**



Source: Screenshot of the ANAP dashboard, World Bank 2019.

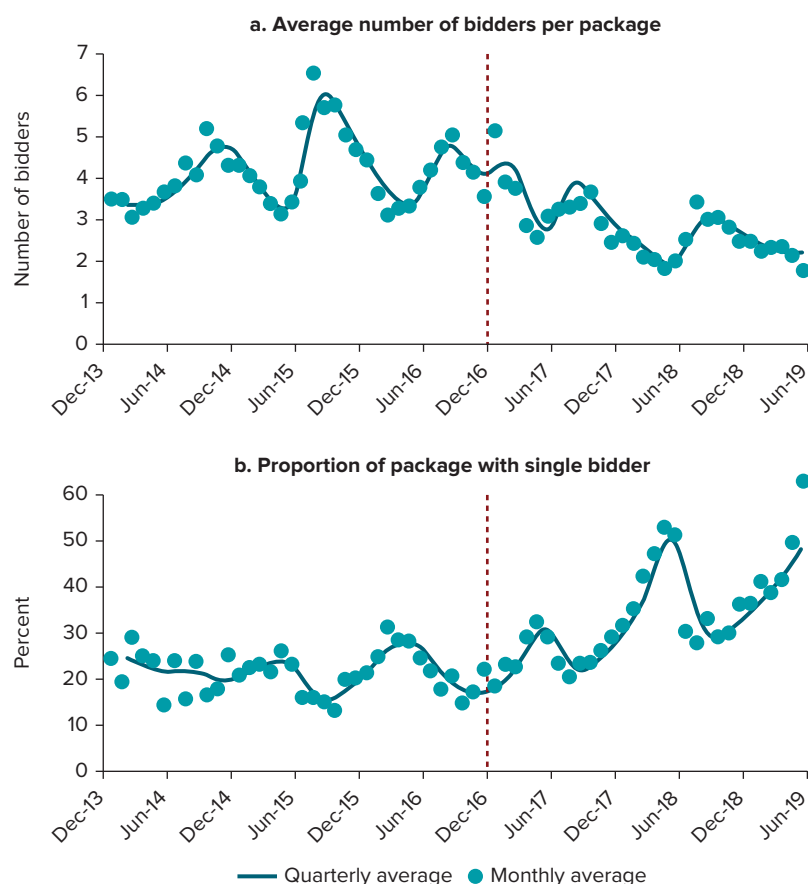
Data analytics is also a useful tool to monitor the consequences of new policies, assess whether they are delivering the expected outcomes, and understand potential trade-offs. For example, in 2020, the World Bank worked on an assessment of the national public procurement system in Bangladesh, the objectives of which were to identify its strengths and weaknesses, formulate appropriate mitigation measures for identified gaps, and develop an action plan for future system development (World Bank 2020). The assessment was built on various data-driven components, such as an analysis of the so-called 10 percent rule (rejecting a tender that is 10 percent below or above the estimated cost) introduced by the government of Bangladesh in December 2016 for the procurement of works using the open tendering method. The primary objective of this policy was to improve the quality of construction works and reduce the risk of cost overruns by restricting bidders from quoting very low prices. However, procuring entity officers largely expressed the opinion that the quality of works had not improved after the introduction of the 10 percent rule, and quantitative analysis of time trends also revealed that this rule had produced undesired consequences, such as decreasing competition (figure 12.5). These empirical insights were instrumental in providing fact-based recommendations to the government about reevaluating the 10 percent rule.

With respect to understanding potential trade-offs, increasing attention toward the multidimensional nature of public procurement implies that policies and strategies should be assessed based on a variety of considerations, including efficiency, integrity, value for money, and socioeconomic and environmental aspects. There are many trade-offs associated with the public procurement function in connection to the private sector and public service delivery, and a comprehensive approach to procurement data analytics allows agencies to correctly assess the potential trade-offs associated with procurement policies and provide complete and accurate policy recommendations. For example, a 2021 World Bank report on the use of framework agreements (World Bank 2021a) shows that in Brazil, the use of framework agreements could reduce unit prices and avoid repetitive processes, but it could also discourage participation by small and medium enterprises (SMEs) and their likelihood of being awarded a contract (table 12.2).<sup>2</sup>

These examples show that quantitative analysis can be quite powerful in identifying key procurement trends in a country and can be foundational in developing and evaluating procurement policies. Given the



**FIGURE 12.5** Assessment of Bangladesh's 10 Percent Rule



Source: World Bank 2020.

**TABLE 12.2** Regression Analysis of Framework Agreements versus Other Open Methods, Brazil

Outcome of interest	Unit price (log)	SME winner
Framework agreements vs. other open methods	-0.0919** (0.0407)	-0.0198*** (0.00582)
Observations	172,605	166,399
R-squared	0.910	0.566

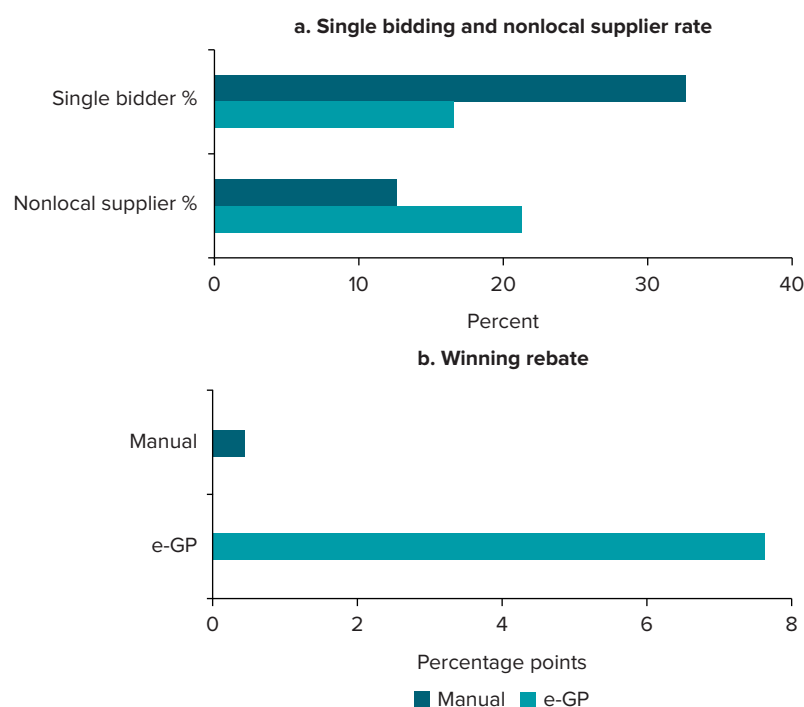
Source: World Bank 2021a.

Note: Model specifications: Comparing FAs and non-FA open methods for the purchase of the same product by the same entity (product—entity FE), with year and quarter FEs. FA = framework agreement; FE = fixed effect; SME = small and medium enterprise. Robust standard errors in parentheses, \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$ .

abundance of microdata in countries with e-GP systems, public procurement can also be an ideal space for implementing impact evaluations of specific procurement policies. Impact evaluations represent one of the most reliable forms of policy assessment because they allow agencies to contrast actual outcomes with counterfactual scenarios by comparing units subjected to a given policy intervention to otherwise similar units that have not yet been treated.

For example, starting in 2011, the government of Bangladesh began the rollout of a comprehensive e-GP system, and the World Bank worked with the Central Procurement Technical Unit to evaluate the impact of the new system on procurement outcomes.<sup>10</sup> The evaluation revealed that the implementation of the e-GP system had led to an improvement in public procurement performance, as demonstrated, for example, by an increase in winning rebates, a decrease in single bidding, and an increase in contracts awarded to nonlocal suppliers (figure 12.6). During the piloting stage, the preliminary results from the evaluation

**FIGURE 12.6 Procurement Outcomes under Manual versus Electronic Government Procurement Systems, Bangladesh**



Source: Turkewitz, Fazekas, and Islam 2020.  
Note: e-GP = electronic government procurement.

helped demonstrate that the new e-GP system was having a good impact on efficiency, transparency, and competition, and this was extremely useful to build consensus and political support around this difficult reform (Turkewitz, Fazekas, and Islam 2020). This example demonstrates the feasibility and usefulness of impact evaluations for navigating the political economy of reforms. Specifically for public procurement, the abundance of data generated by the e-GP system creates a rich space for research because the data already available from the existing e-GP system at the time of piloting and implementing new strategies allow for the tracking of procurement outcomes, from baseline to endline, with no additional costs for data collection.

### Monitoring of Public Procurement by the Public

Public procurement is one of the public administration functions with a prevalence of publicly available data. With the increase in the use of e-GP systems, there is greater potential for increasing transparency and accountability in public procurement processes through monitoring by the public. Open data can be used by civil society, the media, and citizens to acquire information on specific contracts, buyers, or products.<sup>11</sup> Increased transparency and accountability can enable citizen engagement in monitoring public procurement and, therefore, enhance trust between citizens and the state, strengthen the social contract, and improve the quality of contract execution. For example, a citizen engagement project was implemented by the World Bank in collaboration with the Central Procurement Technical Unit in Bangladesh to enable and support the monitoring of civil works projects by local community groups (Hassan 2017). Through the project, “citizen committee” members frequently visited project offices and reported anomalies in civil works projects to the engineer’s office. The project reduced information gaps and increased trust between government officials and local community leaders on civil works projects in their areas, and it also reduced monitoring-related transaction costs through higher citizen engagement.

Making public procurement data available to the public has great potential to increase transparency and accountability. However, even when data are publicly available online, it may be challenging to extract useful

and actionable information from open public procurement data, which requires connecting data sources from different stages of the procurement cycle, constructing relevant indicators, and analyzing microdata for large numbers of tenders and contracts. In light of these challenges, various international organizations have been developing dashboards to ease access to public procurement data for citizens, civil society, and researchers.

For example, in 2021, the World Bank, in collaboration with the Government Transparency Institute, launched a prototype of a global dashboard that provides access to open data from national electronic procurement systems from 46 countries, as well as open data on World Bank- and Inter-American Development Bank-financed contracts for over 100 countries.<sup>12</sup> Similarly, the Opentender dashboard provides access to tender data from Tenders Electronic Daily (TED), which covers 33 EU jurisdictions, including 28 EU member states, Norway, the EU institutions, Iceland, Switzerland, and Georgia.<sup>13</sup> The creation and maintenance of these global public goods, and the use of open public procurement data in general, would be simplified by the adoption of common data standards globally, and the Open Contracting Data Standard from the Open Contracting Partnership provides a promising starting point.<sup>14</sup>

Some governments are also creating public dashboards using their own public procurement data. For example, an intuitive and simple dashboard prepared by the National Informatics Centre in India and hosted on the Central Public Procurement Portal allows users to get some key performance and compliance indicators on public procurement in India.<sup>15</sup> This dashboard not only allows the government to monitor key procurement trends but also reports these indicators to the public for greater transparency and accountability. The public procurement authority in Ukraine has also designed and published a public dashboard to increase transparency and accountability in public procurement.<sup>16</sup> The COVID-19 crisis prompted more countries to increase transparency and enable public scrutiny of purchases made during the health emergency, such as Brazil, Moldova, Lithuania, and South Africa.<sup>17</sup>

## SETTING UP DATA INFRASTRUCTURES AND CAPACITY FOR MEASUREMENT ON PROCUREMENT

### Centralize Public Procurement Data

To ensure that public procurement data are used effectively for decision-making, it is necessary that data are homogeneously collected and maintained across procuring entities and connected to a centralized platform. This is necessary both for external users and researchers as well as for the national agency or central procurement authority. Public procurement data are often decentralized or housed by different institutions responsible for managing different parts of the procurement process, and this may complicate the process of data harmonization and centralization, especially in countries without an e-GP system and where reporting of procurement activities to a central authority is not mandatory or audited. For example, in 2021, a World Bank team conducted a data collection exercise of paper-based procurement records in St. Lucia, Dominica, St. Vincent and the Grenadines, and Grenada to assess public procurement systems in the four countries. The key constraint on data collection was the decentralization of the information among multiple institutions; ultimately, the data had to be collected by enumerators from different procuring entities and national authorities.

Enabling factors for the centralization of public procurement data include legislation, administrative structure, and data infrastructure. Simple mechanisms like an annual data collection exercise at the central level, in which procuring entities send Excel files to a central authority (which audits a sample for data accuracy), can help slowly transfer local data storage mechanisms to more efficient, centralized data management systems. For example, this was recommended in the case of St. Lucia, Dominica, St. Vincent and the Grenadines, and Grenada, with the additional recommendation to conduct regular audits of the quality and accuracy of the data provided by each procuring entity to the central authority. This step can be a key foundation on which an appetite for data literacy and digitalization can be created among governments. In contrast to data sitting in physical files in different procuring entities, a centralized data collection mechanism can allow for easy access to procurement data even in cases where an e-GP system has not yet been implemented.

## **Integrate Data from Various Stages of the Public Procurement and Contract Management Cycle**

Data integration can be an important step in exploring all the stages of the public procurement and contract management cycle. Data integration can be accomplished through two related steps: (1) matching data from various procurement stages and (2) expanding the availability of data to study procurement more holistically. With respect to the first step, public procurement data typically cover the following stages: tendering process, bidding process, bid evaluation, contract award, and contract signing. To meaningfully use this data, it is necessary that the tenders data, lots data, bids data, and contracts data are consistently organized and can be connected (see figure 12.1).

The second step in data integration is expanding the availability of data to cover the full public procurement and contract management cycle, including parts of the process that are not typically included in procurement data, such as data on public procurement planning and budgeting, tender preparation data, contract execution data (for example, data on subcontracting and payments to vendors), complaints data, and proprietor information and beneficial ownership data.

There is great scope for using these additional data sources for procurement data analytics, and some countries are taking steps in this direction. The development of integrated data systems requires the close engagement and partnership of multiple government institutions that house different parts of the procurement and contract management cycle. For example, as part of the design and development of the monitoring mechanism delivered to the ANAP in Romania (see above), the World Bank was able to add data on complaints registered in public procurement processes to the dashboard by leveraging existing data-sharing agreements between the ANAP and the National Council for Solving Complaints. While the establishment of streamlined data management systems is a necessary technical requirement for a data integration process, the most significant constraints often lie in the administrative and bureaucratic structures that may complicate collaboration and data-sharing agreements between different institutions.

## **Data Quality and Completeness**

Data quality and completeness are crucial determinants of the quality of empirical analysis that can be performed on public procurement data. Common issues in public procurement data are noted across countries, both in the data obtained from open sources as well as in the data obtained from governments. Some of these common issues, which are listed in box 12.4, range from missing data to incorrect or ambiguous data structures that restrict or hinder comprehensive empirical analysis.

Some data quality and completeness issues can be mitigated through relatively simple and practical steps by the government. The e-GP system can include automated data quality checks during data entry by procuring entity officers—for example, checking that the procurement process dates follow a logical order and that the contract amounts are within reasonably expected ranges. Detailed audits of the data entered by procuring entity officers may also be conducted regularly to ensure that the official tender and contract documents reflect the data entered into the system. The central procurement authority can also review the data maintained in the procurement system to assess their completeness, especially in light of the compliance and performance indicators the government is interested in monitoring. Last, implementing a fully transactional system that manages the entire procurement process from start to finish and allows multiple government agencies and ministries to engage with different parts of the procurement process may allow for the ideal data integration environment to holistically analyze the full procurement process and all related parts in public administration.

When planning for the public disclosure of procurement data, the same principles of data quality and completeness apply to ensure data transparency and accessibility. In addition, in this case, it is important that the raw data entered into the system are made public, not only the indicators and measures constructed from the administrative microdata. Observations across several countries also show that open data and good policies for data openness and transparency do not necessarily correlate with data quality

## BOX 12.4 Examples of Common Issues with Data Quality and Completeness

- **Missing observations or variables and data errors that pertain to important aspects of the procurement process.** In most countries, electronic government procurement (e-GP) data are not created directly from digitized tenders and contracts but are separately inputted by procuring entity officers. In these cases, the procuring entity officers may still have the option to leave certain data fields blank. This generates gaps in the data and can also indicate strategic behavior by procuring entity officers, who may systematically choose to leave more sensitive data fields blank. Data quality and completeness should be systematically reviewed by a central authority, including for data disclosed to the public.
- **Ambiguity in the level of observation for the data.** Data from different stages of the public procurement cycle (figure 12.1) should be meaningfully connected for analysis through unique identifiers, such as the tender ID or the entity ID. The absence of unique identifiers creates ambiguity in the interpretation of the data and hinders comprehensive empirical analysis. For example, in the case of framework agreements, there can be multiple contracts, buyers, and suppliers under a tendering process, and multiple orders can be associated with the same umbrella contracts. Having a clear and unambiguous data structure is necessary to correctly represent framework agreement processes and enable accurate analysis.
- **Correction of incorrect entries in the e-GP system by entering the entire tendering process again.** This issue is observed in countries where e-GP data are not created directly from digitized tenders and contracts but are separately inputted by procuring entity officers. Possibly because of integrity concerns, some e-GP systems do not allow officers to correct information already entered into the system in the event of data-entry errors. In these cases, the officers' only option is to create a new entry, but the system does not record which entry is correct and which entry is wrong.
- **Poor data integration during transitions from one e-GP system to another.** Throughout the digitalization of public procurement, countries may shift from one e-GP platform to another. For example, Romania transitioned from a platform called SEAP (Sistemul Electronic de Achizitii Publice) to an upgraded platform called SICAP (Sistemului Electronic Colaborativ de Achizitii Publice), and Colombia from a platform called SECOP I to an upgraded platform called SECOP II (Sistema Electrónico para la Contratación Pública). In cases of transition between e-GP systems, it is necessary to ensure that data from both platforms can be integrated and that procuring entity officers cannot enter data for procurement processes in both platforms during the transition.

and completeness. For example, the Open Contracting Data Standard provides guidelines on the effective disclosure of public procurement data to the public, with the ultimate goal of increasing transparency in procurement and allowing analysis of procurement data by a wide range of users. While an increasing number of e-GP systems follow the Open Contracting Data Standard for the public disclosure of procurement data, how well disclosure is implemented largely depends on the quality and completeness of the data made publicly available.

### Building Capacity for Statistical Analysis and a Culture of Data-Driven Policy Making

The adoption of e-GP systems has created a great wealth of data, but it is not obvious that their use and impact are currently being maximized by governments. The development of the capacity for statistical analysis and a culture of data-driven decision-making can help maximize the potential of the microdata available through e-GP platforms. This may include the creation or strengthening of a dedicated statistical office within the public procurement authority.

For example, as part of the design and development of the monitoring mechanism delivered to the ANAP in Romania (see above), the entire monitoring mechanism was created in close collaboration with ANAP staff through weekly capacity-building workshops and meetings to discuss the operational workflow of the monitoring mechanism. This close collaboration and cocreation of the interactive dashboard for visualizing key procurement indicators allowed the government to engage with the data-cleaning and visualization process and built an appetite for data analysis. ANAP staff were provided with the necessary skills and knowledge to edit and develop the code that was used to create the interactive dashboard. Engagements like this allow products like an interactive dashboard to be hosted in a data-curious and analytical environment that builds long-term sustainability through the empowerment of its users.

Beyond statistical capacity and data analytics skills, the proactive use of data and evidence to drive policy-making decisions also requires the necessary organizational culture, institutional arrangements, and incentive systems. For example, data and empirical evidence can be used to improve the performance of procuring entities. This requires the necessary skills and tools to exploit the potential of data analytics, but it also depends on other systemic factors, such as whether and how the performance of procuring entities is evaluated, whether there are consequences of performance evaluations, whether procuring entity officers are incentivized to improve their efficiency and effectiveness, and whether procuring entity officers have space to make discretionary decisions or instead are expected to merely execute regulations. These considerations are related to a broader discussion on management practices in public administration and specifically in procuring entities, and the following section provides more detail on how some of these aspects can be studied empirically.

## **A WHOLE-OF-GOVERNMENT APPROACH: STRATEGIC COMPLEMENTARITIES TO PUBLIC PROCUREMENT DATA**

### **Measuring the Socioeconomic and Environmental Dimensions of Public Procurement**

Increasingly, governments consider using public procurement as a strategic tool to sustain the private sector, especially groups of firms that are traditionally underrepresented in public procurement, such as SMEs and women-owned enterprises (WOEs). Similarly, governments are increasingly adopting green public procurement (GPP) strategies, such as green evaluation criteria, green eligibility criteria, or life-cycle approaches to costing (box 12.5).<sup>18</sup>

However, there is no clear evidence of the best public procurement strategies and policies to achieve these socioeconomic and environmental outcomes. For example, from a theoretical point of view, it is not clear how to incentivize the participation of SMEs in public procurement effectively and efficiently. While this might be achieved through targeted policies (for example, preference policies or set-aside quotas), these policy tools might be distortionary (Medvedev et al. 2021; OECD 2018) or suffer from poor implementation and compliance. Relying on untargeted policies can be an alternative, but it is perhaps a less impactful approach. Two studies conducted on the same preferential treatment program for small firms in California elucidate these potential trade-offs, with Marion (2007) finding that procurement costs are 3.8 percent higher on auctions using preferential treatment and Krasnokutskaya and Seim (2011) finding that those distortionary effects are not huge in comparison to benefits to firm growth. With limited evidence on the impact and trade-offs of these different policy options, there are no clear guidelines on the best strategies to involve SMEs and other underrepresented groups in public procurement.

As another example, some public procurement laws mandate the application of green criteria for bid evaluation, especially in sectors such as transport (for example, types of vehicle and emissions) and construction (for example, construction materials) (Palmujoki, Parikka-Alhola, and Ekroos 2010), but it is unclear what the direct and indirect cost implications of these requirements are. By design, GPP introduces additional laws and regulations, requirements for firms, and more complex criteria for bid evaluation. Therefore, it is natural that there might be concerns about whether GPP compromises the efficiency of public procurement procedures and reduces the attractiveness of public procurement contracts for firms. Providing robust knowledge on the costs and benefits of GPP will support governments in making informed decisions and might remove some of the concerns that prevent broader adoption.



This focus and strategic approach to public procurement requires that public procurement data be expanded to include the necessary information to measure the socioeconomic and environmental dimensions of public procurement, such as by associating an SME tag with bidders and suppliers or by labeling tenders that follow GPP principles. For example, Nissinen, Parikka-Alhola, and Rita (2009) develop a detailed list of environmental indicators to measure GPP, including indicators on product characteristics, policy attached, level of emission of the company, chemistry, and amount of energy used. In practice, across countries, there has been some progress in tagging SME firms—for example, in Croatia, Romania, and Colombia—but very limited progress in GPP (see box 12.5). This impedes advancing the empirical literature on the effectiveness of different policy alternatives, and it also prevents governments and civil society from monitoring the actual use and implementation of GPP legislation.

### BOX 12.5 What We Know about Green Public Procurement

Green public procurement (GPP) is defined by the European Commission (2008) as “a process whereby public authorities seek to procure goods, services and works with a reduced environmental impact throughout their life cycle when compared with goods, services and works with the same primary function that would otherwise be procured.”

GPP can take different forms, and different measurement options should be considered depending on the GPP approach adopted for each specific tender. A first categorization of GPP approaches is as follows (World Bank 2021b):

- **Contract performance clauses** ensure winning suppliers deliver a contract in an environmentally friendly way and continuously improve their environmental performance throughout the contract duration. Examples of these clauses include the requirement to deliver goods in bulk to reduce packaging, the requirement to optimize delivery schedules, and the requirement to recycle or reuse packaging after delivery.
- **Award criteria** can include optional environmental criteria to encourage and reward bidders that propose solutions with improved environmental performance (for example, a higher percentage of recycled content and functional criteria that allow supplier innovation). This approach requires that procuring entities set weights to evaluate the various dimensions of a proposal, such as environmental criteria and price.
- **Qualification criteria and technical specifications** prescribe core environmental criteria that bidders and/or offers must meet to satisfy the requirements of the tender (for example, minimum recycled content or bans on toxic chemicals).<sup>a</sup> For example, supplier-selection criteria aim to ensure that participating bidders have the technical capabilities, ethics, and management processes in place to deliver on the desired environmental outcome. Examples of these criteria are proof of compliance with environmental laws and regulatory standards, the existence of qualified staff with environmental expertise, and environmental certifications.
- **Life-cycle approaches** consider the total cost of ownership (TCO) of a good, service, or work, an estimate that considers not only its purchase price but also the operational and maintenance costs over its entire life cycle. The life-cycle cost (LCC) goes further than the TCO by also taking into account the cost of environmental externalities that can be monetized (for example, greenhouse gas emissions and pollution fees).

Given the speed of innovations in this field, it may be challenging for procuring entities to define appropriate environmental criteria that correspond to current benchmarks and environmental criteria that can be expected of and met by private sector actors. There are various mechanisms that can help procuring entities determine the “environmental friendliness” of a good, service, work, or firm (World Bank 2021b):

*(continues on next page)*

## BOX 12.5 What We Know about Green Public Procurement (*continued*)

- **Ecolabels** are labels of environmental excellence awarded to products and services meeting high environmental standards throughout their life cycle. Ecolabels can be awarded based on third-party certification, supplier claims of environmental conformity, or third-party validation of an environmental product declaration.
- **“Green” product lists** or online databases of preapproved green goods, works, and services can be created by governments and made available to procurers across the government.
- **Framework agreements** can be set up by central procurement authorities to include GPP approaches, making it easier for all procuring entities to purchase green without entering into difficult processes for market analysis, tender design, and bid evaluation.

a. An example of these criteria is detailed by the European Commission on the EU GPP criteria page of its website: [https://ec.europa.eu/environment/gpp/eu\\_gpp\\_criteria\\_en.htm](https://ec.europa.eu/environment/gpp/eu_gpp_criteria_en.htm).

### Linking Public Procurement Data to Other Dimensions of the Public Sector and Public Administration

Public procurement is multidimensional and critically interconnected with other functions of the public sector and public administration. For example, the participation of small firms in the public procurement market may be influenced by the ease of access to finance or by tax subsidies provided to certain disadvantaged firms. Similarly, the administrative burden of public procurement processes may be influenced by the staffing, training, and resources in the local procuring entities. The incentives of participants in a procurement process may be influenced by several factors. A promising area for advancement in public procurement research would be to collect and integrate data from other parts of the public sector, justice, and tax administration to create novel integrated data sets providing a holistic picture of the procurement function. This would provide governments with comprehensive information to develop innovative and impactful procurement strategies, as well as allow researchers to holistically explore the environment within which procurement is conducted.

Many potential data sets could be used to extend the analysis of public procurement through other dimensions of public administration. One example is linking tax registries and public procurement data. Data on tax filings by firms could be useful to characterize the firms operating in public procurement markets—for example, in terms of size—and the link between public procurement and the growth of firms (Ferraz, Finan, and Szerman 2015), as well as to assess the effectiveness of policies that intend to favor the participation of SMEs in public procurement.

Another potential data set is linking public procurement data with audits data. If properly designed, audits can be an effective tool to disincentivize malpractice in public procurement. However, as demonstrated by Gerardino, Litschig, and Pomeranz (2017), the design and targeting of audits can distort incentives for procurement officers. For example, procurement officers may be less likely to use competitive methods if they expect these procedures will be more likely to be audited due to their complexity, or they may be less likely to comply with regulations that are difficult for auditors to monitor, such as the application of preferential policies for SMEs or the application of green award criteria.<sup>19</sup>

Public procurement data can also be complemented with complaints data and judicial data. Box 12.3 discusses the potential for matching public procurement data with judicial data to validate collusion-screening algorithms. Beyond this type of application, there is also space for further research on how performance in public procurement functions is affected by the efficiencies and performance of the judicial sector. Coviello et al. (2018) have demonstrated, in the context of Italy, the implications of inefficient

courts on procurement outcomes, such as longer delays in the delivery of public works, a higher likelihood that contracts are awarded to larger suppliers, and higher shares of payments postponed after delivery. Further studies on the link between public procurement and the justice sector would be necessary to advance our understanding of how these two functions of the state influence each other—for example, whether judicial investigations have an impact on processing and contract execution times, which types of procedures are more likely to result in complaints or investigations, whether the risk of complaints and appeals is a barrier to firm participation, and whether the efficiency of courts has an impact on the propensity of procuring entities to enforce late penalties.

The integration of public procurement into overall public finance management, budgeting, auditing, and service delivery processes has a high potential to lead to better utilization of public resources through better information transmission, standardization, and automation and increased accountability (OECD 2017). Despite this potential, the integration of e-procurement systems into other e-government systems is not yet a common practice. For example, based on a 2016 review of public procurement systems in OECD countries, e-GP systems are most often integrated with business registries (eight countries), tax registries (seven countries), budgeting systems (six countries), and social security databases (six countries) (OECD 2017). Data integration is an area where further work is needed to promote a whole-of-government approach from a data perspective.

### Insights on Public Procurement Data from Survey Data

Along with using administrative data on public sector and public procurement, surveys of procuring entity officers and firms provide important context on the environment in which procurement is conducted. Surveys of procuring entity officers can be used to measure procurement-related information otherwise unavailable in the administrative data, such as time for tender preparation, contract execution quality, and firm performance. For example, in an assessment of the public procurement system in Croatia, the World Bank collected survey responses from procuring entity officers on the quality of delivered goods and services by firms and on contract management deadlines, such as the date of delivery and the final payment amount for contracts. These indicators were not available in the publicly available data in Croatia, and this data collection exercise was successful in identifying constraints during the contract management phase.

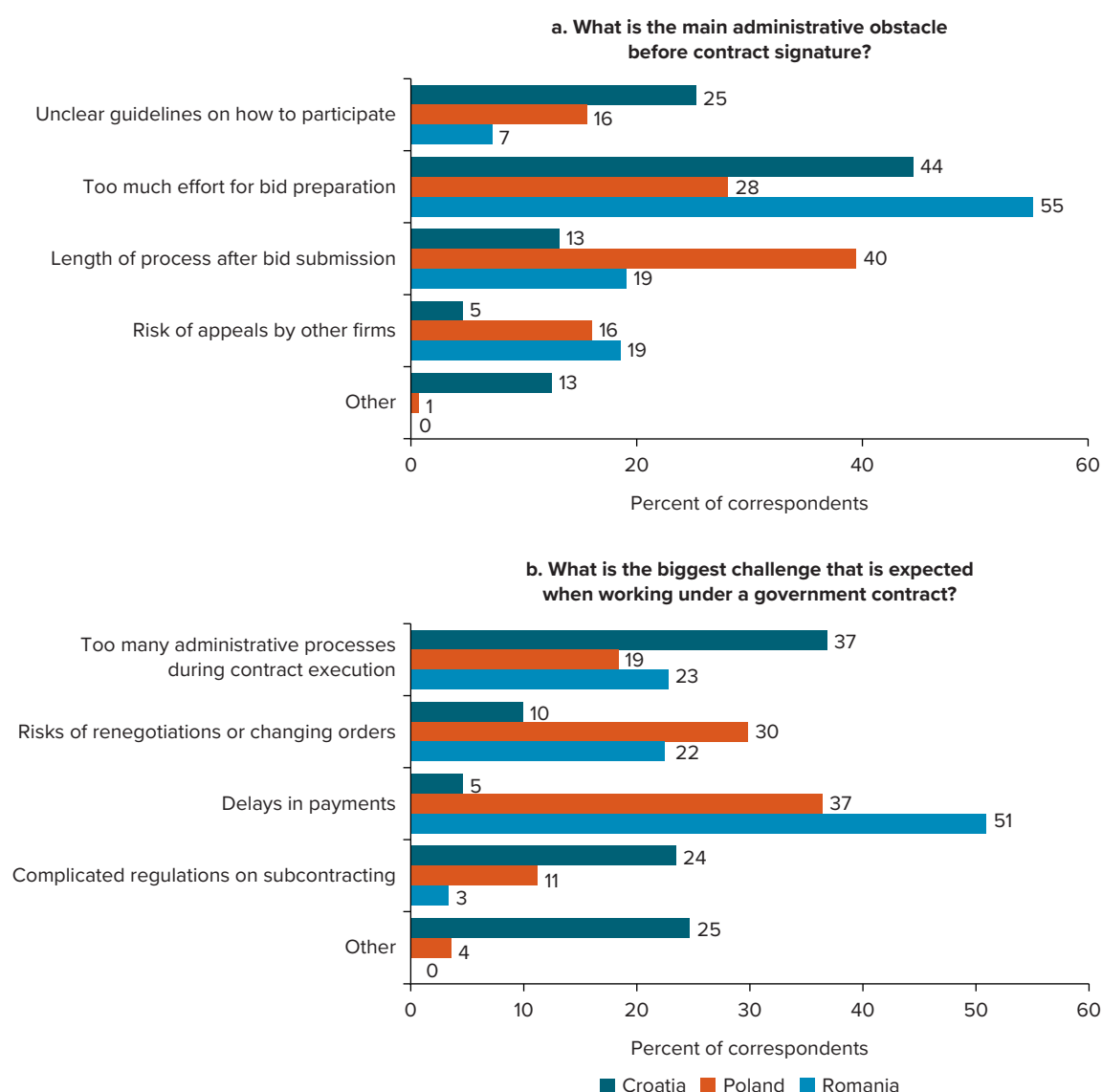
Surveys of procuring entity officers may also help measure the perceptions and behaviors of procuring entity officers with regard to overall organizational management, the administrative burden of conducting and reporting on procurement regulations, human resource management (HRM), roles, and incentives within their teams. For example, the 2021 World Bank report on the use of framework agreements relied on both administrative microdata and survey data.<sup>20</sup> Procuring entity officers in India and Ethiopia were surveyed about the perceived administrative burden of using framework agreements relative to other public bidding methods. While quantitative analysis revealed some savings in price through the use of framework agreements, the survey provided more context about the burden officers might feel when implementing different types of procurement methods. Similarly, several studies on GPP have been conducted through surveys to understand the incentives of procuring entity officers to adopt GPP criteria in the award process, as well as to map the difficulties and challenges entities face with GPP regulations at different stages of the procurement process.

In addition to surveying procuring entity officers, reaching out to firms participating in the public contract market can also provide complementary information for understanding public procurement from the perspective of private sector actors. For example, Uyarra et al. (2014) find that for firms in the United Kingdom, the main barriers to entry into the public procurement market are a lack of interaction with procuring organizations, the low competency of civil servants, and poor management systems, and Knack, Biletska, and Kacker (2019) find that firms are more likely to participate in public procurement in countries where public procurement systems are more transparent and complaint systems are more effective. Surveys of firms can also be a useful tool to analyze special groups of firms (for example, SMEs and WOE) in countries where public procurement data do not allow for the identification of bidder and supplier characteristics

and where procurement data cannot be linked to other administrative data, such as firm registries. In 2021, the World Bank designed a procurement module as part of the Enterprise Surveys to better understand the barriers and challenges experienced by firms with respect to public procurement, and they piloted this module in Romania, Poland, and Croatia. The survey data reveal that the main administrative obstacle to participation in Poland is the length of the process between bid submission and contract signature, while in Croatia and Romania, it is the fact that too much effort is required for bid preparation (figure 12.7a). The biggest challenge when working under a government contract is payment delays in Poland and Romania and the number of administrative processes during contract execution in Croatia (figure 12.7b).

Using survey data in public procurement is relevant from a policy perspective in order to complement administrative data measurements, but it is also relevant from a research point of view. HRM practices,

**FIGURE 12.7 Obstacles and Challenges to Government Contracts, Croatia, Poland, and Romania, 2021**



Source: Original figure for this publication based on microdata from the World Bank Enterprise Surveys Follow-Up on COVID-19 2021, Round 3, for Croatia, Poland, and Romania.

Note: Panel a: Weighted results. Only firms that indicated that administrative procedures before contract signature are an obstacle to attempting to secure a government contract. Panel b: Weighted results. Only firms that indicated that expected challenges during contract execution are an obstacle to attempting to secure a government contract.

attitudes, and motivations in public administration are typically measured through surveys of civil servants. Public procurement can be an ideal area to study the link between these dimensions and outcomes, advancing our understanding of the impact of HRM practices, attitudes, and motivations on performance and compliance.

## CONCLUSION

This chapter has provided an overview of how public procurement data can be used for monitoring and evaluating public procurement, as well as for informing reform efforts and defining new policies and strategies in public procurement. It has included a description of various data analytical tools that can be applied to public procurement, an account of typical challenges encountered in public procurement data and potential solutions, and a discussion of recent innovations, such as the development of interactive dashboards.

The chapter has included various lessons for practitioners and governments on using and analyzing public procurement administrative data, including centralizing public procurement data, integrating data from different procurement stages and from data systems related to other government functions, ensuring data quality and completeness, and building capacity for statistical analysis, such as by creating a dedicated statistical unit in the public procurement authority.

The chapter has also highlighted various areas where there is a need for further development and research, specifically in measuring the quality of contract implementation, integrating public procurement data with other administrative microdata or survey data, measuring GDP, and, more generally, generating robust empirical evidence on effective ways to improve the efficiency, integrity, inclusiveness, and sustainability of public procurement. For example, the World Bank's Governance Global Practice and the Development Impact Evaluation (DIME) Governance and Institution Building unit have been collaborating on a research agenda about the link between public procurement and private sector growth, which includes a series about research projects and data innovations, such as connecting public procurement data, payment data, and tax registry data.<sup>21</sup>

## NOTES

The chapter is based on academic research and operational experience from several World Bank projects that use data analytical tools in the area of public procurement—for example, in Romania (led by Carmen Calin, procurement specialist), Croatia (led by Antonia Viyachka, procurement specialist), and Bangladesh (led by Ishtiaq Siddique, senior procurement specialist). The chapter greatly benefited from comments and inputs by Carmen Calin (World Bank, procurement specialist), Maria Arnald Canudo (consultant, Development Impact Evaluation [DIME] Department), Daniel Rogger (senior economist, DIME), and Christian Schuster (professor, University College London). Stephen Shisoka Okiya (consultant, DIME) provided excellent research assistance.

1. A seminal paper by Bandiera, Prat, and Valletti (2009) demonstrates that in Italy, 83 percent of the total estimated waste in public procurement is due to passive waste caused by inefficiencies related to constraints such as lack of skills, lack of incentives, and excessive regulatory burden.
2. For example, with respect to the United Nations Sustainable Development Goals, public procurement can contribute to increasing access to markets for small and medium enterprises (target 9.3), responsible consumption and production through sustainable public procurement (target 12.7), reducing corruption and bribery (target 16.5), developing effective, accountable, and transparent institutions (target 16.6), and ensuring public access to information (target 16.10). More information about the Sustainable Development Goals is available on the United Nations Commission on International Trade Law website at <https://uncitral.un.org/en/about/sdg>.
3. Further details on the data in figure 12.1 and on the level of e-GP adoption across countries can be found in the World Bank's *Doing Business 2020* data under the topic "Contracting with the Government": <https://archive.doingbusiness.org/en/data/exploretopics/contracting-with-the-government>.

4. Requirements from international organizations or international treaties could be another strategy to incentivize governments to open public procurement data and adopt transparent monitoring and reporting mechanisms. For example, EU member states are mandated to monitor and report key procurement indicators under Directives 2014/23/EU, 2014/24/EU, and 2014/25/EU.
5. The literature has pointed to three different strategies for measurement validity (Adcock and Collier 2001): content validity (the measurement captures the full content of the definition), convergent validity (alternative measures of the same corrupt phenomenon are correlated), and construct validity (well-established empirical relationships are confirmed by the measurement).
6. As noted above, EU member states are mandated to monitor and report key procurement indicators under Directives 2014/23/EU, 2014/24/EU and 2014/25/EU.
7. The PPO website is available at <https://www.uzp.gov.pl/>.
8. The Public Procurement Agency in Bulgaria, the PPO in Poland, the Office for Public Procurement in the Slovak Republic, and the National Agency for Public Procurement (ANAP) in Romania are examples of institutions that conduct audits of compliance and performance monitoring.
9. Deliverable under the World Bank project Framework Agreements for Development Impact: Lessons from Selected Countries for Global Adoption (P173392).
10. Report under the project Impact Evaluation of e-Procurement In Bangladesh (P156394).
11. The role of civil society in monitoring public procurement is widely recognized. For example, within the EU project Integrity Pacts—Civil Control Mechanism for Safeguarding EU Funds, “integrity pacts” are established between a contracting authority and economic operators bidding for public contracts, stipulating that parties will abstain from corrupt practices and conduct a transparent procurement process, and a separate contract with a civil society organization entrusts it with the role of monitoring that all parties comply with their commitments. See the Transparency International website at <https://www.transparency.org/en/projects/integritypacts>.
12. More information about the Government Transparency Institute is available on its website, <http://www.govtransparency.eu/>. The dashboard prototype is available here: <https://www.procurementintegrity.org/>.
13. The Opentender dashboard is available here: <https://opentender.eu/start>.
14. For more information about the Open Contracting Data Standard, see the project website at <https://standard.open-contracting.org/latest/en/>.
15. The India dashboard is available here: <https://eprocure.gov.in/eprocdashboard/KPI.html>.
16. The Ukraine dashboard is available here: <https://bi.prozorro.org/hub/stream/aaec8d41-5201-43ab-809f-3063750dfafd>.
17. On Brazil, see CGU (2020). Moldova’s COVID-19 procurement website can be viewed here: <https://www.tender.health/>. Lithuania’s procurement webpage can be viewed on the Public Procurement Office website at <https://vpt.lrv.lt/kovai-su-covid-19-sudarytos-sutartys>. South Africa’s COVID-19 procurement dashboard can be viewed on the National Treasury website at <http://ocpo.treasury.gov.za/COVID19/Pages/Reporting-Dashboard-Covid.aspx>.
18. Green evaluation criteria can be included in different levels of procurement and in the bidding process by setting technical specifications, specific qualifications, contract requirements, selection criteria, and/or award criteria (Testa et al. 2012).
19. As an example of the former, Gerardino, Litschig, and Pomeranz (2017) investigate the impact of the audit selection process in Chile, using public procurement data from 2011 to 2012. Under the existing audit protocol in that period, open auctions underwent more than twice as many checks as direct contracting. The authors find that, given this protocol, procurement officers shifted toward direct contracting methods and reduced the use of open auctions, especially in procuring entities that experienced more audits and therefore had more opportunities to learn about this targeting design. As an example of the latter, in some countries, procuring entities are required to reserve a given quote of their spending for SMEs, but it is challenging for auditors to monitor compliance with this requirement if public procurement data do not include a tag to identify contracts awarded to SMEs.
20. Deliverable under the World Bank project Framework Agreements for Development Impact: Lessons from Selected Countries for Global Adoption (P173392).
21. See the World Bank project Public Procurement and Firm Behavior (P177551).

## REFERENCES

- Adam, Isabelle, and Mihály Fazekas. 2019. “Big Data Analytics as a Tool for Auditors to Identify and Prevent Fraud and Corruption in Public Procurement.” *European Court of Auditors Journal* 2: 172–80. <https://medium.com/ecajournal/big-data-analytics-as-a-tool-for-auditors-to-identify-and-prevent-fraud-and-corruption-in-public-68184529334c>.
- Adcock, Robert, and David Collier. 2001. “Measurement Validity: A Shared Standard for Qualitative and Quantitative Research.” *American Political Science Review* 95 (3): 529–46. <https://doi.org/10.1017/S0003055401003100>.



- Bandiera, Oriana, Andrea Prat, and Tommaso Valletti. 2009. "Active and Passive Waste in Government Spending: Evidence from a Policy Experiment." *American Economic Review* 99 (4): 1278–308. <https://doi.org/10.1257/aer.99.4.1278>.
- Best, Michael Carlos, Jonas Hjort, and David Szakonyi. 2019. "Individuals and Organizations as Sources of State Effectiveness." NBER Working Paper 23350, National Bureau of Economic Research, Cambridge, MA. <https://doi.org/10.3386/w23350>.
- Bosio, Erica, Simeon Djankov, Edward L. Glaeser, and Andrei Shleifer. 2022. "Public Procurement in Law and Practice." *American Economic Review* 112 (4): 1091–117. <https://doi.org/10.1257/aer.20200738>.
- CGU (Controladoria-Geral da União). 2020. "CGU lança painel para dar transparência a contratações relacionadas à Covid-19." Comptroller General of Brazil, March 7, 2020. <https://www.gov.br/cgu/pt-br/assuntos/noticias/2020/07/cgu-lanca-painel-para-dar-transparencia-a-contratacoes-relacionadas-a-covid-19>.
- Conley, Timothy G., and Francesco Decarolis. 2016. "Detecting Bidders Groups in Collusive Auctions." *American Economic Journal: Microeconomics* 8 (2): 1–38. <https://doi.org/10.1257/mic.20130254>.
- Coviello, Decio, Luigi Moretti, Giancarlo Spagnolo, and Paola Valbonesi. 2018. "Court Efficiency and Procurement Performance." *The Scandinavian Journal of Economics* 120 (3): 826–58. <https://doi.org/10.1111/sjoe.12225>.
- Decarolis, Francesco, Raymond Fisman, Paolo Pinotti, and Silvia Vannutelli. 2020. "Rules, Discretion, and Corruption in Procurement: Evidence from Italian Government Contracting." NBER Working Paper 28209, National Bureau of Economic Research, Cambridge, MA. <https://doi.org/10.3386/w28209>.
- Decarolis, Francesco, and Cristina Giorgiantonio. 2022. "Corruption Red Flags in Public Procurement: New Evidence from Italian Calls for Tenders." *EPJ Data Science* 11: 16. <https://doi.org/10.1140/epjds/s13688-022-00325-x>.
- European Commission. 2008. *Public Procurement for a Better Environment*. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee, and the Committee of the Regions, COM(2008) 400. <https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2008:0400:FIN:EN:PDF>.
- Fazekas, Mihály, Luciana Cingolani, and Bence Tóth. 2018. "Innovations in Objectively Measuring Corruption in Public Procurement." Chap. 7 in *Governance Indicators: Approaches, Progress, Promise*, edited by Helmut K. Anheier, Matthias Haber, and Mark A. Kayser. Oxford: Oxford University Press. <https://doi.org/10.1093/oso/9780198817062.003.0007>.
- Fazekas, Mihály, and Gábor Kocsis. 2020. "Uncovering High-Level Corruption: Cross-National Corruption Proxies Using Public Procurement Data." *British Journal of Political Science* 50 (1): 155–64. <https://doi.org/10.1017/S0007123417000461>.
- Fazekas, Mihály, Salvatore Sberna, and Alberto Vannucci. 2021. "The Extra-Legal Governance of Corruption: Tracing the Organization of Corruption in Public Procurement." *Governance: An International Journal of Policy, Administration, and Institutions* 35 (4): 1139–61. <https://doi.org/10.1111/gove.12648>.
- Fazekas, Mihály, and Bence Tóth. 2018. "The Extent and Cost of Corruption in Transport Infrastructure: New Evidence from Europe." *Transportation Research Part A: Policy and Practice* 113: 35–54. <https://doi.org/10.1016/j.tra.2018.03.021>.
- Fazekas, Mihály, Gavin Ugale, and Angelina Zhao. 2019. *Analytics for Integrity. Data-Driven Approaches for Enhancing Corruption and Fraud Risk Assessments*. Paris: OECD Publishing. <https://www.oecd.org/gov/ethics/analytics-for-integrity.pdf>.
- Ferraz, Claudio, Frederico Finan, and Dimitri Sberman. 2015. "Procuring Firm Growth: The Effects of Government Purchases on Firm Dynamics." NBER Working Paper 21219, National Bureau of Economic Research, Cambridge, MA. <https://doi.org/10.3386/w21219>.
- Gerardino, Maria Paula, Stephan Litschig, and Dina Pomeranz. 2017. "Distortion by Audit: Evidence from Public Procurement." NBER Working Paper 23978, National Bureau of Economic Research, Cambridge, MA. Revised August 2022. <https://doi.org/10.3386/w23978>.
- Hassan, Mirza. 2017. "Citizen Engagement during Public Procurement Implementation in Bangladesh." South Asia Procurement Innovation Awards 2016, World Bank, Washington, DC. <https://wbnpf.procurementinet.org/featured/citizen-engagement-during-public-procurement-implementation-bangladesh>.
- Huber, Martin, and David Imhof. 2019. "Machine Learning with Screens for Detecting Bid-Rigging Cartels." *International Journal of Industrial Organization* 65: 277–301. <https://doi.org/10.1016/j.ijindorg.2019.04.002>.
- Knack, Stephen, Nataliya Biletska, and Kanishka Kacker. 2019. "Deterring Kickbacks and Encouraging Entry in Public Procurement Markets: Evidence from Firm Surveys in 90 Developing Countries." *World Bank Economic Review* 33 (2): 287–309. <http://hdl.handle.net/10986/34863>.
- Krasnokutskaya, Elena, and Katja Seim. 2011. "Bid Preference Programs and Participation in Highway Procurement Auctions." *American Economic Review* 101 (6): 2653–86. <https://doi.org/10.1257/aer.101.6.2653>.
- Marion, Justin. 2007. "Are Bid Preferences Benign? The Effect of Small Business Subsidies in Highway Procurement Auctions." *Journal of Public Economics* 91 (7–8): 1591–624. <https://doi.org/10.1016/j.jpubeco.2006.12.005>.
- Medvedev, Denis, Ramin N. Aliyev, Miriam Bruhn, Paulo Guilherme Correa, Rodrigo Javier Garcia Ayala, Justin Piers William Hill, Subika Farazi, Jose Ernesto Lopez Cordova, Caio Piza, Alena Sakhonchik, and Morten Seja. 2021. *Strengthening World Bank SME-Support Interventions: Operational Guidance Document*. World Bank Report. Washington, DC: World Bank. <http://documents.worldbank.org/curated/en/183521617692963003/Strengthening-World-Bank-SME-Support-Interventions-Operational-Guidance-Document>.

- Mungiu-Pippidi, Alina, and Mihály Fazekas. 2020. "How to Define and Measure Corruption." In *A Research Agenda for Studies of Corruption*, edited by Alina Mungiu-Pippidi and Paul M. Heywood, 7–26. Cheltenham, UK: Edward Elgar. <https://doi.org/10.4337/9781789905007.00008>.
- Nissinen, Ari, Katriina Parikka-Alhola, and Hannu Rita. 2009. "Environmental Criteria in the Public Purchases above the EU Threshold Values by Three Nordic Countries: 2003 and 2005." *Ecological Economics* 68 (6): 1838–49. <https://doi.org/10.1016/j.ecolecon.2008.12.005>.
- OECD (Organisation for Economic Co-operation and Development). 2013. "Ex Officio Cartel Investigations and the Use of Screens to Detect Cartels." Competition Policy Roundtables DAF/COMP(2013)27, Competition Committee, Directorate for Financial and Enterprise Affairs, OECD, Paris. <https://www.oecd.org/daf/competition/exofficio-cartel-investigation-2013.pdf>.
- OECD (Organisation for Economic Co-operation and Development). 2017. *Government at a Glance 2017*. Paris: OECD Publishing. [https://doi.org/10.1787/gov\\_glance-2017-en](https://doi.org/10.1787/gov_glance-2017-en).
- OECD (Organisation for Economic Co-operation and Development). 2018. *SMEs in Public Procurement: Practices and Strategies for Shared Benefits*. OECD Public Governance Reviews. Paris: OECD Publishing. <https://doi.org/10.1787/9789264307476-en>.
- OECD (Organisation for Economic Co-operation and Development). 2021. *Government at a Glance 2021*. Paris: OECD Publishing. <https://doi.org/10.1787/1c258f55-en>.
- OECD and SIGMA (Support for Improvement in Governance and Management). 2019. *Methodological Framework of the Principles of Public Administration*. Paris: OECD Publishing. <https://www.sigmaweb.org/publications/Methodological-Framework-for-the-Principles-of-Public-Administration-May-2019.pdf>.
- Olken, Benjamin A. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 115 (2): 200–49. <https://doi.org/10.1086/517935>.
- Palmujoki, Antti, Katriina Parikka-Alhola, and Ari Ekroos. 2010. "Green Public Procurement: Analysis on the Use of Environmental Criteria in Contracts." *Review of European Community & International Environmental Law* 19 (2): 250–62. <https://doi.org/10.1111/j.1467-9388.2010.00681.x>.
- Singer, Marcos, Garo Konstantinidis, Eduardo Roubik, and Eduardo Beffermann. 2009. "Does e-Procurement Save the State Money?" *Journal of Public Procurement* 9 (1): 58–78. <https://doi.org/10.1108/JOPP-09-01-2009-B002>.
- Testa, Francesco, Fabio Iraldo, Marco Frey, and Tiberio Daddi. 2012. "What Factors Influence the Uptake of GPP (Green Public Procurement) Practices? New Evidence from an Italian Survey." *Ecological Economics* 82: 88–96. <https://doi.org/10.1016/j.ecolecon.2012.07.011>.
- Titl, Vitezslav, and Benny Geys. 2019. "Political Donations and the Allocation of Public Procurement Contracts." *European Economic Review* 111: 443–58. <https://doi.org/10.1016/j.eurocorev.2018.11.004>.
- Turkewitz, Joel, Mihály Fazekas, and Zafrul Islam. 2020. "Case Study 2: e-Procurement Reform in Bangladesh." In *Enhancing Government Effectiveness and Transparency: The Fight against Corruption*, edited by Rajni Bajpai and C. Bernard Myers, 34–39. World Bank Global Report. Washington, DC: World Bank. <http://documents.worldbank.org/curated/en/235541600116631094/Enhancing-Government-Effectiveness-and-Transparency-The-Fight-Against-Corruption>.
- Uyarra, Elvira, Jakob Edler, Javier Garcia-Estevéz, Luke Georgiou, and Jillian Yeow. 2014. "Barriers to Innovation through Public Procurement: A Supplier Perspective." *Technovation* 34 (10): 631–45. <https://doi.org/10.1016/j.technovation.2014.04.003>.
- Wachs, Johannes, Mihály Fazekas, and János Kertész. 2021. "Corruption Risk in Contracting Markets: A Network Science Perspective." *International Journal of Data Science and Analytics* 12: 45–60. <https://doi.org/10.1007/s41060-019-00204-1>.
- World Bank. 2017. *Pakistan—Punjab Land Records Management and Information Systems Project*. ICR00003719. Washington, DC: World Bank. <http://documents.worldbank.org/curated/en/632241498842804246/Pakistan-Land-Records-Management-and-Information-Systems-Project>.
- World Bank. 2019. *Romania—Reimbursable Advisory Services Agreement on Assessment of the Public Procurement System and Further Support to the Implementation of the Public Procurement Strategy: Output 4: Final Version of the Web-Based Guide*. P169141. Washington, DC: World Bank. <https://pubdocs.worldbank.org/en/412981574427978384/RO-TOR-Procurement-SME-2019.pdf>.
- World Bank. 2020. *Assessment of Bangladesh Public Procurement System*. Washington, DC: World Bank. <https://openknowledge.worldbank.org/handle/10986/33882>.
- World Bank. 2021a. *Econometric Analysis of Framework Agreements in Brazil and Colombia*. Washington, DC: World Bank. <https://doi.org/10.1596/36059>.
- World Bank. 2021b. "Green Public Procurement: An Overview of Green Reforms in Country Procurement Systems." Climate Governance Papers, World Bank, Washington, DC. <http://hdl.handle.net/10986/36508>.



## CHAPTER 13

# Government Analytics Using Data on the Quality of Administrative Processes

*Jane Adjabeng, Eugenia Adomako-Gyasi, Moses Akrofi, Maxwell Ampofo, Margherita Fornasari, Ignatius Geegbae, Allan Kasapa, Jennifer Ljungqvist, Wilson Metronao Amevor, Felix Nyarko Ampong, Josiah Okyere Gyimah, Daniel Rogger, Nicholas Sampah, and Martin Williams*

### SUMMARY

This chapter seeks to highlight the value of quantifiable measures of the quality of back-office processes when assessing governments' bureaucratic effectiveness. Conceptually, it defines a framework for understanding administrative *process productivity*. It then presents case studies from the Ghanaian and Liberian civil services, where different measures of internal (within bureaucratic units) versus external (across bureaucratic units) process quality were piloted. Specifically, these pilots sought to assess the feasibility, cost, and scalability of the process measures considered. We explore their correlations with other measures of productivity (for example, financial expenditures and the completion of planned tasks) and the claims and characteristics of civil servants in the surveys we have undertaken.

### ANALYTICS IN PRACTICE

- Many of the activities undertaken by public administrators can be characterized as the application of proper processes and procedures to a project, file, or case. As such, the quality of the processing work undertaken by public officials is an important part of the quality of government activity and a determinant of public sector productivity.

---

Jane Adjabeng, Eugenia Adomako-Gyasi, Moses Akrofi, Maxwell Ampofo, Allan Kasapa, Wilson Metronao Amevor, Felix Nyarko Ampong, Josiah Okyere Gyimah, and Nicholas Sampah are with the Office of the Head of the Civil Service, Ghana. Margherita Fornasari, Jennifer Ljungqvist, and Daniel Rogger (corresponding author) are with the World Bank's Development Impact Evaluation (DIME) Department. Ignatius Geegbae is with the Civil Service Agency, Liberia. Martin Williams is an associate professor in public management at University of Oxford.

- Government analytics systems should include measures of the quality of processing by government officials. This requires a framework for assessing the adherence of administrators' process work to accepted government procedure. An important distinction in the development of a conceptual framework for assessing process quality is whether a process is mainly confined within organizational units (what this chapter calls *internal process productivity*) or is associated with interactions between organizational units (*external process productivity*). Separately, such a framework could be formulated to assess the quality of processes in general within a public service or targeted at domain-specific activities, such as the implementation of an appraisal process.
- By not measuring the quality of process productivity, analytics systems bias measures of the quality of government toward more measurable aspects of government work, such as the activities of frontline officials, and away from the important body of administrative professionals who support them. By taking a narrow approach to defining bureaucratic productivity according to frontline outputs, studies risk missing why a service may be delivered well or poorly.
- Such data can be collected automatically, as part of digitized government work, or manually, by assessors employed to judge process quality in the physical records of projects, files, or cases. Analysts can benchmark the quality of processes evidenced by relevant records in terms of the extent to which they are well organized, logical in flow, and adherent to government procedure; their timeliness with respect to a deadline; or whatever aspect of a process is of interest. To some degree, the digitization of government has supported the improvement of processes—by ensuring that all components of a process are present—but it also facilitates the physical or automated inspection of process quality.
- Measures of process quality in public administration open up three areas of government analytics: assessments of variation in the quality of processes and associated productivity within and across organizations in the same country, comparisons of process quality across countries, and assessments of public sector process quality over time. In this way, government analysts can pinpoint where government procedure is not being adhered to, how different processes relate to public sector productivity, and what dynamics exist across individuals and organizational units.

## INTRODUCTION

The effective delivery of government work requires a long chain of processes undertaken by public officials. From policy development, through budget and planning, to monitoring and evaluation, many of the activities undertaken by public administrators can be characterized as the application of proper processes and procedures to a project, file, or case. This is the back-office administration of public policy. In the case of policy development, for example, proper process in most modern governments would include presenting a balance of evidence on the pros and cons of policy content, ensuring broad-based consultation within (and potentially outside) government, and generating a coherent policy that others in government could follow.

The quality of the processing work undertaken by public officials is thus an important part of the quality of government activity. Given how important these processes are in the chain of delivery for government goods and services—such as the budget and procurement processes—process quality is a key component of public sector productivity.

This chapter articulates a framework for measuring the quality of these processes based on the idea of adherence to accepted government procedure. The rationale of adherence to government procedure may be varied: it may include equity considerations (ensuring all cases are dealt with in a similar way), fiduciary concerns (ensuring resources are utilized for the public good), and legal issues (ensuring that actions are in line with existing laws and the rules of the public service). The quality of government processes can be

measured along these and other margins as a measure of the nature of government functioning. This chapter outlines specific measures of the quality of government processes and discusses their use through two case studies in Ghana and Liberia.

To date, measures of bureaucratic activity and effectiveness have focused on frontline outputs, as described in chapter 29 of the *Handbook* on service delivery indicators (SDI), financial expenditures or the procurement of goods, as described in the *Handbook* chapters on budget and procurement (chapter 12), and the provision of physical infrastructure, as described in the *Handbook* chapter on task completion (chapter 17). This chapter focuses not on the prices, wages, or services achieved by the government but on the quality of the processes applied toward those ends. For example, a procurement officer may gain low prices for a set of goods procured but do so in a way that breaks public service rules and potentially exposes the public sector to unnecessary reputational risks. Similarly, an officer may complete all assigned tasks within deadline and budget but do so in a way that has negative spillovers on other units in the agency.

This approach to understanding the quality of work in the public sector has parallels to aspects of the SDI outlined in chapter 29. For example, in assessing the quality of education, analysts have assessed the extent to which teachers are subject to classroom observation by an independent assessor and are provided feedback on their teaching. This measure does not score the quality of the teaching itself, or even the quality of the feedback. Rather, it assesses the extent to which a process is in place to provide feedback, assuming that feedback is an important part of a quality teaching environment.<sup>1</sup>

Perhaps the closest approach to an assessment of bureaucratic functioning that attempts to measure the quality of government processes is the approach associated with case analysis. As described in chapter 15 on administrative case data, such analysis assesses the quality of responses by public officials to requests for public services, such as in the social security sector, or the fulfillment of public responsibilities, such as the collection of taxes. However, the data collected are almost universally on outcomes of these activities, such as the volume of cases processed in a particular time frame or, conversely, processing speed, prices paid, and so on. A complement to this analysis characterizes the quality of public sector actors' work processes, from the comprehensiveness of records to the quality of the evidence they provide to back up their assertions. So while measures like those in chapter 15 typically characterize the speed of case completion as a positive outcome, a *process productivity* perspective would assess whether sufficient time had been allotted for consultation (such as for advertising a procurement). Little quantitative work has been undertaken on this margin of government activity.<sup>2</sup>

This paucity of preexisting work stands in contrast to the fact that a substantial portion of the work of public administration is best characterized as processing. Rasul, Rogger, and Williams (2021) find that 73 percent of civil service activities in Ghana can be categorized as "processing tasks." The common conception of government work is frequently back-office process work.

The absence of effective measures of the quality of government processes has skewed the focus of public sector studies toward frontline officials and away from the important body of administrative professionals that support them.

Most civil servants play an important role in facilitating the role of frontline staff, by providing the long chain of supporting activities that are at the core of the effectiveness of government. Processing work is a substantial component of this support.

By taking the narrow approach of defining bureaucratic productivity according to frontline outputs, studies also risk missing why a service may be delivered well or poorly. For example, for a citizen to receive a welfare payment, budgetary officers must ensure sufficient funds are available, contracting officers must ensure effective transfer systems to recipients, and accounting officers must ensure a clear paper trail to reduce the diversion of funds. Wrapping the entirety of these activities into a single indicator of payment disbursement does not allow us to uncover which process creates a bottleneck.

Consequently, this chapter seeks to highlight the value of quantifiable measures of this type of back-office, administrative process productivity when assessing governments' bureaucratic effectiveness. It does so by presenting case studies from the Ghanaian and Liberian civil services, where different measures of *internal* (within bureaucratic units) versus *external* (across bureaucratic units) process quality were piloted.



And it considers both the quality of standard work processes (Ghana) and the implementation of a new set of processes related to staff appraisals (Liberia). These pilots introduce concrete ideas for measuring process quality and showcase their feasibility and scalability to entire public services.

Measures of process quality in public administration open up three areas of government analytics. First, we can use such measures to assess variations in process quality and associated productivity within and across organizations in the same country. For example, by using a common assessment of process quality across organizations, we can identify which organizations are appropriately adhering to government procedure across a government. Second, with appropriate caveats, common measures enable comparisons of process productivity across countries. For example, understanding the time it takes for a social sector ministry to provide inputs to the center of government across countries provides microevidence of the relative quality of governance. Finally, given the relative simplicity of these measures, we could collect productivity data on a regular basis and thus provide a more nuanced assessment of public sector capacity over time.

This chapter continues as follows. It begins with an overview of related measures and then presents concrete applications of these ideas in case studies from Ghana and Liberia. It then showcases the results of measurement in these two settings and discusses what we learn about the nature of process quality in the public service.

## CONCEPTUAL FRAMEWORK AND RELATED LITERATURE

Conceptually, the notion that government processes should adhere to particular standards is widespread. Most governments have rules for undertaking (or *processing*) the tasks of public administration that articulate best practices. These best practices almost universally align with themes of completeness, rationality, fairness, and efficiency—all themes extolled by the Weberian school of public administration.

Wilson (1989, 26) argues that understanding public sector productivity, in contrast to that of the private sector, means understanding the processing of tasks. Since the goals of the public sector are too vague to be a useful organizing framework, the public sector must focus on specializing in improved task or process productivity. This reasoning has since been bolstered by a range of authors (for example, Alesina and Tabellini 2007; Dixit 2002).<sup>3</sup>

However, few analysts argue for a coherent notion of government processes as a component of government functioning or of public sector productivity. Yet if government processes mediate the use of inputs to the production function of government, then undertaking them to a high standard would seem to be an output of government work related to functioning and productivity. In relation to ideals of the state, such as the equitable treatment of cases, process may be an end in itself, observed in the capacity of public officials to make coherent decisions.

When public officials make coherent arguments for choosing one policy over another that incorporate relevant information, they improve the quality of government outputs but also characterize government itself. Both of these are public goods of their own type. Similarly, when a manager judges one official eligible for promotion over another using solid evidence of the performance of both officials, the public sector becomes more effective and is characterized as meritocratic. Again, both of these are public goods in distinct ways.

For this reason, whether public officials process government work in the appropriate ways can be studied as a form of public sector productivity: *process productivity*. When the government effectively and efficiently undertakes work according to proper processes, it generates better outputs for the next stage of government work and defines a superior character of government. It is therefore more productive in producing these public goods. Take, for example, a firm that creates parts to sell to other firms that build machines out of those parts: when it does this with a high level of quality and in a reliable way, the parts firm is productive. Likewise, a government organization that undertakes its tasks using proper processes is a more productive institution.

What proper process means will vary by setting and the tasks focused on. However, best practices in government processes often include the clear and complete gathering of evidence and rational decision-making, as well as equity considerations (ensuring all cases are dealt with in a similar way), fiduciary concerns (ensuring resources are utilized for the public good), and legal issues (ensuring that actions are in line with existing laws and the rules of the public service). An example of an approach to assessing the quality of decision-making is the SMART framework, which considers whether relevant elements are specific, measurable, achievable, relevant, and time-bound. This framework will be applied in one of our case studies.

This chapter assesses how an analyst might measure process quality, and thus process productivity, on a large scale (across a substantial portion of units of public administration) using a quantitative approach. Efforts to date to characterize government and its processes have been broad-brush, such as expert assessments of corruption that outline the propensity to circumvent proper processes for personal gain across an administration as a whole.<sup>4</sup> Our focus is on measurement at a granular level, frequently the task, project, or individual level.

This more granular level is the area of measurement for which there is little to no previous work and, as argued in much of the rest of the *Handbook*, where there is the most potential for gains from analysis. For a similar reason, we look for processes that are generally applied across government, rather than a domain-specific set of processes, such as how doctors should treat patients (Bedoya et al. 2017; Daniels et al. 2017; Wafula et al. 2017). However, to provide clarity in the application of our framework to domain-specific settings, our second case study looks at the application of process productivity assessments to a domain-specific activity undertaken by all public servants—performance appraisal.

Empirical assessments of government processes in political science have studied the nature of responses to public information requests (also known as freedom of information requests). By assessing the qualities of government responses, researchers have assessed whether citizens receive a response quickly (Wehner and Poole 2015; Wood and Lewis 2017) and equitably (Berliner et al. 2021; Peisakhin and Pinto 2010). This approach is clearly highly constrained in what it can measure as an external lens with ambiguous links to government functioning.

In the economics literature, Chong et al. (2014) assess the quality of government processes through how quickly misaddressed letters are returned to their original senders. This measure is unrelated to most aspects of government work but can be seen as similar in spirit to the measure we will introduce in this chapter to assess government productivity through how responsive units are to centralized requests for information.

The closest paper to measuring internal process quality in a large-scale, quantitative way is Banerjee et al. (2021), who use retired senior police officers to grade a random set of case files from project police stations. They grade officers on whether scientific techniques were used, the care with which evidence was collected, and so on. Though their focus is explicitly on the clarity of police processes, the approach we elaborate in this chapter is a generalization of their approach.

We distinguish *internal process productivity*, the quality of administrative processes for activities confined within a particular administrative unit, from *external process productivity*, the quality of administrative processes for activities in which units interact. An example of the first is the development of the design of a project in which a unit specializes, while an example of the second is a request for information from one unit by another.

We make this distinction because accountability and professional dynamics vary distinctly between the two cases. Public administration is typically conceptualized around work units organized within a hierarchy. These work units have a degree of flexibility in how they organize their approaches to undertaking government work and implementing process guidelines. However, the head of a unit is responsible for ensuring process quality, as only the head administrator of an organization ensures the organization as a whole adheres to processes. An analogous assessment can be made between organizations and the government as a whole.

Similarly, when communicating within organizational units, different record-keeping formats are required than when communicating between organizations. For this reason, the nature of measurement must vary when analysts are assessing internal versus external conceptions of process productivity.

## EMPIRICAL CASE STUDIES

We study the quality of bureaucratic processes in the public administrations of two West African countries: Ghana and Liberia. These are excellent environments for testing new measures of public service productivity. They are governed by clearly defined and well-structured rules for undertaking government processes. However, similar to many developing countries, the productivity of departments and organizations in these settings varies substantially (Rasul and Rogger 2018; Rasul, Rogger, and Williams 2021). There is mounting evidence that this variation in productivity is also prevalent in the public sectors of wealthier nations (Best, Hjort, and Szakonyi 2017; Fenizia 2022), but the variation we analyze likely subsumes this heterogeneity and is representative of a large portion of the world's public administrations.

In Ghana, we study a representative set of administrative tasks under-taken by the core public administration. In Liberia we focus on process quality in relation to a specific administrative activity: the implementation of a staff appraisal system. We split our efforts into understanding the quality of internal processes, by assessing whether the processing of these tasks adheres to government procedures, and external processes, by assessing how promptly units respond to requests from central agencies. Our discussion of the two case studies thus covers the main features of process productivity outlined in the previous section.

### Institutional Background

Ghana is a lower-middle-income country home to 28 million people, with a central government bureaucracy that is structured along lines reflecting both its British colonial origins and more presidentialist postindependence reforms. Ghana is one of Africa's most democratic countries.

Liberia is a low-income country of nearly 5 million people, with an agency-based administration similar in design to that of the United States. Years of civil war exacerbated recruitment and rewards based on patronage in the service. The resulting bloated workforce, a lack of established processes and procedures—or the presence of overly bureaucratic processes and procedures—and inadequate office resources have delayed and derailed the processing time for needed administrative procedures in the service. Furthermore, while Liberia is Africa's oldest and first modern republic, with a political system heavily influenced by the US Constitution, it has historically been largely characterized by minoritarianism. Democratic and recognized fair elections only commenced in the 21st century. Ghana and Liberia thus represent polities at two ends of Sub-Saharan Africa's distribution of state fragility.

Ghana's civil service consists of 57 central government ministries and departments that primarily perform the core bureaucratic functions of policy making, administration, and service delivery oversight. Ministries and departments are overseen by the Office of the Head of Civil Service (OHCS), which is responsible for personnel management and performance within the civil service. The OHCS coordinates and decides on all hiring, promotion, transfer, and (in rare circumstances) firing of bureaucrats across the service. Similarly, Liberia's Civil Service Agency (CSA) oversees the strategic leadership and management of the country's civil service, formulating and providing guidance on recruitment, personnel management, standards, and performance in civil service institutions. The Liberian service is made up of 31 ministries and agencies, in addition to the country's numerous public autonomous organizations. The architecture of the administration in the two countries has many commonalities.

### Processes under Study

The civil servants we study carry out public administration activities following administrative procedures, which set out guidelines and standards for how to formally proceed with government business. These apply equally across the service and broadly aim to ensure transparency, equity, and efficiency in government business. In both Ghana and Liberia, we seek to assess the efficiency with which civil servants undertake administrative processes. However, the specific processes under study differ.

In Ghana, we focus on an assessment of process quality in core office duties, such as project planning, budgeting, and monitoring. Rasul, Rogger, and Williams (2021) describe the most common types of tasks in Ghana's central government offices. These relate to processing paperwork related to the construction of public infrastructure, such as roads, boreholes, and schools (24 percent of tasks); administrative advocacy (16 percent); and monitoring, review, and auditing (14 percent). The OHCS outlines rules and associated guidelines for Ghanaian civil servants about how to prepare infrastructure or advocacy projects and monitor, review, and audit them according to proper procedures. For this reason, the Ghanaian civil service is characterized by a common set of standards and centrally managed procedures that officials are required to follow when handling administrative files (PRAAD 2007).<sup>5</sup>

In Liberia, we focus on adherence to new processes for performance assessment or “appraisal.” Following the end of Liberia's civil war in 2003, the CSA focused on establishing a more meritocratic civil service by, among other policies, developing a performance management system (PMS) policy (CSA and USAID-GEMS 2016; Forte 2010; Friedman 2012; World Bank 2014). Job descriptions were only recently formulated and formalized across all positions in Liberia's civil service, so an appraisal scheme helps embed them as part of the daily work of public servants.

The PMS is similar in structure to most other performance management schemes in public sectors around the world: in collaboration with their manager, employees commit to a set of performance targets at the start of each annual cycle, which are reviewed and assessed over the cycle, typically twice a year. Managers meet with each of their officers at the start of the cycle to agree on their individual performance targets and how they will be assessed, and they record this information in what we call Form 1. They are then supposed to meet again at midyear, to track progress in achieving individual targets, and at the end of the year, to jointly fill in and discuss a performance diagnostic: Form 2, an updated version of Form 1.<sup>6</sup> Processes are governed by detailed guidelines published by the CSA, which also provides training to managers in how to undertake the process correctly. We focus on the proficiency of individuals and their managers in executing the PMS process.

The PMS has given civil servants who use it better insight into their roles and responsibilities and how these feed into their institutions' overall delivery of public services, but measuring, managing, and rewarding performance remains a challenge. Table 13.1 lists some barriers to ministries' and agencies' effective use of the PMS, as observed by CSA officials.<sup>7</sup>

In addition to the quality of a procedural process as implemented within a unit, the extent to which governments can efficiently manage the communication and coordination of processes across work units is another important measure of the quality of government processes. To assess what we have named external process productivity, we examine the extent to which civil service departments respond to external inquiries. The internal management of the many tasks that civil servants carry out depends on external inputs and consequently requires a chain of activities that span organizational units. We therefore implement a common measurement

**TABLE 13.1 Reasons for Incomplete Adoption or Nonadoption of the PMS Process, Liberia**

Reasons for not adopting the PMS	Reasons for only partly adopting the PMS	Reasons for not filling in the PMS forms properly
HR officers see the PMS as an added burden on their work.	HR officers did not communicate the timeline to staff.	The forms are bulky. The process is paper-based.
The PMS will be used to fire or remove people from their jobs.	Too much of a paper trail.	Some just fill in forms after being coerced and threatened with disciplinary action.
Some institutions struggle to see the benefits of the PMS to them.	Some do not understand what is fully required of them throughout the PMS cycle.	They have not understood the process.
Leadership lack the willpower to adopt the PMS.	Supervisors with more than 10 staff members find the PMS time-consuming.	
The PMS is a CSA-imposed idea.	Staff expect the CSA to provide guidance at every phase of the PMS.	

Source: Original table for this publication.

Note: CSA = Civil Service Agency; HR = human resources; PMS = performance management system.

framework in both Ghana and Liberia that assesses public officials' responsiveness to requests from the central personnel authorities. We track a set of standardized requests relating to annual personnel record updates undertaken by the two institutions of centralized personnel management, the OHCS in Ghana and the CSA in Liberia. Letters requesting information on all officials in an organization were sent to the census of civil service organizations. For example, the central office might request annual updates to the profile of an organization's staff concerning qualifications and training. The aim of such efforts is for the OHCS or the CSA to plan its capacity-building efforts for the next year based on up-to-date information on current capabilities within the public administration. In Ghana, units were asked to provide staff members' names and civil service IDs as well as any training they had received in the past year. In Liberia, units were asked to provide an updated list of the civil service staff currently employed in their team, including staff members' names, payroll IDs, the names of their direct supervisors, and any relevant training undertaken in the past year.

### Assessing the Quality of Processes

In both countries, the processes we study are applicable across all organizations and sectors (though the internal measure in Ghana is general and in Liberia is specific to the appraisal process). This allows us to undertake a common analysis of procedure quality within each public service.

Our approach requires a record of public officials' activities that can be assessed by an independent evaluator. The records in both Ghana and Liberia are dominantly paper-based files that record the "treatment" of projects, files, or cases. The vast majority of such physical files are on-site in a government office. Thus, in the case studies we focus on, we were required to build a team of evaluators that could make physical visits to units to review the government files.<sup>8</sup> To some degree, the digitization of government has supported the improvement of process quality by ensuring that all components of a process required by a procedure are present before the case can be completed. It has also facilitated the sort of inspection and assessment outlined here because enumerators can assess process quality remotely by accessing electronic records, which was not possible in our settings. Besides the ability to access records remotely, however, much of the wider approach described here would be the same in the case of digitized records.

### Internal Process Productivity

The evaluations of internal process productivity we undertook in both countries focused on the completeness of records, their degree of organization, and evidence of transparent, logical, and equitable decision-making. First, we focused on measuring the level to which the principal components of administrative documents adhere to the general filing rules. Second, we examined whether the argument laid out in those documents was complete and consistent. Such an approach accords with the overarching concern of the public service rules in the countries of focus that decisions or activities be clearly documented and indicate a logical and equitable decision-making process. The public service rules of each country set the baseline for measures of how the files should have been completed. The OHCS guided the process of designing an instrument to assess process quality in Ghana, and the *Performance Management Policy Manual* (CSA and USAID-GEMS 2016), along with guidance from the CSA, informed the corresponding instrument in Liberia.

*Completeness* is the level to which the principal components of a file adhere to the general filing rules. In Ghana, the assessment tool collected information on whether the file ladder, folios, memos, minutes, letters, and related documents are compiled correctly, following the public service rules. The file ladder is an important element of a file, summarizing file circulation within an organization and expressing how valuable a file is. According to the general procedure, the file ladder should document all file circulation, specifying the date and the documents involved. To guarantee the accessibility of a file, all documents should be numbered consecutively, starting with folios from the opening of the file to the most recent ones. If actions are required, documents and letters should be minuted, dated, and signed, clearly stating from whom the letters are coming and to whom they are directed. The same procedure is applicable to memos and other relevant records in the file. In addition to dates and signatures, incoming and outgoing correspondence requires



specific stamps: the organizational (incoming) and dispatch (outgoing) stamp. Once a file has been passed on to other record officers or stored in the records office, it should not contain either duplicated and draft documents or misfiled and miscellaneous items. Thus, *completeness* is a catch-all for the general handling of government files, assessing the completeness of the file ladder; the consecutive organization of folios within a file; the availability of minutes, memos, and other relevant documents; and the proportion of incoming and outgoing correspondence with dates, stamps, and signatures.

For the appraisal process in Liberia, we similarly searched for complete sets of PMS documents, with all three forms expected in the annual cycle, that echoed the above considerations regarding completeness. Specifically, we looked at how much information had been entered on the PMS forms and whether the civil servants' listed work objectives were linked to their performance indicators, their performance reports, and their supervisor's feedback.

Beyond completeness, evaluators assessed the quality of content in terms of the overall clarity of the file subject and the decision process. In Ghana, we assessed files along six margins:

- How clear is the background to issues?
- How clear is the outline of courses of action available or taken?
- Is the file organized in a logical flow?
- Are choices based on evidence in the file?
- Is it clear who should take action?
- What proportion of materials have a clear deadline?

In Liberia, we reviewed the extent to which civil servants' objectives and performance indicators follow the required SMART framework: whether relevant elements were specific, measurable, achievable, relevant, and time-bound. We assessed files along six distinct dimensions:

- Are different/unique categories of objectives presented?
- Are these objectives specific/measurable/time-bound?
- Are there associated performance indicators/measures?
- What is the extent and quality of reporting on each of these measures?
- Did the manager give recommendations as to how to meet the objectives?
- Did the manager identify development needs and how they could be met?

We also made note of the scores given by managers in the appraisals to assess whether they were validated by the evidence presented in the appraisal documents and indicated a true distribution across the unit.

Table E.1 in appendix E presents the instrument used in Ghana to measure the quality of general processes in government files. Files were assessed on the following sets of indicators:

- The comprehensiveness of reporting on the activity across the series of tasks (for example, "Where applicable, are minutes, memos and other necessary records present and complete [including from whom, to whom and signature]?")
- The sufficiency of the evidence and rationale following each of the decisions made (following the government's due process) (for example, "How would you characterise the quality of content you have in the file along the following margins? Choices are based on evidence in file.")
- The overall commitment to effective processes of the unit as reflected in the file (for example, "In general, to what extent does the file organisation adhere to government procedure? [Give an overall score from 0 to 100.]").



Table E.2 in appendix E presents the instrument used in Liberia to measure the quality of implementation of the appraisal process. Files were assessed on the following sets of indicators:

- The comprehensiveness of reporting across the series of appraisal forms (for example, “Which forms have been completed for Employee [Name]?”)
- The sufficiency of the evidence and rationale determining each of the appraisal scores given an employee (for example, “Comments are substantive and provide a quality assessment of officer’s contributions [even if discussion is that officer had to do work not in key objectives].”)
- The overall commitment to an effective appraisal process of the unit as reflected in the package of appraisal documents (for example, “When reviewing the whole set of appraisal forms for a unit/all those filled in by an appraiser, were there any of the following discrepancies in the set of appraisal forms for the unit? Objectives are very similar across forms.”).

### **External Process Productivity**

To measure external process productivity, we tracked the timeliness of units’ responses to requests from the centralized service management agency (the OHCS or the CSA) and the completeness and quality of the responses. More specifically, we measured the following:

- The time it took for a unit to respond to the request
- The extent to which all officers on the staff roster for that unit were reported on
- The accuracy of the information (through spot checks, where possible).

In Ghana, the research team tracked request letters from three directorates of the OHCS to public service organizations and the date of delivery of their responses, before and after a clear deadline. The three directorates asked organizations to share five different HR documents: promotion registers, training plans and reports, annual performance reports, the chief director’s (CD) self-assessment report, and a signed head of department/director’s performance agreement. The research team tracked organizations’ internal response time in the execution of a request, recording the period when minutes and memos were executed by schedule officers and the final delivery to the OHCS.

In Liberia, over 400 bureaucratic units and divisions from 28 civil service organizations who were participating in an impact evaluation study were asked to submit personnel files to the CSA. This was done by sending a letter with a set of standardized personnel requests to these study units. The research team then looked at whether the units responded to the request and, if so, what their response time was as a measure of process productivity. The survey firm BRAC assisted the CSA in handing out the letter that communicated this file request and in recording when unit representatives submitted their files in response. Personnel listings were submitted either as hard copies in person or via email to the CSA’s Management Services Directorate.

### **Data Collection**

The exercise to assess internal process productivity in Ghana started in April 2018 and lasted for six months, including a two-month pilot. In total, 763 files were assessed from 55 organizations. Randomly sampling across the four main administrative directorates and technical units, the research team audited files from 256 divisions and units.<sup>9</sup>

In Liberia, enumerators assessed the quality of PMS files completed in 2017–19 for the same 437 units that had participated in an impact evaluation study at that time. All Liberian civil servants were supposed to use the PMS process to track and improve performance management. The enumerators thus assessed whether all staff in each unit had completed the PMS forms each year and, if so, the quality of those forms. These three assessments each took place after the completion of the annual PMS process cycle in December 2017, 2018, and 2019.<sup>10</sup> In total, civil servants employed in 437 units and divisions across 28 organizations were assessed on whether they had completed,

**TABLE 13.2 Completion of PMS Forms, Liberia, 2017–19**

PMS form type	PMS in 2017	PMS in 2018	PMS in 2019
Form 1: Employee performance planning and progress review	1,440	1,655	1,232
Form 2: Employee self-assessment form	774	1,110	600
Form 3: Performance appraisal form	1,297	1,197	547
Individuals with at least one form	2,021	1,587	1,202
Individuals with forms 1 and 3	577	1,090	509
Individuals with all three forms	466	948	498

Source: Original table for this publication.

Note: PMS = performance management system.

in full or in part, the PMS process in 2017–19. Survey data were collected for 7,419 bureaucrats across the three years, whereby 4,810 PMS files were found and could be assessed as a census of available documents (see table 13.2).

The exercise to assess external process productivity in Ghana started in February 2018 and ended in May 2019. In total, 750 letters were tracked during the data collection period sent to 31 ministries and departments in 2018 and 30 ministries and departments in 2019, requesting types of data specific to human resource management (HRM) and policy, planning, monitoring, and evaluation (PPME) organizational divisions. In Liberia, the exercise of requesting and tracking the receipt of personnel files to measure units' responsiveness started on February 24, 2020, and concluded on March 24.

## RESULTS

### Internal Process Productivity in Government

Table 13.3 presents descriptive statistics for the procedural measures of process quality in Ghana, while table 13.4 presents statistics for the quality of the content of assessed files. We can see a substantial number of files were lacking in at least one of our categories, with only 3 percent of files having a complete or near-complete file ladder, 39 percent having close to the required minutes, and 9 percent having sufficient

**TABLE 13.3 Procedural Characteristics of Assessed Files, Ghana**

	(1) File ladder: Completeness	(2) File ladder: Transparency	(3) Folios	(4) Minutes and memos	(5) Incoming letters	(6) Outgoing letters
Proportion of files (0–19%)	0.40	0.70	0.35	0.04	0.06	0.72
Proportion of files (20–39%)	0.33	0.04	0.07	0.04	0.04	0.03
Proportion of files (40–59%)	0.04	0.04	0.10	0.15	0.13	0.03
Proportion of files (60–79%)	0.05	0.04	0.18	0.36	0.32	0.06
Proportion of files (80–100%)	0.03	0.03	0.27	0.39	0.42	0.09
Not applicable	0.12	0.15	0.00	0.01	0.04	0.07
Observations	763	763	763	763	763	763

Source: Original table for this publication.

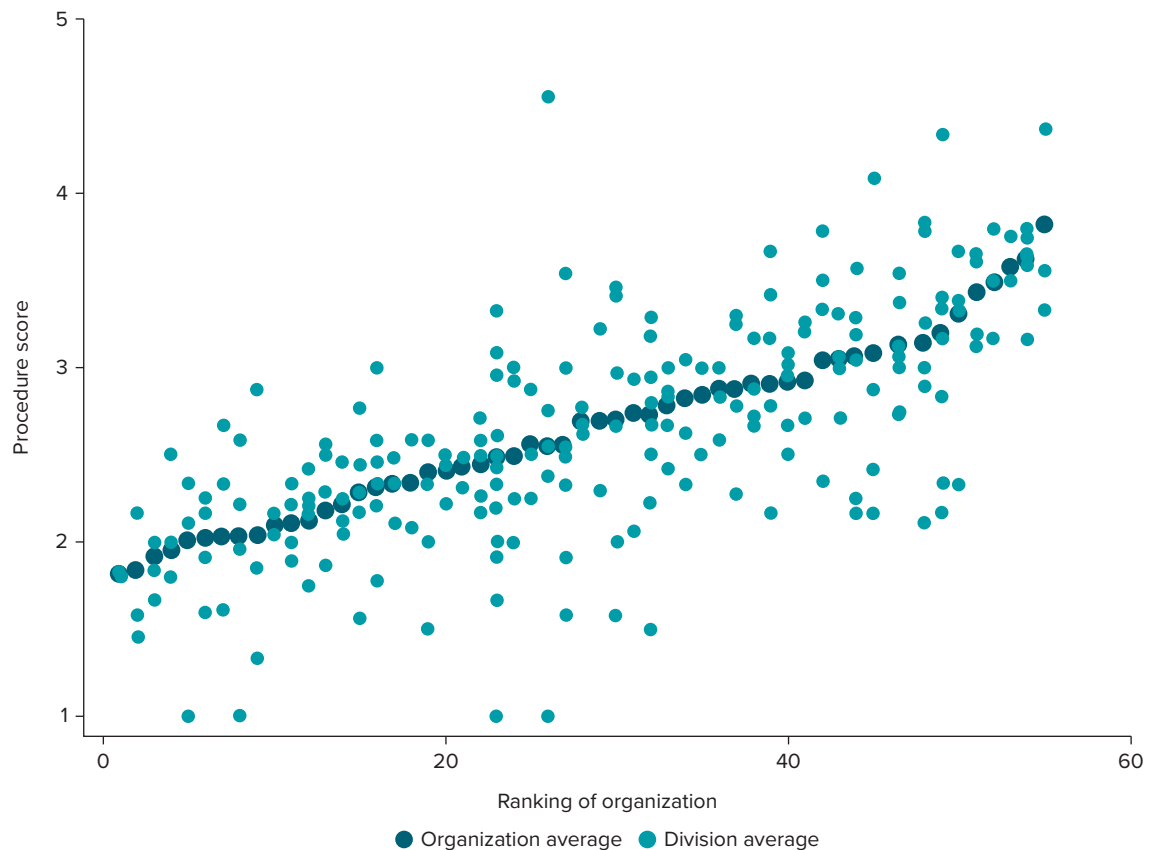
Note: The table reports descriptives of the main dimensions of files' procedural quality. Enumerators were asked to assess files on a Likert scale from 1 to 5, where 1 is "0–19%" and 5 is "80–100%," evaluated on the following margins: "How complete is the file ladder?" (column 1), "Does each step in the file ladder have dates?" (column 2), "Are folios within the file organised and numbered consecutively?" (column 3), "Where applicable, are minutes, memos and other necessary records present and complete (including from whom, to whom and signature)?" (column 4), "What proportion of incoming correspondence has an organisational stamp/date/signature?" (column 5), and "What proportion of outgoing correspondence has a despatch stamp/date/signature?" (column 6).

**TABLE 13.4** Content Characteristics of Assessed Files, Ghana

	(1) Background to issue	(2) Course action	(3) Logical flow	(4) Choices	(5) Action taken	(6) Clear deadline
Score 1	0.00	0.00	0.02	0.02	0.00	0.08
Score 2	0.08	0.07	0.14	0.17	0.07	0.14
Score 3	0.11	0.09	0.22	0.14	0.09	0.08
Score 4	0.60	0.66	0.47	0.54	0.66	0.03
Score 5	0.19	0.16	0.10	0.11	0.16	0.69
Not applicable	0.01	0.02	0.06	0.03	0.02	0.00
Observations	763	763	763	763	763	763

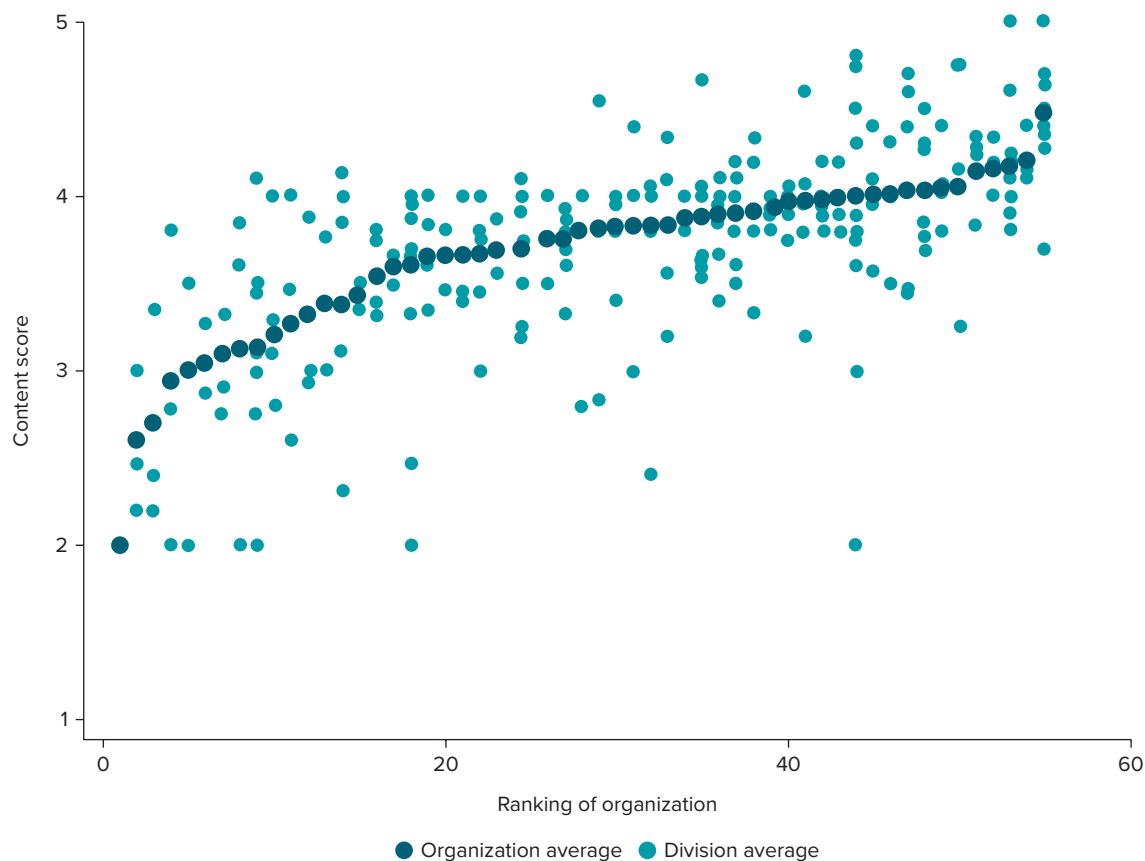
Source: Original table for this publication.

Note: The table reports descriptives on the main dimensions of files' content quality. Enumerators were asked to assess files on a Likert scale from 1 to 5, where 1 is "Very poor" and 5 is "Very good," evaluated on the following margins: "Background to issue" (column 1), "Clearly outlining what courses of action are available or taken" (column 2), "The file is organised in a logical flow" (column 3), "Choices are based on evidence in file" (column 4), and "Clarity on who should take actions at each stage" (column 5). In column 6, enumerators were asked to indicate the proportion of files with a clear deadline.

**FIGURE 13.1** Diversity in Level of Procedural Adherence across Organizations and Divisions, Ghana

Source: Original figure for this publication.

**FIGURE 13.2** Diversity in Content Scores across Organizations and Divisions, Ghana



Source: Original figure for this publication.

copies of outgoing letters. There is substantial room for improvement. Similarly, only 19 percent of files got the highest score in terms of the background they provided to issues, and 10 percent the highest score for logical flow of the argument. In general, the average level of organizational file adherence to public procedure is poor.

Figure 13.1 showcases how process quality varies across Ghanaian organizations. We average the scores for variables shown in table 13.3 into a single index and plot organizational averages of these scores as dark blue dots. There is a substantial degree of variation in the quality of adherence to processes across organizations. We also plot, stacked vertically at the “rank” of each organization, the scores for individual divisions within those organizations as light blue dots. Thus, the dispersion of the light blue dots around the dark blue dots indicates the degree of variation in process quality within an organization. We take a similar approach to the quality of content in figure 13.2, which summarizes an index created using the measures outlined in table 13.4.

We see a relatively high level of variation across organizations but also within organizations, with those in the middle of the distribution having some units whose process productivity is as bad as the average of the worst-performing organizations. At the same time, there is clearly some degree of correlation between an organization’s score and its divisions, indicated by the proximity of the light blue dots to the dark blue ones.

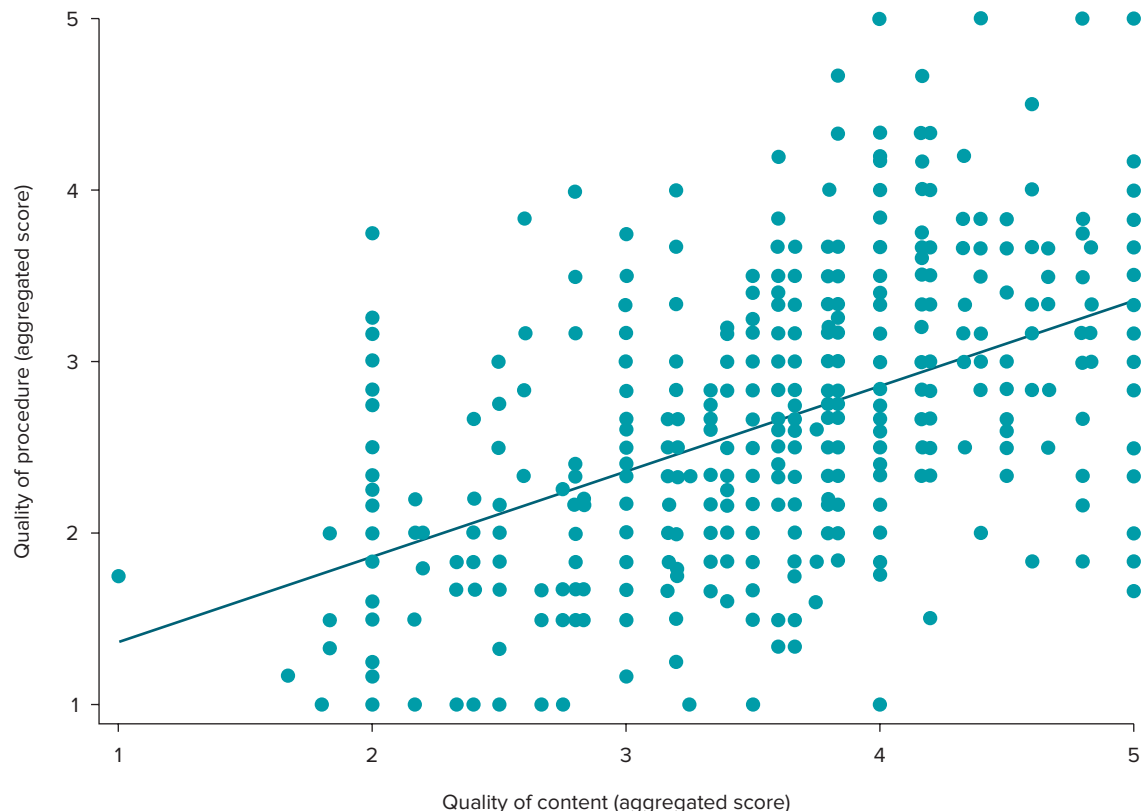
Together, these descriptive statistics tell us that though the general level of process quality is poor, there are some organizations that are able to raise the general standard for processes within their institutions. Though there are still some units that deviate from those practices (either positively or negatively), processes seem to be influenced by organizational practice.

The descriptives also indicate that some areas of process are of higher quality than others. Most of the files presented a blank or not fully complete file ladder, suggesting that organizations were not correctly reporting information on file movement (column 1). About 50 percent of files consecutively numbered folios, while around 40 percent poorly or very poorly organized documents (column 3). A high proportion of memos and minutes on documents were correctly compiled in 75 percent of sampled files (column 4). Looking at correspondence, incoming letters were in general aligned with government procedure, presenting a precise date, a clear signature, and an organizational stamp in 75 percent of cases (column 5). By contrast, outgoing letters were usually poorly compiled: 80 percent of the files show a high percentage of outgoing letters without a dispatch stamp, reflecting an unofficial rule to stamp envelopes rather than letters (column 6).

Likewise, some components of content quality in Ghana fare better than others. About 80 percent of the files had a clear or very clear background to issues (column 1) and clearly outlined what courses of action were available or had been taken (column 2). The files were organized in a logical flow in 57 percent of the cases (column 3), and choices were based on evidence recorded in the documents in 65 percent of the cases (column 4). Documents clearly stated who should take action at each stage in 70 percent of the files in the sample (column 5). On the other hand, the proportion of documents with a very clear deadline is also on the extreme, suggesting either that when documents required a deadline, this was clear, or that some documents did not have a deadline even though required (column 6).<sup>11</sup>

To what extent are those files that adhere to procedures most strongly also those that have a higher quality of content? Figure 13.3 presents a scatterplot (with one dot for each file we study) of the content quality

**FIGURE 13.3** Relationship between Adherence to Procedure and Quality of Content, Ghana



Source: Original figure for this publication.

Note: Each dot represents one file that was studied.

score against procedure quality. We see from the trend line that the relationship is positive, and the correlation is 0.48. However, it is also clear from the figure that there is a high degree of variation, with files that are well organized but with weak arguments, and vice versa.

In the Liberian civil service, only a fifth of public officials in the units assessed went through the PMS, reflecting an even weaker adoption of proper procedure in practice in the service.<sup>12</sup> However, when looking at the units that successfully adopted the PMS practice, on average 68 percent of civil servants working in the unit completed at least one of the PMS steps.<sup>13</sup> Table 13.2 illustrates that most staff who utilized the PMS process together with their supervisor completed their initial work plans and midyear assessments (Form 1) and, to a lesser degree, the end-of-year performance appraisal (Form 3). However, less evidence was found of staff assessing their own performance. This is an important piece of the process to ensure that appraisals are fair and the staff are engaged in and buy into the process. Ultimately, 60 percent or less of those who implemented the PMS did so by completing all three required forms. Overall use of the PMS also appears to have peaked in 2018, then fallen in 2019.<sup>14</sup> Hence, completion rates could improve.

Issues around form completeness further hindered enumerators' ability to assess the quality of the content in the forms found. The proportion of forms in which all compulsory sections had been filled in decreased from 84 percent in 2017 to 58 percent in 2018 and 39 percent in 2019. Furthermore, peaking at 50 percent when assessing 2018 forms, enumerators said that they had all the information they needed to assess quality for only 25 percent of the 2019 forms (see table 13.5). On a positive note, the proportion of files stored without a filing system decreased to just 5 percent of all forms found. Even so, table 13.6 shows how a lack of information in the files; poorly organized and at times missing pages; and, to a lesser extent,

**TABLE 13.5 Sufficiency of Information for Assessing Quality of PMS Forms, Liberia**

Did enumerators have all needed information to assess quality?	PMS in 2017	PMS in 2018	PMS in 2019
Have information needed	743 (37%)	795 (50%)	300 (25%)
Am missing information, but it is not critical to decision on quality	586 (29%)	533 (34%)	521 (43%)
Struggle to make judgment on form quality because of limited information	693 (34%)	259 (16%)	381 (32%)
Observations total	2,027	1,587	1,202

Source: Original table for this publication.

Note: The table shows the number of total observations where true, with the percentage of total observations made in parentheses. Five enumerators refused to answer questions in a survey on the PMS in 2017. PMS = performance management system.

**TABLE 13.6 Challenges in Judging the Quality of PMS Forms, Liberia**

Form quality issues	PMS in 2017	PMS in 2018	PMS in 2019
No challenges encountered	0.43 (0.50)	0.58 (0.49)	0.36 (0.48)
Little information in file	0.42 (0.49)	0.28 (0.45)	0.49 (0.50)
Poorly organized form	0.15 (0.35)	0.20 (0.40)	0.16 (0.37)
Some pages were missing	0.12 (0.33)	0.11 (0.31)	0.11 (0.31)
Poor level of legibility	0.10 (0.30)	0.11 (0.32)	0.10 (0.30)
Lack of coherence	0.09 (0.28)	0.04 (0.19)	0.08 (0.27)
Subject matter difficult to judge	0.07 (0.25)	0.01 (0.09)	0.02 (0.14)
Total observations	2,027	1,587	1,202

Source: Original table for this publication.

Note: The table shows means, with standard deviation in parentheses. PMS = performance management system.



ineligible, incoherent forms made it difficult to assess the PMS files' quality. This goes to show how a process to map and guide staff performance improvement, such as the PMS, is only as valuable as the level of detail and actionable observations recorded in the PMS forms.

Table 13.7 indicates that civil servants and their supervisors got better, at least initially, at developing SMART objectives and targets, which were then followed up on in midyear and end-of-year progress reports—even if their ability to develop time-bound goals could improve. However, table 13.8 suggests that supervisors could improve at providing practical advice and guidance to their staff through constructive feedback on how they could improve.

Looking at how well staff adopted the SMART objectives as one measure of the quality of the PMS process across organizations and units, figure 13.4 shows substantial variation in adoption. We create an index across different measures of the quality of the performance objectives by averaging the number that meet the relevant criteria. As with the Ghana data, we then take averages of those numbers at the unit and organizational levels. Figure 13.4 illustrates how effectively different organizations have implemented the appraisal process, with some organizations articulating their staff's entire work plan in a single objective. Within these organizations, we see substantial variation, dwarfing the variation across organizations. In the case of the Liberian PMS, process productivity seems to be highly influenced by unit staff.

Drilling down into two of the specific features of SMART indicators—the extent to which they are relevant and measurable—we repeat our analysis but restrict it to the proportion of indicators that were deemed relevant and measurable by our assessors. Figure 13.5 shows that there is once again significant variation across organizations but a similar scale of variation across units within organizations. Thus, again we see evidence that factors at the unit level substantially influence the quality of the PMS process.

**TABLE 13.7 Formulating and Reporting on Objectives and Targets, Liberia**

SMART objectives and targets	PMS in 2017	PMS in 2018	PMS in 2019
Percent of objectives that are specific	0.92 (0.21)	0.97 (0.10)	0.94 (0.16)
Percent of objectives that are measurable	0.65 (0.42)	0.74 (0.38)	0.56 (0.43)
Percent of objectives that are timebound	0.32 (0.40)	0.34 (0.41)	0.23 (0.36)
Percent of objectives with progress report (midyear)	0.88 (0.30)	0.95 (0.19)	0.81 (0.38)
Percent of objectives that were met/achieved (midyear)	0.81 (0.34)	0.72 (0.41)	0.61 (0.44)
Percent of targets that relate to objectives	0.92 (0.19)	0.93 (0.18)	0.94 (0.17)
Percent of targets that are measurable	0.41 (0.47)	0.66 (0.43)	0.54 (0.45)
Total observations range	924–1,358	1,394–1,545	1,010–1,165

Source: Original table for this publication.

Note: The table shows the means and standard deviation in parentheses. PMS = performance management system.

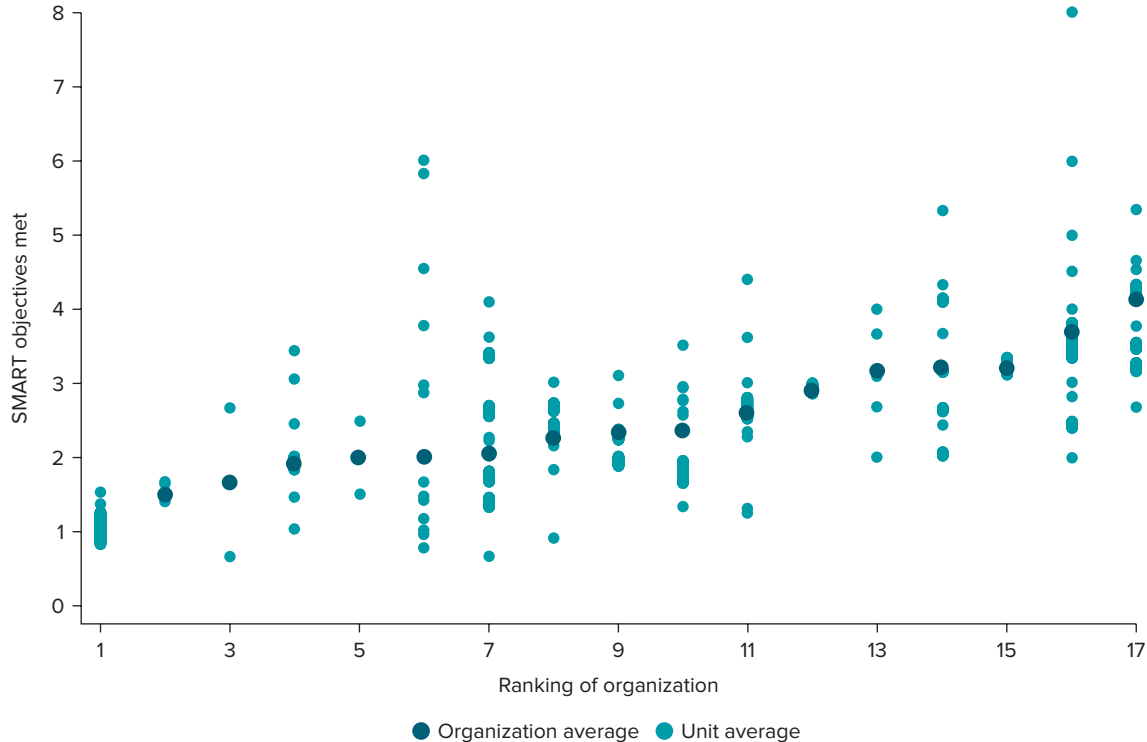
**TABLE 13.8 Quality of Supervisors' Feedback, Liberia**

Quality of feedback	PMS in 2017	PMS in 2018	PMS in 2019
Supervisor gave recommendations on how to meet objective	0.66 (0.48)	0.58 (0.49)	0.40 (0.49)
Supervisor identified development needs of the employee	0.48 (0.50)	0.43 (0.50)	0.41 (0.49)
Supervisor recommended activities to build employee's capacity	0.44 (0.50)	0.41 (0.49)	0.36 (0.48)
All objectives are reported on	0.64 (0.48)	0.72 (0.45)	0.23 (0.42)
All comments are substantive	0.38 (0.49)	0.41 (0.49)	0.27 (0.44)
All comments are constructive	0.25 (0.43)	0.31 (0.46)	0.04 (0.19)
Total observations range	943–1,353	1,069–1,544	510–1,010

Source: Original table for this publication.

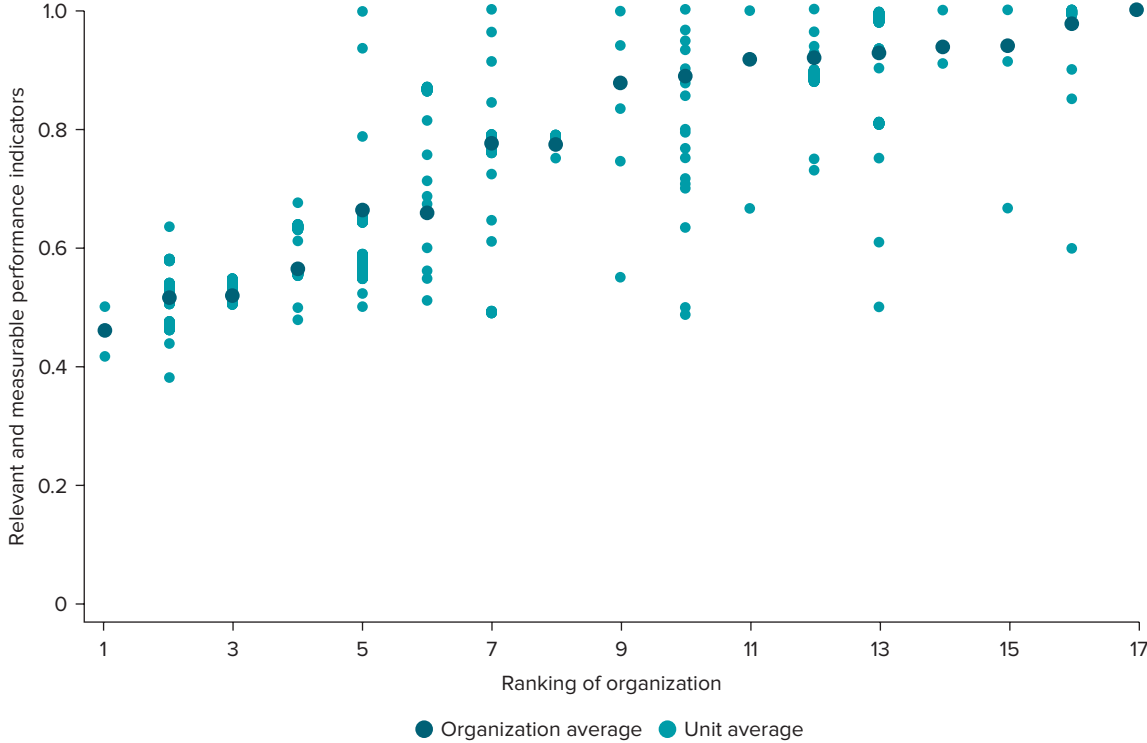
Note: The table shows the means and standard deviation in parentheses. PMS = performance management system.

**FIGURE 13.4** Average Number of SMART Objectives Identified in Appraisal Forms, Liberia



Source: Original figure for this publication.

**FIGURE 13.5** Average Number of Relevant and Measurable Indicators Identified in Appraisal Forms, Liberia



Source: Original figure for this publication.

## External Process Productivity in Government

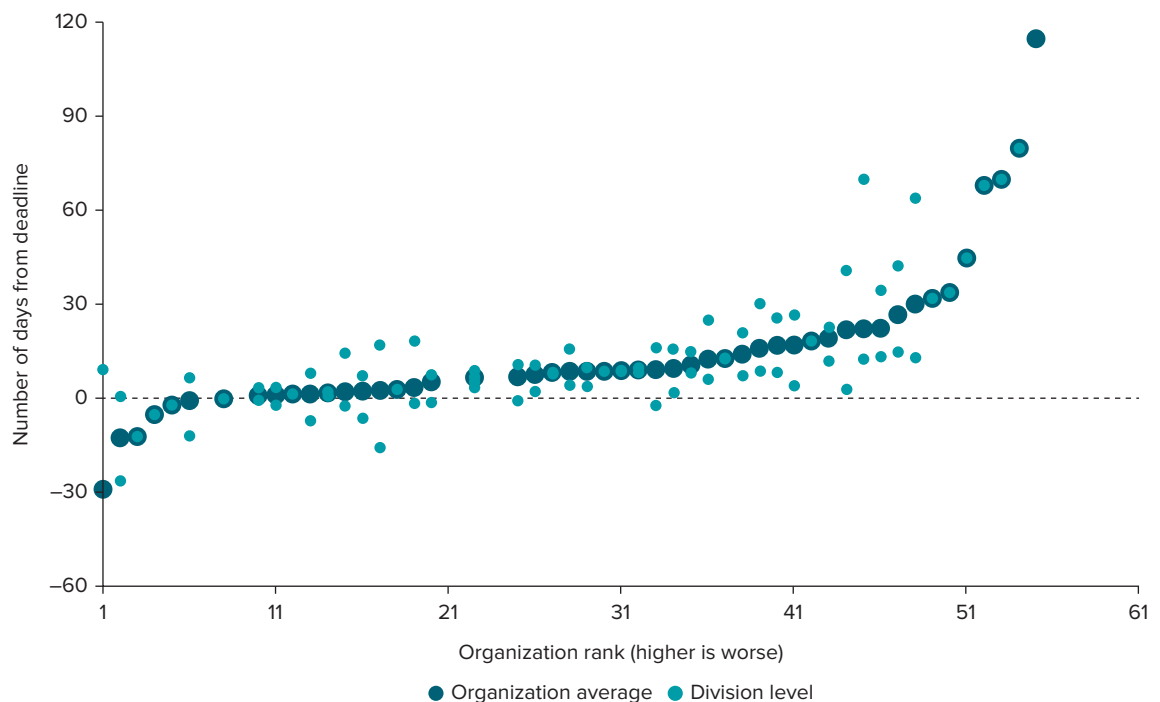
We now turn to the results of our assessments of external process productivity. Figure 13.6 shows the average number of days it took an organization (dark blue dots) to submit a response to the various requests made by the OHCS in 2017 and 2018 ( $y$  axis), with the organizations ranked by overall speed ( $x$  axis). Once again, we also present averages for the units within those organizations (light blue dots stacked vertically at their organization's ranking), but given that many such requests must be sent by the centralized dispatch office of the organization, we see a lot of clustering in the unit averages. A negative number on the  $y$  axis implies that the submission was received before the deadline (represented by 0 on the  $y$  axis).

Perhaps a third of organizations in Ghana's public service who eventually responded met the deadlines set by centralized entities. A minority of organizations were fully unresponsive and thus are not displayed in figure 13.6. However, even among those who eventually responded, perhaps a quarter did so a month or more late. Such delays impact the ability of central organizations to continue activities for which they require external information.

Turning to quality, figure 13.7 shows the completeness of the submissions received by the OHCS. The  $y$  axis displays the proportion of requests for which an organization (dark blue dot) or unit (light blue dot) submitted the required information. Here, organizational and unit averages are less closely related, since central dispatch offices will rarely mediate the quality of submissions. A few ministries, departments, and agencies submitted more than 80 percent of the data requested by the OHCS, and some units submitted all the information, while others submitted less than 20 percent of the information requested. The average level of quality is rather low, with the median organization submitting just over 60 percent of the information requested. All of this has knock-on effects on the capacity of the OHCS to undertake its work.

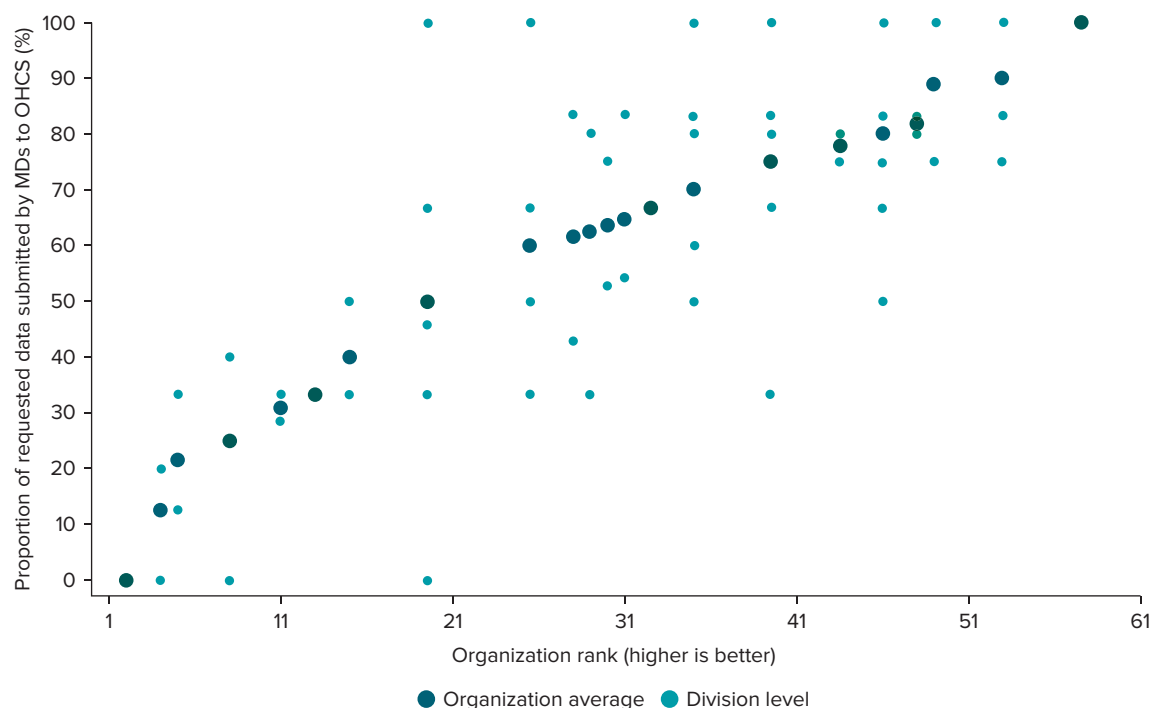
A similar picture is found in Liberia. Though not displayed here, we find similarly low responsiveness to centralized requests, with an even greater number of organizations simply not submitting any response at all.

**FIGURE 13.6** Diversity in Number of Days to Receive Requested Information from Organizations and Divisions, Ghana



Source: Original figure for this publication.

**FIGURE 13.7** Diversity in Proportion of Required Files Submitted by Organizations and Divisions, Ghana



Source: Original figure for this publication.

Note: MDs = ministries and departments; OHCS = Office of the Head of Civil Service.

Of the 348 units across government that we confirmed received the CSA's request, 30 units responded, 21 (70 percent) within the deadline. The quality of those submissions is even more limited, with many containing little to no usable information. Trying to undertake personnel policy making when your colleagues in the rest of the service simply refuse to answer your requests for information must be challenging.

## CONCLUSION

This chapter has put forward a framework for measuring process quality in public administration: identifying evidence of transparent, logical, and equitable decision-making throughout government. Though it is a fundamental part of the activities of the public sector, the quality of public officials' work processes has rarely been measured for government analytics. This drives assessments of government functioning and productivity toward frontline services and limits analysts' capacity to assess where in the long chain of government processes dysfunction might be occurring.

We have distinguished between internal process productivity, the quality of administrative processes for activities confined within a particular administrative unit, and external process productivity, the quality of administrative processes for activities in which units interact. We have made this distinction because accountability and professional dynamics vary distinctly between the two cases but also because appropriate measurement varies as well. We have then applied our framework to two case studies, concerning general government processes in the government of Ghana and the appraisal process in the government of Liberia. We have shown that in these settings, the quality of government processes is generally poor but highly varied, with some organizations and units effectively adhering to government processes and a higher overall quality of administration.

Such measures of process quality in public administration open up three areas of government analytics: assessments of variation in process quality and associated productivity within and across organizations in the same country, comparisons of process quality across countries, and assessments of public sector process quality over time. In this way, government analysts can pinpoint where government procedure is not being adhered to, how different processes relate to public sector productivity, and what the dynamics are across individuals and organizational units.

Strengthening the quality of government processes would require increasing and updating the knowledge of public officials on appropriate ways of handling government work, strengthening senior officers' supervision, and reinforcing their capacity to hold staff to account for poorly adhering to government processes. As the world's public administrations become increasingly digital, the ability to detect substandard processes will become more automated, but the continued assessment of which processes lead to improved productivity will require the use of this information for analysis. We hope this chapter has provided a framework for such work.

## NOTES

The authors gratefully acknowledge funding from the International Growth Centre; Economic Development and Institutions; the World Bank's i2i initiative, Knowledge Change Program, and Governance Global Practice; and the USAID-, Sweden-, and World Bank-sponsored Liberia Public Sector Modernization Project. We thank Nyetuan Mulbah, Francesco Raffaelli, and Andre Cazor Katz for excellent research assistance and the heads of Ghana's and Liberia's civil services under which the work was implemented, Nana Agyekum-Dwamena (Ghana) and Puchu Leonard and James A. Thompson (Liberia). We thank Mrs. Rejoice Dankwa, Mr. Godwin Brocke, Patience Coleman, Stefan Dercon, Erika Deserranno, Aisha Nansamba, Dorothy Kiepeeh, Smile Kwawukume, Vincent Pons, Imran Rasul, and George B. Wah for their guidance. All errors are our own. Finally, this paper was published after the passing of our coauthor, Felix Nyarko Ampompong, and we therefore dedicate the work to him.

1. See the measures under "Instructional leadership" in table 29.3 of chapter 29.
2. It should be noted that some of the indicators of proper procurement and customs procedures are versions of measures of process productivity.
3. Frontier empirical evidence on what bureaucrats do showcased in chapter 17 implies that almost three-quarters of bureaucratic work is related to undertaking bureaucratic processes, such as monitoring, training, and personnel management; financial and budget management; and so forth. It would seem that process productivity is key to the productivity of the public sector.
4. An intermediate approach is Hollyer, Rosendorff, and Vreeland (2017), who use reporting to the World Development Indicators as a measure of government transparency.
5. The OHCS has a Public Records and Archives Administration Department (PRAAD), whose aim is to facilitate and promote good government processes and record-keeping practices across ministries and departments. Officials are trained in relevant processes upon entry to the public service, as well as at regular in-service trainings.
6. At the end-of-year review, employees are supposed to assess their own performance against 10 servicewide standards in what we refer to as Form 2. They are further assessed by their supervisors on these 10 servicewide indicators, as well as on their individual overall performance and behavior in the workplace, in Form 3.
7. Importantly, there have been efforts to engage on the PMS between CSA and ministries or agencies, to train hundreds of supervisors and staff on the PMS cycle, and to assign individuals in each public agency to act as focal points on issues related to the rollout of the PMS. Still, limited political will to adopt the process in a timely manner; its paper-based format; and disconnect from any recognition, rewards, or sanctions system remain persistent challenges.
8. We employed senior and retired civil servants in Ghana to review the extent to which randomly chosen unit files followed appropriate government processes, whereas, in Liberia, this was done by enumerators from an external survey firm.
9. The sampled files were assessed by three assistant management analysts from the Management Services Department of the OHCS. During the piloting period, the tool was adjusted and improved to reflect the records management practices within the Ghanaian civil service. Files in the sample are indicatively opened in 2015, not confidential, and not related to personal or financial subjects.
10. The files were assessed by enumerators from Liberia-based survey firm BRAC.
11. In this case, the tool allowed a "not applicable" option. In 54 percent of the files assessed, documents did not require a specific deadline.

12. With an estimated total workforce of 7,099 in the units assessed, based on 2017 staff lists, only 28 percent, 22 percent and 17 percent of staff had completed at least one of the PMS forms in 2017, 2018, and 2019, respectively.
13. In the units with any adoption in that year, 65 percent, 67 percent, and 71 percent of staff had filled in at least one PMS form in 2017, 2018, and 2019, respectively.
14. A new administration came into office in 2018, and numerous pay reforms that resulted in pay cuts for some in 2018 and 2019 may have impacted civil servants' motivation and prioritization of the PMS process.

## REFERENCES

- Alesina, Alberto, and Guido Tabellini. 2007. "Bureaucrats or Politicians? Part I: A Single Policy Task." *American Economic Review* 97 (1) (March): 169–79. <https://doi.org/10.1257/aer.97.1.169>.
- Banerjee, Abhijit, Raghavendra Chattopadhyay, Esther Duflo, Daniel Keniston, and Nina Singh. 2021. "Improving Police Performance in Rajasthan, India: Experimental Evidence on Incentives, Managerial Autonomy, and Training." *American Economic Journal: Economic Policy* 13 (1) (February): 36–66. <https://doi.org/10.1257/pol.20190664>.
- Bedoya, Guadalupe, Amy Dolinger, Khama Rogo, Njeri Mwaura, Francis Wafula, Jorge Coarasa, Ana Goicoechea, and Jishnu Das. 2017. "Observations of Infection Prevention and Control Practices in Primary Health Care, Kenya." *Bulletin of the World Health Organization* 95 (7) (July): 503–16. <https://doi.org/10.2471/BLT.16.179499>.
- Berliner, Daniel, Benjamin E. Bagozzi, Brian Palmer-Rubin, and Aaron Erlich. 2021. "The Political Logic of Government Disclosure: Evidence from Information Requests in Mexico." *The Journal of Politics* 83 (1): 229–45. <https://doi.org/10.1086/709148>.
- Best, Michael Carlos, Jonas Hjort, and David Szakonyi. 2017. "Individuals and Organizations as Sources of State Effectiveness." NBER Working Paper 23350, National Bureau of Economic Research, Cambridge, MA. <https://ideas.repec.org/p/nbr/nberwo/23350.html>.
- Chong, Alberto, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2014. "Letter Grading Government Efficiency." *Journal of the European Economic Association* 12 (2): 277–99. <https://doi.org/10.1111/jeea.12076>.
- CSA and USAID-GEMS (The Civil Service Agency and the USAID Governance and Economic Management Support Project). 2016. *Performance Management Policy Manual for the Civil Service of Liberia*. Monrovia, Liberia: Civil Service Agency. <https://csa.gov.lr/doc/Performance%20Management%20System%20Manual.pdf>.
- Daniels, Benjamin, Amy Dolinger, Guadalupe Bedoya, Khama Rogo, Ana Goicoechea, Jorge Coarasa, Francis Wafula, Njeri Mwaura, Redemptar Kimeu, and Jishnu Das. 2017. "Use of Standardised Patients to Assess Quality of Healthcare in Nairobi, Kenya: A Pilot, Cross-Sectional Study with International Comparisons." *BMJ Global Health* 2 (2): e000333. <https://doi.org/10.1136/bmjgh-2017-000333>.
- Dixit, Avinash. 2002. "Incentives and Organizations in the Public Sector: An Interpretative Review." *The Journal of Human Resources* 37 (4): 696–727. <https://doi.org/10.2307/3069614>.
- Fenizia, Alessandra. 2022. "Managers and Productivity in the Public Sector." *Econometrica* 90 (3): 1063–84. <https://doi.org/10.3982/ECTA19244>.
- Forte, D. Othniel. 2010. "Civil Service Reform in Post Conflict Liberia." Unpublished manuscript. [https://www.academia.edu/12454955/Civil\\_Service\\_Reform\\_in\\_Post\\_Conflict\\_Liberia](https://www.academia.edu/12454955/Civil_Service_Reform_in_Post_Conflict_Liberia).
- Friedman, Jonathan. 2012. "Building Civil Service Capacity: Post-Conflict Liberia, 2006–2011." *Innovations for Successful Societies*, Princeton University, August 2012. <https://successfulsocieties.princeton.edu/publications/building-civil-service-capacity-post-conflict-liberia-2006-2011>.
- Hollyer, James R., B. Peter Rosendorff, and James Raymond Vreeland. 2017. "Measuring Transparency." *Political Analysis* 22 (4): 413–34. <https://doi.org/10.1093/pan/mpu001>.
- Peisakhin, Leonid, and Paul Pinto. 2010. "Is Transparency an Effective Anti-Corruption Strategy? Evidence from a Field Experiment in India." *Regulation & Governance* 4 (3): 261–80. <https://doi.org/10.1111/j.1748-5991.2010.01081.x>.
- PRAAD (Public Records and Archives Administration Department). 2007. *Records Office Procedures Manual*. Accra: Government of Ghana.
- Rasul, Imran, and Daniel Rogger. 2018. "Management of Bureaucrats and Public Service Delivery: Evidence from the Nigerian Civil Service." *The Economic Journal* 128 (608): 413–46. <https://doi.org/10.1111/eoj.12418>.
- Rasul, Imran, Daniel Rogger, and Martin J. Williams. 2021. "Management, Organizational Performance, and Task Clarity: Evidence from Ghana's Civil Service." *Journal of Public Administration Research and Theory* 31 (2): 259–77. <https://doi.org/10.1093/jopart/muaa034>.



- Wafula, Francis, Amy Dolinger, Benjamin Daniels, Njeri Mwaura, Guadalupe Bedoya, Khama Rogo, Ana Goicoechea, Jishnu Das, and Bernard Olayo. 2017. "Examining the Quality of Medicines at Kenyan Healthcare Facilities: A Validation of an Alternative Post-Market Surveillance Model that Uses Standardized Patients." *Drugs—Real World Outcomes* 4 (1): 53–63. <https://doi.org/10.1007/s40801-016-0100-7>.
- Wehner, Joachim, and John Poole. 2015. "Responsiveness of UK Local Governments to FOIA Requests." LSE Department of Government Working Papers, London School of Economics and Political Science, London.
- Wilson, James Q. 1989. *Bureaucracy: What Government Agencies Do and Why They Do It*. New York: Basic Books.
- Wood, Abby K., and David E. Lewis. 2017. "Agency Performance Challenges and Agency Politicization." *Journal of Public Administration Research and Theory* 27 (4) (June): 581–95. <https://doi.org/10.1093/jopart/mux014>.
- World Bank. 2014. *Liberia—Public Sector Modernization Project*. Washington, DC: World Bank. <http://documents.worldbank.org/curated/en/530841468263672030/Liberia-Public-Sector-Modernization-Project>.

## CHAPTER 14

# Government Analytics Using Customs Data

*Alice Duhaut*

### SUMMARY

Many government agencies have multidimensional missions, in which achieving one objective can reduce the attainment of another organizational objective. This presents particular challenges to government analytics. Incomplete measurement of objectives risks encouraging the attainment of the measured objectives while unknowingly impairing other objectives. This chapter showcases how government analytics can be applied in such contexts, using the example of customs agencies. Customs agencies typically have three core objectives: facilitating trade, collecting revenue, and ensuring the security and safety of the goods entering or exiting the country. Attaining one objective (for example, the greater safety of traded goods) can come at the expense of another (for example, facilitating trade). This puts a premium on the effective measurement of all dimensions of a customs mission, which requires triangulating different data sources. This chapter showcases how this can be done, deriving indicators for trade facilitation (for example, the costs of the process—in particular, time delays), revenue collection (for example, trade volume and revenue collected based on the assessed value), and safety (for example, the number of goods in infraction seized). The chapter also underscores how a wider use of the customs database itself could help measure performance, by combining it with other data collection methods, such as the World Customs Organization (WCO) Time Release Study (TRS) and exciting developments in GPS tracking data.

### ANALYTICS IN PRACTICE

- Government organizations with multidimensional missions—such as customs—typically need to integrate multiple data sources to ensure they measure performance holistically and avoid measuring and focusing on some goals but not others. In customs, the efficiency of the border-crossing process, and the customs agents and other agencies involved in it, should be evaluated with both traditional tools—the World Customs Organization (WCO) Time Release Study Plus (TRS+) and monitoring and evaluation metrics—and new or underused data sources—such as GPS data—to provide a way to reduce

---

Alice Duhaut is an economist in the World Bank's Development Impact Evaluation (DIME) Department.

the cost and increase the frequency of the indicators used to monitor border activities. An important element of the consolidation is to ensure the validity of the data used, match the relevant time stamps to the mapped process, and program indicators and queries to automatize reports.

- Data from different sources are likely to provide a different view, and even different takes, on the process. Measurement validation and triangulation are important components in analyzing customs data. It is thus important to invest in understanding the data routinely collected and to analyze them outside of survey periods. To complement the measures derived from the traditional TRS+, we recommend using customs database data to study time delays under customs' or other border agencies' control and the revenue collected. This requires understanding the full customs process and ensuring entries are not duplicated or incomplete, as might be the case if customs declarations for a shipment can be resubmitted under a different regime (for example, when the importer wants the goods to leave the warehouse and be released).
- Data should be standardized and rendered into reports for easy and fast consumption. Standardization of the extraction process, indicators, questions, and data treatment helps reproduce reports at a high frequency. From user surveys, information on the performance of the customs agent can also be collected.
- Valuation of goods in customs is challenging. To provide a holistic assessment, there are multiple techniques available to measure the value of goods in customs. In particular, comparing the value of goods when they leave a country of origin to their destination may assist in identifying the true value of goods. While valuation is a difficult process, and the World Trade Organization (WTO) rules describe how individual items' values should be evaluated, comparing what is declared at a country's borders to what is declared for similar goods of similar origin in peer countries can provide information on international trade taxes, the duties and excises collected, and the timeliness of the process. This indicator can flag where the value collected at customs is lower than expected.
- Time is an important consideration in customs, but the relevant checkpoints along the customs process against which it is measured must be defined (for example, if the clearance of the goods is considered the endpoint of a time analysis). Time delays can be studied in association with the mapping process to determine the relevant operations: one common operation studied based on Automated System for Customs Data (ASYCUDA) data is the time between assessment and clearance excluding the payment of taxes and duties. This exclusion is important because payment issues can be the cause of a lot of the delays, and such findings would have different policy implications.

## INTRODUCTION

Many government agencies have multidimensional missions, in which achieving one objective can reduce the attainment of another organizational objective. For instance, in some countries, financial regulators are mandated to develop financial services while also protecting consumers, or environmental agencies are mandated to both protect and develop natural resources. Organizations with such multidimensional missions with conflicting goals present particular challenges to government analytics. Incomplete measurement of objectives risks encouraging the attainment of the measured objectives while unknowingly impairing other objectives. Yet different objectives can often only be measured through very different types of data. This chapter showcases how government analytics can be applied in such contexts, using the example of customs agencies. By showcasing the integration of different data sources to measure multidimensional mission attainment holistically, the chapter complements other chapters in *The Government Analytics Handbook* that detail the use of one particular form of data—such as case data in chapter 15 or task data in chapter 17.

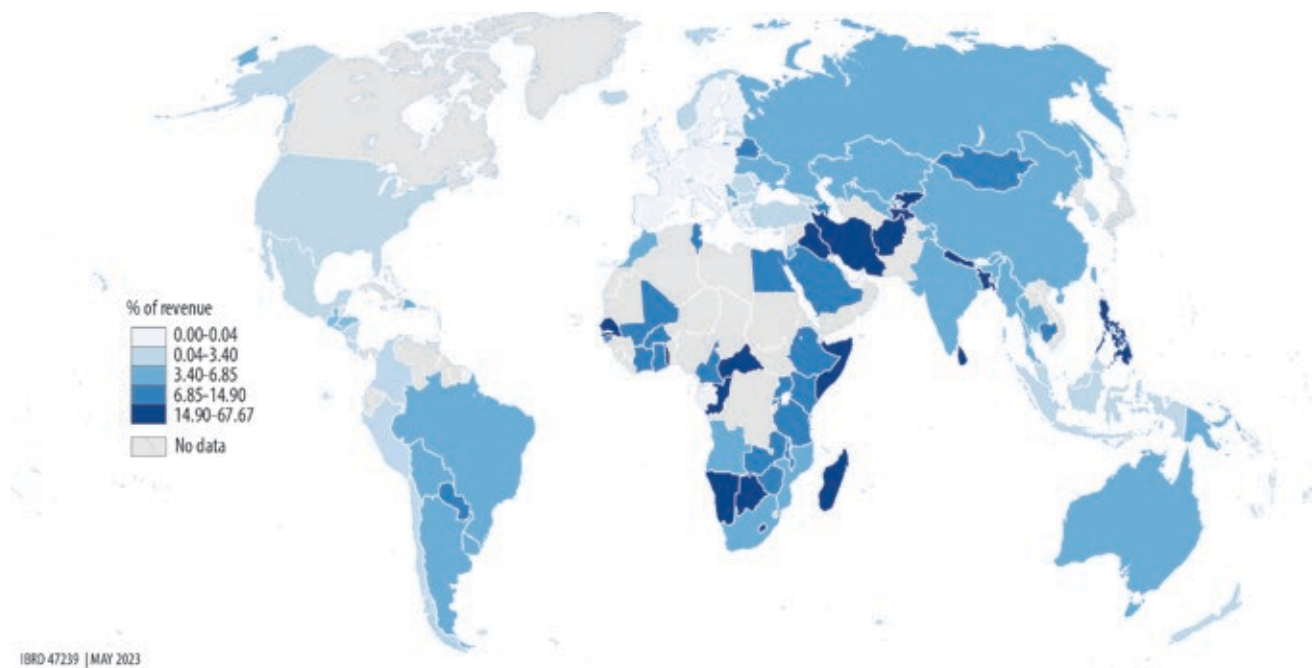
Among agencies tasked with multidimensional missions, customs is arguably a key one. Customs operations are located within the international borders of a country and are responsible for the processing of export and import goods. This includes several steps, from reviewing goods declarations to risk assessments, inspections, clearance, and risk collection. Revenue collection from customs is particularly important in low- and middle-income countries. Frequently, it represents a substantial share of state revenue and, at times, is also used for collecting fees requested by other government agencies.

In high-income countries, customs and other import duties represent, on average, 3.8 percent of state revenue, but, as illustrated in map 14.1, this value rises for countries with lower average incomes. For upper-middle-income countries, it stands at 8.9 percent, for lower-middle-income countries, at 11 percent, and for low-income countries, at 20 percent. For some countries in Sub-Saharan Africa, South Asia, and the Pacific islands, customs and import duties provide over one-third of all tax revenue. In addition to its key role in revenue collection, customs ensures borders' integrity and is the point of entry for goods coming into or going out of the country. For these reasons, the performance of customs operations has substantial implications on the fiscal sustainability and trade engagement of countries. The challenge of improving customs performance can thus be viewed from the vantage point of the following three distinct missions: the facilitation of trade across borders, the collection of revenue, and the protection of the safety of people and the security of goods coming through the borders. Working toward these three missions simultaneously involves trade-offs: making progress toward one goal can undermine the achievement of another. For example, facilitating trade means improving the customs process to reduce its total duration, including inspection and screening times. If the frontline agents were to perform fewer inspections, this would lead to a faster border crossing. However, this would likely have revenue implications, as proper product classification and tax collection would be more prone to errors.

Examples of initiatives undertaken toward those three key objectives, as well as some associated challenges, are illustrated in box 14.1 on the case of Malawi.

This puts a premium on measuring the performance of agencies with multi-dimensional missions—such as customs—holistically to ensure these trade-offs are accounted for. With this in mind, this chapter provides an empirical guide to assessing customs performance across these three objectives, as well as

**MAP 14.1 Customs and Other Import Duties as a Percentage of Tax Revenue**



Source: World Development Indicators (latest available values).

## BOX 14.1 Interactions with Other Agencies: The Case of Malawi

The Republic of Malawi is a landlocked country in southeastern Africa (see map B14.1.1). In Malawi, 14 agencies are present at the border. The agencies perform the inspections related to their missions: for instance, the Malawi Bureau of Standards ensures that the foodstuff coming in to the country respects Malawian standards. Together, these agencies strive to improve performance as related to the three principal objectives of customs operations below:

**Trade facilitation:** Malawian customs agents strive to improve the flow of goods and services across the border. One example of their efforts in this sphere is upgrading to the Automated System for Customs Data (ASYCUDA) World system in 2018. The new system facilitates trading across the border by, among other things, allowing web access for businesses, enabling the round-the-clock submission of customs declarations, and providing customized data extraction features.

MAP B14.1.1 Location of Malawi



Source: World Bank.

**Revenue collection:** Customs and other revenue duties collected in Malawi were equal to MK 88.3 billion in 2019, which represented 8.9 percent of all state tax revenue.

**Protection of the safety of people and security of goods coming through the borders:** The Malawi Revenue Authority restricts the import of certain classes of goods by requiring import licenses. These include military uniforms, ammunition and guns, fertilizer, pharmaceuticals, gold, and several types of foodstuff. In addition, all animals and animal products need to be certified as disease-free. Importation of most types of meat products further requires prior written permission from the Minister of Industry and Trade.

*(continues on next page)*

## BOX 14.1 Interactions with Other Agencies: The Case of Malawi *(continued)*

In the future, the agencies will be connected to the customs database, and customs will perform joint inspections. While the reduction in the number of agencies at the border is likely to reduce the burden on drivers or transporters, customs will also have more responsibilities, and measuring their performance after the reform against the preintervention situation might be complicated. It is thus necessary to create indicators that will reflect the scope of the mission as well as be easy to implement with the existing data.

outlining the diverse data necessary to perform these assessments. There are multiple choices available for practitioners when building both performance indicators and customs databases. Practitioners should prioritize indicators that enable them to accomplish a particular policy objective, while considering the data and human resources constraints they face when developing them. Because modifications in how customs operates affect other policy areas—trade policy and fiscal revenue—any change in how data are ingested and consumed should be coordinated with other government agencies.

The chapter is structured as follows. First, it provides institutional context on how customs operates and the international policy framework governing customs data collection and trade policy. Section 3 reviews the academic literature on customs, emphasizing its role in trade facilitation and fiscal revenue. Section 4 outlines the data infrastructure requirements for analyzing customs performance and generating indicators. Section 5 presents different types of indicators used to measure customs performance, and the final section concludes.

## INSTITUTIONAL CONTEXT

### Customs Process Overview

The customs process follows a linear structure, from the formal declaration of goods to the payment of taxes and exit. The process starts with the submission of a goods declaration to the customs administration by the importer or exporter, or by a broker acting on behalf of the importer or exporter (figure 14.1). This can be done remotely or at the border, depending on the country. The goods declaration usually lists a description of the items, with the classification, weight or quantities, origin, and value of the items in the shipments. Supporting documents, such as an invoice and bill of landing, are submitted along with the declaration. In addition, the customs declaration contains the declarant's assessment of the taxes and duties to be paid.

The next step is risk assessment. This step can take place before the submission is made or when it is made, as soon as the goods arrive at customs. An initial screening is conducted through a customs database risk model, analyzing the risk level of a declaration and issuing recommendations at the product level. The risk department usually issues a color-coded clearance channel, in which the color indicates whether the documents or the goods have to be inspected, sends flags for potential fraud or discrepancies in the declaration, and, potentially, sends comments to help the inspector assess the correct valuation of the shipment.

The customs process then moves to the assessment of the declaration by an inspector. Based on the documentation submitted by the declarant and the diagnostics provided by the risk department, the inspector can overrule the clearance channel recommendations. If the green channel is recommended, nothing happens, and the shipment goes through customs uninspected. If the yellow channel is recommended, the documents submitted along with the goods' declaration are reviewed. If the red channel is recommended, the goods are inspected—either by scanning the container or opening the cargo. Based on the information



**FIGURE 14.1** Diagram of Customs Process



Source: Original figure for this publication.

accumulated, the inspector produces a report on the declaration. The report lists any adjustments to the classification of the goods, the origin, product characteristics or quantity, and, importantly, the value assessed, as well as the taxes and duties to be paid. It can also include penalties to be paid in case of fraud.

In the last two stages, the goods are cleared and the taxes and duties are paid. The goods are released upon proof of payment. The term *clearance* means the accomplishment of all formalities necessary to allow goods to enter home use or to be exported. *Release* means that the goods are physically placed at the disposal of the transporter or importer.

### International Trade Policy Framework

While work to efficiently regulate customs operations is done on the domestic front by customs authorities, international organizations play a significant role as well. Since trade is, by nature, the international flow of goods, multiple international trade policy frameworks have been designed to regulate it and provide guidance for domestic customs authorities. These frameworks have been built and advocated for by a set of international organizations, including the World Trade Organization (WTO), with its rules on customs valuation, the World Customs Organization (WCO), the voice of the international customs community, and the United Nations Conference on Trade and Development (UNCTAD). This section provides practitioners with an overview of these different international trade policy frameworks and agreements.

The WTO trade facilitation agreement (TFA) reached at the 2013 Bali Ministerial Conference includes provisions related to customs operations. Intended to expedite the movement, release, and clearance of goods, the agreement sets up procedures for effective communication between customs authorities and other entities directly involved in customs compliance issues. As a result, all WTO members can benefit from technical assistance and capacity building related to any area of everyday customs work. In particular, the TFA, which finally entered into force in February 2017 after being ratified by two-thirds of WTO members, was followed in July 2014 by the launch of another important tool, the Trade Facilitation Agreement Facility (TFAF). It was the first time in WTO history that the obligation to implement an agreement was linked to the capacity of the country to do so.

The mission of the WTO and other international organizations, such as the WCO, is broad. The WTO and WCO cooperate on a number of initiatives: customs valuation, market access, rules of origin, information technology agreement, and trade facilitation. Among numerous examples of such cooperation is the WTO's Agreement on Customs Valuation, which established the Technical Committee on Customs Valuation under the rule of the WCO. In the area of technical assistance, according to the WTO, the main focus remains on negotiations surrounding technical assistance. Another example is the Harmonized Commodity Description and Coding System, or "Harmonized System," a classification of goods under the lead of the WCO, which the WTO thoroughly follows. Established by the Tokyo Round agreement, the WCO's Technical Committee on Customs Valuation and the General Agreement on Tariffs and Trade (GATT) and WTO Committee on Customs Valuation provide advice and case studies on customs valuation. These international efforts provide a legal framework to regulate customs operations so that each member state determines the value of goods in a *neutral* and *uniform* way.

Historically, general principles for an international system of valuation were established under the GATT Article VII. The agreement sets the actual value of a good, the price at which merchandise is sold under competitive conditions. It was the first agreement for customs valuation that highlighted the importance of competitive conditions for the determination of the sale price and stated that the price under established rules should be related to either comparable quantities or quantities not less favorable. At the same time, the need to simplify and harmonize international trade procedures coexists with growing pressure from the international trading community to minimize the intervention of the government in commercial transactions (Widdowson 2007). WTO rules on customs valuation highlight the discretionary autonomy that customs authorities must retain to fulfill their duties in promoting food safety and security and fighting illegal practices.

Measurement of time as a critical component for efficient customs operations has been dictated by the WCO Time Release Study (TRS) as well. The TRS is a methodology to measure, using data-driven approaches, the time that it usually takes to release cargo. It is a part of the Performance Measurement Mechanism (PMM) thoroughly monitored by WCO. Aimed at data-driven decision-making, the TRS helps customs agencies see opportunities for further improvement of the processes involved in realizing and accepting cargo.

## THE MULTIDIMENSIONAL MISSION OF CUSTOMS AGENCIES

The mission of customs agencies typically translates into three core objectives: trade facilitation, fiscal revenue, and security and food safety. In outlining these objectives, this chapter presents evidence from research exploring how these goals can be pursued, as well as a detailed discussion of their analytical approach. The first cluster of research studies examines the role of customs and nontechnical barriers in trade facilitation. The second subsection provides an extensive discussion of customs as a source of fiscal revenue, with its associated challenges in fighting fraud and illegal practices, such as corruption. The last subsection presents studies that improve our understanding of how customs can promote product safety and ensure security.

### Objective One: The Role of Customs in Trade Facilitation

Scholarly interest in customs research stems from its potential to serve as a tool for trade facilitation. What follows is an overview of the evidence to date. For example, Fernandes, Hillberry, and Alcántara (2021) evaluate Albanian reforms that sharply decreased the number of physical inspections of import shipments. There are clear indications that reduced inspections increase imports substantially. And there is no compelling evidence that the reforms gave rise to evasive behaviors. Similarly, for exports, Martincus, Carballo, and Graziano (2015) focus on time as a critical barrier to trade. Using a unique data set that consists of the universe of Uruguay's export transactions over the period 2002–11, they demonstrate that delays have a substantial negative impact on firms' exports. Furthermore, this effect is more pronounced for newcomers.

A seminal research paper that looks at the measurement of time as an instrumental component for the efficient functioning of customs is by Djankov, Freund, and Pham (2010). The authors examine how time delays affect export volumes. To measure time, the total export delay is considered. This means that the time delay does not include the time spent in a home country, on procedures, or in transit. It consists of the time spent when a container is at the border, transportation from the border to the port, and getting to the ship. The logic is that trade volumes can impact home country trade times; the effect on transit times abroad is likely negligible. Nevertheless, Djankov, Freund, and Pham (2010) estimate a difference gravity equation showing that each additional day a product is delayed prior to being shipped reduces trade by more than 1 percent. Delays have a relatively more significant impact on exports of time-sensitive goods, such as perishable agricultural products. Hence, it is important to measure and study how changes in customs operations can facilitate trade.

## Objective Two: Customs as a Source of Fiscal Revenue

Another key policy objective for customs offices is increasing fiscal revenue. Several studies discuss interventions and propose mechanisms to improve local tax collection practices or incentivize inspectors posted in a given tax collection location. This is not surprising since there is evidence that trade tax revenues collected at the border constitute a large part of GDP, particularly for developing, low-income countries. Baunsgaard and Keen (2010) show, using a panel of 117 countries, that the inability to find alternative sources of revenue may hinder trade liberalization. Results suggest that high-income countries recovered from the revenue they lost during the past wave of trade liberalization, but the same does not apply to emerging markets, where recovery from trade liberalization is weaker.

Another major issue is the presence of tax evasion and corruption in customs administrations. Defining corruption following Bardhan (2006), Dutt and Traca (2010) show that in most cases, corrupt bureaucrats tax trade through either *extortion* or *evasion*. The former refers to a bureaucrat's demanding bribes from exporters for doing his duties, while the latter refers to a situation in which an exporter pays off a public servant to receive preferential treatment, like a lower tariff rate or the lowering of regulatory standards. Evasion may be trade-enhancing in an environment with high tariffs because it allows an exporter to effectively reduce the tariff rate by paying a bribe. However, in order to develop in a sustainable fashion, countries need to combat corruption more efficiently. In particular, developing economies are often in dire need of increasing state fiscal revenue via the rigorous implementation of customs rules, to be able to finance their development policies.

In seeking to increase tax revenues while reducing corruption, researchers and policy makers have been conducting experiments to identify optimal policies (Cantens, Raballand, and Bilangna 2019). One method that is relatively straightforward is mirror analysis, which compares the exports for a given country with the imports for its export client, or vice versa (WCO 2015). This approach is often limited by difficulties in obtaining detailed customs data. When implemented in Madagascar by Chalendard, Raballand, and Rakotoarisoa (2019), this method helped to identify the probability of fraud in the context of customs operations reforms.

Technology can help customs improve its mission while reducing fraud. In a natural experiment in Columbia, Laajaj, Eslava, and Kinda (2019) find that the computerization of imports led to an increase of six log points in the firm's value, with consequences for employment and tax collection. However, Chalendard et al. (2021) show that, through manipulation of the IT system, some customs inspectors and IT specialists were able to manipulate the assignment of import declarations. This was identified by measuring deviations from random assignments prescribed by official rules. Deviant declarations are found to be at greater risk of tax evasion, less likely to be deemed fraudulent, and cleared faster.

Another experiment analyzing policies to curb fraud was conducted in Madagascar (Chalendard et al. 2020). The authors investigated whether providing better information to customs inspectors and monitoring their actions could affect tax revenue and fraud detection. Results from the experiment show that monitoring incentivizes agents to scan more shipments, but they do not necessarily detect more fraud. Relatedly, Khan, Khwaja, and Olken (2019) propose a mechanism to improve the performance of public servants in collecting tax revenue, given their significance in enforcing and determining tax liabilities. Evaluating a two-year field experiment with 525 property tax inspectors in Pakistan, the authors stress the potential of periodic merit-based postings in enhancing bureaucratic performance.

## Objective Three: Security and Food Safety

Customs authorities play an essential role as regulators of food safety and security. Although disruptions in total trade volume due to food safety are relatively rare (Buzby 2003), international organizations such as the WCO assist customs in the event of natural disasters and food crises. In June 2010, the WCO established an ad hoc working group to find ways for customs authorities to quickly react to such emergencies. The WTO, in turn, supports food security practices through the work of its Agriculture Committee and an Agricultural

Market Information System (AMIS) by a recommendation of the UN High-Level Task Force on the Global Food Security Crisis.

The role of customs authorities and their food security practices revolves around two fundamental issues: consumers do not always judge food security properly, and there are substantial differences between countries in terms of the regulation of food safety. The notion of trade security differs considerably in developed countries and developing ones (Diaz-Bonilla et al. 2000). Additionally, there are substantial risks of contamination due to trade. Ercsey-Ravasz et al. (2012) provide evidence that given the international agro-food trade network, the speed of potential contamination is extremely high because it is not possible to track the country of origin of different food products.

Safety is another key concern for customs authorities and is often associated with operations to reduce the illegal trade of products. The academic literature has documented how illegal trade in goods operates. In the European Union, Świerczyńska (2016) provides a list of legal solutions implemented to sustain the twofold goal of customs authorities to combat the illegal trade in goods and, at the same time, decrease control measures that increase the cost of trade. In the Islamic Republic of Iran, Farzanegan (2009) estimates that a penalty rate on smuggling contributed to reducing illegal trade, using historical data from 1970 to 2002.

## DATA REQUIREMENTS

### Data Sources

Before delving into the definition of customs performance indicators, it is useful to explain the data requirements for measuring them. The first and fundamental source of data is the customs database. The most common system in low- and middle-income countries is the ASYCUDA. It is used by 100 countries and territories around the globe. This is a system designed by the United Nations Conference on Trade and Development (UNCTAD). Its purpose is to compile information pertaining to customs declarations, with customs office or border post information, frontline inspectors assigned to the case, potential changes in the clearance channel, irregularities, and final value assessments. In addition, this database lists goods by their characteristics, as well as the taxes and duties due. It was also designed with the goal of generating broad-ranging data for statistical and economic analysis of trade and customs performance. Box 14.2 illustrates the basics of the ASYCUDA's structure.

However, ASYCUDA data are rarely used outside of aggregate statistics of revenue collection. Most of the time, studies of time delays are based on a TRS. A TRS measures the time required for the release and/or clearance of goods, from the time of arrival at the border until the physical release of cargo. A TRS is conducted over a predefined period of time, during which several declarations are followed by the surveyor at some border posts. The surveyor observes all steps until release and makes note of the time spent and the associated costs. As noted by the WCO, the tool is useful to produce a pre-reform benchmark and needs to be repeated often to follow the evolution at a particular border post. However, intercountry comparisons are limited given differences in capacity and infrastructure (WCO 2018).

The information coming from the country databases is usually shared at an aggregated annual level with the UN Statistical Division. This information is treated and aggregated by the Harmonized System, typically using eight- or six-digit codes. The Harmonized System is a standardized classification of traded products based on numerical categories. The system is managed by the WCO and is regularly updated. Each product is described using eight digits.<sup>1</sup> It is used by customs authorities around the world to identify products when assessing duties and taxes and for gathering statistics. The vast range of product categories that customs agents regularly handle is illustrated by figure 14.2. It provides an overview of the total value of imports, classified according to 22 sections of the Harmonized System, across the 50 largest ports of entry in the United States.

## BOX 14.2 ASYCUDA Data Structure

The Automated System for Customs Data (ASYCUDA) database is composed of a series of modules. Each module corresponds to a set of users. The customs broker module gives brokers secure access to the system to fill in a declaration. The customs office module covers declaration processing and is accessible to customs office agents. The accounting module is accessible to auditors only. The operations—registration of the declaration, assignment to an inspector, inspection results, change in value, clearance, and release—all have a time stamp associated with them, but merging this information in one report can be complicated because they are stored in different tables of the relational database.

A typical extract from ASYCUDA data thus contains information on the entry point for a specific declaration, the number of items declared, the agent and importer name, the year, and the registration date (see figure B14.2.1). ASYCUDA data also register *free on board* value—value outside insurance claims and ownership rights on the shipment—and value-added taxes (VAT), duties, and excise values for a chosen declaration, as well as exchange rate information and the currency with which payment for goods has been made (see figure B14.2.2).

**FIGURE B14.2.1 Example of an ASYCUDA Extract: Basic Variables**

1	2	3	4	5	6	7	8	9	10	11	12	13	14
OFFICE	ENTRY	DEC_CODE	AGENT NAME	REGDATE	REGNO	TPIN	IMPORTER NAME	YEAR	ITEMNO	Lane At Select	Current Lane	REGIME	HSCODE
BIR	DED	CA26775	MALAWI AGENT 1	31.01.2022	83	12345678	IMPORTER 1	2020	1	RED	Green	IM4	62034300
BIR	DED	CA26775	MALAWI AGENT 2	01.02.2022	84	12345679	IMPORTER 2	2021	1	RED	Red	IM4	62053010
SWE	DED	CA26776	MALAWI AGENT 3	02.02.2022	85	12345680	IMPORTER 3	2022	1	BLUE	Green	IM4	73261990
BIR	BIR	CA26777	MALAWI AGENT 4	03.02.2022	86	12345681	IMPORTER 4	2022	1	YELLOW	Yellow	IM4	61103000
MUL	DED	CA26778	MALAWI AGENT 5	04.02.2022	87	12345691	IMPORTER 5	2022	1	YELLOW	Green	IM4	62171010
MWA	BIR	CA26779	MALAWI AGENT 6	05.02.2022	88	12345692	IMPORTER 6	2022	1	RED	Green	IM4	87033311

Source: Automated System for Customs Data (ASYCUDA), United Nations Conference on Trade and Development.

**FIGURE B14.2.2 Example of an ASYCUDA Extract: Duty, Excise, and Value-Added Taxes Variables**

33	34	35	36	37	38	39
FOB FCY	CURRENCY	EXCRATE	VDP AMOUN	DUTY	EXCISE	VAT
1000	MWK	1	43718945	435345	0	468396
45000	MWK	1	12843921	435345	0	6849306
134144,85	USD	12,999	2401842	3452	483964	48963
8405	USD	12,999	3234398240	574575	45903	6439634
8405	MWK	1	840399234	8769769	65	84963
8405	GBP	13,888	4820384	769769	872	684396

Source: Automated System for Customs Data (ASYCUDA), United Nations Conference on Trade and Development.

The typical time stamp data are associated with a particular action—such as a change in lane selectivity or in payments due, among others. Linking all the tables, one can extract tailored reports, as in figure B14.2.3, to create indicators of time delays between different actions depending on lane selectivity or type of declaration.

(continues on next page)



## BOX 14.2 ASYCUDA Data Structure (continued)

FIGURE B14.2.3 Example of an ASYCUDA Extract: Time Stamps

	1	2	3	4	5	6	7	8	9	10	11	12	13
	OFFICE	REGDATE	REGNO	REGIME	Lane At Self	Current Lane	VEHICLE	REI VALUE OF DECLARATION	CONTAINER_NUMBER	DOCUMENT	OPERATION	OPERATION	TIME USERNAME
1	BR	02.01.2022	C678	IM4	RED	Green	1234	HIGH VALUE	GFR908432		1 Validate and assess	02.01.2022 16:04	user1_nickname
2	BR	02.01.2022	C679	IM4	RED	Green	1234	HIGH VALUE	GFR908433		2 Request PRN	03.01.2022 16:04	user1_nickname
3	BR	02.01.2022	C680	IM4	RED	Green	1234	HIGH VALUE	GFR908434		3 Payment	04.01.2022 16:04	user1_nickname
4	BR	02.01.2022	C681	IM4	RED	Green	1234	HIGH VALUE	GFR908435		4 Release Order (selectivity)	05.01.2022 16:04	user1_nickname
5	BR	02.01.2022	C682	IM4	RED	Green	1234	HIGH VALUE	GFR908436		5 Control Results	06.01.2022 16:04	user2_nickname
6	BR	02.01.2022	C683	IM4	RED	Green	1234	HIGH VALUE	GFR908437		6 Control Results	07.01.2022 16:04	user2_nickname
7	BR	02.01.2022	C684	IM4	RED	Green	1234	HIGH VALUE	GFR908438		7 Control Results	08.01.2022 16:04	user2_nickname
8	BR	02.01.2022	C685	IM4	RED	Green	1234	HIGH VALUE	GFR908439		8 Clear declaration	09.01.2022 16:04	user1_nickname
9	BR	02.01.2022	C686	IM4	RED	Green	1234	HIGH VALUE	GFR908440		9 System re-route to green	10.01.2022 16:04	user2_nickname
10	BR	02.01.2022	C687	IM4	RED	Green	1234	HIGH VALUE	GFR908441		10 Print Release Order	11.01.2022 16:04	user1_nickname
11	BIA	04.04.2022	C234567	IM4	BLUE	Blue	4321	LOW VALUE			1 Validate and assess	04.04.2022 17:18	user2_nickname
12	BIA	04.04.2022	C234568	IM4	BLUE	Blue	4321	LOW VALUE			2 Request PRN	05.04.2022 17:18	user1_nickname
13	BIA	04.04.2022	C234569	IM4	BLUE	Blue	4321	LOW VALUE			3 Add Scanned Docs	05.04.2022 17:45	user1_nickname
14	BIA	04.04.2022	C234570	IM4	BLUE	Blue	4321	LOW VALUE			4 Post-Entry	06.04.2022 17:00	user1_nickname

Source: Automated System for Customs Data (ASYCUDA), United Nations Conference on Trade and Development.

Another source of data is trader perception surveys. The focus of this type of survey is, as the name suggests, traders, importers, and exporters who directly engage in international trade. For example, traders might think that transport costs not related to border crossing are the most important costs faced when trading across borders, but these costs are unlikely to be shown in regular trade statistics. The burden of import or export certificates and clearance-associated costs is usually not represented either. The issue with these surveys is how to harmonize perception questions across countries to make sure they cover the same issues: what is experienced as a delay might be business as usual in another country, or traders might be reluctant to answer truthfully.

Finally, an emerging source of data is based on GPS trackers. This data source provides an objective time measure for border crossing and also captures the time spent on the road. These data can be used to observe the time spent at the border. Used in conjunction with time stamps, they show what share of time delays is attributable to customs operations as opposed to, for instance, difficulties linked to parking infrastructure. While these data are usually privately collected by firms providing transponders or insurers, some transport corridor authorities or public databases collect and provide these tracking data. One example of such a resource in Southern and Eastern Africa is administered by the World Bank's corridor team.<sup>2</sup>

## PERFORMANCE INDICATORS

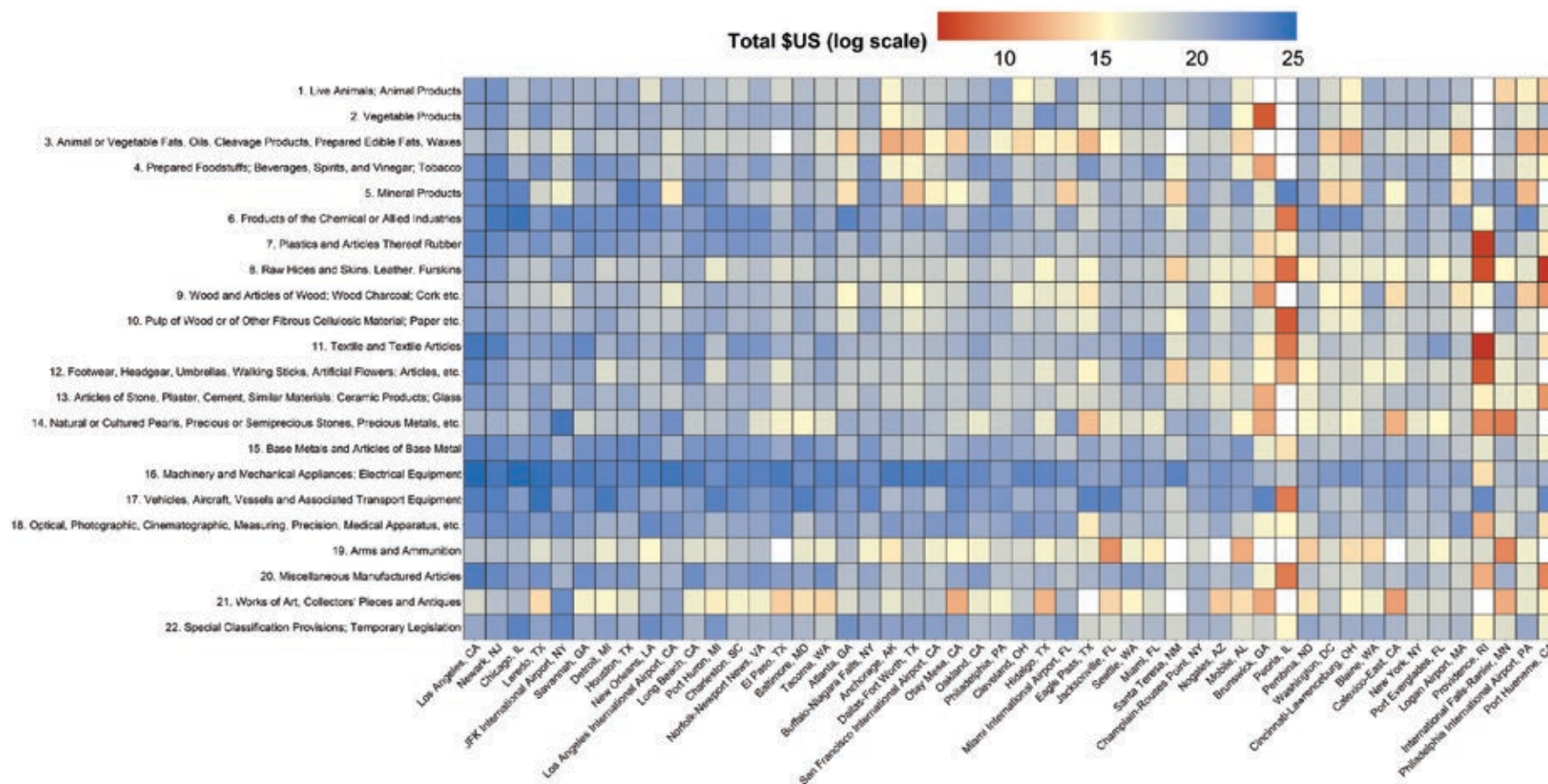
While previous sections have discussed the types of data sources that can be used to measure customs performance, this section describes how customs data can be developed into indicators to measure and further the three key objectives of the multidimensional mission of customs: trade facilitation, revenue collection, and food safety and security.

### Indicators for Trade Facilitation

Indicators related to trade facilitation usually focus on the time spent at the border and for clearance. This is part of the standard assessment of the WCO, the African Customs Union, and the TRS+ implemented by the World Bank. Of course, different border posts and different categories of goods will have different clearance times.



**FIGURE 14.2** Value of Different Product Categories Imported to the United States for 50 Largest Ports of Entry, as Appraised by US Customs and Border Protection



Source: USA Trade Online, US Census Bureau: Economic Indicators Division.

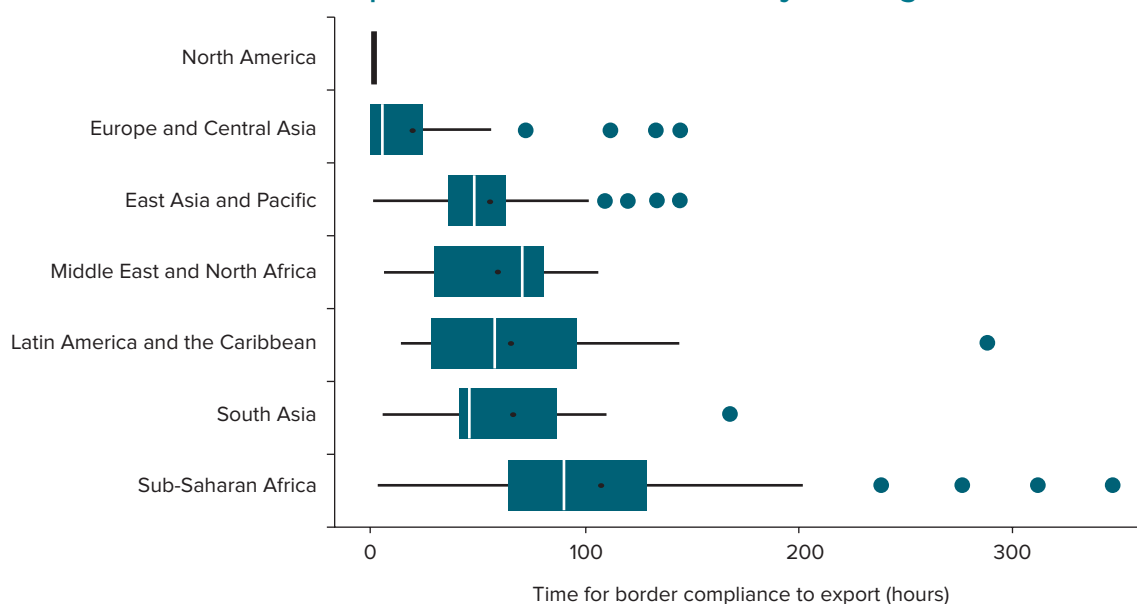
Note: Goods on the y axis are grouped according to the 22 sections of the Harmonized System. Some section labels are shortened due to space considerations. The x axis displays the 50 largest ports of entry in the United States by the total value of all goods imported, in decreasing order.

Figure 14.3 presents an example of a TRS indicator in the form of cross-country and regional disaggregation of border compliance times. Exporters across countries face vastly different times to process through customs. They are close to zero in the Northern American trade involving the United States and Canada, as well as in intra-EU trade. However, they increase more than threefold for Central Asian countries. On the other end of the spectrum, the largest delays are experienced by Sub-Saharan African exporters, where the mean border compliance time is 107 hours, and over 200 hours for several countries in Central Africa.

Not only are these processing times intrinsically heterogeneous, but the data used to measure them also paint a different picture of the customs process. The routinely collected time stamps from the customs database, the ASYCUDA or another, will show the date of the first submission and clearance. However, if the submission is made far in advance—for example, when arriving at the port, while the country itself is still far off—the time will be artificially long. In addition, as mentioned, if other agencies have to clear the goods while under customs custody, the time stamps will reflect a longer process. Indicators should take into account this heterogeneity in measurement approaches.

One possibility is, therefore, to look at the time necessary between the moment the frontline inspector is assigned to the declaration and the moment they clear it. While some agencies may delay the process by requesting additional inspection and clearances, this is less likely to be the case. In the ASYCUDA or other databases, this would correspond to the time difference between the time for assessment and the time at release. An example of such an indicator used for monitoring this time is depicted in figure 14.4. This figure displays the average time between the issuance of a release order and the issuance of a certificate of export at Malaba, on the Northern Corridor between Kenya and Uganda. The TRS follows a declaration at the border from when it is submitted to when the truck arrives and gives a snapshot of the border-crossing process at a moment in time, such that elements related to noncustoms delays can be isolated.

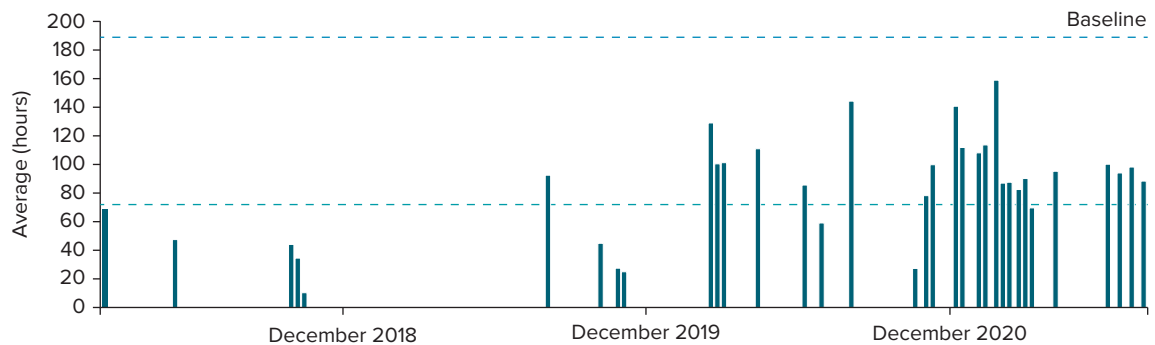
**FIGURE 14.3 Border Compliance Times in Cross-Country and Regional View**



Source: Doing Business database, World Bank.

Note: The component indicator is computed based on the methodology in the Doing Business 2016–20 studies. The boxes in the plot represent the interquartile range (IQR) of the variable—that is, the distance between the 25th and 75th percentiles in the distribution of respective values. The lines in the middle of the box represent the medians, whereas the dots represent means. The time is calculated in hours. The measure includes time for customs clearance and inspection procedures conducted by other agencies. If all customs clearance and other inspections take place at the port or border at the same time, the time estimate for border compliance takes this simultaneity into account.

**FIGURE 14.4** Example of Indicators to Measure Performance: Transit Time on the Northern Corridor



Source: Original figure for this publication.

Finally, the same indicators can be based on surveys of traders to recover their perception of the delays, using a question to estimate how many days it takes between the moment a shipment reaches the border point and when it can be cleared from the border post. In Malawi, a survey is being conducted in this way. Early results show a reported average of two to three days once traders get notified their shipment is at the border.

### Indicators for Revenue Collection

The revenue-collection objective focuses on how much revenue is collected at the border. This is intrinsically difficult to do—see box 14.3 on the problem of valuation—and, therefore, constructing the theoretical revenue that could have been collected requires considerable effort. Hence, this is something that the customs administration rarely does, unless misdeclaration or fraud is discovered. Otherwise, the declared value stays and the revenue collected is assumed to be the revenue that could have been collected by customs. However, not all misdeclaration or fraud is discovered. Hence, assuming that some of the incorrect declarations are missed, it is possible to look at the revenue that could have been collected if the items followed a similar price for other goods of the same class and origin. This is considered one of the acceptable valuation methods by the WTO. While the scholarly literature usually calls this *reference prices*, this clashes with the meaning of the reference prices used by the WTO: it is not an artificial set of prices but a comparison with similar goods' prices.

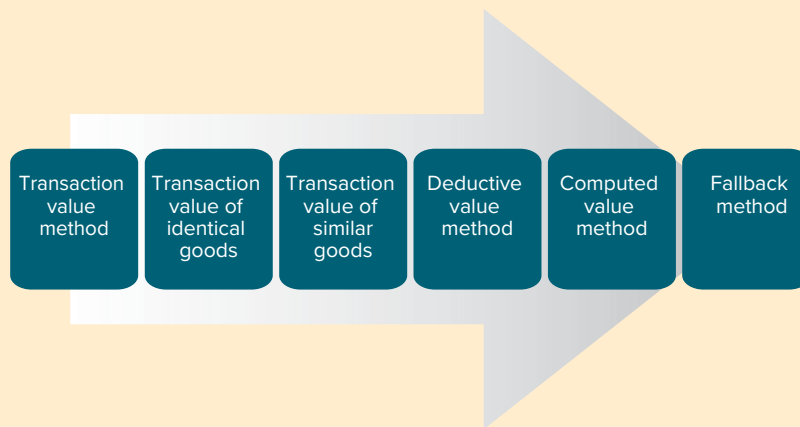
Evaluating the value of an item is intrinsically hard, as the inspector doesn't have precise information on the goods outside of what is listed on the declaration. The WTO agreement establishes rules for the valuation of imported goods that must be applied by all member countries. The WTO mandates using the transaction value supported by invoices and relevant documentation as the assessed value unless there is something missing or suspicion of fraud. In this case, the customs administration is authorized to use other valuation methods. The first method is using the transaction value of identical goods—same goods, same country of origin, same producer, whenever possible. The second method is using the transaction value of similar goods—same function or design, same country of origin, and whenever possible, same producer. Additional methods are outlined in figure B14.3.1. Customs is prohibited from using the same goods produced nationally as a comparison point, and from using arbitrary or fictitious values, such as minimal values or thresholds.

To refine this analysis, it is possible to use it in conjunction with the mirror gap: given the quantities of similar goods declared by the exporting country, how many are missing from the importing country import declarations and vice versa? The quantities declared for import and export in the origin country should be the same. This can give a rough idea of what revenues should be collected—or are missing—on either end. For an example of the use of such data for customs reform, see box 14.4. However, as mentioned earlier, these trade data sets are not updated as frequently as the customs data themselves. Hence, some of these gaps might be an artifact of the data. Another possibility is to reconcile the data at the

### BOX 14.3 The Problem of Valuation

Evaluating the value of an item is intrinsically hard, as the inspector does not have precise information on the goods outside of what is listed on the declaration. The World Trade Organization (WTO) agreement establishes rules for the valuation of imported goods that must be applied by all member countries. The WTO mandates using the transaction value supported by invoices and relevant documentation as the assessed value unless there is something missing or suspicion of fraud. In this case, the customs administration is authorized to use other valuation methods. The first method is using the transaction value of identical goods—same goods, same country of origin, same producer, whenever possible. The second method is using the transaction value of similar goods—same function or design, same country of origin, and whenever possible, same producer. Additional methods are outlined in figure B14.3.1. Customs is prohibited from using the same goods produced nationally as a comparison point, and from using arbitrary or fictitious values, such as minimal values or thresholds.

**FIGURE B14.3.1** World Trade Organization Valuation Methods, Arranged Sequentially



Source: Based on WTO Agreement on Customs Valuation (ACV) of 1994.

individual level, linking exporters' declarations from a country to another country's importers' declarations. This level of analysis can highlight value discrepancies and the potential mistakes or omissions of customs frontline agents.

### Indicators for Food Safety and Security

There is relatively less work on safety because the data are harder to come by. The seized goods could indicate either an increase in customs activity or criminal activity. The TRS and ASYCUDA data can provide a good indication of risk management operations, both in terms of value recovery and physical inspection for safety. In Brazil, for example, the rate of physical inspections performed by customs was found to be around 2 percent during the most recent TRS (Receita Federal do Brasil 2020). However, 12 other government agencies were often involved in the process, granting licenses or permissions necessary for import. Around 60 percent of the declarations required involvement by another agency, whether or not the process required a physical inspection. The delays noted in the TRS process for Brazil thus reflect the need for other agencies' licenses and inspections. Another example is that, for goods under the jurisdiction of health authorities, around one-quarter to a third of the time is actually due to delays in paying the licensing fee.

## BOX 14.4 Information and Customs Performance: The Case of Madagascar

The Republic of Madagascar is an island country lying off the southeastern coast of Africa (see map B14.4.1). In an experiment conducted in Madagascar, Chalendar et al. (2020) measure customs indicators and how they change when customs agents are given additional information. Madagascar is among the countries that rely heavily on customs and other import duties—16.9 percent of the total tax revenue going to Antananarivo proceeds from this source. At the same time, the performance of particular customs inspectors in Madagascar can be highly impactful because each inspector is responsible for a considerable value of import revenues. In the sample of Chalendar et al. (2020), every inspector handles around US\$10 million in import revenues per year. Therefore, ensuring the good performance of its customs officials is a vital interest of the Malagasy authorities.

**MAP B14.4.1** Location of Madagascar



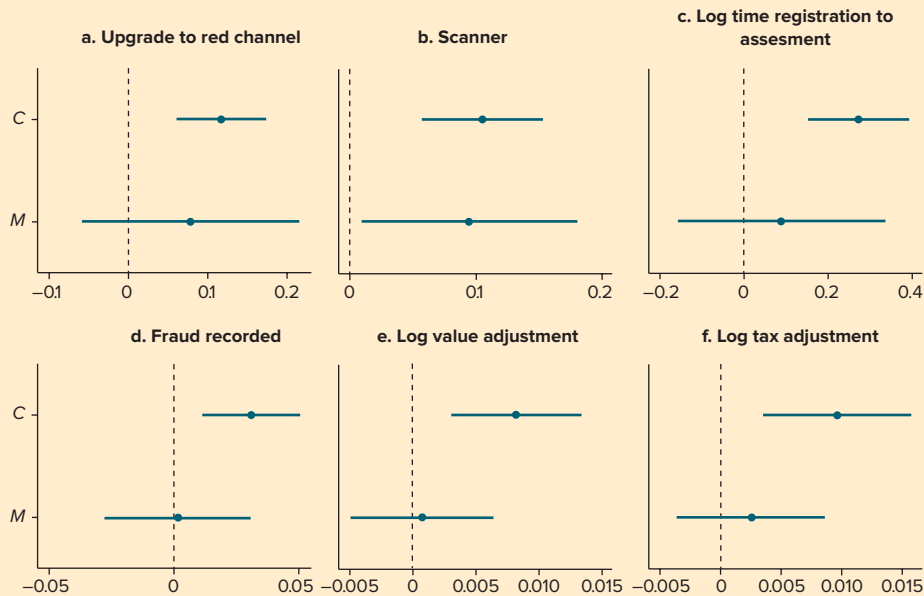
Source: World Bank.

*(continues on next page)*

## BOX 14.4 Information and Customs Performance: The Case of Madagascar (continued)

Chalendard et al. (2020) investigate the role of information provision and monitoring in a randomized setting. One group of officials in their study was provided with a set of detailed risk-analysis comments on high-risk customs declarations (this group is labeled with C in figure B14.4.1). Officials in another group were told they would be more intensively monitored throughout a period of study (this group is labeled with M in the figure). Figure B14.4.1 shows that monitoring has an impact only on the increased frequency of customs officials' scanning containerized goods. In contrast, additional comments about high-risk declarations also lead the officials to more frequently upgrade inspections to the red channel and declare more cases of fraud detection and larger value adjustment. However, this also increases screening times and leads to only small improvements in tax collection, especially for declarations supposed to yield large tax revenues.

**FIGURE B14.4.1** Changes in Malagasy Customs Officials' Performance



Source: Original figure for this publication.

Note: The label C on the y axis indicates the group of inspectors who were provided with comments; the label M indicates the group of inspectors who were told they would be monitored. X-axis units are regression coefficients.

## CONCLUSION

What lessons for practice should be considered by the practitioner interested in exploring customs data for analytics?

First, an initial diagnosis through the TRS can provide a broad overview of the customs process. This can be done either at the beginning of a project or by using baseline data from past exercises. The TRS can provide useful indicators on what part of the clearance process suffers from a bottleneck. This is commonly done by the revenue administration before an overhaul of its process. This can be extended with a trader survey, which asks traders about the most sensitive aspects of the process, which are unlikely to be captured



during the TRS. For example, the issues of speed money—or bribes to speed up the process—or other issues with any of the agencies involved might not be seen by the TRS surveyors but nevertheless influence traders' decisions to import or export.

Second, protocols to ensure data confidentiality while providing external access should be set in place. The anonymity of taxpayers is an important governmental concern, and some administrations are prevented from sharing taxpayer information with third parties. If protocols are set in place, these data can be shared while respecting these anonymity concerns, allowing practitioners and outside researchers to build customs performance indicators and opening the door to further research. These protocols include the deidentification of data whenever possible, such that researchers have access to deidentified tables only. This can be done via the hashing of the tables. Beyond security concerns, ASYCUDA tables might need to be merged and extracted, which can prove challenging in low-capacity settings. A useful solution is to support client engagement by requesting the data needed to build the basic indicators and assemble the data on a safe server. If necessary, the data can be deidentified by the client team based on a hashing code provided by the researcher, a procedure described on the World Bank's Development Impact Evaluation (DIME) Wiki.<sup>3</sup>

Third, stakeholders may resist additional measurement efforts. Some stakeholders may be reticent to use anything other than the TRS, as it is new and requires more effort from the ASYCUDA team. On top of that, while the TRS provides a narrative of the sources of delays, ASYCUDA data offer an often harsher view of the clearance process, as they also include steps that depend on the taxpayer—such as paying taxes. Because the ASYCUDA aggregates so much data, it can incorporate more outliers and influence the mean. This contrasts with the TRS, which is often done in a week, with the inspectors being aware of it. Researchers should thus expect discrepancies with the reported TRS, especially if the survey was done a while back. Thus, triangulating the different sources of data is important, as well as using the TRS results to comment on ASYCUDA-based indicators.

Finally, we suggest first investing in easy-to-produce indicators, such as revenue recovered and revenue recovered compared to similar products of the same type, as well as the easiest types of delays. These indicators should be triangulated with the TRS, if available, or with trader surveys. Further refinement of the indicators could include more precise measures of delays to distinguish tax compliance and the actions of customs, but these should be done once the more foundational indicators are measured and set in place. Of course, as outlined in the introduction, when measuring only select indicators of an organization with a multidimensional mission—such as customs—analysts need to remain cognizant of risks of effort substitution toward measurable indicators and to devise strategies to expand measurement to all core objectives of customs over time.

## NOTES

The author would like to thank Iana Miachenkova for excellent research assistance and acknowledges the support of the Umbrella Trade Trust Fund for the Malawi Trade Facilitation Impact Evaluation.

1. An example of an eight-digit description is 08051000, which corresponds to fresh oranges. Each product belongs, at the broadest level, to one of 22 Harmonized System sections. These are, however, not marked in the product code. Instead, each section is composed of one or more chapters, and the first two digits of the code refer to a specific chapter: in this case, chapter 08: "Edible Fruit and Nuts; Peel of Citrus Fruit or Melons." The next two digits stand for a heading within that chapter: heading 05: "Citrus fruit, fresh or dried." The following two digits stand for subheading 10: "Guavas, mangoes and mangosteens: Oranges." The last two digits can further specify more fine-grain divisions of product category if these exist. In this case, no further specification is indicated by 00.
2. Their website is accessible at <https://www.corridorperformancemonitoringsystem.com/geozone-route-catalogue>.
3. See DIME Wiki, s.v. "De-identification," last modified November 17, 2020, 20:10, <https://dimewiki.worldbank.org/De-identification>.

## REFERENCES

- Bardhan, Pranab. 2006. "The Economist's Approach to the Problem of Corruption." *World Development* 34 (2): 341–48. <https://doi.org/10.1016/j.worlddev.2005.03.011>.
- Baunsgaard, Thomas, and Michael Keen. 2010. "Tax Revenue and (or?) Trade Liberalization." *Journal of Public Economics* 94 (9–10): 563–77. <https://doi.org/10.1016/j.jpubeco.2009.11.007>.
- Buzby, Jean C., ed. 2003. *International Trade and Food Safety: Economic Theory and Case Studies*. Agriculture Economic Report 828. Washington, DC: Economic Research Service, US Department of Agriculture. <https://www.ers.usda.gov/publications/pub-details/?pubid=41618>.
- Cantens, Thomas, Gaël Raballand, and Samson Bilangna. 2019. "Reforming Customs by Measuring Performance: A Cameroon Case Study." *World Customs Journal* 4 (2): 55–74.
- Chalendard, Cyril, Alice Duhaut, Ana M. Fernandes, Aaditya Mattoo, Gaël Raballand, and Bob Rijkers. 2020. "Does Better Information Curb Customs Fraud?" CESifo Working Paper 8371, Munich Society for the Promotion of Economic Research, Munich. <https://doi.org/10.2139/ssrn.3633656>.
- Chalendard, Cyril, Ana M. Fernandes, Gaël Raballand, and Bob Rijkers. 2021. "Corruption in Customs." CESifo Working Paper 9489, Munich Society for the Promotion of Economic Research, Munich. <https://doi.org/10.2139/ssrn.3998027>.
- Chalendard, Cyril, Gaël Raballand, and Antsa Rakotoarisoa. 2019. "The Use of Detailed Statistical Data in Customs Reforms: The Case of Madagascar." *Development Policy Review* 37 (4): 546–63. <https://doi.org/10.1111/dpr.12352>.
- Diaz-Bonilla, Eugenio, Marcelle Thomas, Sherman Robinson, and Andrea Cattaneo. 2000. "Food Security and Trade Negotiations in the World Trade Organization: A Cluster Analysis of Country Groups." TMD Discussion Paper 59, Trade and Macroeconomics Division, International Food Policy Research Institute, Washington, DC. <https://www.ifpri.org/publication/food-security-and-trade-negotiations-world-trade-organization>.
- Djankov, Simeon, Caroline Freund, and Cong S. Pham. 2010. "Trading on Time." *The Review of Economics and Statistics* 92 (1): 166–73. <https://doi.org/10.1162/rest.2009.11498>.
- Dutt, Pushan, and Daniel Traca. 2010. "Corruption and Bilateral Trade Flows: Extortion or Evasion?" *The Review of Economics and Statistics* 92 (4): 843–60. [https://doi.org/10.1162/REST\\_a\\_00034](https://doi.org/10.1162/REST_a_00034).
- Ercsey-Ravasz, Mária, Zoltán Toroczkai, Zoltán Lakner, and József Baranyi. 2012. "Complexity of the International Agro-Food Trade Network and Its Impact on Food Safety." *PLoS One* 7 (5): e37810. <https://doi.org/10.1371/journal.pone.0037810>.
- Farzanegan, Mohammad Reza. 2009. "Illegal Trade in the Iranian Economy: Evidence from a Structural Model." *European Journal of Political Economy* 25 (4): 489–507. <https://doi.org/10.1016/j.ejpoleco.2009.02.008>.
- Fernandes, Ana Margarida, Russell Hillberry, and Alejandra Mendoza Alcántara. 2021. "Trade Effects of Customs Reform: Evidence from Albania." *The World Bank Economic Review* 35 (1): 34–57. <https://doi.org/10.1093/wber/lhz017>.
- Khan, Adnan Q., Asim Ijaz Khwaja, and Benjamin A. Olken. 2019. "Making Moves Matter: Experimental Evidence on Incentivizing Bureaucrats through Performance-Based Postings." *American Economic Review* 109 (1): 237–70. <https://doi.org/10.1257/aer.20180277>.
- Laajaj, Rachid, Marcela Eslava, and Tidiane Kinda. 2019. "The Costs of Bureaucracy and Corruption at Customs: Evidence from the Computerization of Imports in Colombia." Documento CEDE 2019-08, Centro de Estudios sobre Desarrollo Económico, Facultad de Economía, Universidad de los Andes, Bogotá.
- Malawi Revenue Authority. 2019. *Malawi Time Release Study Report 2019*. Lilongwe: Malawi Revenue Authority. <https://www.mra.mw/assets/upload/downloads/MalawiTimeReleaseStudyReport2019FN.pdf>.
- Martincus, Christian Volpe, Jerónimo Carballo, and Alejandro Graziano. 2015. "Customs." *Journal of International Economics* 96 (1): 119–37. <https://doi.org/10.1016/j.jinteco.2015.01.011>.
- Receita Federal do Brasil. 2020. *Time Release Study: June 2020*. Brasília: Receita Federal do Brasil. <https://www.gov.br/receitafederal/pt-br/acao-a-informacao/dados-abertos/resultados/estatisticascomercioexterior/estudos-e-analises/TRSReport.pdf>.
- Świerczyńska, Jolanta. 2016. "The Reduction of Barriers in Customs as One of the Measures Taken by the Customs Service in the Process of Ensuring Security and Safety of Trade." *Studia Ekonomiczne* 266: 212–22.
- WCO (World Customs Organization). 2015. *Tools for Reducing Revenue Risks and the Revenue Gap: (I) Mirror Analysis Guide, Including Case Study (Cameroon)*. Brussels: World Customs Organization. <https://www.wcoesarocb.org/wp-content/uploads/2017/03/11-Mirror-analysis-guide-FINAL-EN.pdf>.
- WCO (World Customs Organization). 2018. *Guide to Measure the Time Required for the Release of Goods*. Version 3. Brussels: World Customs Organization. <https://www.wcoomd.org/-/media/wco/public/global/pdf/topics/facilitation/instruments-and-tools/tools/time-release-study/trs-guideen.pdf?db=web>.
- Widdowson, David. 2007. "The Changing Role of Customs: Evolution or Revolution." *World Customs Journal* 1 (1): 31–37.



## CHAPTER 15

# Government Analytics Using Administrative Case Data

*Michael Carlos Best, Alessandra Fenizia, and Adnan Qadir Khan*

### SUMMARY

Measuring the performance of government agencies is notoriously hard due to a lack of comparable data. At the same time, governments around the world generate an immense amount of data that detail their day-to-day operations. In this chapter, we focus on three functions of government that represent the bulk of its operations and that are fairly standardized: social security programs, public procurement, and tax collection. We discuss how public sector organizations can use existing administrative case data and repurpose them to construct objective measures of performance. We argue that it is paramount to compare cases that are homogeneous or to construct a metric that captures the complexity of a case. We also argue that metrics of government performance should capture both the volume of services provided as well as their quality. With these considerations in mind, case data can be the core of a diagnostic system with the potential to transform the speed and quality of public service delivery.

### ANALYTICS IN PRACTICE

- Governments generate immense amounts of data that detail their day-to-day operations. These data can be repurposed to measure the performance of government agencies. Such data can provide objective comparisons of agency performance, allowing for an assessment of the quality of public administration across jurisdictions, regions, managers, and time.
- Such operational data provide objective records of bureaucratic performance. It is important to construct objective measures of organizational performance and individual performance rather than relying only on subjective evaluations such as performance appraisals.

---

Michael Carlos Best is an assistant professor in the Department of Economics, Columbia University. Alessandra Fenizia is an assistant professor in the Department of Economics, George Washington University. Adnan Qadir Khan is a professor at the School of Public Policy, London School of Economics.

- A prerequisite for constructing a comprehensive measure of performance for a public organization is obtaining a record of all the tasks undertaken by the organization. This may be difficult in practice because government agencies undertake a wide range of tasks, and they may not keep detailed records for all of them.
- One area of government activity where records are objective measures of performance and often relatively comprehensive is case management. Case management data are the records of responses by public officials to requests for public services or the fulfillment of public responsibilities. This chapter argues for the use of administrative data on the processing of cases by public officials as a monitoring tool for government performance and as a core input for government analytics. Relevant measures should capture both the *volume* and *quality* of cases processed.
- To construct an objective measure of performance using case data, one should ensure that cases are comparable to one another. This could entail comparing cases only within a homogeneous category or constructing a metric that captures the complexity of a case. For example, a social security claim that clearly meets the requirements of regulations and does not reference other data systems is a less complicated case to process than one in which there are ambiguities in eligibility and external validation is required. A corresponding metric of complexity might be based on the time spent on an “average” case of that type, allowing for complexity to be defined by the actual performance of public officials.

## INTRODUCTION

In order to implement government policy, the apparatus of the state generates a vast trove of administrative databases tracking the deliberations, actions, and decisions of public officials in the execution of their duties. These data are collected in order to coordinate throughout a large, complex organization delivering a host of services to citizens and to preserve records of how decisions are reached to provide accountability for decisions made in the name of the public.

These data are not, typically, collected for the express purpose of measuring the performance of government officials. But as governments become more and more digitalized, these records contain ever-richer details on the work that is carried out throughout government. This presents an opportunity to repurpose existing data, and possibly extend its reach, to achieve the goal of measuring performance. In turn, such data can then be used to motivate government officials and hold them accountable. Ultimately, a greater ability to *measure* performance can help governments to *monitor* performance. This can improve efficiency in the public sector to deliver more and better services to citizens with the human and material resources the government has available.

Using administrative data has the distinct advantage that the data are already being collected for other purposes. As such, the additional costs of using them to measure performance are largely technical issues surrounding granting access to the data, protecting their confidentiality appropriately, and setting up the information technology (IT) infrastructure to perform statistical analysis on them. These obstacles are typically much simpler to overcome than the obstacles to launching new surveys of public officials or citizens to measure performance.

Set against this advantage, the primary disadvantage of using administrative data to measure performance is that they were not designed to be used for that purpose. As a result, a great deal of careful thought and work must go into how to repurpose the data for performance measurement. This involves thinking carefully about what outputs are being produced, how to measure their quantity and quality, and how to operationalize them within the constraints of the available data. Sometimes, this requires collecting additional data (either through a survey or from external sources) and linking them to the administrative data.

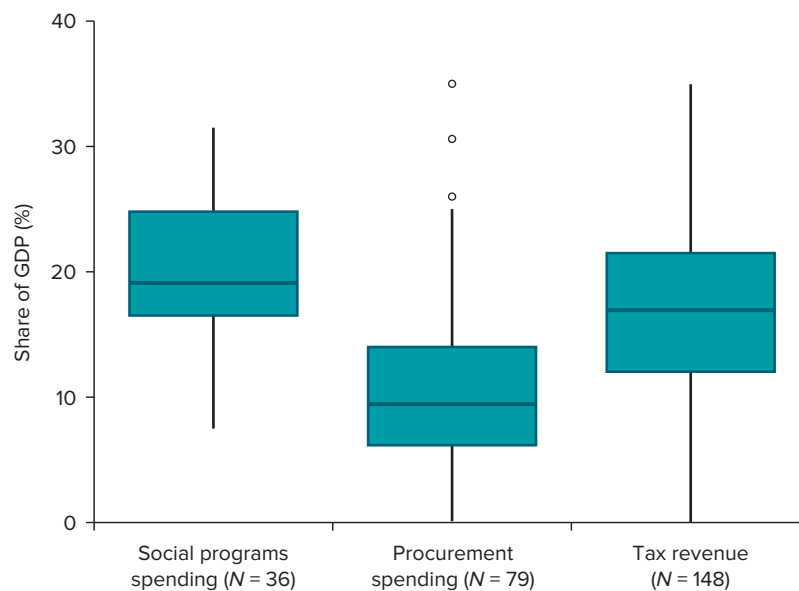
A large share of government operations involve the processing of case files or cases. Case data are the records of responses by public officials to requests for public services or the fulfillment of public responsibilities. A case file is typically a collection of records regarding an application. The nature of the applications varies widely. For one, thousands of claimants file applications every day to receive government services, such as welfare transfers, to gain access to government-sponsored childcare, or to obtain licenses and permits. Public sector organizations around the world initiate auctions to purchase goods and services from private sector suppliers. And millions of citizens and firms all over the globe file taxes every year.

In this chapter, we highlight examples from recent academic work trying to develop new methods to measure performance using administrative data on the processing of government casework. The academic papers provide a window into how similar data from public administrations around the world can be repurposed for analytical purposes.

Our examples cover three important realms of government operations—the delivery of social programs, the collection of taxes, and the procurement of material inputs—that together span a large part of what modern governments do. Figure 15.1 shows that spending on social programs and procurement and tax revenues jointly amount to more than 30 percent of a country's gross domestic product (GDP) on average. While there is some variation in the size of social programming, procurement spending, and tax revenues, these three functions of government represent a large share of government operations in all countries.

Since all governments engage in these activities, exploring potential alternative uses of the data generated in the process is of broad interest. In addition, operations in these areas are usually fairly standardized, tending to boost the quality of related data, which in turn can be used to generate more accurate insights. In all three cases, we highlight the importance of carefully specifying the outputs that are to be measured before undertaking an analysis, as well as how to conceptualize data quality.

**FIGURE 15.1 Cross-Country Scale of the Three Sectors Discussed in the Chapter Relative to National Gross Domestic Product**



Source: Original figure based on data from the Organisation for Economic Co-operation and Development (OECD) (social programs spending), the World Bank Development Indicators (tax revenue), and the World Bank Global Public Procurement Database (procurement spending).

Note: The box represents the interquartile range (IQR)—the distance between the 25th and 75th percentiles in the distribution of each variable. The line in the middle of the box represents the median. Whiskers—that is, the lines extending from the box—represent values lying within 1.5 of the IQR from the median. Outliers lying beyond that range are represented by dots, where one dot represents a country. The value of  $N$  shows the number of country-level observations in each column. GDP = gross domestic product.



We also provide some details on the technical methods used to operationalize these concepts and turn them into concrete performance measures and on how these performance measures are then used in the academic arena. In the conclusion, we discuss how policy makers can use these types of measures in other ways, as well as some important limitations to these approaches. The intention of our exposition of these cases is not to argue that the approach taken in the specific papers we review is optimal for every setting but rather to showcase a way to approach the analysis of government administrative case data.

## CASE DATA IN ADMINISTRATION

### A General Structure for the Analysis of Case Data

Government casework involves a series of standardized elements, each of which can be associated with a measure of the performance of public administration. Casework typically revolves around a set of protocols—perhaps standardized forms that applicants must fill in to apply for social security payments—that make common measures feasible. Cases are processed by government officials, again, frequently in a relatively standardized way.<sup>1</sup> For this reason, measures of performance can be used to judge how efficiently and effectively public officials worked through the relevant protocols. Case data are therefore made up of the records of cases and their processing, including details of the application or case and characteristics that can be analyzed. For example, in electronic case management systems, time and date stamps record exactly when cases were submitted, acted upon by officials, and then resolved. The speed of multiple stages of case processing can thus be easily calculated. Similarly, a decision is often made on a case and a response is sent to the applicant, such as a confirmation to a taxpayer that they have paid their taxes.

To use data on the processing of such cases to monitor and analyze government capabilities, we have to overcome two main challenges. Claims are diverse in how challenging or “complex” the associated case is. A case that involves a claim where a claimant clearly meets the required criteria is less complex than one in which eligibility is ambiguous on one or more margins. In some cases, evaluating the claimant’s eligibility may be fairly straightforward, involving verification of the veracity of a few supporting documents provided by the applicant. In other cases, it may require the officer to request access to a separate archive to pull the claimant’s records.

Thus, we first have to construct a common measure of task complexity that allows us to compare claims of different types. Second, we must ensure that any such measure is not easy to manipulate by government staff and is as objective as possible. For example, to minimize the risk of manipulation of these types of metrics, the tracking of claims should be done by a centralized computer system. Allowing employees to self-report their output and log it onto a computer may leave room for opportunistic behavior aimed at artificially inflating the measure of output. Employees may report processing a higher volume of claims or more complex claims than they actually did. One way around this is to complement electronic records with field observations of a representative sample of tasks at hand that is regularly updated. This approach minimizes the risk that the performance measures become outdated or disentangled from the constantly evolving work environment of public officials.

With these pieces in place, case data can be a source of government analytics. These data can provide objective comparisons of agency performance, allowing for an assessment of the quality of public administration across jurisdictions, regions, managers, and time. Rather than comparing simple output across offices, it is often useful to compare a measure of output per worker (or per unit of time). These measures capture the productivity of the average worker (or the average hour) in each office and are not affected by differences in office size. For instance, larger offices typically process a larger quantity of various cases by virtue of having more workers devoted to back-office operations. However, the fact that larger offices process more cases does not necessarily imply that they are more productive.

A major limitation of evaluating the performance of public sector offices based solely on output or productivity is that these measures reflect production volume and do not capture the quality of the service provided. For example, imagine an official who rubber-stamped applications for a claim. Looking only at

production volumes, the official would seem very productive. However, the officer has de facto awarded welfare transfers to all claimants regardless of their eligibility status. Conditioning on, or including in analysis, a measure of complexity would not adjust for the official's quality of service. Rather, a separate metric related to the quality of decision-making must be constructed to address this concern.

## Extending Analytics Insights

Government agencies can significantly increase the impact of existing administrative data by going beyond a basic analysis of the administrative data they hold. First, they can build assessments of the accuracy of their case data. For example, governments can collect additional data on the accuracy of tax assessment, say, from randomly selected tax units, which will enable them to construct more comprehensive performance measures of tax staff and establish more credible audit and citizen grievance redress mechanisms.

Second, the digitization of case data allows for the use of machine-learning and artificial intelligence algorithms to create better valuation measures, such as to detect clerical and other types of error, flag suspected fraud cases, or classify taxpayer groups in a (more) automated fashion. Further discussion of this topic is provided in chapter 16 of *The Government Analytics Handbook*, and a case study of a similar system is provided in case study 9.2 in chapter 9.

Authorities can also make anonymized case data publicly available, and this increased transparency can enable whistleblowing and peer pressure mechanisms. As one of the following case studies shows, there is precedent for doing this in Pakistan, where the entire tax directory for federal taxes has been published annually for the past decade.

Finally, case data can be integrated with political data to create better measures of politicians' performance at the local government level and thus enhance political accountability. For example, updates to cadastre records, which are crucial for accurate property valuations for tax purposes, were found to be crucially linked to electoral pressures on local officials in Brazil (Christensen and Garfias 2021).

The rest of this paper presents case studies that highlight the analysis and use of case data, focusing on measuring case volume, complexity, and quality, as well as describing ways to strengthen this analysis by linking to other data sources.

## SOCIAL SECURITY CLAIMS DATA

Social security claims data include records relating to old-age programs and social welfare programs, such as unemployment benefits, maternity leave, and subsidies to the poor. Most governments around the world already regularly collect claims data in an electronic format. For this reason, these data can be repurposed to perform quantitative analysis to better understand the performance of the social security system overall, the challenges facing individual public sector offices, and design solutions to address them.

In this section, we discuss a recent academic paper that uses detailed claims data from the Italian Social Security Agency (ISSA) to construct a measure of the performance of public offices and evaluate the effectiveness of ISSA managers. Fenizia (2022) exploits the rotation of managers across sites to estimate the productivity of public sector managers. This study finds significant heterogeneity in the effectiveness of these managers: some managers are very productive and improve the performance of the offices where they work, while others do not. The increase in office productivity brought about by talented managers is mainly driven by changes in personnel practices.

A case in this setting is the process of assessment by a social security officer of the validity of a claim for social security payments to an individual. A key advantage to studying the ISSA is that the tasks employees perform are fairly standardized, and the agency keeps detailed records of all applications and welfare transfers. This allows Fenizia (2022) to construct a comprehensive measure of performance that encompasses all the activities employees perform.

The obvious volume-based measure of productivity in this context is the number of social security claims of a particular type processed by an office in a particular time period divided by the full-time equivalent of workers of that office during that time. Map 15.1 describes how this measure varies across Italian regions, showcasing how such data can be used in government analytics. The figure indicates which regions are more productive than others and thus where investments might be needed in the quality of management or staff.

The first concern with analyzing this sort of data is that some cases may be more complex to process than others. In many settings, it is possible to measure only the output stemming from a subset of activities rather than the associated complexity. In these settings, the measure of performance only reflects the activities being measured and may be harder to interpret. For example, imagine that an agency performs two types of tasks: task A is observable, but task B is not. The measure of performance will reflect only the output from task A. If this measure were to decline over time, this could be driven by a worsening of performance in the agency overall or by the fact that resources had been reallocated from task A to task B. The following section discusses how to construct a measure of complexity using the time spent on an “average” case of a particular type.

The second concern is that production volumes do not reflect the quality of the service provided. After the discussion of complexity, the following section evaluates the strengths and weaknesses of two proxies of quality of service that can be derived from claims data.

## Complexity

Virtually all government agencies that administer old-age and welfare programs process a variety of different claims. While it is relatively straightforward to keep track of the number of incoming and processed claims, it is more challenging to construct a measure of performance for public offices that can be meaningfully compared across sites.

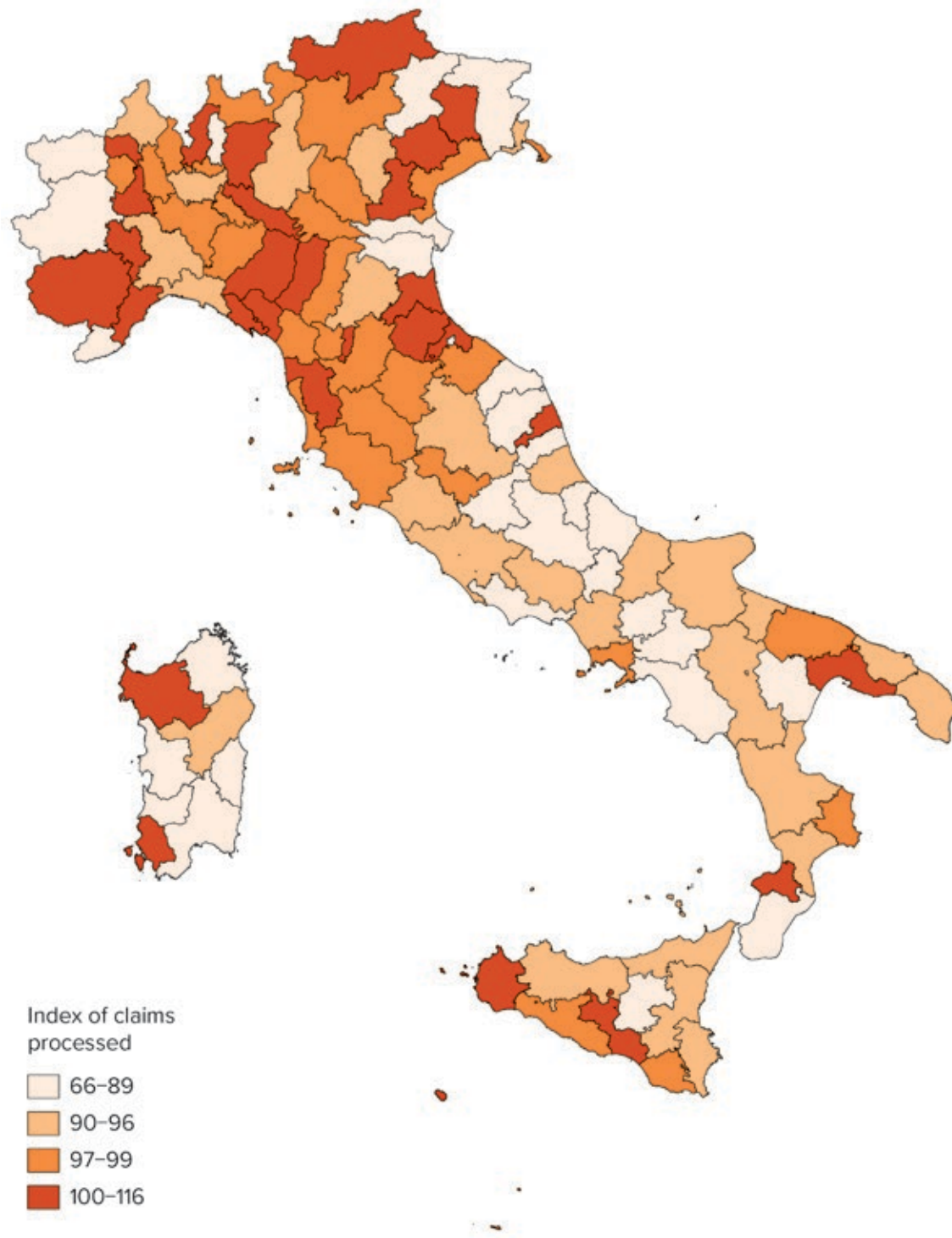
A naive solution might involve counting the number of claims processed by each office. Despite being simple and transparent, this measure suffers from a major draw-back: it does not take into account task complexity. Some claims might be very quick to process, while others might require a lot of time and resources. As mentioned above, in some cases, officers have simply to verify that the documentation provided by the applicant is complete and up-to-date. In other cases, officers may have to acquire further documentation from their internal archives or from other entities. If different offices process a different mix of paperwork, simply counting the number of claims processed would not correctly reflect differences in task complexity across sites. The naive metric would overstate the performance of offices that process simpler claims relative to those that process more sophisticated paperwork.

A solution is to use a complexity-adjusted measure of claims processed. For example, the ISSA constructs a measure of output for public offices that combines the number of claims processed by each site with a measure of their complexity. Specifically, the ISSA has grouped all claim types into more than 1,000 fine categories. Each category is constructed to group highly comparable claims that are equally complex. Each category is assigned a weight representing how much time it should take to process that specific claim type.

Figure 15.2 illustrates the distribution of expected processing time (that is, weights) for the most common types of pensions and welfare transfers. The expected processing time for most pensions ranges between 31 and 38 minutes, with a median of 30 minutes. The expected processing time is more variable for welfare transfers, reflecting the fact that these products are much more heterogeneous. Most of these claims take between 17 and 41 minutes to process, with a median processing time of 28 minutes.

Importantly, the ISSA complexity-adjustment formula uses objective weights as opposed to subjective scores. As part of the ISSA quality control department, there is a team devoted to measuring weights and keeping them up-to-date. To construct the weight for product  $v$ , this team selects an excellent, an average, and a mediocre office and picks a representative sample of product  $v$  claims from each office. Then the team visits each site and records the amount of time each employee took to process each claim. The weight is constructed by averaging all measurements across employees and offices, and it represents the time spent processing an “average” case of that type. The same weights apply to all offices at a given time to ensure that all offices are evaluated using the same standards. Weights can change in response to a technological

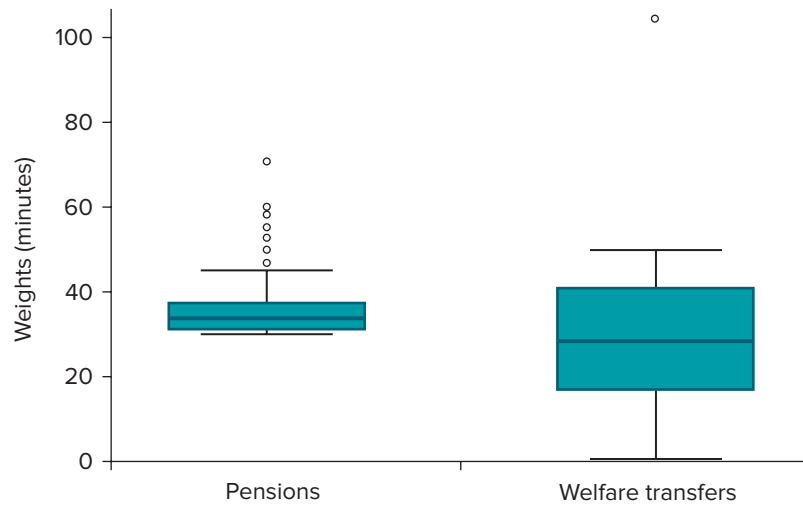
**MAP 15.1** Variations in Productivity of Processing Social Security Cases, Subregions of Italy



Source: Fenizia 2022, using Italian Social Security Agency data.

Note: The key refers to the number of social security claims of a particular type that are processed by an office in a particular time period divided by the full-time equivalent of workers of that office during that time.

**FIGURE 15.2** Expected Processing Time for Most Common Types of Pensions and Welfare Transfers, Italy



Source: Fenizia 2022, based on Italian Social Security Agency data.

Note: This figure illustrates the distribution of the expected processing time (that is, weights) for the most common types of pensions and welfare transfers. The box represents the interquartile range (IQR)—the distance between the 25th and 75th percentiles in the distribution of the weights. The line in the middle of the box represents the median. Whiskers represent values lying within 1.5 of the IQR from the median. Outliers are represented by circles.

improvement, if the time required to process a specific claim shortens, or when the paperwork associated with a claim changes.

The ISSA also ensures that the weights are measured accurately and that there are no opportunities for arbitrage. For example, if processing product  $b$  takes, on average, 10 minutes, and the weight associated with it is equal to 20 minutes, officers have an incentive to process as many  $b$  claims as possible. By doing so, they artificially increase the output of the office. Similarly, if product  $b$  is assigned a weight of 5 minutes when it takes 10 minutes on average to process, officers may be inclined to give priority to other claim types. To minimize arbitrage, the ISSA tracks backlog by product. If the backlog for a given product increases (decreases) across several offices, this may be an indication that the weight associated with it is too low (high). Therefore, the ISSA reevaluates the weights associated with the products that experienced large changes in backlog.

The weights are used to aggregate the number of claims of different types processed by each office  $i$  into a single output measure. The aggregation consists in multiplying the number of product  $v$  claims processed ( $c_{vi}$ ) with their corresponding weight ( $w_v$ ) and then summing across categories as follows:

$$\text{output}_i = \sum_{v=1}^V c_{vi} \times w_v \quad (15.1)$$

This output metric reflects the *theoretical* amount of time that it *should* have taken to process the claims that were effectively processed.

Although the procedure described above is largely specific to the ISSA and its mandate related to social security, similar measures are used in manufacturing firms across the world. These measures are especially popular in the garment sector, where the standard minute value (SMV) has become the standard.

## Quality

In the case of social security claims, a straightforward measure of the quality of service provided is the error rate (that is, the fraction of claims that were processed incorrectly). There are two types of mistakes: a government agency may erroneously give a beneficiary money, or it may erroneously deny a transfer. Keeping track of the errors found when a denied beneficiary files an appeal only catches the latter type of mistake.

This is why, to construct a comprehensive measure of the office error rate and discourage fraudulent behavior, it is paramount to regularly audit a random subset of claims processed by each office.

Agencies may combine the error rate with a second proxy for quality: timeliness in claim processing. While timeliness is an important dimension of the service provided, a drawback of this measure is that it is mechanically correlated with office productivity. In other words, holding constant other office characteristics, offices that process claims quickly are also those that deliver a high level of output.

## Extending Administrative Data

Alternative approaches to measuring the quality of service provided include using subjective customer satisfaction ratings. The main challenge when using customer ratings is that the subset of customers who choose to provide feedback is not representative because customers with more extreme (either positive or negative) opinions are more likely to provide a review (Schoenmueller, Netzer, and Stahl 2020).<sup>2</sup>

This limitation can potentially be overcome by conducting regular surveys of a representative sample of all customers. The US Social Security Administration (SSA) implements a range of such surveys both by phone and in person across different groups of customers (online users of SSA services, callers to the SSA phone number, and visitors to SSA field offices). Although it does not eliminate the possibility that the most (un)happy customers will be more likely to respond to a survey invitation, it does mitigate this concern by targeting a sample of all customers. An indication of average customer satisfaction can also be obtained from surveys conducted by third parties. For example, the different dimensions of services provided by US government agencies are regularly evaluated as one of the topics covered in the American Customer Satisfaction Index (ACSI), which is used to measure the general satisfaction of American customers with various goods and services.

## PROCUREMENT RECORDS

Public procurement—the purchase of goods and services by governments from private sector suppliers—is one of the core functions of the state. Public procurement represents a large portion of governments' budgets and a sizeable fraction of the economy, representing 12 percent of world GDP (Bosio et al. 2022). Public procurement also tends to be a highly technocratic, legalistic process generating large volumes of documents recording every step of the procurement purchase in great detail. These data are generated and recorded as part of the government's procedures in order to uphold the transparency and accountability of the procurement process—core goals of a well-functioning procurement system. However, these same data, either by themselves or in conjunction with additional data, can also be used to measure the performance of the officials and public entities in charge of carrying out procurement.

This section builds on chapter 12 of the *Handbook* to showcase how the indicators outlined in detail there can be considered as individual case data and to showcase the benefits of complementing administrative data with experimental variation. Here, we discuss two recent academic papers that develop methods to use administrative databases on public procurement to construct measures of procurement performance. Best, Hjort, and Szakonyi (2017) use detailed procurement data from Russia spanning all procurement transactions between 2011 and 2016 to construct measures of procurement performance. They show that there are big differences across purchases in how effectively the purchase is carried out, which can be attributed in roughly equal proportions to the effectiveness of the individual civil servants tasked with procurement and the effectiveness of the public entities they represent. They also show how procurement policy can be tailored to the capacity of the implementing bureaucracy in order to offset weaknesses in implementation capacity.

Bandiera et al. (2021) use existing procurement data from Punjab, Pakistan, and supplement it with additional data collected from purchasing offices to construct performance measures. This paper is an example of how a randomized controlled trial (RCT) can be used to complement government administrative data



to better understand the impact of personnel policies and other aspects of public administration. By introducing experiments into government, such initiatives amplify the potential benefits of the analysis of public administration data. Bandiera et al. (2021) show that granting procurement officers additional autonomy to spend public money improves procurement performance, especially when the officers' supervisors caused significant delays in approvals.

## Complexity

A procurement case may be characterized by a differing number of features of the good or service being procured and by a wide range of requirements on those features. For example, the procurement of pencils has far fewer features for the procurement officer to assess than the procurement of a vehicle. For this reason, when comparing the productivity of procurement agents and agencies, it is important to have a measure of the nature of the procurement cases they have to process.

Best, Hjort, and Szakonyi (2017) use publicly available administrative data from Russia to construct measures of performance based on public procurement. Since 2011, a centralized procurement website has provided information to the public and suppliers about all purchases.<sup>3</sup> They use data from this website on the universe of electronic auction requests, review protocols, auction protocols, and contracts from January 1, 2011, through December 31, 2016. The data cover 6.5 million auction announcements for the purchase of 21 million items. However, purchases of services and works contracts are highly idiosyncratic, making comparisons across purchases impossible, so they are dropped from the sample, resulting in a sample of 15 million purchases of relatively homogeneous goods.

To use these data to measure performance, there are two key challenges to overcome. First, the main measure of performance uses prices paid for identical items, requiring precise measures of the items being procured. Second, prices are not the only outcome that matters in public procurement, and so they use administrative data to construct measures of spending quality as well.

The main measure of performance used in Best, Hjort, and Szakonyi (2017) is the price paid for each purchase, holding constant the precise nature of the item being procured. Holding constant the item being procured is crucial to avoid conflating differences in prices paid with differences in the precise variety of item being procured. As described in more detail in appendix F.1, they use the text of the final contracts, in which the precise nature of the good purchased is laid out, to classify purchases, using text analysis methods, into narrow product categories within which quality differences are likely to be negligible.

The method proceeds in three steps. First, the goods descriptions in contracts are converted into vectors of word tokens. Second, they use the universe of Russian Federation customs declarations to train a classification algorithm to assign goods descriptions a 10-digit Harmonized System product code and apply it to the goods descriptions in the procurement data. Third, for goods that are not reliably classified in the second step, either because the goods are nontraded or because their description is insufficiently specific, they develop a clustering algorithm that combines goods descriptions that use similar language into clusters similar to the categories from the second step. Just as in the case of claims data discussed in the preceding section, here it can be seen that the key issue in analyzing case complexity is comparing “apples to apples.” Although many procedures in public administration come with a set of standardized procedures, the actual complexity of each task is highly variable, and, therefore, its accurate evaluation is the key to understanding the performance of public officials. To achieve this, highly detailed metrics might be required. In the case of ISSA claims data, this metric was a continuous weight—time judged as necessary to complete a specific task based on primary data obtained during field observations in various social security offices. In the case of procured goods, the metric used is categorical but narrow enough to avoid classifying goods of a different nature as comparable. It is also not based on field measurements but rather relies on secondary data from descriptions in Russian Federation customs declarations and advanced classification algorithms.

## Quality

Sourcing inputs at low prices is the primary goal of public procurement, but it is not the only outcome that matters.<sup>4</sup> Successful procurement purchases should also be smoothly executed. Contracts should not need to be unduly renegotiated or terminated, and goods should be delivered as specified, without delays. These outcomes reflect the quality of public spending and may conflict with the goal of achieving low prices. If this problem is severe, then it would be misleading to deem purchases effective if they achieve low prices but this is offset by poor performance on spending quality.

To address this, Best, Hjort, and Szakonyi (2017) build direct measures of spending quality by combining a number of proxies for the quality of the nonprice outcomes of a procurement purchase. Specifically, they use six proxies: the number of contract renegotiations, the size of any cost overrun, the length of any delays, whether the end user complained about the execution of the contract, whether the contract was contested and canceled, and whether the product delivered was deemed to be of low quality or banned for use in Russia because it didn't meet official standards.

To summarize spending quality in a single number, they take the six quality proxies and create an index of spending quality  $y_i$  as the average of the six proxies after standardizing each one to have mean zero and standard deviation one, as follows (Kling, Liebman, and Katz 2007):

$$y_i = \frac{1}{6} \sum_{k=1}^6 (y_i^k - \bar{y}^k) / \sigma^k. \quad (15.2)$$

This is done because the proxies are in different units of measurement and because some proxies will be more variable than others. For a deviation in a proxy to be judged as “large,” this approach conditions it on what other deviations we observe for that proxy. For example, there may be many complaints but very few contract cancellations. In that case, one would want to weight a cancellation more heavily than a complaint, in accordance with how rare, and thus significant, a cancellation is. With these measures in hand, Best, Hjort, and Szakonyi (2017) show that there are big differences across purchases in how effectively the purchase is carried out. They also decompose these differences into the part that can be attributed to the effectiveness of the individual public servants working on the purchase and the part that can be attributed to the agency that is receiving the item being purchased. They show that both contribute roughly equally to the differences in effectiveness and that together they explain around 40 percent of the variation in government performance. They also show how these differences in effectiveness contribute to differences in how policy changes manifest in performance outcomes.

They argue that policy that is tailored to the capacity of the implementing bureaucracy can offset overall weaknesses in implementation capacity. The analysis provides an example of how the analytics of public administration can lead to direct implications for the policies that govern it.

## Extending Administrative Data

Existing administrative data can sometimes prove insufficient to measure productivity in public administration, but the required information can nevertheless be obtained by the targeted data collection efforts of governments and researchers. Bandiera et al. (2021) use administrative data from Punjab, Pakistan, to measure procurement performance. In their case, the existing administrative data were not sufficiently detailed to implement their preferred method of performance measurement, and so they worked with the government to design and implement an additional administrative database capturing detailed information about the products being purchased by procurement officers.

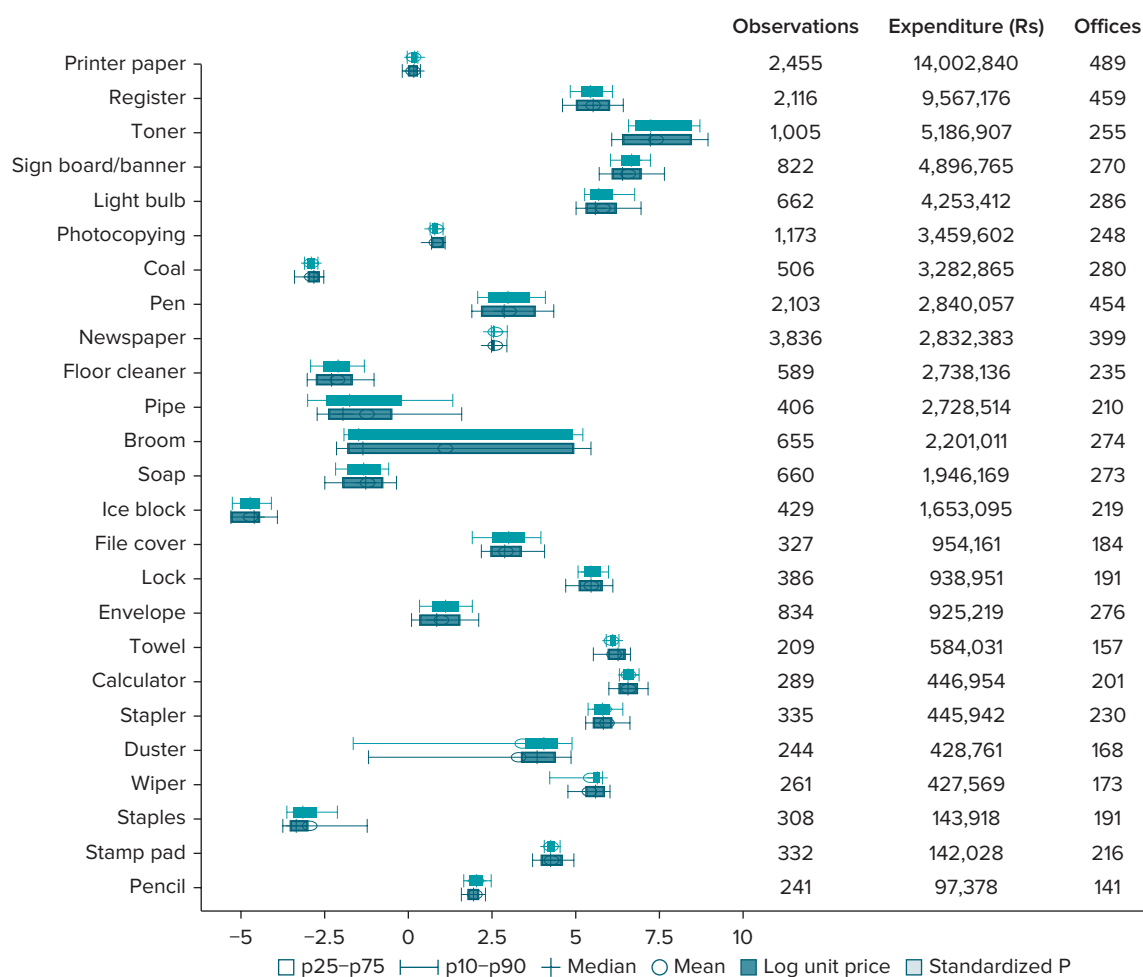
The government of Punjab considers the primary purpose of public procurement to be ensuring that “the object of procurement brings value for money to the procuring agency” (PPRA 2014). In line with this, they developed a measure of bureaucratic performance that seeks to measure value for money in the form of the unit prices paid for the items being purchased, adjusted for the precise variety of the item being purchased.

They proceed in two steps. First, they restrict attention to homogeneous goods for which it is possible to gather detailed enough data to adequately measure the variety of the item being purchased. Second, they partner with the Punjab IT Board to build an e-governance platform—the Punjab Online Procurement System (POPS). This web-based platform allows offices to enter detailed data on the attributes of the items they are purchasing. Over one thousand civil servants were trained in the use of POPS, and the departments they worked with required the offices in the study to enter details of their purchases of generic goods into POPS. To ensure the accuracy of the data, offices were randomly visited to physically verify the attributes entered into POPS and collect any missing attributes required.

After the POPS platform was run for the two-year project and the data the officers entered were cleaned, the analysis data set consists of the 25 most frequently purchased goods—a total of 21,503 purchases. Dropping the top and bottom 1 percent of unit prices results in a data set of 21,183 observations.<sup>5</sup> Figure 15.3 shows summary statistics of the purchases in the POPS data set. The 25 items are remarkably homogeneous goods, such as printing paper and other stationery items, cleaning products, and other office products. While each individual purchase is small, these homogeneous items form a significant part of the procurement: generic goods are 53 percent of the typical office's budget in the sample.

To use these data on prices to measure procurement performance, they again need to be able to compare purchases of exactly the same item. The goods in the analysis are chosen precisely because they are extremely

**FIGURE 15.3 Summary Statistics on the 25 Most Commonly Purchased Goods in the Punjab Online Procurement System, 2014–16**



Source: Bandiera et al. 2021.

Note: The figure displays summary statistics for purchases of the goods in the purchase sample. The figure summarizes the log unit prices paid for the goods, the number of purchases of each good, and the total expenditure on the good (in rupees) in the sample.

homogeneous. Nevertheless, there may still be some differentiation across items, and so Bandiera et al. (2021) use four measures of the variety of the goods being purchased. First, they use the full set of attributes collected in POPS for each good. This measure has the advantage of being very detailed but comes at the cost of being high dimensional. The three other measures reduce the dimensionality of the variety controls. To construct the second and third measures, they run hedonic regressions to attach prices to each of the goods' attributes. They run regressions of the form

$$p_{igto} = \mathbf{X}_{igto} \lambda_g + \rho_g q_{igto} + \gamma_g + \varepsilon_{igto}, \quad (15.3)$$

where  $p_{igto}$  is the log unit price paid in purchase  $i$  of good  $g$  at time  $t$  by office  $o$ ,  $q_{igto}$  is the quantity purchased,  $\gamma_g$  are goods fixed effects, and  $\mathbf{X}_{igto}$  are the attributes of good  $g$ .

The second, *scalar*, measure of goods variety uses the estimated prices for the attributes  $\hat{\lambda}_g$  to construct a scalar measure  $v_{igto} = \sum_{j \in A(g)} \hat{\lambda}_j X_j$ , where  $A(g)$  is the set of attributes of item  $g$ . The third, *coarse*, measure studies the estimated  $\hat{\lambda}_g$ s for each item and partitions purchases into high- and low-price varieties based on the  $\hat{\lambda}_g$ s that are strong predictors of prices in the control group. Finally, the *machine-learning* measure develops a variant of a random forest algorithm to allow for nonlinearities and interactions between attributes that regression (15.1) rules out. Appendix F.2 provides further details. This effort provides a way to homogenize the type and quality of goods on which government analytics can be performed.

## Extending Administrative Data

Extending administrative data does not only imply the collection of further data. Rather, it can imply an extension in the methods used for analysis. A particularly powerful extension is to embed an RCT into data collection. In this way, the data collected reflect groups that have received a policy intervention purely by chance. Comparing measures of case processing between these groups thus allows one to look for differences that are due purely to the policy intervention and not some other mediating factor.

With the above performance measure in hand, Bandiera et al. (2021) perform just such a field experiment in which one group of procurement officers is granted greater autonomy over the procurement process (essentially reducing the amount of paperwork required and streamlining the preapproval of purchases by government monitors), another group is offered a financial bonus based on their performance, and a third group is offered both. By embedding an experiment into their analysis, they find that granting autonomy causes a reduction in prices by around 9 percent, illustrating that in settings where monitoring induces inefficiency, granting frontline public servants more autonomy can improve performance.

## PROPERTY TAX DATA

Taxation is critical for development; however, tax systems throughout the developing world collect substantially less revenue as a share of GDP than their counterparts in the developed world.<sup>6</sup> Weak enforcement, informational constraints, and tax morale provide some explanation. This is also true for property taxes, despite their greater visibility and contribution to local public goods. Khan, Khwaja, and Olken (2016, 2019) describe a long collaboration with the Excise and Taxation Department in Punjab, Pakistan, on different mechanisms for incentivizing property tax collectors—through performance-pay and performance-based postings. Once again, these papers provide insight into how case data, and in this subsection, case data

related to the taxation of individual properties, can be combined with experimental variation to improve the measurement of and insights related to the performance of public administration.

The urban property tax in Punjab is levied on the gross annual rental value (GARV) of the property, which is computed by formula. Specifically, the GARV is determined by measuring the square footage of the land and buildings on the property, and then multiplying by standardized values from a valuation table depending only on property location, use, and occupancy type. These valuation tables divide the province into seven categories (A–G) according to the extent of facilities and infrastructure in the area, with a different rate for each category. Rates further vary by residential, commercial, or industrial status, whether the property is owner occupied or rented, and location. Taxes are paid into designated bank branches.

The Excise and Taxation Department collects regular administrative data. Each quarter, as part of their normal reporting requirements, tax inspectors report their revenue collected during the fiscal year cumulatively through the end of the quarter, which they compile from tax-paid receipts retrieved from the national bank. In addition, they report their total assessed tax base both before exemptions are granted and after exemptions have been granted. These records are compiled separately for current-year taxes and arrears.

In theory, the performance of property tax collectors should be easy to monitor because the key measure of performance, tax revenue, is less subject to measurement issues than other areas of government work. However, in practice, measurement related to the performance of tax inspectors faces many challenges. It is not *ex ante* obvious how much credibility to give to reported tax revenues at the unit level in Punjab, given that the tax department's internal cross-checks are usually run at a higher level of aggregation. Given multiple reporting templates with slightly varying assumptions in use in the province, all officers can overstate the revenues they have generated without their misreporting being effectively detected. Similarly, the continuously evolving environment in which tax collectors operate introduces further complications to understanding relative performance. For example, the boundaries of tax administrative units (called “tax circles” in Punjab) are continuously being changed, and tax circle boundaries do not overlap with the boundaries of political units.

For these reasons, gaining a coherent measure of the taxes collected and the performance of tax officials and agencies can be a challenging task. Since reported tax revenues are a function of the tax base, exemption rate, and collection rate, comparing collection alone is not reflective of performance. Finally, given concerns over multitasking, performance on revenue collection has to be matched with performance on nonrevenue outcomes, especially on the accuracy of tax assessments and citizen/taxpayer satisfaction.

## Complexity

Rather than generating novel measures of complexity or clever systems for categorization, as in the social security and procurement cases, complexity was made more homogeneous in this context by standardizing the reporting templates and matching boundaries. The approach to ensuring a common level of complexity in case data can thus be relatively simple in some settings.

## Quality

In the work in Punjab, to ensure the accuracy of the administrative data unit level, an additional verification program was instituted, involving cross-checking the department's administrative records against bank records. This entailed selecting a subset of circles, obtaining the individual records of payment received from the bank for each property, and manually tallying the sum from the thousands of properties in each circle to ensure that it matched the department total.

The project found virtually no systematic discrepancies between the administrative data received from the department and the findings of this independent verification; the average difference between the independent verification and what the circle had reported revealed underreporting of –0.28 percent, or about zero. In general, if rightly conducted, data diagnostics and audits can ensure the accuracy of administrative

data, help flag issues before policy decisions are based on such data, and align incentives for truthful reporting.

## Extending Administrative Data

Once again, Khan, Khwaja, and Olken (2016) showcase the power of introducing experimentation into government analytics. They ran a large-scale field experiment in which all property tax units in the province were experimentally allocated into one of three performance-pay schemes or a control. After two years, incentivized units had 9.4 log points higher revenue than controls, which translates to a 46 percent higher growth rate. The revenue gains accrued due to a small number of properties that became taxed at their true value, which was substantially more than they had been taxed at previously. The majority of properties in incentivized areas, in fact, paid no more taxes but instead reported higher bribes. The results are consistent with a collusive setting in which performance pay increases collectors' bargaining power over taxpayers, who either have to pay higher bribes to avoid being reassessed or pay substantially higher taxes if collusion breaks down. The paper shows that performance pay for tax collectors has the potential to raise revenues but might come at a cost if it increases the bargaining power of tax collectors relative to taxpayers.

The paper also highlights the limitations of relying on existing administrative data for areas where multitasking can be a concern and where existing systems capture only some aspects of performance—for instance, administrative data usually capture revenue collection but not nonrevenue outcomes, like the accuracy of tax assessments and taxpayer satisfaction. To capture these nonrevenue outcomes, as well as owner and property characteristics to examine any heterogeneous effects, Khan, Khwaja, and Olken (2016) conduct a random property survey.

The survey is based on two distinct samples. The first, the “general population sample,” consists of roughly 12,000 properties selected by randomly sampling five GPS coordinates in each circle and then surveying a total of five (randomly chosen) properties around that coordinate. These properties therefore represent the picture for the typical property in a tax circle. The second sample, referred to as the “reassessed sample,” consists of slightly more than 4,000 properties (roughly 10 per circle) sampled from an administrative list of properties that are newly assessed or reassessed. These properties were then located in the field and surveyed. The purpose of this survey was to oversample the (few) properties that experience such changes each year in order to examine the impacts on such properties separately.

These survey data are used to determine the GARV of the property, which is the main measure of a property's tax value before exemptions and reductions are applied and, unlike tax assessed, is a continuous function of the underlying property characteristics and, hence, much more robust to measurement error. To measure under- or overtaxation, the “tax gap” is determined as

$$\text{Tax Gap} = \frac{(GARV_{\text{Inspector}} - GARV_{\text{Survey}})}{(GARV_{\text{Inspector}} + GARV_{\text{Survey}})}. \quad (15.4)$$

Taxpayer satisfaction is measured based on two survey questions about the quality and results of interactions with the tax department. Accuracy is measured as one minus the absolute value of the difference between the GARV as measured by the survey and the official GARV, as measured from the tax department's administrative records, divided by the average of these two values.

Khan, Khwaja, and Olken (2019), in a subsequent project, examine the impact of performance-based postings in the same setting and rely primarily on administrative data. They propose a performance-ranked serial dictatorship mechanism, whereby public servants sequentially choose desired locations in order of performance. They evaluate this using a two-year field experiment with 525 property tax inspectors. The mechanism increases annual tax revenue growth by 30–41 percent. Inspectors who the model predicts face high equilibrium incentives under the scheme indeed increase performance more. These results highlight the potential of periodic merit-based postings in enhancing bureaucratic performance.<sup>7</sup>



## CONCLUSION

In this chapter, we have discussed how public sector organizations can use administrative data to construct measures of performance across three important realms of government operations: the delivery of social security programs, the procurement of material inputs, and tax collection. Agencies whose primary work consists of processing claims can use their existing records to construct a measure of the volume of services provided (that is, a complexity-adjusted index of claims processed) and proxies for the quality of service (that is, the error rate and timeliness in claim processing). Similarly, government organizations purchasing goods and services can leverage their existing procurement records to construct two measures of performance: the price paid for homogeneous goods and an index of spending quality that combines information on the number of contract renegotiations, cost overrun, the length of delays, complaints, contract cancellations, and whether the product delivered did not meet minimum quality standards. When the administrative data are not sufficiently detailed, governments can develop a platform that standardizes the procurement process and collects the underlying data. Finally, taxation authorities can construct reliable measures of tax revenue by standardizing the process through which tax collectors report the taxes they have collected and instituting a set of automatic checks to ensure data accuracy.

Better measures of performance may help governments improve the effectiveness of public service provision. For example, policy makers can use these performance measures to identify the best-performing offices, learn about “best practices,” and export them to the underperforming sites. Government agencies can also use these metrics to identify understaffed sites and reallocate resources toward them. Moreover, governments can *monitor* the performance of public offices and intervene promptly when a challenge arises. Finally, they can use these measures to design incentive schemes aimed at improving public service provision.

Administrative records typically include large amounts of data, and performing statistical analyses on them involves some practical challenges. First, not all public sector organizations employ workers who have the technical skills to repurpose data for performance measurement and carry out the statistical analyses. This challenge can be addressed by partnering with external researchers experienced in this area. Second, governments should take all necessary steps to protect data confidentiality when granting access to their internal records. This may involve anonymizing data to protect the identity of the subjects being studied, transferring data through secure protocols, and ensuring that data are stored on a secure server. In some cases, government organizations may also invest in their own IT infrastructure, such as a large server to store data and a set of workstations through which researchers can access anonymized administrative records.

The approaches described in this chapter have the potential to promote evidence-based policy making within government organizations, resulting in more effective public service provision. An example of such impacts comes from the tax analytics work described in this chapter. Over the course of the research collaborations discussed, the Punjabi tax authorities began to digitize and geocode unit data at the property level. This database is now being regularly updated. Tax notices are now issued through an automated process, supporting tax staff still responsible for field work and for updating property status—for example, covered area, usage (residential, commercial, or industrial), and status (owner-occupied or rented)—and for providing the information relevant for deciding on exemptions. This reduces the human interface between tax collectors and taxpayers. It allows for more sophisticated analysis and data visualization conducted at more granular levels—for example, at neighborhood levels—in real time. The data are now being used by the Urban Unit in Pakistan, different government agencies, and by analysts to address a range of policy questions.

## NOTES

1. Many governments put effort into standardizing case data to increase their capacity to undertake analytics. For example, a number of countries have introduced the Standard Audit File for Tax (SAF-T) for all taxpayers, a protocol for the data collected on each case (OECD 2017).
2. To evaluate the performance of government agencies, it is also important to account for the fact that many government agencies also have front-office operations. Measuring productivity in any customer-facing setting is challenging. While some agencies use customer ratings, the ISSA measures front-office output using the inputs—the amount of time employees spend on front-office duties. Thus, the measure bluntly captures the value of staffing the office without adjusting for the number of customers served or the complexity of their demands. An agency may also consider constructing a measure of front-office operations analogous to the one used for claim processing. The additional challenge is that allowing front-office employees to self-report their output may incentivize employees to misreport the activities they undertake.
3. The website can be accessed at <http://zakupki.gov.ru/>.
4. Article 1 of Federal Law 94 (FZ-94), which transformed Russia's public procurement system in 2005, declares the aim of procurement to be the "effective, efficient use of budget funds." The law also introduced minimum price as the key criterion for selecting winners for most types of selection mechanisms (Yakovlev, Yakobson, and Yudkevich 2010).
5. The majority of these outliers are the result of officers adding or omitting zeros in the number of units purchased.
6. According to 2018 World Bank data, tax revenue as a share of gross domestic product stood at 11.4 percent in lower-middle-income countries, compared to 15.3 percent in high-income countries.
7. In ongoing work with the tax authorities and the local government, Khwaja et al. (2020) examine strengthening the social compact between citizens/taxpayers and the government by linking the (property) taxes citizens pay with the services they receive at the neighborhood level. Combining administrative data from tax and municipal agencies at the neighborhood level provides local-level measures of variation in public service provision, tax and fiscal gap, administrative performance, and sociopolitical dynamics.

## REFERENCES

- Bandiera, Oriana, Michael Carlos Best, Adnan Qadir Khan, and Andrea Prat. 2021. "The Allocation of Authority in Organizations: A Field Experiment with Bureaucrats." *Quarterly Journal of Economics* 136 (4): 2195–242. <https://doi.org/10.1093/qje/qjab029>.
- Best, Michael Carlos, Jonas Hjort, and David Szakonyi. 2017. "Individuals and Organizations as Sources of State Effectiveness." NBER Working Paper 23350, National Bureau of Economic Research, Cambridge, MA. <https://doi.org/10.3386/w23350>.
- Bosio, Erica, Simeon Djankov, Edward Glaeser, and Andrei Shleifer. 2022. "Public Procurement in Law and Practice." *American Economic Review* 112 (4): 1091–117. <https://doi.org/10.1257/aer.20200738>.
- Christensen, Darin, and Francisco Garfias. 2021. "The Politics of Property Taxation: Fiscal Infrastructure and Electoral Incentives in Brazil." *The Journal of Politics* 83 (4): 1399–416. <https://doi.org/10.1086/711902>.
- Fenizia, Alessandra. 2022. "Managers and Productivity in the Public Sector." *Econometrica* 90 (3): 1063–84. <https://doi.org/10.3982/ECTA19244>.
- Khan, Adnan Q., Asim I. Khwaja, and Benjamin A. Olken. 2016. "Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors." *The Quarterly Journal of Economics* 131 (1): 219–71. <https://doi.org/10.1093/qje/qjv042>.
- Khan, Adnan Q., Asim I. Khwaja, and Benjamin A. Olken. 2019. "Making Moves Matter: Experimental Evidence on Incentivizing Bureaucrats through Performance-Based Postings." *American Economic Review* 109 (1): 237–70. <https://doi.org/10.1257/aer.20180277>.
- Khwaja, Asim Ijaz, Osman Haq, Adnan Qadir Khan, Benjamin Olken, and Mahvish Shaukat. 2020. *Rebuilding the Social Compact: Urban Service Delivery and Property Taxes in Pakistan*. 3ie Impact Evaluation Report 117. New Delhi: International Institute for Impact Evaluation (3ie). <https://doi.org/10.23846/DPW1IE117>.
- Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica* 75 (1): 83–119. <https://doi.org/10.1111/j.1468-0262.2007.00733.x>.
- OECD (Organisation for Economic Co-operation and Development). 2017. *The Changing Tax Compliance Environment and the Role of Audit*. Paris: OECD Publishing. <https://doi.org/10.1787/9789264282186-en>.

- PPRA (Punjab Procurement Regulatory Authority). 2014. *Punjab Procurement Rules 2014*. No. ADMN (PPRA) 10–2/2013. Lahore: Government of the Punjab Services General Administration Department. <https://ppra.punjab.gov.pk/system/files/Final%20Notified%20PPR-2014%20%28ammended%20upto%2006.01.2016%29.pdf>.
- Schoenmueller, Verena, Oded Netzer, and Florian Stahl. 2020. “The Polarity of Online Reviews: Prevalence, Drivers and Implications.” *Journal of Marketing Research* 57 (5). <https://doi.org/10.1177/0022243720941832>.
- Yakovlev, Andrei, Lev Yakobson, and Maria Yudkevich. 2010. “The Public Procurement System in Russia: Road toward a New Quality.” 4th International Public Procurement Conference, Seoul, Republic of Korea, August 26–28. <http://ippa.org/images/PROCEEDINGS/IPPC4/01ComparativeProcurement/Paper1-9.pdf>.

## CHAPTER 16

# Government Analytics Using Machine Learning

*Sandeep Bhupatiraju, Daniel Chen, Slava Jankin, Galileu Kim, Maximilian Kupi, and Manuel Ramos Maqueda*

### SUMMARY

The use of machine learning offers new opportunities for improving the productivity of the public sector. The increasing availability of public sector data and algorithmic approaches provides a conducive environment for machine learning for government analytics. However, the successful deployment of machine-learning solutions requires first developing data infrastructure of the required quality to feed these algorithms, as well as building the human capital necessary to develop them. Ethical principles regarding the use of machine-learning technologies must be defined and respected, particularly for the justice system. This chapter provides an overview of potential applications of machine learning in the public sector and in the justice system specifically, as well as the necessary steps to develop them sustainably and ethically. It then analyzes the case of machine-learning deployment in India to illustrate this process in practice.

### ANALYTICS IN PRACTICE

- Machine learning is fundamentally a methodological approach: it defines a performance indicator and uses collected data to train an algorithm to improve this indicator. Because of this relatively broad definition, machine learning includes different algorithms and may be applied in a variety of domains, from payroll fraud detection to court rulings. This flexibility requires practitioners to make key design decisions: what kind of performance indicator will be used? What training data and algorithm will be deployed? These decisions may substantially alter the machine-learning algorithm's results. Making these decisions thus requires close collaboration between machine-learning engineers, domain experts, and the agencies that will use the technology.

---

The authors' names are listed alphabetically. Sandeep Bhupatiraju is a consultant in the World Bank's Development Impact Evaluation (DIME) Department. Daniel Chen is a senior economist at DIME. Slava Jankin is a professor of data science and public policy at the Hertie School. Galileu Kim is a research analyst at DIME. Maximilian Kupi is a PhD candidate at the Hertie School. Manuel Ramos Maqueda is a research analyst at DIME.

- Machine learning can leverage large amounts of administrative data to improve the functioning of public administration, particularly in policy domains where the volume of tasks is large and data are abundant but human resources are constrained. Governments generate large amounts of administrative data on an almost-daily basis, but these data are seldom used to improve the production function of public administration. At the same time, civil servants are constrained in the amount of time they can dedicate to complete tasks—as well as in the amount of information they have readily available. Machine learning can process large amounts of administrative data, structuring them around performance indicators that algorithms are well suited to optimize. For example, machine-learning algorithms can be trained, using procurement data, to predict whether new, incoming contracts are irregular or not, at a scale and speed which far exceed human capacity.
- While machine learning can offer efficiency gains in public administration, governments need to be aware of their role in generating and using measurements in public administration, as well as their ethical responsibilities. Machine-learning algorithms require extensive data and measurements on both citizens and civil servants, but these data are often collected without their consent. As a result, governments should be transparent about how these data are being used to enhance public administration and how these technologies are being used to affect public administration. Care should be taken not to reproduce biases, such as racial or gender discrimination, in the machine-learning algorithms.
- To fully reap the benefits of machine learning, governments must undertake long-term investments in data infrastructure and human capital. Before machine learning is implemented, investments in data infrastructure must take place. Data quality needs to be improved and data pipelines must be developed to train the algorithm. Additionally, specialized machine-learning engineers must be hired and trained to implement the technology in the public sector. These investments are costly and require long-term planning: governments should not expect machine-learning technologies to be developed overnight.
- Machine-learning experts should collaborate with subject matter experts to guide how machine-learning technology will benefit the civil servants who will use it. Besides the technical knowledge to develop and operate machine-learning technologies, substantial levels of domain and political expertise, as well as awareness about potential ethical and legal pitfalls, are necessary to ensure the effective use of a machine-learning solution. For example, if the machine-learning technology is meant to assist judges in reducing racial bias, judicial experts and judges themselves should be consulted to ensure that relevant performance indicators and data are used. Including civil servants in the development of the solution also facilitates the adoption of the new technology.
- Machine learning is not a panacea, and practitioners should be aware of the limitations of the approach to fully leverage its benefits. Algorithms are limited to what is measurable by data, and performance indicators may reflect the bias of machine-learning engineers and the data themselves. Improving a particular performance indicator may not necessarily be the best way of achieving a policy goal. As a result, machine-learning applications should not be considered as a substitute for policy making but as a tool to complement and enhance decisions made by government agencies and their civil servants.

## INTRODUCTION

Machine learning is a discipline that focuses on the development of computer systems (machines) that, through the analysis of training data, can improve their performance (learn) (Jordan and Mitchell 2015). Recent advances in data collection and processing power have expanded opportunities for machine-learning applications in a variety of fields. Advances in machine learning have brought tangible benefits in the worlds of business and medicine and in large-scale systems more generally (Brynjolfsson and McAfee 2017). However, this growth in opportunity has often led to excessive optimism about what machine learning can

accomplish, as well as a tendency to downplay the potential steep costs of deploying machine-learning technologies (Chen and Asch 2017).

We start our discussion by offering a general definition of machine learning. What distinguishes machine learning from other methodological approaches is the definition of a learning problem under a statistical framework. Following Jordan and Mitchell (2015, 225), we define a learning problem as a “problem of improving some measure of performance when executing some task, through ... training experience.” The following example illustrates a machine-learning approach. Suppose a government is interested in reducing irregularities in its payroll system. One measure of performance would be the proportion of irregular paychecks correctly identified by the machine. The training experience—or data—would be a collection of paychecks manually classified by payroll analysts (see case study 2 in chapter 9). The learning problem would thus define a statistical model that learned how best to predict irregular paychecks by being trained on historical payroll data.<sup>1</sup>

In this chapter, we discuss applications of machine learning to public administration. We outline the data infrastructure and human capital requirements for developing machine-learning applications, as well as potential complementarities with public servant surveys. As noted in chapter 9 of *The Government Analytics Handbook*, the foundational step for any form of data analytics—including machine learning—is the development of a robust data infrastructure. We also highlight ethical concerns regarding the development and deployment of machine-learning applications, which relate directly to the discussion in chapter 6. Despite our focus on machine learning, we consider the broader shift to a data-driven and statistically informed culture—regardless of the implementation of algorithms—to be often already sufficient for bringing substantial benefits to public service delivery. These benefits include organizational changes, data literacy, and performance monitoring. With the transition to data-driven policy making, machine-learning applications are a natural next step in government analytics, automatically leveraging data to improve the performance of public administration through well-defined performance metrics.

Following this broader discussion of machine learning in public administration, we focus on machine-learning applications in justice systems. Within public administration, the justice system generates a large number of case rulings linking legal cases and actors (training data) to ideally fair rulings (performance). Thus, a potential machine-learning learning problem is: can we identify and reduce racial bias in court rulings by training an algorithm on a collection of case rulings? Instead of defining what a fair ruling is, we might define what an unfair ruling is. For instance, the decision of the judge should not be influenced by extraneous factors, such as the time of day or the race of the defendant. By identifying cases in which the decision has been influenced by such biases, a machine-learning model can potentially identify and ultimately prevent unfair rulings. This approach thus allows machine learning to improve the quality and fairness of judicial decisions (Ramos-Maqueda and Chen 2021). Despite this promise, limited data literacy and statistical training inhibit applications of data analytics in general—and machine learning specifically—in the judiciary.

Machine learning is not a panacea: it requires significant investments before any of its benefits come to fruition. As we detail throughout this chapter, machine-learning applications require investments in data infrastructure and the development of the human capital necessary to develop and deploy machine-learning algorithms. Undertaking these investments—and often long cycles of development—requires resources and long-term horizons before the benefits of the approach become apparent. Additionally, ethical concerns regarding embedded racial or gender biases in training data highlight how technologies can inadvertently reproduce the same human biases they were designed to eliminate. Thus, initial optimism regarding the revolutionary potential of machine-learning approaches should be balanced by a recognition of their limitations (Cross 2020).

Practitioners have much to gain from deploying machine learning in public administration. Defining performance metrics and automating the training of algorithms through large-scale data can improve the functioning of public administration—particularly when oriented toward well-defined tasks with an abundance of quantitative data. Performance improvement can come simply from embedding a data-driven approach to government functioning. There is not always a need for sophisticated approaches in machine learning to make progress. In fact, machine learning generally comes only after more basic steps have been taken on data management and analytics in public administration, as highlighted in other chapters of the *Handbook*.



This chapter is structured as follows. Section 1 describes applications of machine learning in public administration. Section 2 then outlines a road map to applying machine learning in the public sector, focusing on data infrastructure requirements and human capital needs. Section 3 shifts our focus from public administration in general to the justice system. In doing so, it highlights applications of machine learning in the justice system, as well as the data infrastructure and human capital required to implement them. Section 4 presents a case study from India illustrating machine-learning approaches to justice in practice. Section 5 moves beyond descriptive analysis to outline how machine learning can be used to assist causal inference. Finally, we conclude.

## MACHINE LEARNING FOR PUBLIC ADMINISTRATION

The use of machine learning is spreading across many functional areas of public administration. While European Union (EU) governments focus on service delivery and public engagement, other areas, such as internal management and law enforcement, are progressively being targeted for the deployment of machine-learning solutions to increase their efficiency and effectiveness (Misuraca and van Noordt 2020). The applications are diverse, from detecting COVID-19 outbreaks to simulating the impact of changes in macroeconomic policy. Machine-learning applications thus provide novel ways for governments to use their data to improve public administration. Chapter 15 of the *Handbook* highlights how machine learning can be used to detect similarities between goods in public procurement.<sup>2</sup>

The use of machine learning provides a few advantages compared with more standard analytical approaches. Standard data analytics provides the analyst with tools bounded by the analyst's capacity to investigate connections between variables in the data—often the coefficients in a regression specification. However, in many public administration settings, the analyst is confronted with factors—individual or organizational—that may influence a policy outcome without the analyst's knowledge. Machine learning enables the exploration of relationships between variables in a principled, and often unsupervised, way. However, causality in machine learning is a relatively recent development, and it presents considerable challenges.<sup>3</sup> Potential applications, therefore, focus less on causal interventions or experiments and more on solutions that, based on given data, best perform an accurate prediction.

The primary focus of this section is on applications of machine learning for administrative data.<sup>4</sup> However, governments may leverage public servant surveys to complement this analysis, particularly for personnel data. For example, a government may be interested in better understanding job satisfaction and how it relates to staff turnover. While a machine-learning analysis could be useful in identifying potential patterns in civil service exit from the full population of interest as a function of demographics (sex, age, education, or race), it might not provide much information about the attitudes of the staff who are at risk of exiting. A public servant survey provides a complement for answering this kind of “why” question, but it may not be large enough to find general patterns in the first place—particularly if it is not linked to administrative data on exits, which is often difficult to do. Thus, machine-learning applications on human resources data outperform surveys at identifying certain kinds of patterns, but they need to be complemented by surveys explaining these patterns and highlighting potential interventions that might address problems.

### Machine-Learning Applications

Applications of machine learning can be subsumed under the following three categories.

#### Detection and Prediction

Machine learning can help policy makers detect and predict destructive events, improving the design and implementation of adequate policy measures. This is the largest application of machine-learning approaches,

addressing issues such as COVID-19 outbreaks, fake news, hate speech, tax fraud, military aggression, terrorist activity, cyberattacks, natural disasters, street crime, and traffic congestion—to mention only a few. While the detection and prediction of destructive events is only the first step toward effective government intervention, it is an important instrument for effective policy making.

For example, in Delhi, over 7,500 CCTV cameras, automatic traffic lights, and 1,000 LED signs are equipped with sensors and cameras that collect traffic data, which a machine-learning system processes into real-time insights. Local authorities can then decide how to balance traffic flow in real time and identify traffic patterns and congestion trends in order to plan for the long-term mitigation of traffic problems (Devanesan 2020). Besides these benefits, which are geared toward improving general traffic flow, these systems are also used by the Delhi Police to track and enforce traffic violations, such as speeding or illegal parking (Lal 2021).

More generally within public administration, machine-learning approaches have been used to improve the machinery of government itself. For example, chapter 15 of the *Handbook* highlights how machine-learning techniques have been applied to detect corruption in public procurement. One prominent example of this application is the use of decision tree models (random forest and gradient boosting machine) to detect the presence of Mafia activity in procurement contracts in Italy (Fazekas, Sberna, and Vannucci 2021). In Brazil, federal agencies have deployed machine learning to detect evidence of corruption in federal transfers to municipal governments, as well as in irregularities in paychecks issued to civil servants, as described in case study 2 in chapter 9.<sup>5</sup>

### **Simulation and Evaluation**

Simulating and evaluating the impact of future policy measures is another widespread application area for machine learning. Simulating the potential costs of a policy measure against its expected benefits has become an increasingly relevant tool for governments. For example, in the United States, a simulation known as the National Planning Scenario 1 allows policy makers to simulate what might happen if Washington, DC, were subject to a nuclear attack (Waldrop 2018). Whether policies are designed to stimulate the economy or to contain the spread of a virus, simulation and evaluation provide valuable insight to policy makers before implementation, allowing them to choose which policies maximize intended effects.

### **Personalization and Automation**

Machine learning can also be applied to the personalization and automation of government processes and services. For example, policy makers may customize digital government services for parents to every life stage of their newborn child or tailor the provision of health care services to each patient's particular needs. Additionally, the automation of repetitive tasks leaves more time for public servants to do other tasks. All in all, these novel technologies may help governments be more efficient in their use of time and increase their responsiveness to citizens' needs.

A medical example illustrates this approach. There has been growing interest within the US federal government in using machine learning to improve public health outcomes. A series of pilots to develop such machine-learning solutions have been rolled out. These include the prediction of potential adverse drug reactions using medical reports, the classification of whether a child is likely to have autism based on medical records, and the prediction of unplanned hospital admissions and adverse events (Engstrom et al. 2020, appendix). Another study has found, through the application of machine-learning techniques, that physicians overtest low-risk patients but simultaneously undertest high-risk patients (Mullainathan and Obermeyer 2022).

### **Practical Steps for Machine Learning in Public Administration**

The implementation of machine learning in public administration comprises two key steps. The first is building a high-quality data infrastructure to feed the necessary training data to the machine-learning algorithm. Because public administration data infrastructures are often developed without machine-learning

applications in mind, adaptation is often necessary. New data pipelines need to integrate public sector information systems that previously operated in isolation, such as public procurement and budget data. Data standardization practices, such as ensuring that variables in different data tables are named consistently, and other quality checks need to be in place to ensure that the data fed to the machine-learning system are accurate and comprehensive.

Another key step is developing the human capital necessary to deploy machine learning. Before fulfilling the promise of automated, self-learning algorithms, a team of human developers is necessary to set the system in place. In fact, the entire pipeline, from data infrastructure to the training of the algorithm to disseminating actionable insights for policy makers, has to be designed by humans. Having an in-house team capable of developing and maintaining machine-learning applications is crucial. Continuous collaboration between the machine-learning implementation team and policy colleagues who will use its insights ensures that applications are adapted for and stay relevant to public administration's needs.

In the following sections, we dig deeper into these steps. In so doing, we highlight examples of strategies to ensure that both the data infrastructure and the human capital requirements are in place to deploy machine learning in public administration.

## Public Sector Data Infrastructure for Machine Learning

Machine learning requires large volumes of data. These data should be of high quality: they should be comprehensive, covering all measurements necessary for the algorithm, and complete, reducing to the extent possible any gaps in measurement that may arise. A robust data infrastructure ensures that these two principles are respected and is a prerequisite for any machine-learning application. The implementation process may require upgrading legacy information systems or integrating new systems into old ones to process the resulting, often large, data sets. Practitioners may benefit from referring to the *Handbook's* wider discussion of how to reform data infrastructure for analytical insights; this discussion provides lessons that apply to machine-learning settings as well (chapter 9).

Some types of data structure may be more amenable to machine learning than others. Machine-learning applications often require structured data—with well-defined formats and measurements—so policy areas that traditionally deal with structured data, such as finance or budget data, lend themselves particularly well to it. (For an overview of using budgetary data for analytics, see chapter 11.) At the same time, governments produce unstructured data—which lack a predefined data format—such as written documents, meeting recordings, and satellite imagery. To take full advantage of this range of data, practitioners should develop a flexible storage solution that accommodates different types of data. This flexibility should be complemented by thorough documentation of data collection and standardization practices, as well as by measures to ensure compliance with data security regulations, such as the EU's General Data Protection Regulation (GDPR).

The deployment of this data infrastructure requires long-term, costly investment. Data engineers and information technology technicians should partner with the machine-learning implementation team to define data requirements, identify relevant variables, and connect the machine-learning applications to the data infrastructure. The development of a robust data infrastructure is foundational for the effective deployment of machine learning in public administration and should always precede it. Since the data infrastructure is embedded within public administration, its development requires careful coordination between machine-learning engineers, data engineers, and their institutional counterparts who own permissions to the data. Data should be integrated across government agencies to ensure that the largest pool of data is made available for training the application. Open communication between teams and agencies is therefore key.

## Human Capital Requirements for Machine Learning in Public Administration

A sustainable machine-learning application is often best achieved by building on in-house human capital. This ensures that the developed solutions are in line with existing government regulations and that policy choices are encoded faithfully. Furthermore, in-house machine-learning experts will be more likely to

possess the necessary subject and political expertise required to implement machine-learning solutions in a policy area. Finally, even if an agency decides to rely on external service providers, a certain level of embedded expertise is required to know what is technically possible and feasible, as well as to make informed judgments about the quality of contractor-provided solutions.

To build the necessary skills infrastructure in government organizations, it is first necessary to understand what competencies are needed for machine-learning developers. Naturally, knowledge about machine-learning and deep-learning algorithms is necessary. Beyond this basic knowledge, methods for dealing with large-scale data and databases in general and knowledge about distributed computing systems are also key. To successfully develop, deploy, and operate machine learning in government, familiarity with human-centered design and acquaintance with the legal and ethical frameworks in public administration are important. Finally, policy-area expertise and knowledge about governance and policy making in general enable machine-learning applications to be anchored in the operational needs of government.

Integrating the necessary skills infrastructure within government organizations often proves to be difficult. Hence, governments should follow one or more of the following best practices. Machine-learning talent does not usually follow the classical tenure path of public sector officials. Lateral entries or dedicated programs that allow entries for a limited amount of time can be effective methods for attracting these specialists into government offices. Furthermore, adapting job-classification schemes to include machine-learning-related job categories and increasing salaries and career prospects to better compete with comparable private sector job placements are advisable strategies. Ultimately, it is important to raise awareness among the target talent group about the motivating challenges (for example, social impact) and rewarding benefits (for example, job stability and work-life balance) of public sector work.<sup>6</sup>

Once in government, machine-learning experts' work can benefit greatly from exchange and knowledge sharing with colleagues. Establishing so-called communities of practice to cross knowledge boundaries within and across agencies can help gain legitimacy in relation to relevant stakeholders and foster collaboration among different agencies. Including nontech colleagues in these communities can also ensure that machine-learning applications are developed in a user-friendly manner and integrate well into the daily activities of public servants.<sup>7</sup> Another often-applied practice is the establishment of excellence centers that offer research, support, and training services and help agencies stay on the cutting edge of machine-learning technology. Finally, open communication, such as through blogs or dedicated events, can help other departments take notice and learn from each other's experiences.

Collaborations with external experts and research institutions can be another effective approach to bringing external expertise into a specific project while maintaining control and monitoring quality. Besides concrete project collaborations, establishing academic partnerships, like mobility or internship programs for the temporary assignment of personnel between government agencies and universities or research centers, can help institutionalize such collaborative efforts. Also, tailoring the machine-learning-related educational offerings of partner academic institutions to the particular needs of government organizations can be a viable way to ensure an inflow of machine-learning talent. Finally, building and sustaining intersectoral and interdisciplinary networking initiatives focused on the use of machine learning in government can help establish collaborations and foster learning and exchange.

## Ethical Considerations for the Deployment of Machine Learning

Ethical considerations should be at the forefront of machine-learning deployment in public administration, and of analytical applications more broadly.<sup>8</sup> The social contract between governments and citizens differs substantially from the one that private sector companies have with their customers. Citizens or civil servants rarely have a choice about whether to share their data with the government. This makes data security and privacy particularly sensitive because most machine-learning applications require substantial amounts of data for appropriate training. On top of more general regulations, like the GDPR, ensuring the responsible usage and sharing of data, potentially by applying adequate anonymization techniques, should be a priority for governments to ensure citizens' trust (for a more extensive discussion, see chapter 28).

Another factor that can inhibit citizens' trust stems from the rare position of governments in relation to machine-learning technologies. Governments unify the roles of user and regulator in a single entity. This makes the public sector's use of machine learning a particularly delicate target of public scrutiny. Cases in which government machine-learning systems violate citizens' rights, like the recent case of the Dutch automated surveillance system for detecting welfare fraud, pose serious threats to citizens' trust (see box 16.1). It is therefore necessary to faithfully encode legal and political choices and ensure compliance with international regulatory frameworks to ensure ethical machine-learning applications in the public sector.

Applications of machine learning in government must consider that citizens often rely only on the government for public services like social security. This is a particular challenge to using machine learning in settings where the algorithm must make a choice. For instance, regarding social security systems, an algorithm might decide whether a citizen is eligible for a particular government benefit. In this situation, the algorithm would have to compare what would happen if the citizen were granted the benefit versus if the citizen were not. The algorithms that underlie this decision-making have to make assumptions about what would happen in each scenario, and the usefulness of the final decision depends on how appropriate these underlying assumptions were. If a citizen were denied a benefit due to an algorithm's decision, who would hold the algorithm accountable?

### **BOX 16.1 The Precedent Case of the Netherlands' Automated Surveillance System**

On February 5, 2020, the District Court of The Hague ruled that SyRI (Systeem Risico Indicatie), a machine-learning application used by the government of the Netherlands to detect welfare fraud, violated Article 8 of the European Convention on Human Rights (ECHR)—that is, the right to respect for private and family life. This case is one of the first times a court has stopped a government's use of machine-learning technologies on human rights grounds and is thus considered to offer an important legal precedent for other courts to follow.

SyRI was designed to prevent and combat fraud in areas such as social benefits, allowances, and taxes by linking and analyzing data from various government and public agencies and generating personal risk profiles. It was deployed by the Minister of Social Affairs and Employment upon the request of various public agencies, including, among others, municipalities, the Social Insurance Bank, and the Employee Insurance Agency. The system mainly used a neighborhood-oriented approach, meaning it targeted specific neighborhoods where the linked data indicated an increased risk of welfare fraud.

Although the Court agreed with the government of the Netherlands that the fight against fraud is crucial and thus that employing novel technologies offering more possibilities to prevent and combat fraud generally serves a legitimate purpose, it ruled that the way SyRI was operated did not strike a "fair balance" between social interests and violation of the private lives of citizens, as required by the ECHR. In particular, the Court stated that due to the lack of insight into the risk indicators and the operation of the risk model, the system had violated the transparency principle and that discrimination or stigmatization of citizens in problem areas could not be ruled out.

The ruling, which led to the immediate halt of SyRI and caused public uproar far beyond the Netherlands, is a telling example of the potential negative consequences of applying machine-learning systems for government purposes without adequately addressing their potential ethically adverse side effects.

Often, modeling assumptions are not directly testable and hence require a substantial level of expertise over both what assumptions the algorithm is making and the suitability of those assumptions for a given public sector setting (Athey 2017). Public policy making through machine learning therefore raises important ethical questions. Choices may be made on behalf of government officials about citizen outcomes by machines they do not fully understand. For this reason, there is tension between the use of machine-learning technology to improve public administration and the oversight required to ensure that its use is in accordance with ethical principles. This tension becomes particularly salient when the use of previous administrative data for algorithm training introduces human biases into the system. Not uncovered, these biases can lead to “discrimination at scale” in sensitive areas such as racial profiling.

Finally, most applications of machine learning for government purposes are not static and should be adapted to evolving understandings of ethical principles. For example, algorithms for detecting fraud need to be constantly updated or retrained to address new forms of misconduct uncovered by agency employees and avoid an excessive focus on past forms of misconduct. Without such updating, algorithms may be biased toward past versions of criminal conduct. Constant updating by consulting domain experts and ethical advisors is necessary to ensure the effectiveness and ethical compliance of machine-learning technologies in government.

## MACHINE LEARNING FOR JUSTICE

We now turn our focus to applications of machine learning in the justice system. The justice system is an institutional setting with high-frequency data, well-documented cases, extensive textual evidence, and a host of legal actors. As such, it is a useful setting within which to explore the use of machine learning for administration in the public service. An example of a core analytical question in justice is how the characteristics of judges impact judicial outcomes, such as rulings. This is a formulation of a wider question about how the individual characteristics of public officials impact the quality of public services provided by the government. It is a question that the analytics of public administration can investigate with the right measurement, data infrastructure, and skills for analysis.

Significant progress has been made in answering this question using machine learning (Chen 2019a, 2019b; Rigano 2018). In the United States, machine learning is already used in processing bail applications, DNA analysis of crimes, gunshot detection, and crime forecasting (Epps and Warren 2020; Rigano 2018). The large volume of data from surveillance systems, digital payment platforms, newly computerized bureaucratic systems, and even social media platforms can be analyzed to detect anomalous activity, investigate potential criminal activity, and improve systems of justice. For example, after the January 6, 2021, riots at the US Capitol, investigators used machine-learning-powered facial-detection technologies to identify participants and initiate prosecutions (Harwell and Timberg 2021). Machine-learning systems can also reduce the barriers to accessing courts by providing users with timely information directly, rather than through lawyers. Sadka, Seira, and Woodruff (2017, 50) find that providing information to litigants in mediation reduces the overconfidence of litigants and nearly doubles the overall settlement rate, but this only occurs when litigants are informed directly rather than through their lawyers.

The application of machine-learning systems to justice systems is useful because slight tendencies in human behavior can have significant impacts on judicial outcomes. A growing body of work demonstrates how small external factors, most of which participants are unaware of, can leave their mark on the outcomes of legal cases. Analysis of courts in the US, France, Israel, the United Kingdom, and Chile, for example, has found that in various settings, the tone of the words used in the first 3 minutes of a hearing, the incidence of birthdays, the outcomes of sporting events, and even the time of day of a hearing or a defendant’s name can affect the outcome of a case (Chen 2019a). An analysis of 18,686 judicial rulings by the 12 US circuit courts (also known as courts of appeals or federal appellate courts), collected over 77 years, illustrates that judges demonstrate considerable bias before national elections (Berdejó and Chen 2017). Similarly, there is



new evidence on sequencing matters in high-stakes decisions: decisions made on previous cases affect the outcomes of current cases even though the cases have nothing to do with each other. For instance, refugee asylum judges are two percentage points more likely to deny asylum to refugees if their previous decision granted asylum (Chen, Moskowitz, and Shue 2016).

Given the abundant evidence of how bias shapes decisions made by officials in the justice system, machine-learning methods can identify these sources of bias and signal when they shape judicial outcomes. The subtlety of different forms of bias requires an approach that searches through a very large number of relationships to detect their wider effects, an approach for which machine learning may be well suited. This can result in a more streamlined system and a reduction in backlog. Such tools can identify discrimination and bias even when these are not evident to the participants in the courts themselves, thereby strengthening the credibility of the judiciary (Bhushan 2009; Galanter 1984; Kannabiran 2009). Moreover, as large backlogs of cases are a significant problem for the efficiency of the judiciary in developing countries, interest is growing in performance metrics that will improve the functioning of the judiciary.

The adoption of machine-learning systems, however, is not an easy-to-implement solution, in particular for the justice system. Despite the growing availability of judicial data, it is first necessary to process these data in a way that is amenable to machine learning. This requires the integration of data from different sources, the processing of textual data into quantifiable metrics, and the definition of indicators for learning tasks that reflect either performance objectives or the operationalization of the concepts of fairness and impartiality. This is not an easy task: it requires substantial investments in data infrastructure and human capital, as well as building a conceptual framework. Therefore, to implement machine-learning algorithms successfully, justice systems need to acquire and train teams of machine-learning engineers, subject matter experts, and legal actors to develop machine-learning algorithms that are operationally relevant. These considerations are similar to the ones highlighted in the broader consideration of public administration.

Finally, there are ethical concerns regarding machine-learning applications for judicial outcomes. Practitioners and citizens may raise questions regarding the interpretability of algorithms because technological sophistication creates a “black box” problem: increases in technology’s sophistication make its operation less interpretable (Pasquale 2015). The challenge of interpretability also raises concerns about accountability and oversight for these systems. Furthermore, the gap between those who can and those who cannot access and understand these technologies exacerbates existing social divisions and intensifies polarization. For all these reasons, machine-learning tools should be seen as complements to rather than substitutes for human decision-making, in particular for institutions that make life-altering decisions, such as the judiciary.

## Judicial Data Infrastructure for Machine Learning

It is increasingly recognized that “the world’s most valuable resource is no longer oil, but data” (*Economist* 2017). Like oil, raw data are not valuable in and of themselves; their value arises when they are cleaned, refined, processed, and connected to other databases that allow for the generation of insights that inform decision-making. This is particularly true in the field of machine learning, which requires large amounts of data to build accurate predictive models that provide information on the process, behaviors, and results of any indicators of interest.

Judiciaries collect vast amounts of data daily. Despite the availability of data, judicial data have rarely been analyzed quantitatively. In recent years, the transition from paper to e-filing and case management systems has facilitated the systematic analysis of massive amounts of data, generating performance metrics that can be used to evaluate courts and justice actors. Furthermore, with advances in machine learning, natural language processing (NLP), and processing power, these data create valuable opportunities to apply machine-learning models to evaluate and improve justice systems (Ramos-Maqueda and Chen 2021). Nonetheless, the extent to which countries can utilize novel approaches in machine learning and data analytics will depend on the available data infrastructure. The question, then, is which data do (and should) judiciaries collect?

In the justice domain, an integrated justice system brings together data on each case and connects these data with information on the actions and decisions at each case milestone. For instance, this includes

information on the case filing, initial decisions, hearings, rulings, and sentences for each case. These data should also relate to potential appeals in order to help understand the evolution of the case. By implementing NLP on the text of case filings or judicial decisions, judiciaries can automate the revision of case filings or identify relevant jurisprudence for judges and court actors, for instance. Beyond information on the justice process itself, judiciaries will gain valuable insights from connecting these data with other information, such as human resources data, information on recruitment, or data from judicial training, to understand how best to select, train, and motivate judges depending on their background and experience.

To evaluate the impact of machine-learning interventions in justice, judiciaries should ideally collect information from other agencies involved in the justice process, as well as the economic outcomes of the parties involved. In criminal justice, an interoperable data ecosystem will connect judicial data with data from the prosecutor's office, the police, and the prisons, which will enable judiciaries to understand where the case has come from and the implications of judicial sentences. In civil cases, this may include economic data on citizens and firms who participate in the justice system, such as tax data, social insurance data, or procurement data. This way, judiciaries will be able to evaluate the impact of machine-learning applications not only on the judicial process but also on the lives of citizens and the financial status of firms that use the justice system.

In addition to the external and internal databases, it is also recommended that judiciaries carry out surveys that complement administrative information with the experience of the parties and employees involved in the justice system. Administrative data will not capture important elements of user or employee satisfaction, for instance, so survey data are a necessary complement for understanding the impacts of any new machine-learning models. We also recommend surveying those who are not necessarily part of the justice system—but who could be potential users in the future—through legal needs assessments.

Finally, there are additional complexities to developing data ecosystems for machine learning. Data must be of high quality, and large volumes need to be collected and stored to make them amenable to artificial intelligence (AI) algorithms, which, in general, presuppose big data. This requires data extraction, transformation, and loading (ETL) processes designed to support AI pipelines and, in most use cases, dedicated data engineers to maintain them.

## Human Capital Requirements for Machine Learning in Justice

Justice systems often have limited in-house access to the necessary human capital to develop machine-learning applications. Judicial officers are rarely experts on data analysis—which is seldom part of their training—and machine-learning engineers generally lack the domain expertise necessary to understand the functioning of the law. In courts without sufficient human capital to take advantage of available data, training public servants in even simple data analysis skills may be a valuable long-term investment for improving the functioning of courts. Nevertheless, the development of machine-learning approaches may remain out of reach for public officials whose training does not include statistical modeling or data engineering.

An alternative approach is relying on nongovernmental organizations, international organizations, or even private companies to develop machine-learning applications. An example of this approach is the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) tool in the United States, which is an algorithm that generates recidivism risk scores to aid judges in their rulings. However, judiciaries should consider the long-term sustainability of an outsourced solution as well as ethical concerns. COMPAS has been the target of controversy due to its proprietary algorithm and the inability of public officials and citizens to understand how it operates under the hood.<sup>9</sup> Additionally, a reliance on external contractors often substitutes for in-house development of the necessary human capital to develop machine-learning technologies, reproducing external reliance on nonjudicial actors for both the maintenance and expansion of machine-learning solutions.

Whether in-house or externally sourced, the forms of human capital required for machine-learning applications are diverse and costly. The implementation team should include machine-learning engineers, software developers to code the user interface, data engineers to develop the data infrastructure, legal

experts, and project managers to communicate the judiciary's needs to the implementation team. Because each component of the project relies on the others—there can be no user interface without a data infrastructure to feed it—the team should ensure that their timelines are aligned. A sufficient budget should be allocated to the project to cover the team's time for both the implementation and the monitoring of the technology after the first version of the application is developed.

## Ethical Concerns

Developers of machine-learning applications should carefully consider the ethical implications of their use by judicial actors. Only technologies that aid human decision-making, rather than replacing it, should be adopted in the courts. There are multiple reasons for this recommendation. As noted earlier, algorithms have the “black-box” problem of interpretability—that is, it is not easy to trace the output of complex algorithms to the data inputs themselves. Additionally, biases in the decisions of judicial actors are reflected in the algorithm's training data and may be encoded into the algorithm itself. Thus, using machine-learning algorithms to inform judicial decisions without critical oversight raises the risk of replicating these biases elsewhere in the system. The inherent choice of performance metrics can also reinforce existing biases by decision-makers within the system. Addressing these issues requires a participatory and deliberative approach to the design, implementation, and evaluation of machine-learning technologies.

A reasonable demand to guarantee trust and fairness is that algorithms be interpretable. A judge may request the reason why a particular decision has been recommended by the algorithm. This transparency enhances judges' trust in the technology and allows for disagreement with its recommendations. Given the complexities of working with machine-learning algorithms, any rollout must be preceded by a phase of comprehensive study and rigorous testing of the systems themselves. Randomized controlled trials that carefully estimate the causal impacts of the adoption of these algorithms to properly evaluate their costs and benefits are essential. A carefully constructed trial can provide important benchmarks on cost, efficiency, user satisfaction, and impact on key performance metrics, all essential for a justice system to credibly serve citizens.

## CASE STUDY: MACHINE LEARNING FOR JUSTICE SYSTEMS IN INDIA

This case study illustrates how machine learning was implemented in the national justice system of India by the Data and Evidence for Justice Reform (DE JURE) team. Due to India's large population and volume of cases, justice officials are often unable to effectively manage cases in a timely fashion. India has just 19 judges per million people and 27 million (2.7 crore) pending cases (Amirapu 2020; Chemin 2012; Kumar 2012). To address this, the Indian justice system has made considerable advances in the adoption of information technology and has released large volumes of data to court users and encouraged them to use electronic systems. Yet legislative, institutional, and resource constraints have limited the full impact of these advances (Amirapu 2020; Damle and Anand 2020).

In this section, we describe how the DE JURE team implemented machine-learning applications in India. We first highlight the data infrastructure requirements for implementing machine-learning applications, as well as how these applications could enhance the functioning of the justice system.

### Judicial Data Infrastructure in India

In the past 15 years, considerable efforts have been made to adopt and deploy information technology systems in the courts of India. One of the most significant projects, the e-Courts project, was first launched in 2005 by the Supreme Court of India through the National Policy and Action Plan for Implementation of Information and Communication Technology (ICT) in the Indian Judiciary. The e-Courts initiative introduced technology into Indian courts in a variety of ways.

The most obvious feature of the system was the deployment of technology within courtrooms. Judges were provided with LCD touch screens, screens and projectors were connected via a local network to disseminate information to lawyers, and electronic boards at the courts displayed the queue of case numbers scheduled for hearing on a particular day. Outside the courtroom, e-filing procedures were established, and a data management architecture was created that ranged from the scanning of old cases into the electronic system to the creation of digital archives. The ICT plan also established direct electronic communication with litigants and an online case management system.

These investments eventually paved the way for the creation of the National Judicial Data Grid (NJDG), a database of 27 million cases that allows court users to view the status of pending cases and access information on past hearings. For the DE JURE team's goal to implement machine-learning tools, the most significant resources were the digital archives of cases. The team was able to scrape these publicly available digital archives to construct an e-Courts district court data set of 83 million cases from 3,289 court establishments.<sup>10</sup> They were able to curate details, like the act under which the case was filed, the case type (criminal or civil), the district where it originated, the parties to the case, and the history of case hearings, in a manner that made the data amenable to large-scale analysis.

A wider data ecosystem has been created by joining additional sources to the case data, including the following:

- **Data on judges:** To better understand the impact of specific judges—their identity, training, and experience—the team constructed a database of judges for the courts of India. They began this task by extracting data from editions of *The Handbook on Judges of the Supreme Court of India and the High Courts*, released by the Supreme Court of India, and appending to it information from various High Court websites. So far, the team has assembled detailed information for 2,239 judges from the handbooks for the years 2014–20. Most notably, 93.5 percent of these judges are men and 6.5 percent are women, and their range of experience covers a period spanning approximately 70 years.
- **Database of Central Acts:** This auxiliary data set is intended to give a definitive list of standardized act names. This could then be used to standardize the act names appearing in various cases. This would allow the team to analyze all cases filed under a given act. The team has, for example, examined all cases related to the Water Act of 1974 and found a total of 978 such cases at the Supreme Court and High Courts of India. The list of central (federal) acts can be viewed on the Legislative Department website of the Ministry of Law and Justice. There is currently no centralized source for all state legislation: this needs to be obtained from state websites separately.
- **Other administrative data:** Data on other institutions can be linked to the judicial data at the district as well as the state level. For example, data on Indian banks and their branches are available through the Reserve Bank of India. This database contains information on names, locations, license numbers, license dates, addresses, and other unique identifiers. The team has scraped, cleaned, and organized these data for further analysis. The database contains about 160,000 entries. The unique identifiers and location information allow the team to merge these data with banks appearing in litigation in courts that are present in the e-Courts databases. Merging these data with the legal data allows the team to examine a variety of interesting questions about the development of financial markets in an area, participation in the justice system, and the impacts of legal rulings.

The quality of these data varies significantly: there is no nationally standardized system for defining variables or reporting on them. For instance, in some states, the legal act name and section numbers are well delineated, but in other states, this is not the case. This makes it difficult to compare individual case types across courts and states (Damle and Anand 2020). There are no standardized identifiers within the data to follow a case through its potential sequence of appeals in higher courts. In a similar vein, there is no easy way to track a criminal case from its entry into the system as a first information report (FIR) to its exit as a judgment. There are inconsistencies in identifying information about participants, their attributes, and the types of laws or acts that cases relate to. There are also issues of incorrect reporting and spelling mistakes.

## Machine-Learning Applications in the Courts of India

The quality of data in India's justice system is often compromised: case data are incomplete or litigants' identities are not registered. The DE JURE team has constructed a robust data pipeline to collect often-incomplete judicial data, as well as machine-learning tools to clean and prepare them for analysis. In this section, we contextualize the problem: how data quality issues in judicial data manifest themselves in India. In the following section, we describe the solution: how machine-learning tools have been designed to enhance the quality of judicial data for analysis.

Legal data released by the Indian judiciary are voluminous, messy, and complex (Damle and Anand 2020). The typical case has clear tags for some key dates (filing date), the key actors (petitioner, respondent, and judges) and the court name, but information about the type of case, the outcome of deliberations, and pertinent acts cited is often not clearly identifiable in the textual body of the judgment. Cleaning and preprocessing the data is critical for any form of analysis, especially for supervised algorithms trained on these data. Traditional empirical legal studies have typically addressed this issue by relying on small-scale data sets in which legal variables are manually coded and the scope of inference is related to a small body of legal cases pertinent to a single issue (Baxi 1986; Gauri 2009; Krishnaswamy, Sivakumar, and Bail 2014).

These traditional approaches are unable to keep up with the incoming volume of cases. In this context, machine-learning tools provide an alternate approach to detecting errors or gaps in the data and correcting them in an automated fashion. Using machine learning, it is possible to infer the identities of participants even when these data are not registered. Additionally, laws used as precedents for a ruling can be identified through text analysis. Beyond the data quality itself, machine-learning approaches can help identify biases and discrimination in judges' rulings.

### *Inference about the Identities of Participants*

Some databases of judgments provide no identifying information about participants in the cases themselves. To better understand who participates in the courts, the team first extracts litigant names from the raw text of the judgments and then uses matching algorithms to identify the type of litigant (individual, company, or state institution). Classifying participants can be challenging. If the identification exercise involves government agencies, it is first necessary to compile all the different agencies of the state and national governments. Manually tagging entities is prohibitively time-consuming, but the existence of latent patterns in the names makes this fertile ground for machine-learning applications.

The machine-learning application relies on similarity across names for participants that belong to the same "group" to classify a particular name as belonging to that group. Using prelabeled data—individual name A belongs to group B—the machine-learning algorithm can extrapolate to unlabeled data, where an individual's name is available but not their group. Some obvious groups of interest are gender, caste, and religion, which are not recorded in judicial data but are available in other data sources. Another group of interest may be whether a participant is a government agency or not. We focus here on people's first and last names, for illustration.

The team first formats individuals' names to ensure that each individual can be identified by an honorific title, a first name, and a last name. Honorifics, such as Shri, Sri, Smt., Mr., Mrs., and Ms., enable the algorithm to directly identify an individual's gender. To extend this classification to names without an honorific, the team trains an algorithm on a publicly available corpus of labeled common Indian first names. Training this algorithm, often referred to as training a classifier, is the process by which the algorithm learns patterns within the data related to a group. Here, these patterns are the statistics of co-occurrence of letters in names, the lengths of names, and other features that allow the algorithm to determine whether a name indeed belongs to a particular group: in this case, a gender.<sup>11</sup>

These algorithms formalize intuitive notions of why a name belongs to a given group by identifying frequently occurring patterns within names associated with that group. Caste assignment is more complicated because the same last name could be associated with multiple caste groups. The name "Kumar," for example, could belong to a person belonging to the Scheduled Castes, the Scheduled Tribes, or the category "Other."



In the case of such names, the team generates distributions of the last name across the different caste categories. They then use this distribution to generate a prediction and combine this with the predictions of other models to ensure a robust prediction. They assign a caste to each household based on a simple majority vote among these models.

### **Identification of Laws and Acts**

Legal texts in India's justice system do not currently employ a standardized citation style for referring to acts or laws. For example, the Hindu Marriage Act may be referred to in a variety of ways, such as "u/s 13 clause 2 of Hindu Marriage Act," "u/s 13(b) Hindu Marriage Act," or "u/s 13 of Hindu Marriage Act 1995." Again, machine-learning tools can be used to address this issue.

In this project, the DE JURE team uses a set of tools that create mathematical representations of the text in the form of vectors. Term frequency-inverse document frequency (TF-IDF) is one popular method for representing a string of words as a numerical score that reflects how frequently a word is used within a given text and how infrequently it appears in the corpus. Applying this to act names, the team uses different clustering algorithms to group particular act citations based on how similar they are numerically. This approach groups the underlying act-name data in a manner that best preserves the coherence within groups (a particular act name) and the distance across groups to make the classification.

The identification of specific acts and how often they are cited opens new opportunities for legal analysis. The team can, for example, compare the different types of acts that are cited in the different courts within India's justice system. It can allow researchers and practitioners to identify the real-time evolution of legal citation—and legal thought—as judges refer to these acts.

### **Using Descriptive Analysis and Machine Learning to Identify Discrimination and Bias**

Bias and discrimination can occur in different areas of policy making, particularly when civil servants exercise discretion, such as in judicial rulings. Judges may favor or deny due process to plaintiffs depending on their ethnic or gender identity, undermining the rule of law and the right to impartial judgment. This challenge is, of course, not unique to judiciaries. A broad academic literature has demonstrated that bias in a human decision-maker can have conscious as well as unconscious drivers and may manifest in complex ways and in a variety of contexts that can be difficult to prove (Banerjee et al. 2009; Bertrand and Mullainathan 2004; Ewens, Tomlin, and Wang 2014; Kleinberg et al. 2018). In other settings, such as labor markets and educational institutions, algorithms—rules that take "inputs" (like the characteristics of a job applicant) and predict some outcome (like a person's salary)—have been shown to create new forms of transparency and to serve as methods for detecting discrimination (Kleinberg et al. 2020; LeCun, Bengio, and Hinton 2015).

Machine-learning algorithms can identify these biases and forms of discrimination, enabling governments to measure their prevalence and design policy changes to reduce them. In the courts of India, algorithms could help judges make critical decisions about cases (for example, dismissals or bail applications) and reduce bias in their rulings. Building such algorithms requires a rich data set that includes litigant characteristics (caste and gender), lawyer characteristics, court characteristics, case details (filing details and evidence provided), and case outcomes (such as the granting of bail or the dismissal of a case). A machine-learning engineer would develop a "learning procedure" that would aim to provide a predicted outcome from a broad range of inputs and modeling approaches, such as a neural network (Dayhoff 1990). These models differ from traditional statistical methods, such as linear regression, which are more deductive (presuming a linear fit between a few sets of variables) than inductive (allowing the data to report the best fit between a large set of variables).

These insights could be invaluable not only within the courtroom itself but also in judicial education. Experiments are currently underway, in the Judicial Academy of Peru, for example, to assess methods to improve case-based teaching by using the history of a judge's past decisions, which can reveal potential bias or error. The data are also suitable for creating personalized dashboards and interfaces that provide judges,



mediators, and other decision-makers with real-time information about their own performance relative to their own previous decisions and those of others who are comparable (Kling 2006). This information can be used to augment the capabilities of judges and lawyers, increase their productivity, and reduce potential bias in their decisions.

## BEYOND DESCRIPTIVE ANALYSIS: IMPACT EVALUATION IN THE JUSTICE SYSTEM

Moving beyond descriptive analysis and more correlational analysis of data, an underexplored field in the justice system is policy experiments for impact evaluation. Legal scholars and judges have long debated the merits of implementing various laws and regulations and have justified their arguments with theories about the effects of these legal rules. This situation resembles the field of medicine a century ago: before clinical trials, medical research focused on theoretical debates rather than rigorous causal evidence.

A growing body of empirical research now demonstrates that causal inference is possible in judicial studies. For example, in situations where cases are randomly assigned to judges, the random assignment itself can be used as an exogenous source of variation to evaluate the impact of judicial decisions. Since judges do not choose their cases, observed rulings reflect the judge's personal characteristics (ideological preferences or biases) and features of the case rather than the judicial process as a whole. Additionally, informational treatments can have an impact on the behavior of judges, improving their performance (box 16.2).

### BOX 16.2 Leveraging Data and Technology to Improve Judicial Efficiency in Kenya

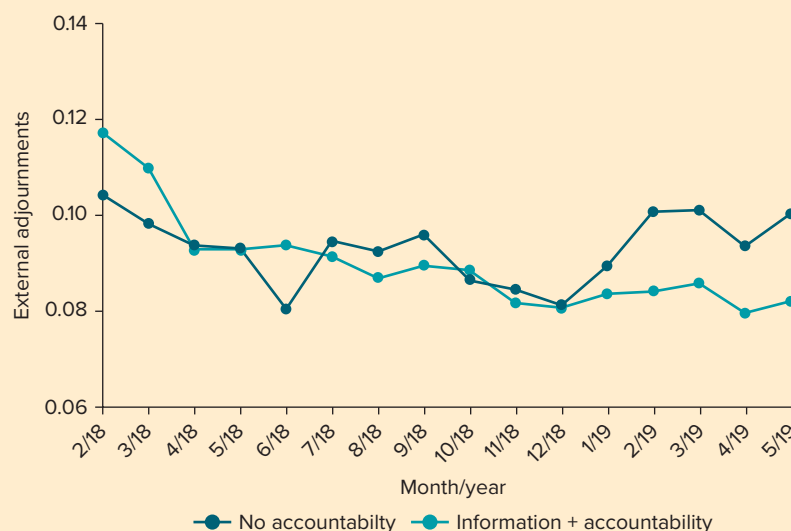
In partnership with the Judiciary of Kenya and McGill University, the World Bank's Development Impact Evaluation (DIME) Data and Evidence for Justice Reform (DE JURE) team has been leveraging underutilized administrative data to improve judicial efficiency. Through its case management system, the Judiciary of Kenya collects large amounts of administrative data on the characteristics of cases, the dates of hearings and reasons for adjournments, and other important metrics of court performance. These data are readily available for understanding and designing interventions to address challenges to the efficient delivery of justice, such as adjournments of hearings, which cause large delays in court proceedings. Despite the richness of these data, they were not being used for decision-making. DIME and the Judiciary of Kenya decided to leverage these data systems to design an algorithm identifying the greatest sources of delay for each court and presenting recommended actions. The team included performance information in a one-page feedback report. Then it studied whether this simplified, action-oriented information could reduce adjournments and improve judicial performance.

In a randomized controlled trial across all 124 court stations in Kenya, the team compared the impact of sharing the one-page feedback reports only with judges and supervisors to the impact of sharing them with Court User Committees as well, the latter acting as an additional accountability mechanism (figure B16.2.1). The team found that the one-page feedback report with the accountability mechanism reduced the number of adjournments by 20 percent over a four-month period and increased the number of cases resolved (Chemin et al. 2022). The conclusion was that the report was more effective when both tribunals and Court User Committees received it. Thus, sharing performance information with courts may be effective to improve efficiency, but it is particularly effective when this information is also shared with civil society and court stakeholders. This study served as proof of concept that utilizing data to provide information to judicial actors can reduce adjournments and increase the speed of justice, which have a downstream impact on the economic outcomes of citizens and firms.

*(continues on next page)*

## BOX 16.2 Leveraging Data and Technology to Improve Judicial Efficiency in Kenya (continued)

**FIGURE B16.2.1** Impact of One-Page Feedback Reports on Case Delays in Kenya



Source: DE JURE, World Bank.

Randomly assigning cases to judges predicted to be harsh or lenient allows researchers to identify the long-run causal impacts of the length of sentences (Dobbie and Song 2015). To identify the causal effect of a sentence length of eight months or eight years, a randomized controlled trial would need to randomize the sentence, which is impossible. However, assigning a defendant to a judge predicted to assign an eight-month sentence or to another judge predicted to assign an eight-year sentence allows researchers to identify the causal impact of sentence length on subsequent life outcomes.

The same conceptual framework can examine the causal effects of debt relief on individuals' earnings, employment, and mortality (Sampat and Williams 2019). This causal approach sheds light on the impact these judicial decisions can have on individuals' welfare outcomes. By applying machine learning to infer the bias, lenience, and ideological preference of a judge, researchers can identify the causal effect of these variables on the judicial system and the life outcomes of those affected by judges' decisions.

## CONCLUSION

In this chapter, we argue that machine learning is a powerful tool for improving public administration in general and the justice system in particular. Machine learning, at its core, emphasizes a methodological approach centered around a learning problem: defining indicators and using evidence to improve them. Under the umbrella of this methodological approach, multiple applications are available to tackle key issues in public administration. Algorithms can be written to draw inferences about the identities of participants and study the deliberative processes they employ within courtrooms. Machine-learning tools can also convert a high volume of textual data to

numerical estimates that can be used for understanding the processes and outcomes of different types of case data, including public procurement, taxes, and the systems of justice themselves.

These tools, however, have several limitations and requirements that need to be addressed before they can be effectively deployed in the courts. At the very outset, there are significant issues related to the privacy of personally identifiable information, security, and the control of legal data. Next, the algorithms require data preprocessing, training on large, high-frequency data sets, and iterative refinement with respect to the actual use cases in which they are deployed. This requires strong pilot programs that are studied as part of randomized controlled trials. Insights on data privacy and costs as well as outcomes require that these pilots be constructed on a reasonable scale.

Public administration officials often execute a range of tasks, from the ordinary to the complex, such as the adjudication of a trial. The smart application of machine learning can enhance levels of automation, productivity, and the level of information extracted from data generated in the public sector. If done right, it can help reduce noise—and this can be one step toward aiding the impersonal execution of tasks, reducing bias, enhancing predictability, and making decision rules more transparent. But none of these outcomes can be presupposed from the machine-learning approach: they depend on the ethical framework and operational relevance underlying its implementation. Machine-learning practitioners are therefore advised to take necessary precautions and develop solutions that are accountable to the public and useful for government officials.

## NOTES

1. Machine learning is therefore a methodological approach anchored in a learning framework. It is not a radical departure from classical approaches to statistics and, in fact, often builds on canonical models (such as linear or logistic regressions), nor is it exempt from well-known challenges, such as bias and model misspecification.
2. The machine-learning application is part of a broader study on bureaucratic allocation available in Bandiera et al. (2021).
3. For a discussion of causality in machine learning, see Schölkopf (2022).
4. Public servant surveys, due to their smaller sample frames and limited applications in prediction, are rarely used directly for machine-learning applications.
5. Federal transfers have been analyzed using machine learning since 2018 (CGU 2018).
6. The Inter-American Development Bank has written extensively on the topic. See chapter 3 of Porrúa et al. (2021).
7. For a concrete example of this approach, see case study 2 in chapter 9 of the *Handbook* on the Brazilian federal government's experience with the development of machine learning for payroll analysis).
8. For an overview, see chapter 6 of the *Handbook*.
9. For a discussion, see Yong (2018).
10. The e-Courts data are public and can be accessed via the district court websites, the e-Courts Android/iOS app, or the district court services webpage at [https://services.ecourts.gov.in/ecourtindia\\_v6/](https://services.ecourts.gov.in/ecourtindia_v6/).
11. To reduce model overfitting, we use the majority vote from multiple trained classifiers, including a logistic regression model and a random forest classifier to make predictions on gender. A logistic regression models the probability of a binary outcome or event. A random forest classifier will use decision trees (nested if-then statements) on features of the data to make the prediction. We have also made predictions of religion and caste using similar approaches. Muslims can be recognized in the data through the distinctiveness of Muslim names: common names such as Khan and Ahmed can easily be assigned and coded, but for others, we utilize the occurrence of specific letters (such as *q* and *z*) through appropriate classifiers to identify additional names.

## REFERENCES

- Amirapu, Amrit. 2020. "Justice Delayed Is Growth Denied: The Effect of Slow Courts on Relationship-Specific Industries in India." *Economic Development and Cultural Change* 70 (1): 415–51. <https://doi.org/10.1086/711171>.
- Athey, Susan. 2017. "Beyond Prediction: Using Big Data for Policy Problems." *Science* 355 (6324): 483–85. <https://doi.org/10.1126/science.aal4321>.

- Bandiera, Oriana, Michael Carlos Best, Adnan Qadir Khan, and Andrea Prat. 2021. "The Allocation of Authority in Organizations: A Field Experiment with Bureaucrats." *The Quarterly Journal of Economics* 136 (4): 2195–242. <https://doi.org/10.1093/qje/qjab029>.
- Banerjee, Abhijit, Marianne Bertrand, Saugato Datta, and Sendhil Mullainathan. 2009. "Labor Market Discrimination in Delhi: Evidence from a Field Experiment." *Journal of Comparative Economics* 37 (1): 14–27. <https://doi.org/10.1016/j.jce.2008.09.002>.
- Baxi, Upendra. 1986. *Towards a Sociology of Indian Law*. New Delhi: Satvahan.
- Berdej6, Carlos, and Daniel L. Chen. 2017. "Electoral Cycles among US Courts of Appeals Judges." *The Journal of Law and Economics* 60 (3): 479–96. <https://doi.org/10.1086/696237>.
- Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* 94 (4): 991–1013. <https://doi.org/10.1257/0002828042002561>.
- Bhushan, Prashant. 2009. "Misplaced Priorities and Class Bias of the Judiciary." *Economic and Political Weekly* 44 (14): 32–37. <https://www.jstor.org/stable/40278698>.
- Brynjolfsson, Erik, and Andrew McAfee. 2017. "The Business of Artificial Intelligence." *Harvard Business Review*, July 18, 2017. <https://hbr.org/2017/07/the-business-of-artificial-intelligence>.
- CGU (Controladoria-Geral da Uni6o). 2018. "Intelig6ncia Artificial Analisar6 Presta6o de contas em Transfer6ncias da Uni6o." Comptroller General of Brazil, October 23, 2018. <https://www.gov.br/cgu/pt-br/assuntos/noticias/2018/10/inteligencia-artificial-analisara-prestacao-de-contas-em-transferencias-da-uniao>.
- Chemin, Matthieu. 2012. "Does Court Speed Shape Economic Activity? Evidence from a Court Reform in India." *The Journal of Law, Economics, & Organization* 28 (3): 460–85. <https://doi.org/10.1093/jleo/ewq014>.
- Chemin, Matthieu, Daniel L. Chen, Vincenzo Di Maro, Paul Kimalu, Momanyi Mokaya, and Manuel Ramos-Maqueda. 2022. "Data Science for Justice: The Short-Term Effects of a Randomized Judicial Reform in Kenya." TSE Working Paper 22-1391, Toulouse School of Economics, Toulouse.
- Chen, Daniel L. 2019a. "Judicial Analytics and the Great Transformation of American Law." *Artificial Intelligence and Law* 27: 15–42. <https://doi.org/10.1007/s10506-018-9237-x>.
- Chen, Daniel L. 2019b. "Machine Learning and the Rule of Law." In *Law as Data: Computation, Text, and the Future of Legal Analysis*, edited by Michael A. Livermore and Daniel N. Rockmore, 433–41. Santa Fe, NM: Santa Fe Institute Press. <https://doi.org/10.37911/9781947864085.16>.
- Chen, Daniel L., Tobias J. Moskowitz, and Kelly Shue. 2016. "Decision Making under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires." *The Quarterly Journal of Economics* 131 (3): 1181–242. <https://doi.org/10.1093/qje/qjw017>.
- Chen, Jonathan H., and Steven M. Asch. 2017. "Machine Learning and Prediction in Medicine—Beyond the Peak of Inflated Expectations." *New England Journal of Medicine* 376: 2507–09. <https://doi.org/10.1056/NEJMp1702071>.
- Cross, Tim. 2020. "Artificial Intelligence and Its Limits: An Understanding of AI's Limitations Is Starting to Sink In." *Economist*, Technology Quarterly, June 13, 2020. <https://www.economist.com/technology-quarterly/2020/06/11/an-understanding-of-ais-limitations-is-starting-to-sink-in>.
- Damle, Devendra, and Tushar Anand. 2020. "Problems with the e-Courts Data." National Institute of Public Finance and Policy Working Paper 314, National Institute of Public Finance and Policy, New Delhi, India. [https://www.nipfp.org.in/media/medialibrary/2020/07/WP\\_314\\_\\_2020.pdf](https://www.nipfp.org.in/media/medialibrary/2020/07/WP_314__2020.pdf).
- Dayhoff, Judith E. 1990. *Neural Network Architectures: An Introduction*. New York: Van Nostrand Reinhold.
- Devanesan, Joe. 2020. "AI-Powered Traffic Management Is Slashing Asia's Congestion Problem." *Techwire Asia*, August 28, 2020. <https://techwireasia.com/2020/08/ai-powered-traffic-management-is-slashing-asias-congestion-problem>.
- Dobbie, Will, and Jae Song. 2015. "Debt Relief and Debtor Outcomes: Measuring the Effects of Consumer Bankruptcy Protection." *American Economic Review* 105 (3): 1272–311. <https://doi.org/10.1257/aer.20130612>.
- Economist*. 2017. "The World's Most Valuable Resource Is No Longer Oil, but Data." May 6, 2017. <https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data>.
- Engstrom, David Freeman, Daniel E. Ho, Catherine M. Sharkey, and Mariano-Florentino Cu6llar. 2020. *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies*. Washington, DC: Administrative Conference of the United States. <https://www.acus.gov/research-projects/agency-use-artificial-intelligence>.
- Epps, Willie J. Jr., and Jonathan M. Warren. 2020. "Artificial Intelligence: Now Being Deployed in the Field of Law." *The Judges' Journal* 59 (1): 16–39. [https://www.americanbar.org/groups/judicial/publications/judges\\_journal/2020/winter/artificial-intelligence-now-being-deployed-the-field-law/](https://www.americanbar.org/groups/judicial/publications/judges_journal/2020/winter/artificial-intelligence-now-being-deployed-the-field-law/).
- Ewens, Michael, Bryan Tomlin, and Liang Choon Wang. 2014. "Statistical Discrimination or Prejudice? A Large Sample Field Experiment." *The Review of Economics and Statistics* 96 (1): 119–34. <http://www.jstor.org/stable/43554917>.

- Fazekas, Mihály, Salvatore Sberna, and Alberto Vannucci. 2021. "The Extra-Legal Governance of Corruption: Tracing the Organization of Corruption in Public Procurement." *Governance: An International Journal of Policy, Administration, and Institutions* 35 (4): 1139–61. <https://doi.org/10.1111/gove.12648>.
- Galanter, Marc. 1984. *Competing Equalities: Law and the Backward Classes in India*. Oxford: Oxford University Press.
- Gauri, Varun. 2009. "Public Interest Litigation in India: Overreaching or Underachieving?" Policy Research Working Paper 5109, World Bank, Washington, DC. <https://doi.org/10.1596/1813-9450-5109>.
- Harwell, Drew, and Craig Timberg. 2021. "How America's Surveillance Networks Helped the FBI Catch the Capitol Mob." *Washington Post*, April 2, 2021. <https://www.washingtonpost.com/technology/2021/04/02/capitol-siege-arrests-technology-fbi-privacy/>.
- Jordan, M. I., and T. M. Mitchell. 2015. "Machine Learning: Trends, Perspectives, and Prospects." *Science* 349 (6245): 255–60. <https://doi.org/10.1126/science.aaa8415>.
- Kannabiran, Kalpana. 2009. "Judicial Meanderings in Patriarchal Thickets: Litigating Sex Discrimination in India." *Economic and Political Weekly* 44 (44): 88–98. <https://www.jstor.org/stable/25663738>.
- Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein. 2018. "Discrimination in the Age of Algorithms." *Journal of Legal Analysis* 10: 113–74. <https://doi.org/10.1093/jla/laz001>.
- Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein. 2020. "Algorithms as Discrimination Detectors." *Proceedings of the National Academy of Sciences* 117 (48): 30096–100. <https://doi.org/10.1073/pnas.191279011>.
- Kling, Jeffrey R. 2006. "Incarceration Length, Employment, and Earnings." *American Economic Review* 96 (3): 863–76. <https://www.jstor.org/stable/30034076>.
- Krishnaswamy, Sudhir, Sindhu K. Sivakumar, and Shishir Bail. 2014. "Legal and Judicial Reform in India: A Call for Systemic and Empirical Approaches." *Journal of National Law University Delhi* 2 (1): 1–25. <https://doi.org/10.1177/2277401720140101>.
- Kumar, Vandana Ajay. 2012. "Judicial Delays in India: Causes & Remedies." *Journal of Law, Policy & Globalization* 4: 16–21. <https://www.iiste.org/Journals/index.php/JLPG/article/view/2069>.
- Lal, Niharika. 2021. "How Traffic Cameras Issue E-Challans." *Times of India*, April 17, 2021. <https://timesofindia.indiatimes.com/city/delhi/how-traffic-cameras-issue-e-challans/articleshow/82103731.cms>.
- LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521: 436–44. <https://doi.org/10.1038/nature14539>.
- Misuraca, Gianluca, and Colin van Noordt. 2020. *AI Watch: Artificial Intelligence in Public Services*. EUR 30255 EN. Luxembourg: Publications Office of the European Union. <https://doi.org/10.2760/039619>.
- Mullainathan, Sendhil, and Ziad Obermeyer. 2022. "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care." *The Quarterly Journal of Economics* 137.2 (May): 679–727. <https://doi.org/10.1093/qje/qjab046>.
- Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.
- Porrua, Miguel, Mariano Lafuente, Edgardo Mosqueira, Benjamin Roseth, and Angela María Reyes, eds. 2021. *Digital Transformation and Public Employment: The Future of Government Work*. Washington, DC: Inter-American Development Bank. <https://publications.iadb.org/publications/english/document/Digital-Transformation-and-Public-Employment-The-Future-of-Government-Work.pdf>.
- Ramos-Maqueda, Manuel, and Daniel L. Chen. 2021. "The Role of Justice in Development: The Data Revolution." Policy Research Working Paper 9720, World Bank, Washington, DC. <https://openknowledge.worldbank.org/handle/10986/35891>.
- Rigano, Christopher. 2018. "Using Artificial Intelligence to Address Criminal Justice Needs." National Institute of Justice, October 8, 2018. <https://nij.ojp.gov/topics/articles/using-artificial-intelligence-address-criminal-justice-needs#citation--0>.
- Sadka, Joyce, Enrique Seira, and Christopher Woodruff. 2017. "Overconfidence and Settlement: Evidence from Mexican Labor Courts." Unpublished manuscript. [http://www.enriqueseira.com/uploads/3/1/5/9/31599787/overconfidence\\_and\\_settlement\\_preliminary.pdf](http://www.enriqueseira.com/uploads/3/1/5/9/31599787/overconfidence_and_settlement_preliminary.pdf).
- Sampat, Bhaven, and Heidi L. Williams. 2019. "How Do Patents Affect Follow-On Innovation? Evidence from the Human Genome." *American Economic Review* 109 (1): 203–36. <https://doi.org/10.1257/aer.20151398>.
- Schölkopf, Bernhard. 2022. "Causality for Machine Learning." In *Probabilistic and Causal Inference: The Works of Judea Pearl*, edited by Hector Geffner, Rina Dechter, and Joseph Y. Halpern, 765–804. New York: Association for Computing Machinery. <https://doi.org/10.1145/3501714.3501755>.
- Waldrop, M. Mitchell. 2018. "Free Agents." *Science* 360 (6385): 144–47. <https://doi.org/10.1126/science.360.6385.144>.
- Yong, ed. 2018. "A Popular Algorithm Is No Better at Predicting Crimes Than Random People." *Atlantic*, January 17, 2018. <https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/>.

## CHAPTER 17

# Government Analytics Using Data on Task and Project Completion

*Imran Rasul, Daniel Rogger, Martin Williams, and  
Eleanor Florence Woodhouse*

### SUMMARY

Much government work consists of the completion of tasks, from creating major reports to undertaking training programs and procuring and building infrastructure. This chapter surveys a range of methods for measuring and analyzing task completion as a measure of the performance of government organizations, giving examples of where these methods have been implemented in practice. We discuss the strengths and limitations of each approach from the perspectives both of practice and research. While no single measure of task completion provides a holistic performance metric, when used appropriately, such measures can provide a powerful set of insights for analysts and managers alike.

### ANALYTICS IN PRACTICE

- Much government activity can be conceived as discrete tasks: bounded pieces of work with definite outputs. Public sector planning is often organized around the achievement of specific thresholds; the completion of planning, strategy, or budgetary documents; or the delivery of infrastructure projects. *Task completion* is a useful conception of government activity because it allows analysts to assess public performance in a standardized way across organizations and types of activity.
- Assessing government performance based solely on the passing of legislation or the delivery of frontline services misses a substantial component of government work. Using a task completion approach pushes analysts to better encapsulate the breadth of work undertaken by public administration across government. It thus pushes analysts to engage with the full set of government tasks.

---

Imran Rasul is a professor in the Department of Economics, University College London. Daniel Rogger is a senior economist in the World Bank's Development Impact Evaluation (DIME) Department. Martin Williams is an associate professor in the Blavatnik School of Government, University of Oxford. Eleanor Florence Woodhouse is an assistant professor in the Department of Political Science and School of Public Policy, University College London.



- A task completion approach also allows for the investigation of which units and organizations are most likely to initiate, make progress on, and complete tasks. Though not a full picture of government work—it is complementary to the analysis of process quality or sector-specific measures of quality, for example—it allows for a rigorous approach to comparisons frequently made implicitly in budgetary and management decisions.
- Collecting data across projects on determinants of progress, such as overruns, and matching them to input data, such as budget disbursements, allows for a coherent investigation of the mechanisms driving task progress across government or within specific settings.
- Attempting to assess task completion in a consistent way across government is complicated by the fact that tasks vary in nature, size, and complexity. By collecting data on these features of a task, analysts can go some way toward alleviating concerns over the variability of the tasks being considered. For example, analysis can be undertaken within particular types of task or size, and complexity can be conditioned on in any analysis. An important distinction in the existing literature is how to integrate the analysis of tasks related to the creation of physical infrastructure and tasks related to administration.

## INTRODUCTION

A fundamental question for government scholars and practitioners alike is whether governments are performing their functions well. What these functions are and what performing “well” means in practice are complex issues in the public sector, given the diverse tasks undertaken and their often indeterminate nature. Despite the importance of these questions, there is little consensus as to how to define government effectiveness in a coherent way across the public service or how to measure it within a unified approach across governments’ diverse task environments (Rainey 2009; Talbot 2010). Such considerations have practical importance because government entities, such as political oversight or central budget authorities, frequently have to make implicit comparisons between the relative functioning of public agencies. For example, when drawing up a budget, public sector managers must make some comparison of the likely use of funds across units and whether these funds will eventually result in the intended outputs of those units, however varied the tasks are in scope. From an analytical perspective, the more comprehensive a measure of government functioning, the greater the capacity of analytical methods to draw insights from the best-performing parts of government.

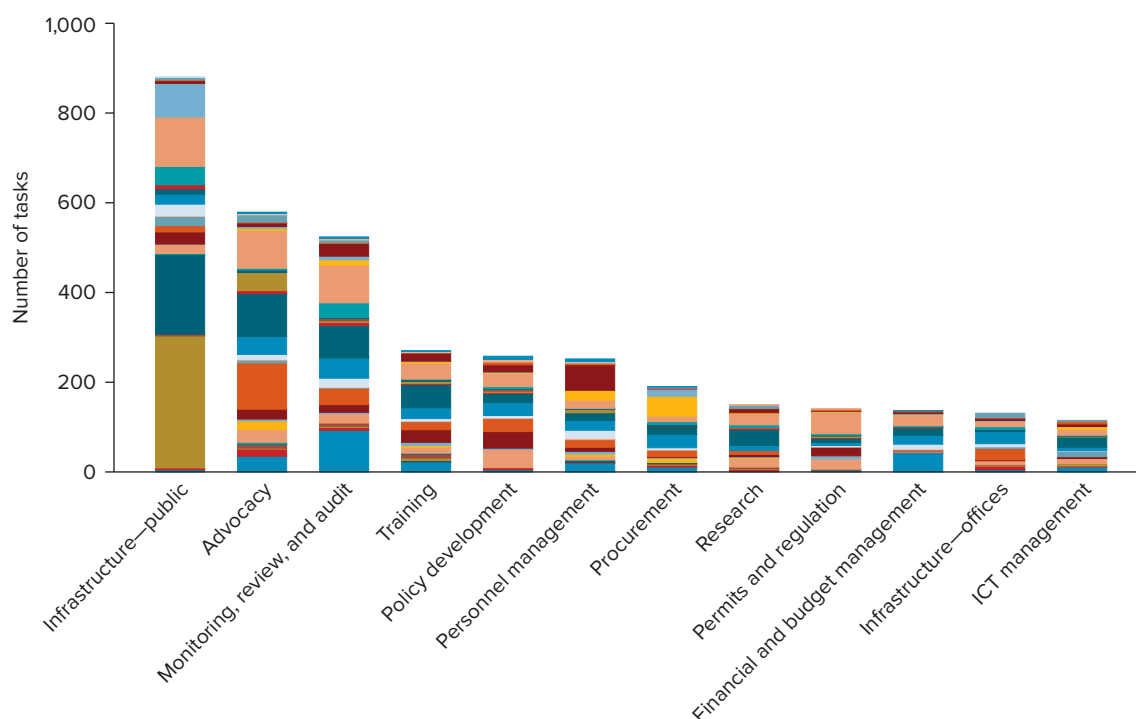
Much government activity can be conceived as discrete *tasks*: bounded pieces of work with definite outputs. Public sector planning is often organized around the achievement of specific thresholds; the completion of planning, strategy, or budgetary documents; or the delivery of projects. Government *projects* are also often conceived as bounded activities with definite outputs but frequently encompass multiple tasks within a wider conception of completion. *Task completion* (or project completion) is thus a useful conception of government activity—however that activity is conceived—because it allows analysts to assess public performance in a standardized way across organizations and types of activity. This kind of assessment contrasts with continuous regulatory monitoring, the assessment of the stability of citizens’ access to frontline services, and the equity of activities related to redistribution, which are better understood as the evaluation of ongoing processes. In this chapter, we propose a way to leverage data on task completion to assess the government’s effectiveness across its diverse task environments and learn from related analysis. We argue that by utilizing a unified framework for task completion, analysts can assess whether a government or government agency “does well what it is supposed to do, whether people. . . work hard and well, whether the actions and procedures of the agency and its members help achieve its mission, and, in the end, whether it actually achieves its mission” (Lee and Whitford 2009, 251; paraphrasing Rainey and Steinbauer 1999).

Task comparison is useful for three core reasons. First, task completion is a concept that can be applied across much government work, thus allowing for a broad consideration of government functioning.

We believe that by working with a task completion framework, analysts can gain a fuller and more accurate picture of the functions of government that reflects the full range of government activities—from human resource management to policy definition, infrastructure planning and implementation, service delivery, and audit and evaluation. We know little about the full distribution of tasks that public administrators undertake. As shown in figure 17.1, the few studies that do apply a task completion framework find that administrators undertake a vast range of activities—from advocacy to auditing and monitoring to planning—that go well beyond infrastructure and service delivery, the activities that are usually considered in the academic literature.

Figure 17.1 displays the frequency of the most prevalent tasks undertaken by Ghanaian public officials in their daily duties. Infrastructure provision for the public (rather than upgrading government facilities) is the most common activity, partly motivating our particular attention to it in this chapter. However, the figure indicates the broad diversity of tasks undertaken by the public service. The distinct colors in each bar of the histogram indicate different organizations undertaking that type of task. Thus, it is clear that each type of task is undertaken by many different organizations. Second, a task completion framework pushes analysts to think carefully about the characteristics of each of the tasks assessed. We define projects above as collections of tasks; an obvious question is how to apply boundaries to tasks or projects uniformly across government. There is very limited research on the characteristics of the tasks undertaken by public administrators and how to assess whether they are being undertaken adequately. The task completion framework pushes analysts to think in detail about the activities that administrators engage in and how successfully they do so. That is to say, they must think not only about whether a bridge is completed but what the full conception of the bridge project is, whether the bridge was of a complex design that was hard to implement, whether the quality of the implementation of the bridge is adequate, whether it was completed within a reasonable time frame given the complexity of the project, and so on.

**FIGURE 17.1 Task Types across Organizations**



Source: Rasul, Rogger, and Williams 2021.

Note: The task type classification refers to the primary classification for each output. Each color in a column represents an organization implementing tasks of that type, but the same color across columns may represent multiple organizations. Figures represent all 30 organizations with coded task data. ICT = information and communication technology.

Third, by creating comparators from across government, an integrated measurement approach yields analytical benefits that more than make up for the losses from abstraction for many types of analysis. Analysts can investigate the determinants of successful task completion from a large sample, with varying management environments, buffeted by differential shocks, and on which a greater range of statistical methods can be effectively applied. The task completion framework has the advantage of capturing a wide range of activities and being comparable across departments. Thus, the framework can pool task types to allow analysts to draw conclusions about government effectiveness more broadly, rather than, for example, allowing only for inferences about a specific type of task (for example, the delivery of a specific service, such as passport processing times, versus a range of government tasks that are indicative of administrative effectiveness more broadly). By leveraging these data—on the nature, size, and complexity of tasks—analyses can be undertaken within particular types of tasks—for example, distinguishing between the completion of physical and nonphysical outputs. Taking the analysis a step further, having measures of task completion that are comparable across teams or organizations can enable researchers to identify the determinants of task completion—although this entails its own methodological challenges and is beyond the scope of this measurement-focused chapter.

Much of the existing literature that seeks to describe how well governments perform their functions focuses on upstream steps in the public sector production process (such as budgetary inputs or processes), which are useful for management (but not so relevant to the public) (Andrews et al. 2005; Lewis 2007; Nistotskaya and Cingolani 2016; Rauch and Evans 2000). Or it focuses on final outcomes (such as goods provided or services delivered), which are relevant to the public (but not always so useful for management) (Ammons 2014; Boyne 2003; Carter, Klein, and Day 1992; Hefetz and Warner 2012). The completion of tasks and projects falls between these two approaches: it is useful for management and relevant to the public. The task completion framework helps analysts address the gap between inputs and final outcomes in terms of how they measure government performance. It gives analysts a way to engage with the full distribution of government tasks and to assess the characteristics of the tasks themselves. The task approach outlined in this chapter is closely aligned with the discussion in chapter 15 of *The Government Analytics Handbook*. There, the relevant task is the processing of an administrative case. Clearly, there are other types of tasks in government, and this chapter aims to present a framework that can encapsulate them all. Given the scale of case processing in government, however, chapter 15 presents a discussion specific to that type of task. Related arguments can be made for chapter 12 on procurement, chapter 14 on customs, and chapter 29 on indicators of service delivery. Across these chapters, the *Handbook* provides discussions of the specific analytical opportunities afforded by different types of government activity. These chapters contain some common elements, such as discussions of some form of complexity in relation to the task under focus. This chapter showcases the considerations required for an integrated approach across task types.

Through the task completion framework, we aim to encourage practitioners and scholars alike to conceive of government activity more broadly and to leverage widely available data sources, such as government progress reports or independent public expenditure reviews, to do so. As well as being widely available, these kinds of objective data are highly valuable because they usually cover a wide range of different task types performed by numerous different agencies and departments.

This chapter continues as follows. First, we conceptualize government work as task completion. Second, we use tasks related to the creation of physical infrastructure to illustrate the task completion framework. Third, we show how the framework applies to other types of tasks. Fourth, we explore how to measure task characteristics (considering the complexity of tasks and their ex ante and ex post clarity). Fifth, we discuss key challenges in integrating these measures with each other and into management practice. Finally, we conclude.

## CONCEPTUALIZING GOVERNMENT WORK AS THE COMPLETION OF TASKS

Much of the literature in public administration has focused on how to measure government effectiveness by relying on the tasks of single agencies (Brown and Coulter 1983; Ho and Cho 2017; Lu 2016), on a set of

agencies undertaking the same task (Fenizia 2022), or on a broad conception of the central government as a single entity (Lee and Whitford 2009). These approaches limit analysis to a single conception of government effectiveness, which in the case of a single agency or sector, can be precisely defined. However, almost by definition, this limits analysis to a subset of government work and thus raises concerns over what such analysis tells us about government performance as a whole or how performance in one area of government affects other areas.

In addition to measuring government effectiveness on the basis of a partial vision of government, many studies that have sought to investigate government effectiveness have relied on perception-based measures of effectiveness, based on the evaluations of either government employees or external stakeholders and experts (Poister and Streib 1999; Thomas, Poister, and Ertas 2009; Walker et al. 2018). Such measures are frequently available only at an aggregate or even country level because of how distant these individuals are from actual government tasks, and they frequently assess not the outputs of those tasks directly but some perception of “general effectiveness.”<sup>1</sup>

Objective measures of government functioning have frequently been eschewed because of obstacles related to data availability, their purported inability to capture the complexity of government work, or conflicting understandings of what effectiveness means. However, many government agencies produce their own reports on the progress they have made across the full distribution of their work. Similarly, agencies often have administrative data on the totality of their activities that provide quantities related to the complexity of task completion that can be repurposed for analytics. These data are collected for management and reporting purposes as part of the daily duties of agency staff. These reports frequently contain characteristics of the tasks undertaken and progress indicators outlining how far tasks have progressed. These reports can be the basis of an integrated analysis of government functioning.

For tasks related to physical infrastructure and administration, analysts can use quantities from these reports, or similar primary data collection, to conceptualize government work in a unified task completion framework. The following discussion of the strengths and limitations or challenges of such an approach focuses on a small set of research papers that have applied a task completion framework to the assessment of government functioning. It thus aims to illustrate the utility of the task completion framework rather than being in any way comprehensive. Where relevant, we provide a number of examples of how public officials have taken a similar approach.

We rely on two simple definitions throughout the chapter. First, a *task* is the bounded activity for which a given organization, team, or individual in the government is responsible. Second, an *output* is the final product a government organization, team, or individual delivers to society. An output is the result of a successful task. In government performance assessment, outputs are defined as “the goods or services produced by government agencies (e.g., teaching hours delivered, welfare benefits assessed and paid).”<sup>2</sup> An example of a government task might be developing a draft competition policy or organizing a stakeholder meeting to validate the draft competition policy (Rasul, Rogger, and Williams 2021, appendix). The corresponding outputs would be the draft competition policy itself and the holding of the stakeholder meeting. These tasks are usually repeated and are completed within varying time frames, depending on the complexity and urgency of the activity at hand.

More granular guidance on how to define a task is challenged by the fact that the appropriate conception of a task will vary by the focus of the analysis. However, to illustrate common conceptions, some examples from the analyses that will be discussed in this chapter include the design, drilling, and development of a water well (including all taps linked to a single source of water); the design, construction, and finishing of a school; the renovation of a neighborhood sewage system; a full maintenance review and associated activities, such as resurfacing, to bring a road up to a functioning state as determined by local standards; the development of a new public health curriculum for primary school students; and the updating of a human resources management information system with current personnel characteristics for all health-related agencies.

By conceiving all government activity as consisting of tasks with intended outputs, analysts can construct a standardized measure of government performance and can gather multiple tasks together to assess government performance across teams within an organization, across organizations, and over time.

Government performance can be defined as the frequency with which particular government actors are able to produce outputs from corresponding tasks. We now turn to considerations in the definition of a task or project and an output in the case of physical and nonphysical outputs.

## Physical Outputs

We first consider a task completion framework as it pertains to the accomplishment of physical infrastructure, or, more precisely, tasks relating to the production of physical outputs. In lower-middle-income countries in particular, the noncompletion of infrastructure projects is a widespread and costly phenomenon, with recent estimates suggesting that over one-third of the infrastructure projects started in these countries are not completed (Rasul and Rogger 2018; Williams 2017).

We focus on task completion measures developed from coding administrative data that are at least somewhat comparable across organizations and can be implemented at scale, rather than on performance audits of specific programs (for example, by national audit offices or international financial institutions' internal performance reports) or on the evaluation of performance against key performance indicators (for example, in leadership performance contracts or through central target-setting mechanisms). Many governments or government agencies have infrastructure-project-tracking databases (either electronic or in paper-based files). These records may be for implementation management, for budgeting and fiduciary reasons, or for audit and evaluation. These databases keep records of how far physical projects have been implemented relative to their planned scope.

For example, in Nigeria, Rasul and Rogger (2018) use independent engineering assessments of thousands of projects from across the government implemented by the Nigerian public service to assess the functioning of government agencies. They complement this with a management survey in the agencies responsible for the projects and examine how management practices matter for the completion rates of projects. The analysis exploits a specific period in the Nigerian public service when “the activities of public bureaucracies were subject to detailed and independent scrutiny” (2) and a special office was set up to track the quality of the project implementation of a broad subset of government activities. This was due to an effort by the presidency to independently verify the status of many of the public infrastructure projects funded by the proceeds of debt relief and implemented by agencies across the federal government. The records of this tracking initiative allowed the authors to quantify both the extent of project implementation and the assessment of the quality of the public goods provided.

A second application of the task completion framework to an empirical setting examining physical outputs is provided by Williams (2017), who collects, digitizes, and codes district annual progress reports in Ghana. These reports, which are written annually by each district's bureaucracy and submitted to the central government, include a table listing basic information about projects that were ongoing or active during the calendar year. Such reports are widely produced but not frequently available in a digital format or used for government analytics. The potential of these data for useful insights into government performance is great. Williams uses the reports on physical projects to examine the determinants of noncompletion, presenting evidence that corruption and clientelism are not to blame but rather a dynamic collective action process among political actors facing commitment problems in contexts of limited resources.

Similarly, Bancelari (2022) uses district administrative data on sewerage projects in Peru to explore the social costs of unfinished projects. She uses a combination of mortality statistics, viability studies, annual budget reports on sewerage projects (which allow her to identify unfinished and completed projects), spatial topography data, and population data in order to provide evidence that infant mortality and under-five mortality increase with increases in unfinished sewerage projects. She also finds that mayors who are better connected to the national parliament are able to complete more projects.

Beyond using administrative data, analysts have also undertaken primary fieldwork to explore the completion of physical projects. For example, Olken (2007) uses various surveys on villages, households, individuals, and the assessments of engineering experts to investigate the level of corruption involved in building roads in Indonesia. Olken is able to produce a measure of corruption in terms of missing expenditures by



calculating discrepancies between official project costs and an independent engineer's estimate of costs defined by the survey responses. Primary field activity also allows analysts to undertake randomized controlled trials of potential policies to improve government functioning. In the case of Olken (2007), randomized audits of villages are used to estimate the effect of top-down monitoring on the quality of government outputs: in this case, the building of roads. Such a research design and measure are highly valuable and capture a very important feature of government activity, although they come at a high cost in terms of the resources needed to capture these government tasks.

Other papers have studied the maintenance rather than the construction of physical outputs. In these cases, task completion is the effective continuation of physical outputs. Once again using primary fieldwork to collect required data, Khwaja (2009) uses survey team site visits and household surveys to measure the maintenance of infrastructure projects in rural communities in northern Pakistan (Baltistan) as a form of task completion. Maintenance here is measured through surveys of expert engineers who assess the maintenance of infrastructure projects in terms of their physical state (that is, how they compare to their initial condition), their functional state (that is, the percentage of the initial project purpose satisfied), and their maintenance-work state (that is, the percentage of required maintenance that needs to be carried out). Khwaja (2009) uses these data to examine whether project design can improve collective success in maintaining local infrastructure. The paper presents within-community evidence that project design makes a difference to maintenance levels: "designing projects that face fewer appropriation risks through better leadership and lower complexity, eliciting greater local information through the involvement of community members in project decisions, investing in simpler and existing projects, ensuring a more equitable distribution of project returns, and emulating NGOs can substantially improve project performance even in communities with low social capital" (Khwaja 2009, 913).

We have seen several examples of "government analytics" that seek to measure the completion rate of tasks related to the provision (or maintenance) of physical outputs. From Nigerian federally approved social sector projects, such as providing dams, boreholes, and roads, to Indonesian road building, analysts have defined measures of task completion based on physical outputs. The analysis has used administrative data, existing household surveys, and primary fieldwork (sometimes in combination with one another) to generate insights into the determinants of government functioning.

These papers measure task completion in a series of different ways that all aim to capture the underlying phenomenon of what share of the intended outputs are completed. But there are important commonalities to their approaches. First, the definition of a task or project is determined by a common, or consensus, engineering judgment that crosses institutional boundaries. Thus, though a ministry of urban development may bundle the creation of multiple water distribution points, the building of a health center, and road repaving into a single "slum upgrading" project, the analysts discussed above split these groupings into individual components that would be recognizable across settings, and thus across government. A water distribution point will be conceived as a discrete task whether it is a component of a project in an agriculture, education, health, or water infrastructure project. The wider point is that an external conception of what makes up a discrete activity, such as the common engineering conception of a water distribution point, provides discipline on the boundaries of what is conceived as a single task for any analytical exercise.

Second, within these conceptions of projects, an externally valid notion of completion and progress can be applied. For example, the threshold for a water distribution point is that it produces a sufficient flow of water over a sustained period for it to be considered "completed." Williams (2017) uses the engineering assessments included in administrative data to categorize projects into bins of "complete" (for values such as "complete" or "installed and in use") or "incomplete" (for values such as "ongoing" or "lintel level"). Rasul and Rogger (2018) use engineering documents specific to each project to define a percentage scale of completion for each project allowing for a more granular measure of task progress, mapping them along a 0–1 continuum. Thus, highly varied project designs are mapped into a common scale of progress by consideration of the underlying production function for that class of infrastructure. What constitutes a halfway point in the development of a water distribution point and a dam will differ, but both can be feasibly assessed as having a halfway point.

Third, notions of scale or complexity can be determined from project documentation, providing a basis for improving the credibility of comparisons across tasks. As will be discussed in section three, there is little consensus about how to proxy such complexity across tasks. The literature on complexity in project



management and engineering emphasizes the multiple dimensions of complexity (Remington and Pollack 2007). This can be seen as a strength, in that a common framework for coding complexity can be flexibly adapted to the particular environment or analytical question. In the above examples, planned (rather than expended) budget is frequently used as one way to proxy scale and complexity. The challenge is that the planned budget may already be determined by features related to task completion. For example, the history of task completion at an agency may influence contemporary budget allocations.

For this reason, physical infrastructure tasks can be conceptualized and judged by external conceptions and scales that discipline the analysis. A strength of these measurement options is that they offer a relatively clear, unambiguous measure of task completion. Fundamentally, generating a sensible binary completion value requires understanding how progress maps onto public benefit (for example, an 80 percent finished water distribution point is of zero public value). With this basic knowledge across project types, task completion indicators can be computed for the full range of physical outputs produced by government.<sup>3</sup>

However, this type of task completion framework measurement also comes with limitations. It is easier to measure completion than quality with these types of measures. Quality is typically multifaceted, such that it is more demanding to collect and harmonize into an indicator that can be applied across project types. In Rasul and Rogger (2018), assessors evaluate the quality of infrastructure projects on a coarse scale related to broad indicators that implementation is of “satisfactory” quality relative to professional engineering norms. Analysis can then be defined by whether tasks are, first, completed, and second, completed to a satisfactory level of quality. Administrative progress reports vary in their information content but tend to assume quality and focus on the technical fulfillment of different stages in the completion process.

One way to gain information on quality is to undertake independent audits or checks, though these tend to be highly resource intensive relative to the use of administrative data. For example, Olken (2007, 203) relies on a team of engineers and surveyors to assess the quality of road infrastructure, who “after the projects were completed, dug core samples in each road to estimate the quantity of materials used, surveyed local suppliers to estimate prices, and interviewed villagers to determine the wages paid on the project.” From these data, Olken constructs an independent estimate of the quality of each road project.

Some conceptions of quality go as far as the citizen experience of the good or service or how durable or well managed it is. Rasul and Rogger (2018) also include assessments of citizen satisfaction with the project overall as determined by civil society assessors, but such data are almost never available in administrative records and have to be collected independently.

There are also issues pertaining to the reliability and interpretation of task completion that are worth highlighting. First, doubts may be raised when the progress reports that act as the foundation for task completion assessments are provided by the same public organizations that undertake the projects themselves (see the discussion in chapter 4). For this reason, they may not constitute reliable measures of progress, or at least may be perceived as unreliable. The problem is whether organizations can be considered reliable in their assessments of their own work. Measures of task progress sourced from administrative data must thus be used with care and, ideally, validated against a separate (independent) measure of progress. A good example of this comes from Rasul, Rogger, and Williams (2021), who match a subsample of tasks from government-produced progress reports to task audits conducted by external auditors in a separate process.<sup>4</sup> Such validation exercises can be very helpful in providing evidence that the measures produced by government organizations on their own performance are credible, thus salvaging an important source of data that might otherwise be deemed unusable.

Additionally, *noncompletion* can mean different things depending on how the timeline of infrastructure procurement, construction, and operation is organized. This is especially clear in the case described by Bancalari (2022), where it is hard to establish whether the effect uncovered is an effect of noncompletion or delays and cost overruns in delivery.<sup>5</sup> It can be hard to distinguish noncompletion (a project will remain unfinished) from delays (a project will be completed but is running over schedule). Here, the point in time when one decides to measure completion and the initial time frame set for a given task become important and can affect how one interprets task noncompletion.

Finally, a separate issue pertains to whether tasks are completed as planned, not simply whether they are completed. The existing literature from management studies has mostly focused on overruns, delays, and

over-estimated benefits rather than on noncompletion per se (Bertelli, Mele, and Whitford 2020; Post 2014). This body of literature tends to focus on the service and goods delivery side of government rather than on the full range of government activities. However, it is an important complement to the task completion framework precisely because it focuses on whether the tasks governments undertake are being completed *and* are being completed in the time frame and up to the standard that they were planned for. For example, a vast body of literature emphasizes the value-for-money or cost calculations of infrastructure projects rather than the efficiency or effectiveness of the processes via which they are delivered (for example, Engel, Fischer, and Galetovic 2013). Scholars such as Flyvbjerg (2009, 344) have argued that the “worst” infrastructure gets built because “ex ante estimates of costs and benefits are often very different from actual ex post costs and benefits. For large infrastructure projects the consequences are cost overruns, benefit shortfalls, and the systematic underestimation of risks.”

## Nonphysical Outputs

Now we turn to the task completion framework as it applies to the production of nonphysical outputs. Examples of nonphysical outputs are auditing activities, identifying localities where infrastructure is required, raising awareness about a given social benefit scheme, or planning for management meetings. These types of task, in short, involve government activities that pertain to the less visible side of government: not delivery in the form of physical goods or services but the planning, monitoring, information sharing, reviewing, and organizational tasks of government.

Rasul, Rogger, and Williams (2021) use administrative data on the roughly 3,600 tasks that civil servants undertook in the Ghanaian civil service in 2015. The data on these tasks are extracted from quarterly progress reports and represent the full spectrum of government activities. As can be seen from figure 17.1, a large proportion of these tasks are related to nonphysical outputs. For each type of task, in relation to both physical and nonphysical outputs, the researchers identify a scheme by which to judge task completion by allocating a threshold of progress to represent completion for each task type.

Rasul, Rogger, and Williams (2021) also collect data on the management practices under which these tasks are undertaken via in-person surveys with managers covering six dimensions of management: roles, flexibility, incentives, monitoring, staffing, and targets. Together, the task and management data allow for an assessment of how public sector management impacts task completion, allowing for the comparison of the effect of management practices on the same tasks across different organizations. Their data demonstrate, first, that there is substantial variation in task completion across types of task and across civil service organizations. Second, there is also substantial variation in the types of management practice that public servants are subject to across organizations, and the nature of management correlates significantly with task completion rates.

Integrating the analysis of tasks related to both physical and nonphysical outputs allows for a broad assessment of government functioning, encompassing the many interactions between tasks of different natures. Such a holistic approach also enables the assessment of tasks with different underlying characteristics, which has long been identified as a core determinant of government performance.

Rasul, Rogger, and Williams (2021) are interested in exploring whether different management techniques are differentially effective, depending on the clarity of the task in project documents. They build on the literature arguing that where settings involve intensive multitasking, coordination, or instability, management techniques using monitoring and incentive systems are likely to backfire. The question, as they put it, harking back to the Friedrich vs. Finer debate (Finer 1941; Friedrich 1940), is “to what extent should [civil servants] be managed with the carrot and the stick, and to what extent should they be empowered with the discretion associated with other professions?” (Rasul, Rogger, and Williams 2021, 262). Their central finding is that there are “positive conditional associations between task completion and organizational practices related to autonomy and discretion, but negative conditional associations with management practices related to incentives and monitoring” (Rasul, Rogger, and Williams 2021, 274).<sup>6</sup> The authors distinguish between government tasks with high and low ex ante and ex post clarity. Incentives and monitoring-intensive management approaches are hypothesized (and found) to be more effective when ex ante task clarity is high

(and ex post task clarity is low), whereas autonomy and discretion-intensive management approaches are relatively more effective when ex ante task clarity is low (and ex post task clarity is high).

The main contribution of Rasul, Rogger, and Williams (2021) to the discussion of this chapter is providing a holistic, output-based organizational performance metric. However, their approach also takes a holistic account of the multifarious nature of management practices in government and showcases the value of combining such data. The authors “conceptualize management in public organizations as a portfolio of practices that correspond to different aspects of management, each of which may be implemented more or less well. Bureaucracies may differ in their intended management styles, that is, what bundle of management practices they are aiming to implement, and may also differ in how well they are executing these practices” (262). That is, there is a combination of both intent and implementation when it comes to management practices that may affect the effectiveness of an organization. The task completion framework, with its focus on both the breadth of activities that government bodies undertake and on the detail of the characteristics of government tasks, represents an important stepping stone toward a more holistic and realistic understanding of government work and effectiveness.

A separate body of literature that brings together tasks and projects of distinct types into a single analytical framework is the literature on donor projects. For example, using data on the development projects of international development organizations (IDOs)—specifically, eight agencies—including project outcome ratings of holistic project performance, Honig (2019) investigates the success of IDO projects according to internal administrative evaluations. The success ratings are undertaken by IDO administrators, who employ a consistent underlying construct across different IDOs, with an OECD-wide standard in place. These ratings are combined with a host of other variables capturing various features of the projects (for example, their start and end dates, whether there was an IDO office presence in situ, what the sector of the project is, etc.).

Honig (2019, 172, 196) uses “variation in recipient-country environments as a source of exogenous variation in the net effects of tight principal control” to find that “less politically constrained IDOs see systematically lower performance declines in more unpredictable contexts than do their more-constrained peers.” That is to say that monitoring comes with costs in terms of reducing the ability of agents to adapt, particularly in less predictable environments.

Similarly, Denizer, Kaufmann, and Kraay (2013, 288) leverage a data set of over 6,000 World Bank projects (over 130 developing countries) to “simultaneously investigate the relative importance of country-level ‘macro’ factors and project-level ‘micro’ factors in driving project level outcomes.” The authors leverage Implementation Status Results Reports completed by task team leaders at the World Bank, which report on the status of the projects, as well as Implementation Completion Reports, which include a “subjective assessment of the degree to which the project was successful in meeting its development objective” (290), plus more detailed ex post evaluations of about 25 percent of projects, in order to assess project outcomes. They find that roughly 80 percent of the variation in project outputs occurs across projects within countries, rather than between countries, and that a large set of project-level variables influence aid project outputs.

A related but separate body of literature considers nonphysical task completion by frontline delivery agents. For example, using the case of the Department of Health in Pakistan, Khan (2021) undertakes an experiment in which he randomly emphasizes the department’s public health mission to community health workers, provides performance-linked financial incentives, or does both. He measures task completion through a combination of internal administrative data on service delivery and outputs, gathered as part of routine monitoring processes, and household surveys of beneficiaries. Mansoor, Genicot, and Mansuri (2021), instead, use the case of the agriculture extension department in Punjab, Pakistan, to measure both objective task completion and supervisors’ subjective perception of performance. They measure this through a combination of household surveys and data from a mobile phone tracking app that frontline providers use to guide and record their work.

Analogous to the physical outputs case, then, to apply a task completion framework to tasks related to nonphysical outputs, we require common definitions of tasks that cross institutional boundaries, externally valid notions of completion and progress, and notions of scale or complexity. Such external standards for what completion and quality look like across institutions are rare, but they do exist in some fields, such as health care (see the example of the joint health inspection checklist in Bedoya, Das, and

Dolinger [forthcoming]). Creating an analogous approach to these issues for tasks related to nonphysical outputs ensures comparability with tasks related to physical outputs. However, they are also valid pillars for analysis even within the set of tasks related to nonphysical outputs only.

For many tasks related to nonphysical outputs, there are, in fact, natural conceptions of task and output. For example, a curriculum development project is only complete once the curriculum is signed off on by all stakeholders, and an infrastructure monitoring program is only complete when a census of the relevant infrastructure has been completed. Similarly, such an approach can be developed for measures of progress. The curriculum development will typically be broken down into substantive stages in planning documents, and each of these stages can be assigned a proportion of progress. In the infrastructure monitoring case, a simple proportion of infrastructure projects assessed, perhaps weighted by scale or distance measures, seems fitting. Not all cases will be so clear-cut. To identify a consensus definition of task by task type that could apply across institutional boundaries, Rasul, Rogger, and Williams (2021) employ public servants at a central analytics office (in the Ghanaian case, this was the Management Services Department) to agree on relevant definitions using data from across government. As will be seen below, this team also defines measures of complexity relevant across the full set of tasks, including (as mentioned above) clarity of design. Decisions as to how to define task completion will be influenced by, but then very much influence, the approach to data collection. Table 17.1 summarizes the approaches analysts have taken to measuring task completion for physical and nonphysical outputs.

While we have focused our discussion mainly on research-oriented examples of measuring task completion, there are also examples of government organizations' use of task completion measures for tasks related to physical and nonphysical outputs—with varying degrees of formality. For example, the United Kingdom Infrastructure and Projects Authority conducts in-depth annual monitoring of all large-scale projects across UK government departments—235 as of 2022—and publishes an annual report with a red/amber/green project outlook rating (IPA 2022). At the other end of the formality and resource-intensiveness spectrum, in their engagement with the government of Ghana in 2015–16 in the course of conducting fieldwork, Rasul, Rogger, and Williams (2021) found that Ghana's Environmental Protection Agency tallied the percentage of outputs completed by each unit in their quarterly and annual reports for internal monitoring purposes. In between these two examples, the Uganda Ministry of Finance and the International Growth Centre (IGC) have partnered to apply Rasul, Rogger, and Williams's (2021) coding methodology (supplemented with qualitative interviews) to monitor the implementation progress of 153 priority policy actions across government and examine the determinants of their completion (Kaddu, Aguilera, and Carson n.d.). And of course, as argued above, many if not most government organizations do some form of task or output completion measurement in the course of their own routine reporting—despite most not taking the next step of using these data for formal analytical purposes.

**TABLE 17.1** Selected Measures of Task Completion

Task type	Potential data sources and measurement methods	Selected examples
Physical tasks	<ul style="list-style-type: none"> <li>• Site visits by expert teams</li> <li>• Site visits by survey teams</li> <li>• Compilation from other secondary sources (for example, media or project reports)</li> <li>• Administrative data from periodic reports</li> </ul>	Olken (2007); Rasul and Rogger (2018) Khwaja (2009) Flyvbjerg, Skamris Holm, and Buhl (2002); Williams (2017) Bancalari (2022)
Nonphysical tasks	<ul style="list-style-type: none"> <li>• Surveys of beneficiaries or citizens</li> <li>• Tracking app used by frontline personnel</li> <li>• Administrative data from periodic reports</li> <li>• Administrative data from internal management monitoring sources</li> <li>• International donor project evaluation reports</li> </ul>	Khan (2021) Mansoor, Genicot, and Mansuri (2021) Rasul, Rogger, and Williams (2021) Mansoor, Genicot, and Mansuri (2021), Khan (2021) Denizer, Kaufmann, and Kraay (2013), Honig (2019)

Source: Original table for this publication.

Applying the task completion framework to nonphysical outputs comes with its challenges and limitations, building on those noted above for physical outputs. The issues pertaining to assessing the quality of the implementation of tasks related to nonphysical outputs are twofold. First, establishing how to assess quality is not straightforward, and second, the nature of a task can render the difficulty of assessing quality differentially complex. For instance, if the task one is measuring is the completion of a bridge, one first has to establish the criteria that dictate whether it can be considered a high- or low-quality bridge, whereas if one is also considering nonphysical outputs, such as the development of an education strategy, then one faces a potentially even greater challenge in defining what “high-quality” means for such a project (see Bertelli et al. [2021] for a discussion of this).

There are certain types of task, in short, for which establishing objective benchmarks is more difficult than for others. It does not seem like too much of a leap, for example, to hypothesize that the nonphysical tasks we have considered in this section might frequently be more complex to benchmark in terms of quality than the physical outputs we described earlier.

This difficulty creates discontinuity in measurement quality across physical and nonphysical goods, which, in turn, raises the issue of the potential endogeneity of task and output selection. That is to say, out of the universe of possible government tasks, the types of tasks we are best able to measure may be correlated with particular outputs. This could provide us with a distorted image of the types of tasks that are conducive to producing certain outputs.

## MEASURING TASK CHARACTERISTICS

As we outline in the introduction to this chapter, a task completion framework is helpful to analysts in two main senses. First, it pushes analysts to better encapsulate the breadth of work undertaken by public administration across government. Second, it encourages them to think carefully about the characteristics of the tasks themselves. In this section, we will focus on the latter feature of a task completion framework: how to measure task characteristics.

There are, naturally, a plethora of government task characteristics on which one could focus. Here, we will focus on several of the most relevant characteristics from the perspective of implementation. We concentrate on implementation because it has been the focus of the literature on task completion and because it is of direct relevance to the work of practitioners, the intended audience of this chapter.

We start by considering task complexity. When examining government outputs and their relationship to phenomena such as management practices, government turnover, or risk environment, it is often important to understand their relationship with project or task complexity (Prendergast 2002). This is because the complexity of the task will frequently be strongly correlated with variables such as time to completion, total cost, the likelihood of delays, and customer satisfaction, which might be of interest to scholars or practitioners interested in task completion. Table 17.2 summarizes how the analysts described in this paper have attempted to implement measurement of complexity, as well as how authors have measured two further important features of government tasks to which we will turn next, visibility and clarity.

Rasul and Rogger (2018, 12), in their study of public services in the Nigerian civil service, create complexity indicators that capture “the number of inputs and methods needed for the project, the ease with which the relevant labour and capital inputs can be obtained, ambiguities in design and project implementation, and the overall difficulty in managing the project.” They are thus able to condition on the complexity of projects along these margins when exploring the relationship between managerial practices and project completion rates. However, such an approach does not account for the fact that worse-performing agencies may be assigned easier (less complex) tasks in a dynamic process over time. So in background work for the study, Rasul and Rogger assess the extent to which there was sorting of projects across agencies by their level of complexity, a task only feasible with appropriate measures. They do not find any evidence of such sorting.



**TABLE 17.2** Selected Measures of Task Characteristics

Task or project characteristic	Potential data sources and measurement methods	Selected examples
Complexity	<ul style="list-style-type: none"> <li>Expert data coding from site visits</li> <li>Semi-expert data coding from administrative reports</li> <li>International donor project evaluation reports</li> </ul>	Khwaja (2009); Rasul and Rogger (2018); Rasul, Rogger, and Williams (2021); Denizer, Kaufmann, and Kraay (2013)
Visibility	<ul style="list-style-type: none"> <li>Project-level data from infrastructure database assembled from governmental and financial sources</li> </ul>	Woodhouse (2022)
Clarity (ex ante and ex post)	<ul style="list-style-type: none"> <li>Semi-expert data coding from administrative reports</li> </ul>	Rasul, Rogger, and Williams (2021)

Source: Original table for this publication.

Khwaja (2009, 915), instead, captures project complexity by creating an index that measures whether “the project has greater cash (for outside labor and materials) versus noncash (local labor and materials) maintenance requirements, . . . the community has had little experience with such a project, and . . . the project requires greater skilled labor or spare parts relative to unskilled labor for project maintenance.” In this way, he is able to distinguish group-specific features—such as social capital—from features of task design—such as degree of complexity—in order to better understand their relative importance to one another.

Denizer, Kaufmann, and Kraay (2013) also consider complexity in their study of how micro (project-level) or macro (country-level) factors are correlated with aid project performance, albeit as a secondary focus. Using three proxies for project complexity (the extent to which a project spans multiple sectors, a project’s novelty, and the size of the project), they find “only some evidence that larger—and so possibly more complex—projects are less likely to be successful. On the other hand, greater dispersion of a project across sectors is in fact significantly associated with better project outcomes, and whether a project is a ‘repeater’ project or not does not seem to matter much for outcomes” (Denizer, Kaufmann, and Kraay 2013, 302).

Given, then, that the issue of accounting for complexity is widespread and often relies upon assessments that are not anchored to an external concept or measure of what complexity is, what are some of the ways that analysts can validate their measures of complexity? Rasul, Rogger, and Williams (2021), in their construction of a measure of the complexity of the tasks being undertaken by Ghanaian civil servants, ensure that coding is undertaken by two independent coders because the variables they measure require coders to make judgment calls about the information reported by government agencies. They also implement reconciliation by managers in cases where there are differences between coders. Discussion between coders and managers about how they see different categories or levels of complexity can be a good way to iron out differences in the measurement of complexity.

Another way to ensure consistency in measuring complexity can be to randomly reinsert particular tasks into the set of tasks being assessed by the coders to check whether they award the same complexity score to identical tasks. This is something that Rasul and Rogger (2018) do in their construction of a measure of task complexity completed by the Nigerian civil service. Rasul and Rogger (2018) also assess the similarity of scores between their two coders and leverage the passing of time to get one of the coders to recode a subsample of projects from scratch (without prompting) to assess the consistency of coding in an additional way.

In a similar spirit, audits of coding can be an effective way to validate a measure of complexity, albeit a costly one. For example, Rasul, Rogger, and Williams (2021, 265) use an auditing technique to check the validity of their measure of task completion; they “matched a subsample of 14% of tasks from progress reports to task audits conducted by external auditors through a separate exercise.” Although this technique was applied to task completion, a similar method could easily be used to validate a complexity measure in many contexts; if there are data available on the technical complexity of a task (for example, from engineers or other field specialists), such assessments could be used to check a subsample of the analyst’s own evaluations of complexity. Rasul and Rogger (2018), for example, work with a pair of Nigerian engineers to get them to assess the complexity of government tasks according to five dimensions.



Another salient feature of government tasks is how easy it is to define a given task and to evaluate whether and when it has been completed. This feature is related to, but conceptually separate from, the *complexity* of the task. Rasul, Rogger, and Williams (2021) call this feature *ex ante* and *ex post task clarity*. According to their definition, bureaucratic tasks are “*ex ante* clear when the task can be defined in such a way as to create little uncertainty about what is required to complete the task, and are *ex post* clear when a report of the actual action undertaken leaves little uncertainty about whether the task was effectively completed” (Rasul, Rogger, and Williams 2021, 260).

Task clarity is an important characteristic to consider, especially in relation to management practices, because the types of management strategy that one wishes to implement may be heavily influenced by the types of task that they govern. Indeed, Rasul, Rogger, and Williams (2021, 260) hypothesize, and find evidence, that “top-down control strategies of incentives and monitoring should be relatively more effective when tasks are easy to define *ex ante* because it is easier to specify what should be done and construct an appropriate monitoring scheme.” On the other hand, they also theorize (and, again, find evidence) that “empowering staff with autonomy and discretion should be relatively more effective when tasks are unclear *ex ante*, as well as when the actual achievement of the task is clear *ex post*” (260).

The clarity of task definition is thus also important to take into consideration when exploring questions pertaining to the management of public administration. The degree to which a task is easy to describe and evaluate has a significant bearing on the types of management strategy that make sense to employ when undertaking that task. Task clarity can also impact a number of other features of government work, such as the level of political and citizen support it enjoys—with simpler, more visible projects tending to garner more interest from politicians and support from citizens (Mani and Mukand 2007; Woodhouse 2022)—or the degree to which a task is subject to measurement or performance-pay mechanisms.

Task clarity is important to measure for its potential interactions with the concepts of effort substitution and gaming (Kelman and Friedman 2009). If performance measures are applied only to those tasks that are *ex ante* and *ex post* clear, such tasks may be prioritized to the detriment of others because they are subject to measurement or because bureaucrats seek to “game” the system by focusing their attention on improving statistics relating to their performance but not their actual performance. As we have seen in the work of Honig (2019) and Khan (2021), it is especially in complex, multidimensional task environments where granting autonomy or discretion to bureaucrats can have beneficial results. In short, thinking about the nature of the task at hand and its interaction with features such as the management practices being adopted and individual behavioral responses on the part of public servants and politicians is highly important if one wants to get to the bottom of “what works” in government.

## DISCUSSION: KEY CHALLENGES

The previous sections have reviewed the scattered and relatively young literature on the systematic measurement of task and project completion in government organizations. The measurement methods and data sources identified hold great promise for practitioners and researchers but also present a number of conceptual and practical challenges. While we have discussed some of these above in relation to specific papers or measurement methods, in this section, we briefly highlight some cross-cutting issues for measurement and analysis as well as for integration into management practice and decision-making.

The first challenge is determining what a *task* is. At the beginning of this chapter, we defined outputs as the final products delivered by government organizations to society and tasks as the intermediate steps taken by individuals or teams within government to produce those outputs. We characterized both as discrete, bounded, and clearly linked to each other. While this is conceptually useful and can serve as a guide for measurement, it is also a profound simplification of the messy, interlinked, and uncertain reality of work inside most government organizations. Indeed, the research insights produced by several of the studies we have discussed emphasize that the ambiguity, complexity, and interconnection of tasks and bureaucratic actions

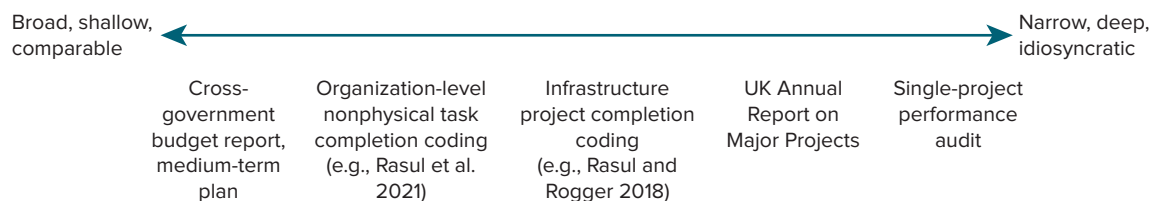
often mean that simplistic management efforts do not produce their anticipated effects. Analysts interested in measuring task completion must thus strike a difficult balance between identifying distinct tasks, projects, and outputs in order to measure their completion and simultaneously calibrating their analysis and inference to capture the nuances of the effective performance of these tasks.

A second and related challenge is drawing appropriate inferences from measures of task completion, which, in itself, is just a descriptive fact of the level of task performance. On its own, measuring task completion does not diagnose the causes of task (in)completion, predict future levels of performance, pinpoint needs for improvement, or measure the performance of the individual personnel responsible for a task (since factors outside their control may also matter). It does, however, provide a foundation upon which to conduct further analysis along these lines. Indeed, for most of the studies cited above, the measurement of task completion simply provides a dependent variable for analysis of a diverse range of potential factors and mechanisms. This chapter has focused mainly on the measurement of this dependent variable; linking it to causes and consequences requires additional analysis, which will differ in its aims and methods depending on an analyst's purposes.

A third challenge relates to integrating the measurement of task completion into practice and management—that is, taking action based on it. One main challenge relates to the well-known potential for gaming and distorting effort across multiple tasks (Dixit 2002; Propper and Wilson 2003), exemplified by “Goodhart’s Law”: “any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes” (Goodhart 1984). In other words, it may well be possible to accurately measure task completion in government organizations, but using these measures for the purpose of management—particularly if it involves benefits or consequences for the actors involved—risks undermining the validity of the measures and their linkage to bureaucratic performance. See the discussion in chapter 4. While some strategies can be put in place to mitigate such effects (for example, data quality audits or measuring multiple dimensions of bureaucratic performance), these are nearly always imperfect. Analysts should thus seek to innovate in measuring task completion as a means of improving understanding while being cautious and selective in how they use it to guide management actions.

A final consideration in deciding what tasks to measure and how is the trade-off between prioritizing breadth and comparability, on the one hand, and specificity and depth, on the other. Figure 17.2 illustrates this trade-off. In general, task completion measures that are widely applicable across the whole of government will naturally tend to be less specific to (and hence less informative of) the performance of any given unit or task. An example of this might be the type of data contained in a government’s annual report, budget execution report, or multiyear plan, which usually cover the whole of government activity but do so at a relatively shallow level. At the other extreme, researchers or practitioners can gather a great deal of information about the completion of a specific task, as a performance audit might do. This gives a very informative picture of the completion of that particular task but permits little comparison across tasks or units. In between, one can locate the various measurement options we have discussed in this chapter. For example, Rasul and Rogger’s (2018) project completion data set focuses on physical infrastructure projects, which are likely to be more comparable to each other and across organizations than Rasul, Rogger, and Williams’s (2021) data set of both physical and nonphysical outputs—but at the cost of less comprehensive coverage of government activity. The optimal place on this spectrum for any given measure of task completion naturally depends on

**FIGURE 17.2 A Spectrum of Task Completion Measures, with Selected Examples**



Source: Original figure for this publication.

the analytical purpose for which it is being created. From the standpoint of advancing measurement, the aim is to find ways to surmount this trade-off by increasing both the comparability and the rigor of task completion measures.

## CONCLUSION

We conclude by returning to the question with which we opened: how do we know if governments are performing their functions well? In this chapter, we have sought to describe and demonstrate how to apply the task completion framework in order to answer precisely this question. The framework conceives government activity in such a way as to allow analysts to assess public performance in a standardized manner across organizations and types of activity. As such, it gives us a fuller and more accurate picture of government work, forces us to think more carefully about the characteristics of the tasks that different agencies perform, and facilitates comparison of performance on a large sample that spans many types of organizations.

We have applied the framework to different categories of tasks in order to illustrate both its strengths and its limitations. In the case of tasks related to physical outputs, we have shown how data such as engineering assessments, annual progress reports, and budget reports can be merged with other data, such as management or user surveys, to provide a hitherto-inaccessible vision of the extent of project implementation and the quality of the work undertaken.

Much of this work relies, at least partly, on data that already exist but have to be digitized or rendered usable in some other way. The existence of objective, external benchmarks—produced, for example, by experts such as infrastructure engineers—means that the development of projects of many different types can be mapped onto a comparable continuum. The strength of the evaluation of physical outputs is that analysts can produce a meaningful measure of completion that gives the user some sense of how task completion maps onto public benefit. However, the weakness of the approach, as applied to physical outputs, is that the quality of task completion is often overlooked because it rests upon more complex, multifaceted assessments that are difficult to harmonize into a single indicator. Moreover, the reliability of such measures may be called into question where completion rates are reported by the same organizations that undertake the tasks themselves (although this can be counteracted to some degree if external audits of task reports are available to validate the measure).

In the case of nonphysical outputs (such as auditing, planning, or awareness-raising activities), we have demonstrated how data may come from existing sources, such as progress reports, that need to be digitized or processed to be used for analysis. The strength of extending task completion assessments to nonphysical outputs is that this provides a much richer and fuller picture of the activities that governments engage in and allows for meaningful comparisons across departments. However, the task completion framework as applied to nonphysical outputs also suffers from the same potential misreporting concern associated with physical outputs and comes with additional challenges in terms of how to measure the quality of the tasks being completed. The challenges of measuring quality are distinct from those for physical outputs, in that quality is not necessarily overlooked but is more difficult to define. For example, how do you assess the quality of a health strategy objectively and in such a way that it is comparable with, for example, education strategies or fiscal strategies?

The task completion framework, in short, moves us in the right direction when it comes to measuring the performance of governments in a way that takes into account the full breadth of government activity. However, there is much room for improvement when it comes to the measurement of the quality of the provision of both physical and nonphysical outputs. For physical outputs, expert benchmarks are often taken at face value without critical engagement with what the index or evaluation actually captures; whereas, for nonphysical outputs, benchmarks are often nonexistent, with no way to anchor quality assessments that makes them comparable across organizations. This is where we see the frontier in terms of the measurement of government performance; we need to expand the application of the task

completion framework and complement this with greater attention to how technical benchmarks are used in the measurement of physical outputs and the development of workable benchmarks for the measurement of nonphysical outputs.

## NOTES

The authors gratefully acknowledge funding from the World Bank's i2i initiative, Knowledge Change Program, and Governance Global Practice. We are grateful to Galileu Kim and Robert Lipinski for helpful comments.

1. See, for instance, the World Bank's World Governance Indicators, available at <https://info.worldbank.org/governance/wgi>, and the Millennium Challenge Corporation scorecards—for example, on the website of the Millennium Challenge Coordinating Unit for Sierra Leone, <http://www.mccu-sl.gov.sl/scorecards.html>.
2. *Outputs* are not to be confused with *outcomes*, or “the impacts on social, economic, or other indicators arising from the delivery of outputs (e.g., student learning, social equity).” *OECD Glossary of Statistical Terms*, s.vv. “output,” “outcome” (Paris: OECD Publishing, 2022), <http://stats.oecd.org/glossary>.
3. Such indicators do not rely upon subjective citizen-survey responses, which are limited by their reliance on human judgment and prey to multiple biases and recall issues (Golden 1992), both from the researcher designing the study and the experts or citizen respondents evaluating the government.
4. No evidence was found that completion levels differed significantly across auditors and agencies, with 94 percent of completion rates being corroborated across coding groups (Rasul, Rogger, and Williams 2021, 265).
5. The measure of unfinished projects is a “combination of projects still underway (on time or delays) and abandoned (temporarily or indefinitely) in a given district” (Bancalari 2022, 10).
6. It is important to note that their findings are relative to one another—that is, “organizations appear to be overbalancing their management practice portfolios toward top-down control measures at the expense of entrusting and empowering the professionalism of their staff” (Rasul, Rogger, and Williams 2021, 261).

## REFERENCES

- Ammons, David N. 2014. *Municipal Benchmarks: Assessing Local Performance and Establishing Community Standards*. 3rd ed. London: Routledge.
- Andrews, Rhys, George A. Boyne, Kenneth J. Meier, Laurence J. O'Toole Jr., and Richard M. Walker. 2005. “Representative Bureaucracy, Organizational Strategy, and Public Service Performance: An Empirical Analysis of English Local Government.” *Journal of Public Administration Research and Theory* 15 (4): 489–504. <https://doi.org/10.1093/jopart/mui032>.
- Bancalari, Antonella. 2022. “Can White Elephants Kill? Unintended Consequences of Infrastructure Development in Peru.” IFS Working Paper 202227, Institute for Fiscal Studies, London. <https://ifs.org.uk/publications/can-white-elephants-kill-unintended-consequences-infrastructure-development>.
- Bedoya, Guadalupe, Jishnu Das, and Amy Dolinger. Forthcoming. “Randomized Regulation: The Impact of Minimum Quality Standards on Health Markets.” Working paper, World Bank, Washington, DC.
- Bertelli, Anthony Michael, Eleanor Florence Woodhouse, Michele Castiglioni, and Paolo Belardinelli. 2021. *Partnership Communities*. Cambridge Elements: Public and Nonprofit Administration. Cambridge: Cambridge University Press.
- Bertelli, Anthony Michael, Valentina Mele, and Andrew B. Whitford. 2020. “When New Public Management Fails: Infrastructure Public-Private Partnerships and Political Constraints in Developing and Transitional Economies.” *Governance: An International Journal of Policy, Administration, and Institutions* 33 (3): 477–93. <https://doi.org/10.1111/gove.12428>.
- Boyne, George A. 2003. “What Is Public Service Improvement?” *Public Administration* 81 (2): 211–27. <https://doi.org/10.1111/1467-9299.00343>.
- Brown, Karin, and Philip B. Coulter. 1983. “Subjective and Objective Measures of Police Service Delivery.” *Public Administration Review* 43 (1): 50–58. <https://doi.org/10.2307/975299>.
- Carter, Neil, Rudolf Klein, and Patricia Day. 1992. *How Organisations Measure Success: The Use of Performance Indicators in Government*. London: Routledge.
- Denizer, Cevdet, Daniel Kaufmann, and Aart Kraay. 2013. “Good Countries or Good Projects? Macro and Micro Correlates of World Bank Project Performance.” *Journal of Development Economics* 105: 288–302. <https://doi.org/10.1016/j.jdeveco.2013.06.003>.

- Dixit, Avinash. 2002. "Incentives and Organizations in the Public Sector: An Interpretative Review." *Journal of Human Resources* 37 (4): 696–727. <https://doi.org/10.2307/3069614>.
- Engel, Eduardo, Ronald Fischer, and Alexander Galetovic. 2013. "The Basic Public Finance of Public-Private Partnerships." *Journal of the European Economic Association* 11 (1): 83–111. <https://www.jstor.org/stable/23355049>.
- Fenizia, Alessandra. 2022. "Managers and Productivity in the Public Sector." *Econometrica* 90 (3): 1063–84. <https://doi.org/10.3982/ECTA19244>.
- Finer, Herman. 1941. "Administrative Responsibility in Democratic Government." *Public Administration Review* 1 (4): 335–50. <https://doi.org/10.2307/972907>.
- Flyvbjerg, Bent. 2009. "Survival of the Unfittest: Why the Worst Infrastructure Gets Built—And What We Can Do about It." *Oxford Review of Economic Policy* 25 (3): 344–67. <https://doi.org/10.1093/oxrep/grp024>.
- Flyvbjerg, Bent, Mette Skamris Holm, and Søren Buhl. 2002. "Underestimating Costs in Public Works Projects: Error or Lie?" *Journal of the American Planning Association* 68 (3): 279–95. <https://doi.org/10.1080/01944360208976273>.
- Friedrich, Carl J. 1940. "Public Policy and the Nature of Administrative Responsibility." In *Public Policy: A Yearbook of the Graduate School of Public Administration, Harvard University* 1: 1–20.
- Golden, Brian R. 1992. "The Past Is the Past—Or Is It? The Use of Retrospective Accounts as Indicators of Past Strategy." *Academy of Management Journal* 35 (4): 848–60. <https://doi.org/10.2307/256318>.
- Goodhart, Charles A. E. 1984. "Problems of Monetary Management: The UK Experience." In *Monetary Theory and Practice: The UK Experience*, 91–121. London: Red Globe Press. <https://doi.org/10.1007/978-1-349-17295-5>.
- Hefetz, Amir, and Mildred E. Warner. 2012. "Contracting or Public Delivery? The Importance of Service, Market, and Management Characteristics." *Journal of Public Administration Research and Theory* 22 (2): 289–317. <https://doi.org/10.1093/jopart/mur006>.
- Ho, Alfred Tat-Kei, and Wonhyuk Cho. 2017. "Government Communication Effectiveness and Satisfaction with Police Performance: A Large-Scale Survey Study." *Public Administration Review* 77 (2): 228–39. <https://doi.org/10.1111/puar.12563>.
- Honig, Dan. 2019. "When Reporting Undermines Performance: The Costs of Politically Constrained Organizational Autonomy in Foreign Aid Implementation." *International Organization* 73 (1): 171–201. <https://doi.org/10.1017/S002081831800036X>.
- IPA (Infrastructure and Projects Authority). 2022. *Annual Report on Major Projects 2021–22*. Reporting to Cabinet Office and HM Treasury, United Kingdom Government. London: IPA. [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/1092181/IPAAR2022.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1092181/IPAAR2022.pdf).
- Kaddu, M., J. Aguilera, and L. Carson. n.d. *Challenges to Policy Implementation in Uganda (Review of Policy Implementation in Uganda)*. London: International Growth Centre, London School of Economics and Political Science.
- Kelman, Steven, and John N. Friedman. 2009. "Performance Improvement and Performance Dysfunction: An Empirical Examination of Distortionary Impacts of the Emergency Room Wait-Time Target in the English National Health Service." *Journal of Public Administration Research and Theory* 19 (4): 917–46. <https://doi.org/10.1093/jopart/mun028>.
- Khan, Muhammad Yasir. 2021. "Mission Motivation and Public Sector Performance: Experimental Evidence from Pakistan." Working paper delivered at 25th Annual Conference of the Society for Institutional Organizational Economics, June 24–26, 2021 (accessed February 8, 2023). <https://y-khan.github.io/yasirkhan.org/muhammadyasirkhanjmp.pdf>.
- Khwaja, Asim Ijaz. 2009. "Can Good Projects Succeed in Bad Communities?" *Journal of Public Economics* 93 (7–8): 899–916. <https://doi.org/10.1016/j.jpubeco.2009.02.010>.
- Lee, Soo-Young, and Andrew B. Whitford. 2009. "Government Effectiveness in Comparative Perspective." *Journal of Comparative Policy Analysis* 11 (2): 249–81. <https://doi.org/10.1080/13876980902888111>.
- Lewis, David E. 2007. "Testing Pendleton's Premise: Do Political Appointees Make Worse Bureaucrats?" *The Journal of Politics* 69 (4): 1073–88. <https://doi.org/10.1111/j.1468-2508.2007.00608.x>.
- Lu, Jiahuan. 2016. "The Performance of Performance-Based Contracting in Human Services: A Quasi-Experiment." *Journal of Public Administration Research and Theory* 26 (2): 277–93. <https://doi.org/10.1093/jopart/muv002>.
- Mani, Anandi, and Sharun Mukand. 2007. "Democracy, Visibility and Public Good Provision." *Journal of Development Economics* 83 (2): 506–29. <https://doi.org/10.1016/j.jdeveco.2005.06.008>.
- Mansoor, Zahra, Garance Genicot, and Ghazala Mansuri. 2021. "Rules versus Discretion: Experimental Evidence on Incentives for Agriculture Extension Staff." Unpublished manuscript.
- Nistotskaya, Marina, and Luciana Cingolani. 2016. "Bureaucratic Structure, Regulatory Quality, and Entrepreneurship in a Comparative Perspective: Cross-Sectional and Panel Data Evidence." *Journal of Public Administration Research and Theory* 26 (3): 519–34. <https://doi.org/10.1093/jopart/muv026>.
- Olken, Benjamin A. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 115 (2): 200–49. <https://doi.org/10.1086/517935>.
- Poister, Theodore H., and Gregory Streib. 1999. "Performance Measurement in Municipal Government: Assessing the State of the Practice." *Public Administration Review* 59 (4): 325–35. <https://doi.org/10.2307/3110115>.
- Post, Alison E. 2014. *Foreign and Domestic Investment in Argentina: The Politics of Privatized Infrastructure*. Cambridge, UK: Cambridge University Press.



- Prendergast, Canice. 2002. "The Tenuous Trade-Off between Risk and Incentives." *Journal of Political Economy* 110 (5): 1071–102. <https://doi.org/10.1086/341874>.
- Propper, Carol, and Deborah Wilson. 2003. "The Use and Usefulness of Performance Measures in the Public Sector." *Oxford Review of Economic Policy* 19 (2): 250–67. <https://doi.org/10.1093/oxrep/19.2.250>.
- Rainey, Hal G. 2009. *Understanding and Managing Public Organizations*. 4th ed. New York: John Wiley & Sons.
- Rainey, Hal G., and Paula Steinbauer. 1999. "Gallopers: Developing Elements of a Theory of Effective Government Organizations." *Journal of Public Administration Research and Theory* 9 (1): 1–32. <https://doi.org/10.1093/oxfordjournals.jpart.a024401>.
- Rasul, Imran, and Daniel Rogger. 2018. "Management of Bureaucrats and Public Service Delivery: Evidence from the Nigerian Civil Service." *The Economic Journal* 128 (608): 413–46. <https://doi.org/10.1111/econj.12418>.
- Rasul, Imran, Daniel Rogger, and Martin J. Williams. 2021. "Management, Organizational Performance, and Task Clarity: Evidence from Ghana's Civil Service." *Journal of Public Administration Research and Theory* 31 (2): 259–77. <https://doi.org/10.1093/jopart/muaa034>.
- Rauch, James E., and Peter B. Evans. 2000. "Bureaucratic Structure and Bureaucratic Performance in Less Developed Countries." *Journal of Public Economics* 75 (1): 49–71. [https://doi.org/10.1016/S0047-2727\(99\)00044-4](https://doi.org/10.1016/S0047-2727(99)00044-4).
- Remington, Kaye, and Julien Pollack. 2007. *Tools for Complex Projects*. Aldershot, UK: Gower.
- Talbot, Colin. 2010. *Theories of Performance: Organizational and Service Improvement in the Public Domain*. Oxford: Oxford University Press.
- Thomas, John Clayton, Theodore H. Poister, and Nevbahar Ertas. 2009. "Customer, Partner, Principal: Local Government Perspectives on State Agency Performance in Georgia." *Journal of Public Administration Research and Theory* 20 (4): 779–99. <https://doi.org/10.1093/jopart/mup024>.
- Walker, Richard M., M. Jin Lee, Oliver James, and Samuel M. Y. Ho. 2018. "Analyzing the Complexity of Performance Information Use: Experiments with Stakeholders to Disaggregate Dimensions of Performance, Data Sources, and Data Types." *Public Administration Review* 78 (6): 852–63. <https://doi.org/10.1111/puar.12920>.
- Williams, Martin J. 2017. "The Political Economy of Unfinished Development Projects: Corruption, Clientelism, or Collective Choice?" *American Political Science Review* 111 (4): 705–23. <https://doi.org/10.1017/S0003055417000351>.
- Woodhouse, Eleanor Florence. 2022. "The Distributive Politics of Privately Financed Infrastructure Agreements." Unpublished manuscript.





The background features a series of orange squares arranged in a grid-like pattern that curves upwards from the left. Overlaid on this is a stream of binary code (0s and 1s) in a light orange color, also curving upwards from the left. A solid orange horizontal band spans the width of the page, serving as a backdrop for the 'PART 4' text.

## **PART 4**

# Government Analytics Using Public Servant Surveys



## CHAPTER 18

# Surveys of Public Servants

## The Global Landscape

*Ayesha Khurshid and Christian Schuster*

### SUMMARY

Governments around the world increasingly implement surveys of public servants to better understand—and to provide evidence to improve—public administration. As context for the subsequent chapters in *The Government Analytics Handbook* on surveys of public servants, this chapter reviews the existing landscape of governmentwide surveys of public servants. What concepts are measured in these surveys? How are these concepts measured? And what survey methodologies are used? Our review finds that while governments measure similar concepts across surveys, the precise questions asked to measure these concepts vary, as do survey methodologies—for instance, in terms of sampling approaches, survey weights, and survey modes. The chapter concludes, first, that discrepancies in survey questions for the same concepts put a premium on cross-country questionnaire harmonization, and it introduces the Global Survey of Public Servants (GSPS) as a tool to achieve harmonization. Second, the chapter concludes that methodological differences across surveys—despite similar survey objectives—underscore the need for stronger evidence to inform methodological choices in surveys of public servants. The remaining chapters in this part focus on providing such evidence.

### ANALYTICS IN PRACTICE

- Surveys of public servants have been implemented by an increasing number of countries in the last two decades. They tend to measure similar concepts, focusing on a core set of employee attitudes (such as job satisfaction or engagement), on the one hand, and management practices (such as the quality of leadership), on the other.
- Despite measuring similar concepts, questionnaires across surveys of public servants are not harmonized: different governments use different measures for the same concepts.

---

Ayesha Khurshid is a consultant in the World Bank's Development Impact Evaluation (DIME) Department. Christian Schuster is a professor at University College London.

- Despite having similar aims, methodologies for surveys of public servants vary across countries—for instance, in terms of sampling approaches, survey weighting, survey populations, survey modes, and response rates achieved.
- Differences in survey methodology underscore the importance of robust evidence to ensure good-practice methodologies in surveying public servants, the topic of the remainder of this part.

## INTRODUCTION

Understanding government and providing actionable data and evidence to public sector managers to improve the machinery of government requires microdata about government institutions (chapter 2). Surveys of public servants are one such microdata source. Many key features of the environment of public servants cannot be measured efficiently through other (administrative data) mediums. For example, how public servants are managed, their motivations, and their behaviors are all phenomena internal to an official's lived experience. Management quality is fundamentally an experienced interaction that can often only be measured by employees' or managers' reports of it. Public employees' motivations are difficult to observe outside of their own expressions of them. Thus, self-reporting through surveys becomes the primary means of measurement for many aspects of officialdom and, as detailed elsewhere in *The Government Analytics Handbook*, of the public sector production function (see chapter 1).

This section of the *Handbook* provides frontier evidence on key choices in public servant surveys—from the appropriate survey mode (chapter 19), to determining sampling sizes (chapter 21), questionnaire design (chapters 20 and 22), and the effective reporting of survey results (chapter 25). To contextualize the chapters in this section, this introductory chapter provides an overview of the state of play in public servant surveys around the world.

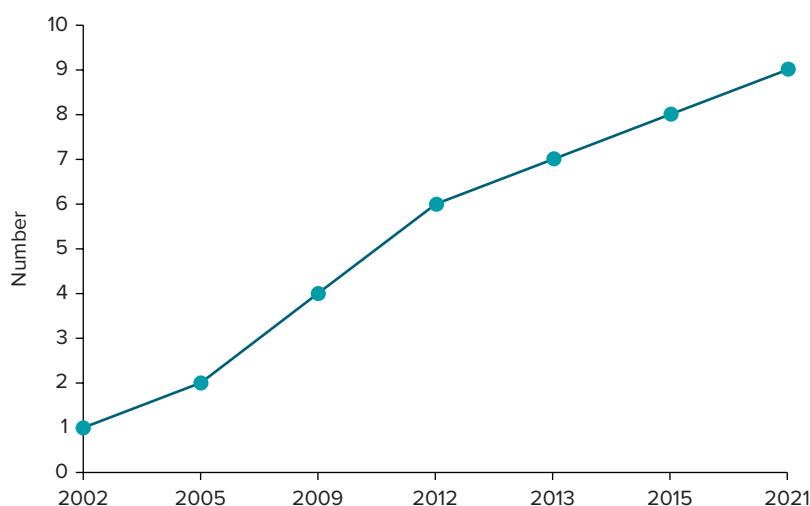
To present the state of play in this field, we review the existing landscape of regular, governmentwide employee surveys—that is, surveys that are run on a regular (annual or biannual) basis with repeated measurements (on at least three previous occasions) for a central government. We thus focus this chapter on surveys that are institutionalized as measurement and management instruments in governments. This contrasts with other reviews—in particular, Organisation for Economic Co-operation and Development (OECD 2016)—which comprise ad hoc, non-central-governmentwide surveys with varying content and methodologies.<sup>1</sup>

The first introductory takeaway from this review is that surveys of public servants have recently become more popular with governments. As illustrated in figure 18.1, the number of countries undertaking governmentwide employee surveys has increased continuously over the last decade, reaching nine countries in 2021. (We might, of course, underestimate the number of institutionalized surveys of public servants outside the English-speaking world, so this number is a lower bound.)

As detailed in table 18.1, all countries for which we were able to review and validate the implementation of regular surveys of public servants belong to the OECD (though some, such as Colombia, are recent OECD joiners). While most of these countries have been implementing institutionalized surveys for over a decade, countries such as New Zealand have only begun the exercise in recent years. All countries implement their surveys annually except Ireland and Canada, which implement their surveys every two years.

This chapter will provide an overview of the key features of these surveys, in part to contextualize the remainder of this section of the *Handbook*, which will provide novel empirical evidence on the design, implementation, and dissemination of public servant surveys. The chapter will first review what established surveys of public servants measure. Subsequently, it will look at survey methodologies across countries: how are surveys implemented (for instance, in terms of sampling and response rates)?

**FIGURE 18.1** Countries with Regular, Governmentwide Employee Surveys, Cumulative Count, 2002–21



Source: Original figure for this publication.

**TABLE 18.1** Countries with Regular, Governmentwide Employee Surveys, 2002–22

Country	Survey title	Undertaken since	Latest year	Frequency
Australia	Australian Public Service Employee Census	2012	2022	Annual
Canada	Public Service Employee Survey	2005	2020	Biannual
Colombia	Survey of the Institutional Environment and Performance in the Public Sector [Encuesta sobre ambiente y desempeño institucional nacional]	2009	2021	Annual
Ireland	Civil Service Employee Engagement Survey	2015	2020	Biannual
Korea, Rep.	Public Service Life Survey	2013	2021	Annual
New Zealand	Te Taunaki Public Service Census	2021	2021	Annual
Switzerland	Staff Survey of the Federal Administration [Enquête auprès du personnel de l'administration fédérale]	2012	2021	Annual
United Kingdom	Civil Service People Survey	2009	2021	Annual
United States	Federal Employee Viewpoint Survey	2002	2021	Annual

Source: Original table for this publication.

## A REVIEW OF CONCEPTS AND MEASURES IN EXISTING SURVEYS OF PUBLIC SERVANTS

To understand the key concepts for measurement when governments undertake surveys of their employees, we summarize a review by Meyer-Sahling et al. (2021) of the concepts measured in six of the government employee surveys outlined above. This review comprises the United States' Federal Employee Viewpoint Survey, Canada's Public Service Employee Survey, the United Kingdom's Civil Service People Survey, the Australian Public Service (APS) Employee Census, Colombia's Survey of the Institutional Environment and Performance in the Public Sector, and Ireland's Civil Service Employee Engagement Survey. The focus



of the review is on measurement in the last year before the COVID-19 pandemic, as the pandemic led to an exceptional focus on teleworking—rather than the implementation of the regular annual survey—in a number of countries.

Meyer-Sahling et al. (2021) frame their review within a production function of the public service (analogous to the production function presented in chapter 1 of the *Handbook*) that outlines how the productivity of public services depends on the quality and quantity of outputs relative to inputs. Inputs include staff (that is, public servants) and other resources. Inputs are converted to public sector outputs and outcomes by management practices and public or organizational policies. Whether inputs are effectively converted to outputs is moderated by exogenous factors (such as the political environment) and mediated by the attitudes and behaviors of public servants.

Surveys of public servants can be used to shed light on different components of this public service productivity chain. As detailed by Meyer-Sahling et al. (2021), surveys of public servants are particularly suitable for measuring management practices and complementary inputs, on the one hand (for example, employees' perception of the quality of leadership in their organization), and public employees' attitudes and behaviors, on the other (for example, their work motivation). These parts of the public sector production function often cannot be recorded through administrative data in a valid way. Thus, self-reporting through surveys becomes the primary measurement tool.

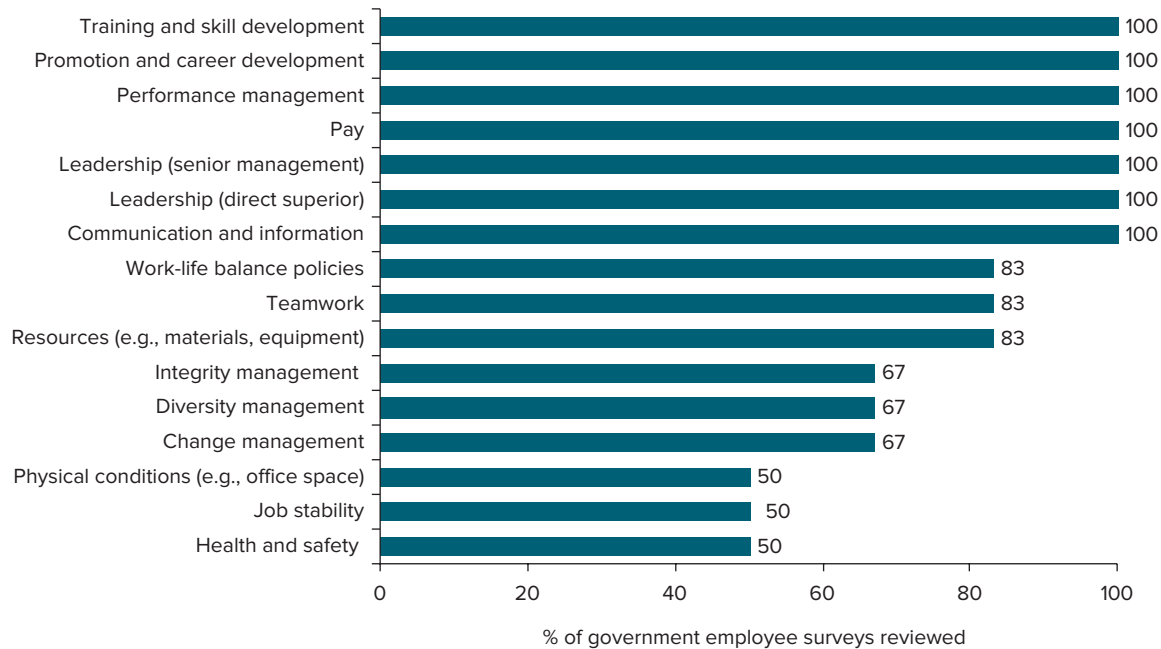
Which areas of management practice, on the one hand, and employee attitudes, on the other, do existing surveys of public servants primarily measure? By classifying topics in the six countries, seven broad areas of management practices are measured across all government employee surveys reviewed: leadership (by both the direct superior and senior management), performance management, pay, training and skills development, promotion and career development, and communication and information to employees. Three further areas—practices to foster work-life balance, teamwork, and the sufficiency of resources (for example, equipment)—were measured in all but one employee survey. As figure 18.2 shows, these 10 management areas are thus plausibly core to (almost) all government employee surveys.

Looking next at employee attitudes, the review finds that government employee surveys also measure an overlapping set of core employee attitudes and behaviors. As illustrated in figure 18.3, all reviewed government employee surveys measure the organizational commitment of public employees, their engagement with their jobs, and their perception of their workloads and work-life balance. Moreover, four additional concepts—job satisfaction, career/turnover intentions, integrity, and innovation attitudes—are measured in all but one of the government employee surveys. These six attitudes and behaviors are thus plausibly core to (almost) all government employee surveys.

Thus, governments measure similar concepts across many of their employee surveys. (Of course, governments also add idiosyncratic modules that are of particular interest to them in any given year, such as remote work during the COVID-19 pandemic.) This plausibly reflects an interest in a similar set of core management practices and employee attitudes and behaviors to improve public sector performance. At the same time, as outlined below, the exact wording of measures for the same concept frequently differs across countries (as does the precise coverage of a concept—for instance, whether pay is measured in relation to performance, satisfaction, fairness, or other pay-related factors), which is a core rationale for harmonizing this wording through the Global Survey of Public Servants (GSPS) (see below).

Two caveats regarding these conclusions about commonality are due. First, the review's coverage extends to OECD countries. In countries of the Global South, other concepts, such as meritocracy, politicization, and corruption, are often central to the (non)functioning of the public sector and might thus deserve greater pride of place in surveys of public servants (Meyer-Sahling et al. 2021). Second, some recent surveys have shifted toward a greater focus on directly actionable survey questions—for instance, to check for good practice in performance evaluations or onboarding procedures and showcase where basic practices are not in place (see Fukuyama et al. 2022). That most existing governmentwide employee surveys are silent on these topics suggests that focusing on more actionable survey questions is one margin for improving many existing questionnaires.

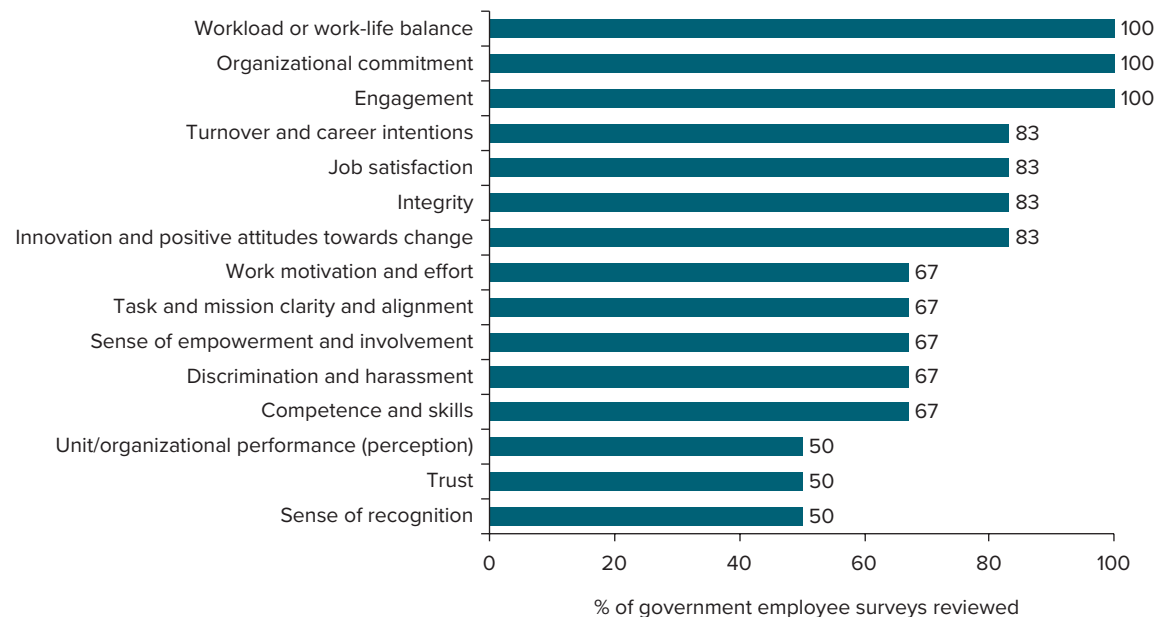
**FIGURE 18.2 Management Practices Measured in Government Employee Surveys**



Source: Meyer-Sahling et al. 2021.

Note: Only concepts covered in at least half of the surveys reviewed are shown.

**FIGURE 18.3 Employee Attitudes and Behaviors Measured in Government Employee Surveys**



Source: Meyer-Sahling et al. 2021.

Note: Only concepts covered in at least half of the surveys reviewed are shown.

## METHODOLOGIES IN SURVEYS OF PUBLIC SERVANTS

Having reviewed the content of existing governmentwide employee surveys, in this section, we will review their methodologies. How are respondents sampled by governments? Are surveys conducted online, on paper, in person, or by phone? How long are public servant survey questionnaires? What response rates are achieved and how are survey weights constructed to enhance representativeness? The remaining chapters in the public servant survey section of the *Handbook* provide novel empirical evidence to enable governments and practitioners to make evidence-based choices in response to these and other methodological questions, along the decision tree in survey design, implementation, and reporting detailed in chapter 1. To contextualize these empirical and methodological chapters, the remainder of this section briefly reviews practices and methodological choices in existing governmentwide employee surveys. Table 18.2 summarizes the findings from this comparison.

### Survey Mode

One of the first methodological choices in public servant surveys is the enumeration method, or survey mode. Different survey modes come with different response biases to questions and different overall response rates.

All nine government surveys reviewed in table 18.1 were implemented online, using an invitation link sent to public servants through email or shared through the administration's intranet. Additionally, to enhance accessibility (for instance, for staff with difficulty accessing or completing an online survey), Colombia, Switzerland, the UK, and a few Australian agencies offered their surveys in a paper format, while New Zealand offered its survey through paper and telephone upon request.

Field experimental evidence from the *Handbook* suggests—albeit based on data from Romania only—that these diverging survey modes do not substantially impact aggregate estimates at the national level (see chapter 19). They do, however, affect the comparability of findings across organizations, among other things (see chapter 19). Governments that offer varying survey modes should thus be careful when comparing the scores of organizations if some implement the survey primarily online while others implement it primarily on pen and paper.

## SURVEY POPULATION

Across the nine surveys reviewed, the survey population generally consists of central-government civil servants, although the extent to which public sector organizations and employee contracts outside the (legally defined) civil service are covered varies—for instance, in other branches of government or frontline services.

For the UK government employee survey, all public servants from 101 agencies are eligible, excluding the Northern Ireland Civil Service, the NHS (which conducts its own survey), and frontline officials (for instance, police officers and teachers) (Cabinet Office 2020). The US survey invites all federal, nonseasonal, and permanent public servants (including all full- and part-time employees) in 82 executive branch agencies to participate (OPM 2020).

The Australian survey includes all employees from 101 agencies. While agencies set their own eligibility requirements, it generally excludes public servants on leave during the survey and those with a short tenure in the agency (Australian Public Service Commission 2021). Similarly, in Colombia, all public servants working in Bogotá with a tenure of more than six months at the central level of the executive, legislative, and judicial powers and in the headquarters of the regional autonomous corporations and public universities (200 agencies) are eligible to participate in the survey (DANE 2020).

**TABLE 18.2 Methodological Choices in Public Servant Surveys**

Country	Survey mode(s)	Survey population	Sampling	Response rate <sup>a</sup> (%)	Survey weighting <sup>b</sup>	Questionnaire length (number of questions) <sup>c</sup>
Australia	Primarily online with some agencies offering a paper-based option	All regular employees from 101 agencies with sufficient tenure in their agency	Census	77	No weights applied	112
Canada	Online	All paid employees in 90 core agencies	Census	61	Nonresponse weights applied	112
Colombia	Primarily online with a paper-based option	All employees in 200 agencies with a tenure of at least six months working in Bogotá and in the headquarters of regional autonomous corporations and public universities	Census for smaller agencies and stratified sampling for larger agencies	96	Nonresponse weights applied	65
Ireland	Online	All employees in 50 agencies in Ireland and those based abroad	Census	65	—	112
Korea, Rep.	Online	All employees from central administrative agencies and metropolitan governments	Sampled survey using multistage stratification and probability-proportional-to-size sampling	—	—	48
New Zealand	Primarily online with a paper-based and telephone option	All employees in 36 agencies and those based abroad, excluding the NZCIS and the GSCB <sup>d</sup>	Census	63	—	61
Switzerland	Primarily online with a paper-based option	All monthly paid employees (excluding parliamentary services and the Public Ministry of the Confederation and the courts)	Census every three years with a sampled survey in all other years	71	—	24
United Kingdom	Primarily online with a paper-based option	All employees from 101 agencies (excluding the Northern Ireland Civil Service, the National Health Service, and frontline officials)	Census	62	No weights applied	72
United States	Online	All permanently employed and nonseasonal federal employees in 82 agencies	Census every few years (2012, 2018, 2019, and 2020) with a sampled survey using stratified sampling in other years	44	Nonresponse weights applied	101

Source: Original table for this publication.

Note: The table displays “—” wherever information was unavailable to the authors.

a. Response rates are presented for the latest year for which data and/or results were available. The response rate for the Korean survey was unavailable.

b. Information about nonresponse weights in Canada, Ireland, Republic of Korea, New Zealand, and Switzerland was, unfortunately, unavailable.

c. Questionnaire lengths were reviewed for the last year before the COVID-19 pandemic.

d. New Zealand Security Intelligence Service (NZCIS); Government Communications Security Bureau (GSCB).

The Irish survey targets all public servants from 50 agencies in Ireland and those based abroad (Department of Public Expenditure and Reform 2020). Similar to Ireland, the New Zealand survey includes all public servants working in 36 public service agencies and those based overseas, apart from the New Zealand Security Intelligence Service (NZSIS) and the Government Communications Security Bureau (GSCB) (both of which conduct their own surveys) (Research New Zealand 2021). While limited information is available on the Korean survey, its target population includes all public servants from central administrative agencies and metropolitan governments (Korea Institute of Public Administration 2021).

The Canadian survey has the most flexible eligibility criteria: all indeterminate, term, seasonal, casual, and student employees in 90 core public administration agencies are eligible (excluding ministers' exempt staff, private contractors and consultants, and employees on unpaid leave) (Government of Canada 2022). Similarly, the Swiss survey population consists of all permanent staff that are paid monthly but excludes public servants working in the parliamentary services, the Public Ministry of the Confederation, and the courts (OFPER 2022).

## Sampling Design

Approaches to sampling across countries vary, ranging from census to random, ad hoc, and stratified sampling. Australia, New Zealand, and the UK adopt a census approach in which all eligible public sector employees are invited to participate in the survey (Australian Public Service Commission 2021; Cabinet Office 2020; Research New Zealand 2021). Canada's Public Service Employee Survey and Ireland's Civil Service Employee Engagement Survey are also based on a census approach, albeit one with an open link offering less control over who responds (Department of Public Expenditure and Reform 2020). In Canada, public sector organizations reach out to their staff to complete the survey, but the government also makes the survey available online for anyone who decides they fit the eligibility criteria (Government of Canada 2022).

The US government Federal Employee Viewpoint Survey uses stratified randomized sampling approaches for most years but conducts a census every few years (2012, 2018, 2019, and 2020), in order to update sampling frames, with the survey link sent to all eligible respondents (OPM 2020). Similarly, Switzerland conducts a census every three years (2014, 2017, and 2020) and a sampled survey in other years (OFPER 2022). Colombia's public servant survey, in turn, uses a mixed approach: for larger organizations, a stratified sampling approach is used, while for smaller organizations (with fewer than 110 employees), a census is taken to protect anonymity. For larger organizations, the sampling frame is stratified by organization and hierarchy, and public servants are selected to participate using simple random sampling within strata (DANE 2020).

Meanwhile, the Republic of Korea adopts a sampling approach for all annual surveys. Approximately 4,000 respondents are sampled each year each using multistage stratification and probability-proportional-to-size sampling to ensure the representativeness of the sample (Korea Institute of Public Administration 2021).

As detailed later in this section of the *Handbook* (chapter 25, census approaches offer the advantage of sufficient response numbers to provide unit-level management reports based on survey results, even at more disaggregated levels. The UK government, for instance, produces over 12,000 management diagnostics or reports based on its results. At the same time, census sampling approaches are costly in terms of the opportunity cost of staff time spent on completing the survey. As detailed in chapter 20, the appropriate sampling approach thus depends on the types of inference one seeks to draw from the data. Chapter 20 offers a sampling tool to allow governments to estimate appropriate sample sizes based on the types of inference and benchmarking exercises they wish to make with the data. Interestingly, existing government approaches to sampling respondents in public servant surveys do not seem to be (explicitly) based on such a data-driven approach to sampling, suggesting that the potential to optimize sampling in surveys of public servants remains.

## Response Rates and Nonresponse Weighting

Beyond their sampling approaches, surveys of public servants across governments also differ in response rates and their approaches to correcting for nonresponse bias. As detailed in table 18.2, survey response rates

vary from 44 percent in the US to 96 percent in Colombia. In Colombia, the national statistical office (DANE) conducts the survey, and statistics legislation mandates that sampled respondents complete the survey. In the remaining countries, participation in the survey is voluntary, leading to relatively lower response rates.

To enhance the likelihood that the final sample is representative of the target population of public servants, Canada, Colombia, and the US apply nonresponse weights.<sup>2</sup> Canada uses nonresponse weights to enhance the representativeness of occupational groups in each agency (Statistics Canada 2018). To construct nonresponse weights, the US survey uses subagency identifier, supervisory status, gender, minority status, age, tenure, full- or part-time status, and location from administrative data (OPM 2020). The Colombian survey, in turn, uses nonresponse weights based on the same variables as in its sampling approach—for example, hierarchical level or the institution a respondent works for (DANE 2020).

The Australian survey checks for the representativeness of respondents across age, gender, state or territory, and classification. As survey respondents do not significantly differ from the survey population in these characteristics in the Australian case, the Australian survey does not use nonresponse weights (Australian Public Service Commission 2021). Similarly, the UK does not apply nonresponse weights to the final set of respondents.

Evidence from elsewhere in the *Handbook* suggests that the effect of nonresponse weights (constructed from demographic information) on national-level averages in particular is relatively limited, at least in the country studied in the chapter (chapter 19). This is good news for cases, like the UK, where governmentwide demographic information to construct weights is in limited supply. At the same time, some nonresponse weights are straightforward to construct for all governments—for instance, weights to correct for differential response rates in institutions of differential size. They thus merit consideration where not currently applied.

## QUESTIONNAIRE LENGTH

Beyond these differences in nonresponse weights, surveys of public servants also differ in questionnaire design, including length. In the last year before the COVID-19 pandemic, questionnaire lengths varied significantly.<sup>3</sup> Ireland and Canada implemented the longest public servant survey, with 112 questions, followed by the Australian and US surveys (100 questions each). Switzerland implemented the shortest, with 24 questions. Colombia and New Zealand (each approximately 60 questions) and Republic of Korea (48 questions) sat in between.

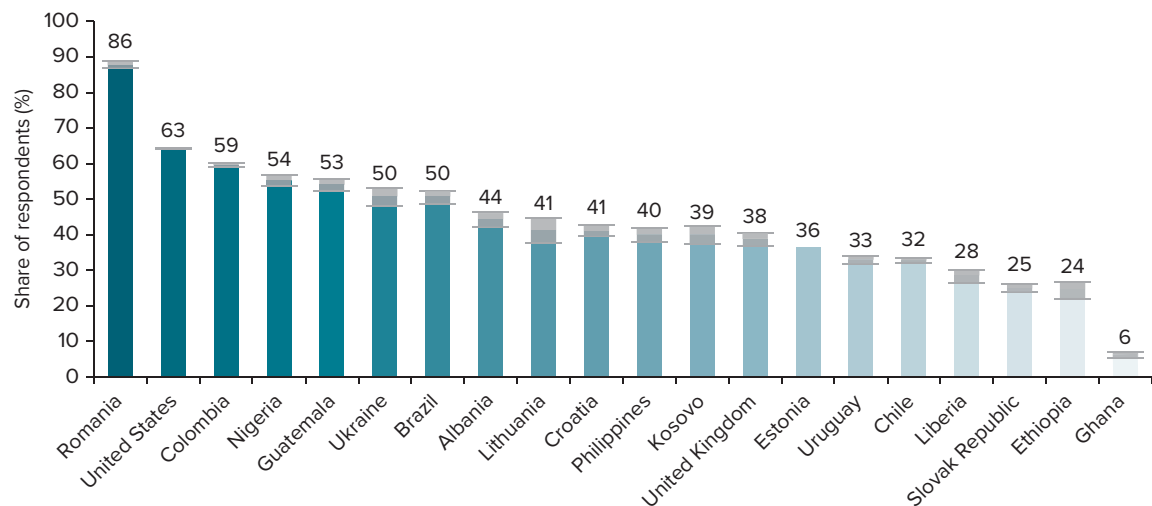
Longer questionnaires can generate survey fatigue, with potentially greater item nonresponse and survey dropout (Liu and Wronski 2017). For instance—though this is merely suggestive—the correlation coefficient between response rates and questionnaire length in eight of the nine countries reviewed is  $r = -0.29$ .<sup>4</sup> Question design can potentially mitigate such nonresponse. Chapter 22 of the *Handbook* assesses how to phrase questions so as to minimize item nonresponse.

## THE GLOBAL SURVEY OF PUBLIC SERVANTS AS AN INSTRUMENT FOR CROSS-COUNTRY SURVEY HARMONIZATION

As this chapter has illustrated, governments often use dissimilar questions and methodologies to measure similar concepts. As a result, even though governments measure similar concepts, they cannot benchmark themselves against other governments on these concepts. This puts a premium on evidence-based, cross-country harmonization of survey questionnaires and methodologies to further the degree of consistency in measurement across surveys of public servants.



**FIGURE 18.4** Share of Public Servants Satisfied with Their Pay and/or Total Benefits, Various Countries



Source: Fukuyama et al. 2022.

Note: Years of measurement vary by country. Colors denote the extent of job satisfaction, with darker shades signifying greater job satisfaction. The gray vertical bars denote 95% confidence intervals.

The GSPS was created with this objective in mind and, more broadly, to encourage the adoption of surveys of public servants by governments, good practice in public servant survey design and implementation, and the collection of cross-country and cross-institution data on public servants in governments around the world (Fukuyama et al. 2022). The aim is to increase the volume, quality, and coherence of survey data on public administration over time. The GSPS is the product of a consortium of researchers and practitioners from Stanford University, University College London (UCL), the University of Nottingham, and the World Bank.

To facilitate the harmonization of survey questions and methodologies for surveys of public servants, the GSPS presents existing questions and methods in an accessible form and provides methodological evidence on the efficacy of these questions and methods. It presents a core module of questions as a proposal for inclusion in independent surveys of public servants and publishes detailed guidance on the implementation of the core module to ease the comparison of any individual survey results with other surveys (Meyer-Sahling et al. 2021). This ensures that the data collected on public servants are comparable across independent data collection exercises.

Figure 18.4 provides an example of the type of comparison possible through the GSPS initiative, benchmarking governments on the percentage of public servants satisfied with their pay and/or total benefits. The GSPS enables governments to understand strengths and areas for development for their civil service in global comparative terms, although, as chapter 24 shows empirically, care needs to be taken when comparing responses across countries for culturally contingent concepts in particular. In figure 18.4, for instance, it is striking how differentially satisfied public servants are with their pay in countries at roughly similar levels of development, such as in the US federal government (63 percent satisfied with their pay) and the UK civil service (36 percent satisfied with their pay). This kind of comparison can help governments understand strengths and areas for development.

## CONCLUSION

The number of governments implementing governmentwide surveys of public servants has increased continuously in the last two decades, though many countries have yet to implement or institutionalize the implementation of employee surveys. Our review has shown that surveys of public servants in governments

are similar: they tend to measure similar concepts, focusing on a core set of employee attitudes (such as job satisfaction or engagement), on the one hand, and on management practices (such as the quality of leadership), on the other. They are thus implemented with a comparable set of measurement objectives.

At the same time, surveys across governments differ in the methodologies used and the precise measures applied to measure concepts. In terms of methodology, the review has found that surveys differ in key aspects: sampling approaches, survey weighting, survey populations, survey modes, questionnaire length, and response rates achieved. Some of these differences may stem from differences in practical or legal constraints. For instance, the civil service agency (or other entity) in charge of conducting the survey may not have a mandate for personnel management beyond the core civil service, complicating extending the survey coverage beyond the core civil service. And a central human resources management information system with demographic data about civil servants to construct survey weights may or may not be available, as detailed elsewhere in the *Handbook* (chapter 9). Some of the differences, however—for example, in sampling approaches and survey modes—are arguably due to limited methodological evidence on governmentwide surveys of public servants. The remaining chapters of this section of the *Handbook* address part of this void and can help governments make more evidence-based methodological choices in surveys of public servants. The GPS builds on this evidence to offer governments a globally comparable set of survey questions and methodologies.

In short, the global landscape of surveys of public servants holds much promise for the future. An ever-increasing number of governments are implementing surveys, better evidence for methodological choices in surveys of public servants is becoming available, and the GPS amplifies opportunities for global benchmarking.

## NOTES

1. In line with the varying terminology used by different governments conducting such surveys, we use the terms “public servant surveys” and “government employee surveys” interchangeably.
2. Information about nonresponse weights in Ireland, Republic of Korea, New Zealand, and Switzerland was, unfortunately, unavailable.
3. The pandemic led to a number of additional pandemic and remote-work-related questions in the surveys that would ordinarily not be asked, thus reducing the generalizability of comparisons of questionnaire length during the pandemic.
4. Response rates were unavailable for the Korean survey.

## REFERENCES

- Australian Public Service Commission. 2021. *Australian Public Service Employee Census: Explanatory Guide 2021*. Canberra: Australian Public Service Commission, Australian Government. <https://www.apsc.gov.au/initiatives-and-programs/workforce-information/aps-employee-census-2021#downloads>.
- Cabinet Office. 2020. *Civil Service People Survey 2020: Technical Guide*. London: Cabinet Office, United Kingdom Government. [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/977279/Civil\\_Service\\_People\\_Survey\\_2020-\\_Technical\\_Guide.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/977279/Civil_Service_People_Survey_2020_-_Technical_Guide.pdf).
- DANE (Departamento Administrativo Nacional de Estadística). 2020. *Metodología general encuesta sobre ambiente y desempeño institucional—EDI*. DSO-EDI-MET-001, version 4. Bogotá: Dirección de Metodología y Producción Estadística (DIMPE), Departamento Administrativo Nacional de Estadística, Government of Colombia. <https://www.dane.gov.co/index.php/estadisticas-por-tema/gobierno/encuesta-sobre-ambiente-y-desempeno-institucional-nacional-edi>.
- Department of Public Expenditure and Reform (Department of Public Expenditure, National Development Plan Delivery and Reform). 2020. *Civil Service Employee Engagement Survey*. Dublin: Department of Public Expenditure and Reform, Government of Ireland. <https://www.gov.ie/en/collection/5e7009-civil-service-employee-engagement-survey/#2020>.

- Fukuyama, Francis, Daniel Rogger, Zahid Husnain, Katherine Bersch, Dinsha Mistree, Christian Schuster, Kim Sass Mikkelsen, Kerenssa Kay, and Jan-Hinrik Meyer-Sahling. 2022. *Global Survey of Public Servants Indicators*. <https://www.globalsurveyofpublicservants.org/>.
- Government of Canada. 2022. "About the 2022/2023 Public Service Employee Survey." Government of Canada, October 21, 2022 (accessed March 27, 2023), <https://www.canada.ca/en/treasury-board-secretariat/services/innovation/public-service-employee-survey/2022-23/about-2022-23-public-service-employee-survey.html>.
- Korea Institute of Public Administration. 2021. *2021 Public Service Life Survey*. Seoul: Government Data Research Center, Korea Institute of Public Administration (accessed July 1, 2022), <https://www.kipa.re.kr/site/kipa/sta/selectReList.do?seSubCode=BIZ017A002>.
- Liu, Mingnan, and Laura Wronski. 2017. "Examining Completion Rates in Web Surveys via Over 25,000 Real-World Surveys." *Social Science Computer Review* 36 (1): 116–24. <https://doi.org/10.1177/0894439317695581>.
- Meyer-Sahling, Jan, Dinsha Mistree, Kim Mikkelsen, Katherine Bersch, Francis Fukuyama, Kerenssa Kay, Christian Schuster, Zahid Hasnain, and Daniel Rogger. 2021. *The Global Survey of Public Servants: Approach & Conceptual Framework*. Global Survey of Public Servants. <https://www.globalsurveyofpublicservants.org/about>.
- OECD (Organisation for Economic Co-operation and Development). 2016. *Engaging Public Employees for a High-Performing Civil Service*. OECD Public Governance Reviews. Paris: OECD Publishing. <https://doi.org/10.1787/9789264267190-en>.
- OFPER (Office Federal du Personnel). 2022. *Aperçu des résultats de l'enquête 2021 auprès du personnel*. Bern: OFPER, Federal Council, Switzerland (accessed July 1, 2022), <https://www.epa.admin.ch/epa/fr/home/themes/politique-du-personnel/enquete-aupres-du-personnel.html>.
- OPM (Office of Personnel Management). 2020. *2020 OPM Federal Employee Viewpoint Survey: Technical Report*. Washington, DC: US Office of Personnel Management, United States Government. <https://www.opm.gov/fevs/reports/technical-reports/technical-report/technical-report/2020/2020-technical-report.pdf>.
- Research New Zealand. 2021. *Technical Report: Te Taunaki Public Service Census 2021*. Report prepared for the Public Service Commission [Te Kawa Mataaho], New Zealand Government. Wellington: Research New Zealand. <https://www.publicservice.govt.nz/research-and-data/te-taunaki-public-service-census-2021/>.
- Statistics Canada. 2018. "Public Service Employee Survey (PSES): Detailed Information for 2017." Surveys and Statistical Programs, Definitions, Data Sources and Methods, Statistics Canada. <https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=384108>.

## CHAPTER 19

# Determining Survey Modes and Response Rates

## Do Public Officials Respond Differently to Online and In-Person Surveys?

*Xu Han, Camille Parker, Daniel Rogger, and Christian Schuster*

### SUMMARY

Measuring important aspects of public administration, such as the level of motivation of public servants and the quality of management they work under, requires the use of surveys. The choice of survey mode is a key design feature in such exercises and therefore a key factor in our understanding of the state. This chapter presents evidence on the impact of survey mode from an experiment undertaken in Romania that varied whether officials were administered the same survey face-to-face or online. The experiment shows that at the national level, the survey mode does not substantially impact the mean estimates. However, the mode effects have a detectable impact at the organizational level as well as across matched individual respondents. Basic organizational and demographic characteristics explain little of the variation in these effects. The results imply that survey design in public service should pay attention to survey mode, in particular in making fine-grain comparisons across lower-level units of observation.

### ANALYTICS IN PRACTICE

- Most governments—and many researchers—running surveys of public officials do so online. This reduces cost, increases flexibility, and theoretically reduces biases, such as those induced by respondents' notions of socially desirable answers.
- However, online surveys tend to have lower response rates than other survey modes and a greater degree of exit before surveys are completed, leading to different samples of respondents. This raises the concern

---

Xu Han was a consultant at the World Bank. Camille Parker is an economist at the United States Agency for International Development. Daniel Rogger is a senior economist in the World Bank's Development Impact Evaluation (DIME) Department. Christian Schuster is a professor at University College London.

that the data resulting from online surveys are not a valid representation of the population—in this case, the entire public administration.

- This chapter presents evidence from a randomized controlled trial that compares face-to-face and online survey responses. Our intention is to showcase an approach to measurement validation that can be followed by other survey teams for understanding the validity of their analyses.
- We show that the mean difference between online and face-to-face responses across all officials, which we call the *national level*, is between 0.17 and 0.35 standard deviations. Such an effect is of a similar magnitude to moving from 4.4 to 4.5 on a 1–5 scoring system (for example, “strongly disagree” to “strongly agree”) on one of the aggregate variables we study. Thus, in surveys with similar mode effects, measurement mode is unlikely to make a qualitative difference to conclusions when reporting at the national level so long as such small deviations are not overanalyzed.
- At the organizational level, the modal difference across all questions is roughly consistent with the country-level average. However, several organizations exhibit a modal difference of over one standard deviation. Given the lack of objective benchmarks, we interpret sensitivity to mode as indicative of underlying measurement issues. Problems arising from sensitivity to measurement are particularly acute when ranking organizations, with mode effects having substantial impacts on the ordering of organizations. This evidence casts doubt on the validity of organization-level ranking that does not appropriately address these measurement concerns.
- At the individual level, the mode effects remain significant and substantial for most of the outcomes. We see that the survey mode effects persist across individuals matched using propensity score matching (PSM) as well across different groups, like managers and nonmanagers, although some groups appear more sensitive to survey mode than others. This evidence places a burden of proof on survey analysts to demonstrate the validity of presenting data at the unit or individual level.
- A common approach to correcting online surveys is to use survey weights. In our experiment, we find little evidence that survey weights reduce the sensitivity of results to the measurement approach.
- Identifying organizations and individuals particularly susceptible to mode effects would allow for a significant reduction in aggregate mode effects. This might be pursued through a small, face-to-face survey across organizations, upon which estimates of individual mode responses could be based.

## INTRODUCTION

Measuring many aspects of public servants and their working lives is difficult. Management quality is frequently experienced rather than recorded in administrative data. Public employees’ motivations are difficult to observe outside of their own expressions of their motives. Thus, self-reporting through surveys becomes the primary means of measurement for many aspects of officialdom. Externally sourced measures, perhaps from administrative data, are simply unable to record features of these important variables.

Survey design is therefore an important mediator in our understanding of the state. This part of *The Government Analytics Handbook* assesses how to determine the particular content of a survey of public servants from multiple angles. This chapter focuses on a key aspect of survey administration: whether the survey is conducted online or in person (that is, face-to-face). Though there are other modes of survey delivery, from periodic text message surveys to laboratory-in-the-field games, the debate in this context typically concerns these two forms, which will therefore be our focus (Haan et al. 2017; Heerwegh and Loosveldt 2008; Kaminska and Foulsham 2014; Tourangeau and Yan 2007).

Public servant surveys run by governments are typically carried out online, with a small or nonexistent proportion of staff allowed to use a paper form or speak to an enumerator directly (for a review of the most prominent such surveys, see chapter 18). This is done predominantly for cost reasons, but online surveys enjoy several advantages. They enable researchers to rapidly collect large amounts of data and can be quickly and flexibly deployed across a range of organizational contexts.

However, this reliance on online surveys is based on the rarely tested assumption that online surveys are able to provide valid and reliable data. This assumption may be incorrect for several reasons: online surveys often suffer from low response rates, potentially undermining the representativeness of the respondent group (Cornesse and Bošnjak 2018). Online surveys are also associated with higher levels of survey drop-off and item nonresponse than other survey modes (Daikeler, Bošnjak, and Lozar Manfreda 2020; Heerwegh and Loosveldt 2008; Peytchev 2009). The resulting higher levels of missing values may undermine the validity and reliability of the data (Baumgartner and Steenkamp 2001; Jensen, Li, and Rahman 2010; Podsakoff et al. 2003).

Face-to-face surveys can be a viable alternative to online survey data collection. Many microempirical studies, in which the individual is taken as the unit of observation, prefer to administer surveys in person. Although they consume significantly more time and resources than online surveys, face-to-face surveys tend to report significantly higher response rates and lower rates of breakoff, and they can be substantially longer without respondent exit. Talking to someone in person is a fundamentally more engaging activity than filling in a form on the screen, enabling a wider range of data to be collected from a single interview.<sup>1</sup> It is therefore possible that the final set of responses collected from an online survey will come from a different effective sample than would be the case in the face-to-face mode (see, for example, Couper et al. 2007).

We turn now from respondents to the answers they provide. A key feature of online surveying is that it distances the respondent from an enumerator. This potentially reduces social-desirability bias arising from a respondent's inclination to answer in a way that may be demanded by the social features of a face-to-face survey (Heerwegh 2009; Newman et al. 2002; Tourangeau and Yan 2007). An online survey is also relatively consistent in its delivery of a survey to respondents, while individual enumerators may not be.

Despite the potential reduction in social-desirability bias (Ye, Fulton, and Tourangeau 2011), online surveys may introduce other biases—for example, those derived from a lack of comprehension of the question. Where enumerators can provide clarifications, online surveys typically do not have that option, nor is it likely to be regularly used by respondents. It has also been shown that the online survey respondents engage in a larger degree of *satisficing*—that is, they more often respond “I don’t know,” skip questions, choose neutral response options, etc. to minimize the cognitive burden of responding (see, for example, Heerwegh and Loosveldt 2008; Krosnick and Presser 2010; see section two below for further discussion). Whereas the desire to satisfice is also present in face-to-face surveys, an experienced enumerator might probe respondents to, for example, think for a while about a question rather than saying “I don’t know.” Therefore, another concern is that even comparable samples of respondents may provide different responses if surveyed using different survey modes.

A series of trade-offs therefore characterizes the choice between online and face-to-face survey modes. Conceptually, there may be differences in what sample of respondents each mode attracts and how the mode affects the responses they provide. Practically, the costs and feasible lengths of the two approaches differ. While researchers and research communities typically have strong beliefs about which approach optimally resolves this tension, there is little to no rigorous empirical evidence on this subject in the field of public administration.<sup>2</sup>

The nature of public administration, with its hierarchical and bureaucratic communication norms, potentially implies a substantial survey mode effect. For example, written communication at work, such as filling in an online form or survey, may be regarded very differently by a public official and a private citizen. On the other hand, a 1-hour meeting to discuss public service life is similar to many of the meetings public officials have in a day. Findings from other sectors may therefore not be externally valid in a public administration setting.



What, therefore, are public sector managers or researchers to do in collecting survey data from public servants? This question is complicated by the fact that many features of public administration, as noted above, cannot be definitively validated outside of survey data. It can be argued that the appropriate conception of management is the individual employee's specific experience of it. Thus, objective data for the purpose of benchmarking the two most common survey modes are absent for many topics. The answer to the question may also vary across topics, individuals, and settings, such that an effective answer must go beyond a simple comparison of aggregate means to understand what quantities are most affected by survey mode.

While the existing literature is an obvious foundation for our analysis, our aim in this chapter is to investigate the robustness of survey results to survey mode within a public administration setting. Given the difficulties of generating objective benchmarks for many of the topics we study, our interpretation of this robustness is used as an indicator of the validity of the underlying responses. Where feasible, we also investigate the organizational and individual determinants of mode effects, with the aim of better understanding which groups or organizations may be most impacted by differences in survey mode.

Our intention in this chapter is to showcase to survey managers and related stakeholders an approach to testing the robustness of survey responses to survey mode. We provide evidence from a single experiment to illustrate our approach, but in doing so, we provide some of the first experimental evidence on the impacts of survey mode in public administration. As such, this chapter hopes to provide frontier evidence from a single setting and a framework for investigating these issues in other surveys.

The rest of this chapter proceeds as follows. Section two outlines the existing literature on survey mode effects and how it relates to the public administration setting. A major gap in the literature on mode effects in surveys of public servants is the absence of an experimental comparison between the two modes. We address this gap through a field experiment with 6,037 public servants in 81 government institutions in Romania, in which we randomly assign each official to complete either a face-to-face or an online survey. The survey's content replicates that found in typical government employee surveys, covering both employee attitudes and management practices. By studying survey responses across the two modes with a high degree of heterogeneity in response rates, we can disentangle survey mode effects at the point of response from nonresponse bias due to the lower take-up of online compared to face-to-face surveys. Given the frontier nature of this empirical evidence, sections three to five investigate the impacts of survey mode within this data set. Section six discusses the implications of our findings for the implementation of public servant surveys and further research.

## LITERATURE REVIEW

The existing literature on survey mode effects in general finds that the survey mode has significant impacts on the robustness of survey estimates across three primary dimensions: response rates, survey breakoff, and survey responses.

### Response Rates

Much of the existing research on survey modes has focused on the difference in response rates between modes. In general, online surveys have been found to have significantly lower response rates compared to all other survey modes, including face-to-face (Biemer et al. 2018; Lozar Manfreda et al. 2008; Shih and Fan 2008). While not specific to public administration, a recent meta-analysis conducted by Daikeler, Bošnjak, and Lozar Manfreda (2020) summarizes the results of 114 experimental studies conducted among many different populations (students, the general public, businesses, and employees), on diverse topics (public opinion, technology, lifestyle, job, etc.), by various sponsors (academic, governmental, and commercial),

both with and without participation incentives, and with varying recruitment strategies, prenotification methods, and solicitation methods. They found that in aggregate, online surveys have response rates that are 12 percentage points lower than all other survey modes.<sup>3</sup>

Those who do respond to online surveys tend to differ from respondents to other survey modes across several demographic characteristics, spurring concerns over the representativeness of online samples. For instance, several studies have found that online survey respondents tend to be younger and more educated than face-to-face survey respondents (Braekman et al. 2020; Couper et al. 2007; Duffy et al. 2005). A recent meta-analysis suggests that online surveys are associated with higher nonresponse biases than other survey modes (Cornesse and Bošnjak 2018). It is also worth noting that differences between respondents and nonrespondents are attributed more to the noncoverage of some population subgroups in the sample frame than to the nonresponse of people invited to participate in surveys (Couper et al. 2007). Online surveys of public servants are more likely to have a complete sample frame and, therefore, are less susceptible to nonresponse biases than online surveys of general populations.

Within public administration, there is a high level of heterogeneity in terms of response rates to existing large-scale, online public administration surveys in Organisation for Economic Co-operation and Development (OECD) countries. As shown in table 18.2 in chapter 18, while some large-scale public administration surveys, such as the survey administered in Colombia, enjoy response rates around 80 or 90 percent, others, such as the Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS) in the United States, have struggled to bring their response rates above 50 percent and have been experiencing a steady decline in overall response rates in the past five years.<sup>4</sup> Public administration surveys in non-OECD countries exhibit similarly heterogeneous response rates, ranging from 11 percent in Brazil to 47 percent in Albania. Troublingly, despite these surveys' importance in shaping public administration organizations' priorities as they relate to hiring, employee engagement, and performance management, among other topics, the question of whether declining response rates to online surveys present a threat to the overall validity of inferences about public officials drawn from the data has not been extensively studied in the public administration literature.

While response rates for country surveys tend to remain relatively consistent at the national level over time, there is a high degree of variation in survey response rates at the organizational level. For example, in the 2019 FEVS, response rates within US government organizations ranged from 86 percent to just 27 percent. While research on survey response rates in public administration is limited, the research that does exist posits several potential explanations for this variation at the organizational level. Some researchers have argued that low employee morale in certain agencies may contribute to declining response rates (de la Rocha 2015). Others, while *not* explicitly studying survey response, have found a positive relationship between voluntary behavior (such as taking a survey) and employee engagement levels (Rich, Lepine, and Crawford 2010), suggesting that organizations with higher levels of employee engagement may also experience higher response rates to employee surveys. Similarly, public employees with strong public service motivation or organizational commitment have been found to be more willing to perform extra-role tasks, including filling out surveys (Moynihan and Pandey 2010; Newell et al. 2010). Other researchers have identified links between response rates and individuals' attitudes toward the survey's sponsor institution. For instance, in a study of university students, Spitzmüller et al. (2006) find that survey nonrespondents are less likely to believe that their university values their contributions or cares about their well-being.

These differences between online respondents and nonrespondents to government surveys suggest that variation in response rates may significantly impact the degree to which online surveys provide unbiased estimates of public employees' perceptions and behaviors. In addition, the proclivity of managers and researchers to compare survey responses across organizations or other subgroups means that variation in response rates may lead to the comparison of differential subgroups of staff (Groves 2006). The self-selection issues in public administration surveys are less of a concern in the face-to-face mode because most surveys of this type record response rates close to 100 percent. For example, the Romanian face-to-face survey analyzed here collected responses from 3,316 out of 3,592 sampled individuals,

yielding a response rate of 92 percent. Similar surveys in different settings give comparably high response rates: for example, Guatemala (96 percent) and Ethiopia (94 percent). Assuming successful random sampling, the almost-perfect response rate minimizes any issues arising from differences between survey respondents and nonrespondents in the face-to-face mode.

## Survey Breakoff

Beyond impacting survey estimates through differential response rates, the survey mode can also impact survey estimates through different rates of breakoff. Overall, online surveys are associated with significantly higher rates of survey breakoff because they are generally less able to maintain respondents' interest and attention throughout the duration of the survey (Galesic 2006; Haan et al. 2017; Heerwegh and Loosveldt 2008; Kaminska and Foulsham 2014; Krosnick and Presser 2010; Peytchev 2009; Steinbrecher, Roßmann, and Blumenstiel 2015). This threat of breakoff can be significant: meta-analyses of the issue have found online surveys experience breakoff rates between 16 and 34 percent (Lozar Manfreda and Vehovar 2002; Musch and Reips 2000).

The ability to maintain respondents' interest throughout the survey varies depending on several survey design features, including the presence of long blocks of questions and the overall time it takes to complete the survey (Galesic 2006; Peytchev 2009; Steinbrecher, Roßmann, and Blumenstiel 2015). Many of the demographic characteristics associated with survey response are also associated with higher levels of survey breakoff, with younger, more educated respondents generally being more likely to exit an online survey before completing it (Peytchev 2009). We provide more information on this in chapter 22.

Within the public administration sector, the issue of survey breakoff has not been extensively studied, and statistics on survey breakoff in major public administration surveys, such as the FEVS, are generally not made publicly available. In the 2019 survey of the Australian Public Service, approximately 92.5 percent of respondents who began the survey completed it, for a breakoff rate of 7.5 percent (N. Borgelt, Australian Public Service Commission, pers. comm., June 24, 2020). Consistent with the survey research literature, breakoff was the highest among long blocks of matrix-style questions and questions involving a reasonably high cognitive load (such as a question asking respondents how many sick days they had taken over the last 12 months) (Peytchev 2009; Steinbrecher, Roßmann, and Blumenstiel 2015; Tourangeau, Conrad, and Couper 2013).

This evidence implies a similar concern as the above for valid inference. Comparisons of questions with higher and lower rates of breakoff may differ simply due to the subgroups that respond to them and are thus vulnerable to endogenous selection concerns. If the most self-motivated individuals are more likely to respond to motivation questions, then an analysis of these variables relative to management variables may incorrectly imply the relative importance of self-motivation over management. Once again, this issue is often minimized by a face-to-face survey interview. Such settings make the survey process more engaging to the respondent and add a possible social cost to ending the interview midstream, as this might be seen as "impolite" to the enumerator (Peytchev 2006).

## Survey Responses

Finally, a substantial portion of the existing survey research literature has focused on the degree to which survey modes may impact the magnitude of survey responses directly. In general, online survey respondents tend to exhibit lower levels of motivation to answer survey questions and often pay less attention when answering questions compared to face-to-face respondents (Kaminska and Foulsham 2014; Krosnick 1991). Several studies have found that online surveys are associated with higher rates of satisficing behaviors, including selecting "I don't know" or "N/A" response options, providing less differentiation across groups of responses, and providing more neutral responses (for example, "Neither agree nor disagree" or "Neutral") than face-to-face surveys (Duffy et al. 2005; Haan et al. 2017; Heerwegh and Loosveldt 2008). Online surveys

are also more likely to produce noncontingent responses (NCR), wherein there is a substantial difference between survey items that are expected to be highly correlated with each other (Heerwegh and Loosveldt 2008; Krosnick and Presser 2010). These kinds of responses imply that respondents may have simply selected answers at random or read through survey items carelessly in order to quickly complete the survey (Anduiza and Galais 2017). Taken together, these satisficing behaviors can reduce the validity and reliability of online responses (Baumgartner and Steenkamp 2001; Podsakoff et al. 2003).

At the same time, however, the existing literature suggests that online surveys may be better at eliciting candid responses to sensitive questions. Because online surveys provide respondents with a higher level of anonymity than face-to-face surveys, online survey respondents tend to be more likely to respond truthfully to questions related to socially sensitive topics (Gnambs and Kaspar 2015; Kays, Gathercoal, and Buhrow 2012; Tourangeau and Yan 2007). In the context of public administration, these findings suggest that online surveys may be particularly advantageous when measuring sensitive topics, such as ethics violations, turnover, or evaluations of organizational performance. However, the applicability of these findings to public administration has not been rigorously studied, and there is limited knowledge about the relevance of survey mode on the validity of data collected through these studies.

## A SURVEY MODE EFFECTS EXPERIMENT

We address a number of these gaps in the existing literature on mode effects through a field experiment with 6,037 public servants in 81 government institutions in Romania. We randomly assigned each target respondent to complete either a face-to-face or an online survey covering several topics typical of public administration surveys: recruitment, performance appraisal, turnover, dismissal, salary, motivation, goal-setting, leadership, and ethics.<sup>5</sup>

### How Does the Survey Mode Impact Response Rates?

Our face-to-face survey has high response rates across most government institutions, with an average of 92.5 percent, while our online response rate—consistent with other online government employee surveys—varies widely across government institutions and ranges from a maximum value of 100 percent (5 organizations) to a minimum of 0 percent (13 organizations). For the purposes of this analysis, we remove both face-to-face and online observations from organizations who declined to participate in the online survey, as well as organizations with online response rates of less than 5 percent.<sup>6</sup> After this removal, the sample comprises of 4,819 public servants in 50 government institutions. Figure 19.1 presents the remaining heterogeneity in organizational response rates, with an average response rate across organizations of 86.2 percent in the face-to-face mode and 53.8 percent in the online mode.

We use heterogeneity in online response rates across organizations to disentangle survey mode effects at the point of response from nonresponse bias due to lower take-up of online surveys. By comparing questions in high online-response organizations with their face-to-face equivalents, we can abstract from selection bias. By comparing bias across the full sample, we can investigate the role of response rate in question differences.<sup>7</sup>

### How Does the Survey Mode Affect the Distribution of Respondent Characteristics?

Table 19.1 shows the results of *t*-tests conducted between the online and face-to-face survey samples across several key demographic groups. Given that our face-to-face survey is a representative sample from staff lists and has a high average response rate, it can be seen as a reflection of the true distribution of characteristics of

**FIGURE 19.1** Online and Face-to-Face Survey Response Rates, by Organization

Source: Original figure for this publication.

**TABLE 19.1** Balance in Demographic Characteristics between Surveys

Variable	N	(1) Face-to-face sample mean [SE]	N	(2) Online sample mean [SE]	T-test difference (2)–(1)
Age	2,137	45.804 [0.191]	2,682	45.392 [0.167]	–0.412
Years worked in position	2,137	7.423 [0.136]	2,682	8.029 [0.132]	0.607***
Years worked in organization	2,137	11.565 [0.174]	2,682	10.828 [0.149]	–0.737***
Years worked in public administration	2,137	14.719 [0.175]	2,682	13.893 [0.154]	–0.826***
Employee status (1 = Civil servant)	2,137	0.873 [0.07]	2,682	0.91 [0.006]	0.037***
Gender (1 = Male)	2,137	0.31 [0.01]	2,682	0.26 [0.008]	–0.051***
Highest level of education attained: less than college (1 = Yes)	2,137	0.033 [0.004]	2,682	0.04 [0.004]	0.006
Highest level of education attained: undergraduate degree (1 = Yes)	2,137	0.474 [0.011]	2,682	0.433 [0.01]	–0.041***
Highest level of education attained: master's degree (1 = Yes)	2,137	0.453 [0.011]	2,682	0.481 [0.01]	0.028*
Highest level of education attained: PhD (1 = Yes)	2,137	0.035 [0.004]	2,682	0.037 [0.004]	0.001

Source: Original table for this publication.

Note: The values displayed for t-tests are the differences in means between the two survey modes (face-to-face and online).

Significance level: \* = 5 percent, \*\* = 1 percent, \*\*\* = 0.1 percent.

public servants. Thus, differences between the two reflect a deviation of the online survey from a representative sample.

Consistent with the existing literature, we find many statistically significant (at the 1 percent level) deviations from the population's values in the sample of online survey respondents. Most noticeably, 31 percent of face-to-face survey respondents are male, compared to only 26 percent female.<sup>8</sup> They are also relatively

less educated, with 47.4 percent having an undergraduate degree and 48.8 percent having a Master's degree or PhD, whereas, for online respondents, these numbers stand at 43.3 percent and 51.8 percent, respectively.<sup>2</sup> Moreover, 87.3 percent of face-to-face respondents are civil servants (as opposed to contractors), compared to 91 percent of online respondents. We also find statistically significant differences in average tenure, but being below one year, these differences appear to be of limited magnitude.

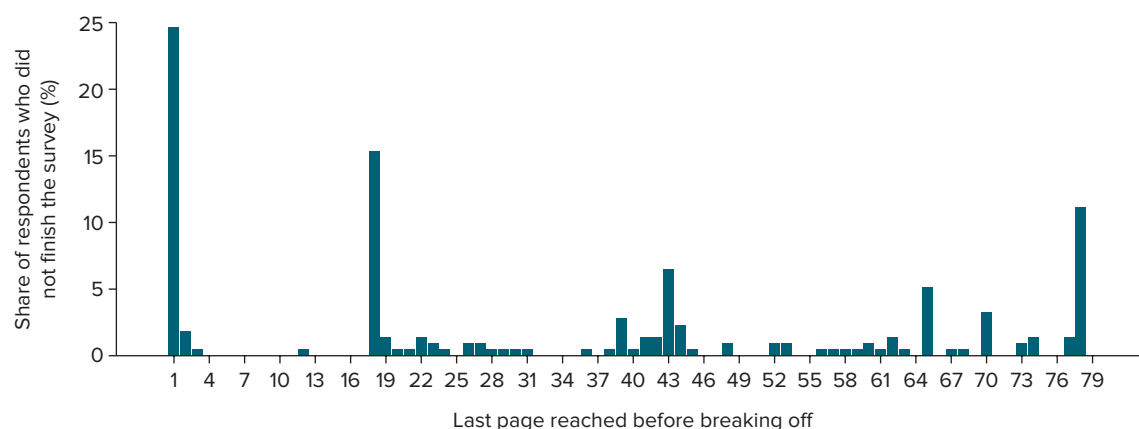
Overall, our data reflect the frequent finding that face-to-face and online samples differ along a range of margins. As many of these variables, like gender, education, and contract status, can affect survey responses, table 19.1 provides an initial rationale to look deeper into the differences between modes in the Romania survey.

## How Does the Survey Mode Affect Survey Breakoff and Item Nonresponse?

Our online survey also exhibits considerably higher levels of survey breakoff than the face-to-face survey. While the breakoff rate for the face-to-face survey is almost zero, the breakoff rate for the online survey is approximately 10 percent (see figure 19.2 below, as well as figure G.2 in appendix G for the breakoff pattern by mode). While many major civil service surveys do not generally publicize their levels of survey breakoff, the evidence that does exist suggests that the breakoff rate in our survey is, generally speaking, consistent with similar public administration surveys and lower than average for surveys in general. For example, in 2019, the Australian Public Service Employee Census had a breakoff rate of 7.5 percent in its online survey (N. Borgelt, Australian Public Service Commission, pers. comm., June 24, 2020). Overall, online surveys of the general population experience an average breakoff rate of 16–34 percent (Lozar Manfreda and Vehovar 2002; Musch and Reips 2000), which suggests that civil servants are more likely to complete a survey once started.

Interestingly, as shown in figure 19.2, the largest proportion (just under a quarter of the total) of survey breakoff in the online survey occurred on the first page, suggesting that encouraging individuals to start the survey is the biggest hurdle to obtaining a complete response.<sup>10</sup> Survival analysis conducted on the profile of individuals who dropped out of the survey (using a Cox-Weibull hazard model) finds that demographic characteristics are poor predictors of breakoff. Only the age variable appears to have a relatively consistent impact on breakoff, with individual age, as well as average age at the organization as a whole, increasing the chances of respondents' finishing the survey (for a full summary of findings, see appendix G, table G.2).

**FIGURE 19.2** Online Survey Breakoff, by Page



Source: Original figure for this publication.

Note: Minimum page = 1; maximum page = 79.



In addition to analyzing the individuals who dropped out of the survey, we also examine the profile of those who dropped out of the survey and returned to complete it later. Overall, 326 individuals dropped out of the online survey and returned to complete it later.<sup>11</sup> The vast majority of these individuals (80 percent) returned to the survey within one day of exiting it. However, several individuals did not return to the survey for several weeks, suggesting that subsequent reminders to complete the survey may have spurred them to revisit it.<sup>12</sup> There are no notable demographic differences between these individuals and the broader survey sample.

Table 19A.3 also shows that even the individuals who do not exit the online survey altogether are less likely to provide responses. The online mode of delivery is associated with all types of item nonresponse, with individuals being more likely to say “I don’t know,” to refuse to respond, and to skip questions. Chapter 22 discusses in greater detail the issues and determinants of item nonresponse, so here we only note that apart from larger survey nonresponse, differential demographic characteristics, and higher breakoff rate, the rate at which respondents omit particular questions should also be on the radar of researchers using online surveys, as this value is significantly larger than in equivalent face-to-face surveys.

## SURVEY MODE EFFECTS ON THE VALIDITY AND RELIABILITY OF DATA

As seen above, online surveys have lower response rates, attract a nonrepresentative sample of the survey population, and suffer from survey exit more frequently than face-to-face surveys. This suggests that the *process* of responding to an online survey differs from the process of responding to a face-to-face one. But the critical question is whether any of this matters for the *measurement of outcomes* that the surveys yield. Since we undertake a randomized controlled trial that exogenously separates individual respondents into in-person and online enumeration modes, we can compare the results reached by these two methods to investigate the validity and reliability of the corresponding data. These are clearly the two most important outcomes of any change in measurement approach.

As described above, assessing which survey is best able to reflect the underlying truth is complicated by the fact that the survey mode impacts responses directly as well as through sample selection. Since we are dealing with concepts such as management and motivation that are difficult to proxy with objective data in public administration settings, our focus is on investigating the scale and determinants of any difference in the quantities the two modes yield. We interpret significant changes in question outcomes as implying vulnerability to measurement outcomes, thereby undermining the robustness of our estimates from any single approach.

### Does the Survey Mode Make a Difference to Question Values?

In order to ascertain the degree to which the survey mode impacts survey estimates, we undertake an analysis with respect to the mean mode difference in survey question responses. We average the responses into three indexes: management, motivation, and ethics. In all three cases, higher index values indicate more-positive, or “desirable,” traits, like exemplary leadership, job satisfaction, and aversion to bribe-taking.<sup>13</sup> The management index presents the average of a series of survey items related to managerial practices and performance management. The motivation index shows the average of survey items related to employees’ levels of motivation and engagement in their work. Finally, the ethics index aggregates the average of survey items related to employees’ perception of the prevalence of ethics violations in their organization. These dimensions reflect three of the most commonly investigated areas of public sector life in public servant surveys (see figures 18.2 and 18.3 in chapter 18).

In all instances, we compare the survey mode effects by calculating the mean response from the online survey minus the mean response from the face-to-face survey. A negative mean difference thus implies that the face-to-face survey produces higher average estimates (that is, more-positive responses) than the

**TABLE 19.2 Mean Modal Difference, by Level of Analysis**

	Mean	Minimum	Maximum	p25	p50	p75
<i>(1) National level</i>						
Management index	-0.239					
Motivation index	-0.350					
Ethics index	-0.208					
<i>(2) Organizational level</i>						
Management index	-0.331	-1.925	0.978	-0.617	-0.258	0.081
Motivation index	-0.308	-1.194	0.831	-0.660	-0.348	-0.039
Ethics index	-0.171	-1.430	1.099	-0.401	-0.134	0.121
<i>(3) Individual level</i>						
Management index	-0.242	-4.600	3.864	-1.196	-0.255	0.692
Motivation index	-0.312	-8.365	5.611	-1.247	-0.312	0.623
Ethics index	-0.196	-7.879	7.879	-0.563	0.000	0.000

Source: Original table for this publication.

Note: Panel (1) shows the full-sample differences in the means of the indexes between the online and face-to-face survey modes ( $\hat{x}_{online} - \hat{x}_{f2f}$ ). Panel (2) calculates these differences at the level of each organization and summarizes their values for mean level

$\left( \left[ \frac{1}{50} \right] \left[ \sum_{org=1}^{50} [\hat{x}_{org,online} - \hat{x}_{org,f2f}] \right] \right)$  and other key distribution statistics. Panel (3) shows the distribution of differences in index values

between individuals matched on the following variables: organization, job tenure, organization tenure, public administration tenure, pay grade, employee status (civil servant vs. contractual staff), age, gender, and education level. Propensity score matching estimators impute the missing potential outcomes for each treated subject by using the average of the outcomes of similar subjects that receive the other treatment. Observations are matched using nearest-neighbor matching and the probability of treatment is calculated using a logit model. In the case of a tie, observations are matched with all ties with the corresponding difference averaged out.

online survey. For ease of interpretation and unless otherwise indicated, the differences are presented in terms of z-scores, so coefficients are in standard deviations.<sup>14</sup> Table 19.2 presents the mean survey mode effects across statistics calculated at the national, organizational, and individual levels. These three levels are discussed in turn below.

### Country-Level Quantities

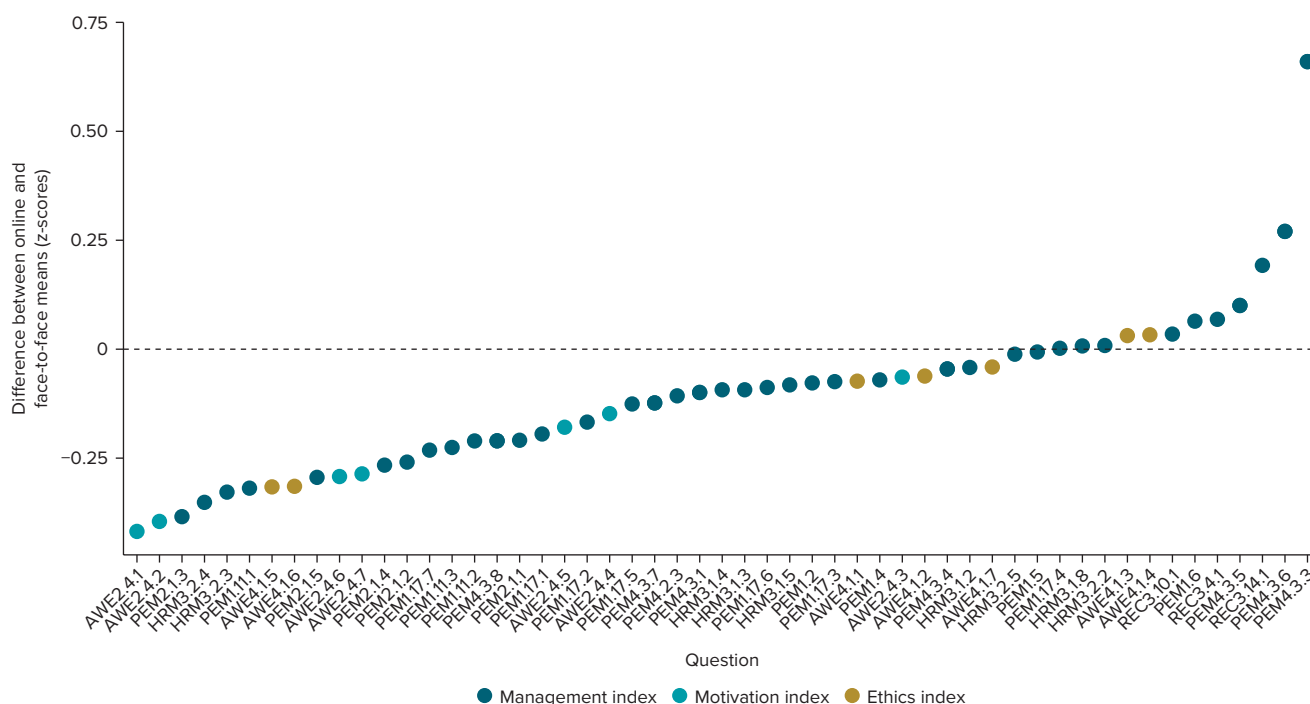
At the national level (panel 1 of table 19.2), we calculate the mean difference across all civil servants as the average score of the index in the online sample minus the average score on the same index in the face-to-face sample.

We see that the differences range from -0.208 for the ethics index, through -0.239 for the motivation index, to -0.350 for the management index. All of the average modal differences are negative, implying that the estimates produced by face-to-face surveys are, on average, higher and therefore point toward more-positive, or “desirable,” responses than those produced by online surveys.

The effect size of these differences on a 1–5 Likert scale is moving the average around 0.1 higher for the face-to-face sample than the online sample. Thus, the evidence from this experiment is that survey mode effects are small for most questions in data aggregated across all respondents. Reporting at this level seems relatively robust to the mode of data collection.

The average survey mode effects are an artifact of the survey mode effects associated with the particular questions composing a given index. Figure 19.3 presents survey mode effects by question item across all items included in the three indexes outlined above.<sup>15</sup> The survey mode effects vary considerably among individual question items for each index. Some items within each of the indexes are more sensitive to survey mode effects than others (Braekman et al. 2020; Gnambs and Kaspar 2015; Ye, Fulton, and Tourangeau 2011).

**FIGURE 19.3** Average Item Difference, by Survey Topic



Source: Original figure for this publication.

Note: For the question text, see table G.8 in appendix G. AWE = attitudes and work environment; HRM = human resources management; PEM = performance management; REC = recruitment.

For instance, while the ethics index as a whole exhibits significant negative survey mode effects, at the item level, two items (“How frequently do employees in your institution observe unethical behavior among colleagues?” and “How frequently do employees in your institution report colleagues for not behaving ethically?”) appear highly sensitive to survey mode, with mean differences of  $-0.40$  and  $-0.37$  standard deviations, respectively. The three other items that compose the ethics index (“How frequently do employees accept gifts or money from companies?,” “How frequently do employees accept gifts or money from citizens?,” and “How frequently do employees pressure other employees not to speak out against unethical behavior?”) all have mean mode differences close to zero.<sup>16</sup>

At the national level, all of the mode effects exhibited in figure 19.3 are within relatively limited thresholds. Even for topics such as ethics, we find limited average mode effects across the population.

### Organization-Level Quantities

At the organizational level (panel 2 of table 19.2), we calculate the mean difference as the average difference in online and face-to-face scores across each organization. For example, an organization’s management index score as determined by the results of the face-to-face survey is subtracted from an organization’s management index score as determined by the results of the online survey. These differences within organizations are then averaged to produce the mean difference in index scores. Other statistics relating to the distribution of scores across organizations are also shown.

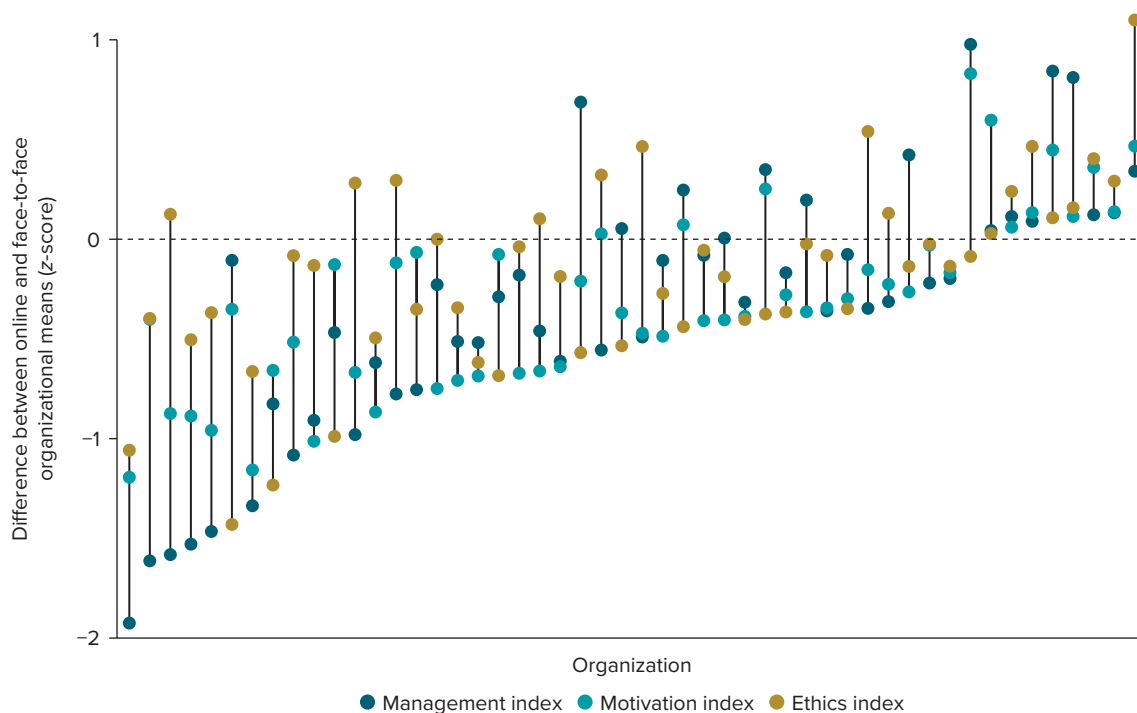
The *average* coefficients at the organizational level are not unlike those at the national level (perhaps naturally, since we are now simply producing a weighted correspondence of the national statistics). The change relative to the national level is the largest for the management index, where the mode difference increases by 38 percent. Still, the overall magnitude and direction of mean mode differences point us to the same conclusion of more-negative responses in the online mode.

However, we also see a high degree of heterogeneity in mode effects across organizations, implying that organizational characteristics may mediate respondents' experience of the survey and its mode of delivery. As shown in figure 19.4, organizations present highly varied responses to the mode of measurement. For instance, while the average mode difference across organizations for the management index is 0.331 standard deviations, seven organizations display differences above one standard deviation between the survey modes on that index. Given that the difference between organizations scoring the lowest and the highest on the management index is just above two standard deviations, this value implies a considerable impact of the survey mode on respondents *within* some organizations. Comparably large differences for some organizations are also observed for other indexes. Figure 19.4 further confirms that the survey mode effects differ across topics, as some organizations have largely different mode effects depending on the index chosen.<sup>17</sup>

Thus, in statistics produced at the organizational level, we start to see substantial effects of the mode of measurement, especially for a subportion of our sample. Ordinary least squares (OLS) regressions examining the relationship between the aggregate mean difference and organizational characteristics, such as organization size, gender composition, and average age, provide little evidence of the determinants of mode effects. This suggests that it is organizational characteristics typically unobservable in a public officials survey that are driving survey mode effects (for a full summary of results, see appendix G, table G.4).

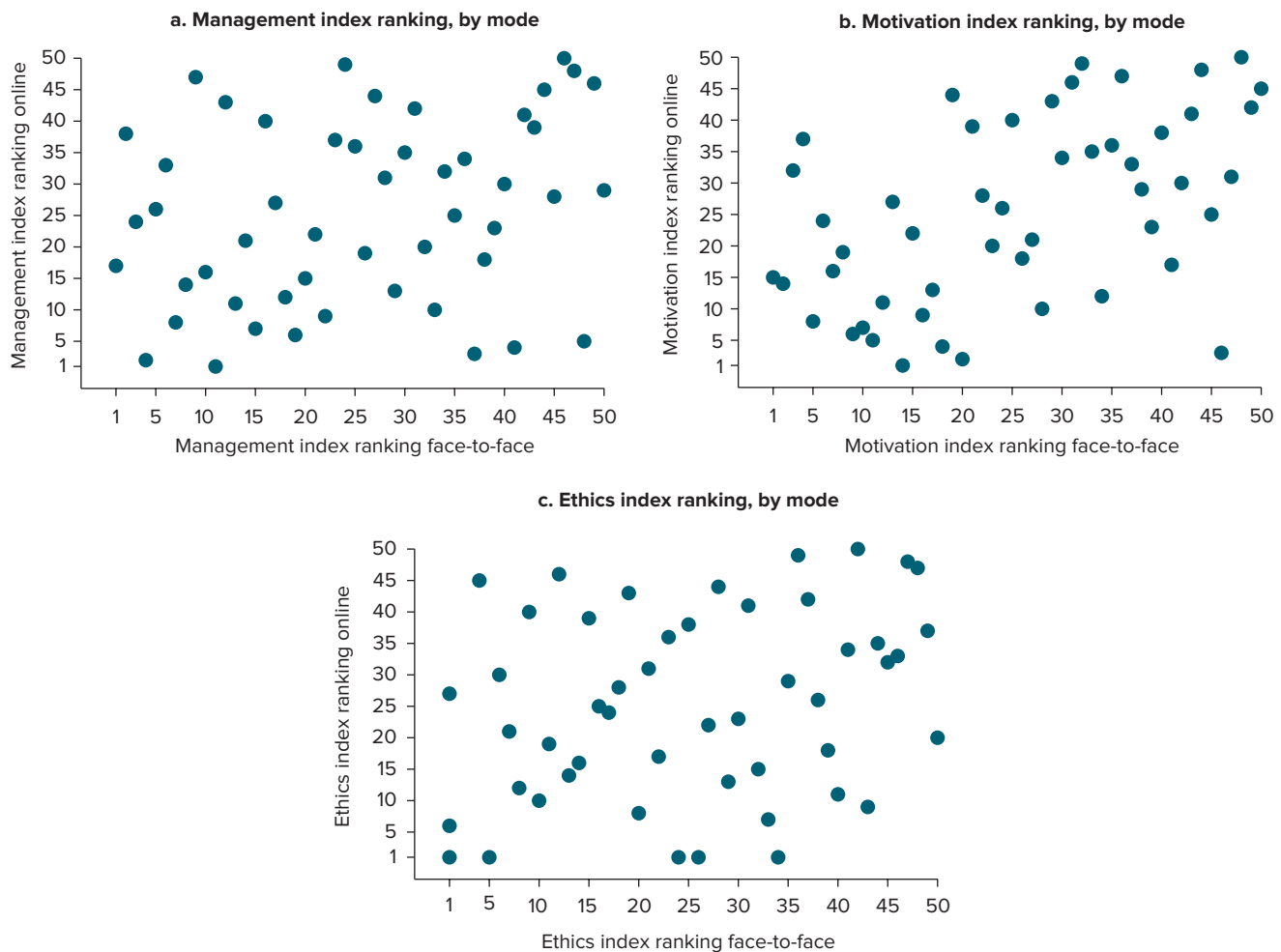
Building on the discussion in chapter 20, these heterogeneous mode effects at the organizational level are of particular concern to policy makers if they intend to present survey results as organizational rankings. Specifically, we find that the rank of a public sector organization (that is, its place on a list of organizations sorted in descending order of the value of a given index) as determined by the online survey correlates only poorly with its rank as determined by the face-to-face survey, across all three indexes.<sup>18</sup> Figure 19.5 plots organizations' ranks according to the face-to-face (*x* axis) and online

**FIGURE 19.4 Average Modal Difference, by Organization**



Source: Original figure for this publication.

**FIGURE 19.5 Organization Rankings, by Index and Mode**



Source: Original figure for this publication.

(y axis) surveys for the three indexes we focus on.<sup>19</sup> The low rank correlation between the two modes of measurement implies that such rankings are highly sensitive to measurement effects. The correlation coefficient is highest for the motivation index (coef. = 0.494,  $p$ -value = 0.00), followed by the ethics index (coef. = 0.270,  $p$  = 0.060) and the management index (coef. = 0.264,  $p$  = 0.063).

Looking at the quintile distribution of organizations across modes is even more suggestive. Out of 50 organizations included in the sample, two-thirds or more are in a different quintile when comparing face-to-face and online rankings. For the management index, 37 organizations change quintile, depending on which mode we use to rank the organizations. For the motivation index, this value is 33, and for the ethics index, it is 38 organizations.

All this suggests that benchmarking public sector organizations using employee survey results—a practice currently undertaken by several major public administration surveys—can be highly dependent on methodological choices like survey mode. These are rarely explicitly discussed in this context yet largely shape these rankings. Changes in the relative ranking of organizations may very likely be due to measurement rather than real changes in the underlying variables. As hinted at by the analyses above, this may be a concern not only regarding an organization's specific place in a ranking but also its broader position in the overall distribution of scores.

## Individual-Level Quantities

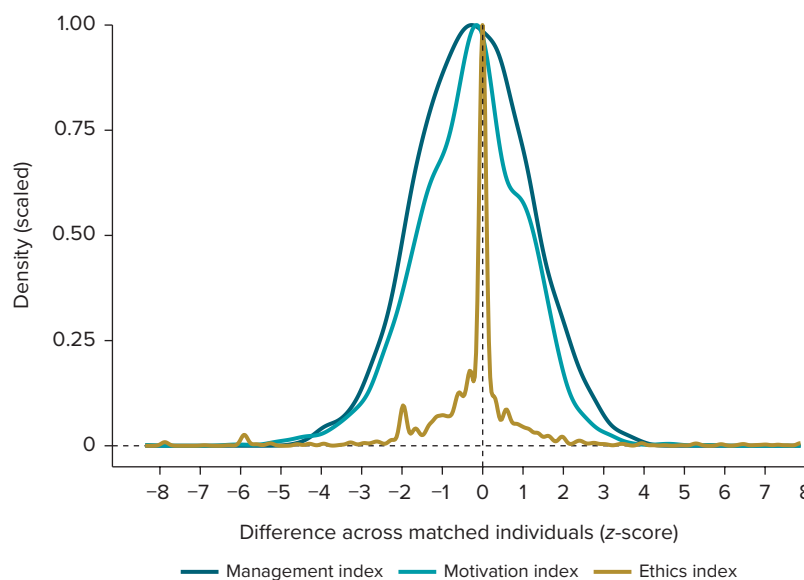
Showing the summary statistics at the individual level, as in panel 3 of table 19.2, requires matching the respondents on their observable characteristics. We use PSM to address the concern of selection bias in who chooses to respond to online surveys (Tourangeau, Conrad, and Couper 2013). PSM employs a logit model to evaluate respondents' likelihood of being in the treatment group—that is, in the online survey mode. PSM is based on the assumption that individuals with comparable observable demographic characteristics (see the note to table 19.2) should, on average, provide comparable answers. If the only meaningful difference left between matched individuals is their treatment status, then any differences in the outcomes of interest should be attributable to it. In using a PSM approach to compare survey modes, we follow earlier examples in the literature that similarly use PSM to adjust for self-selection into an online survey mode (Lee 2006; Lugtig et al. 2011). Moreover, as demonstrated in table 19.1, our experiment shows moderate signs of imbalance on key demographic items. Therefore, PSM can be seen as an additional robustness check, which ensures that these demographic imbalances between treatment arms do not taint our results.

The values shown in panel 3 of table 19.2 are calculated by taking each treated (online mode) individual and his or her index score and subtracting from it the corresponding index scores of the matched respondent(s) from the face-to-face mode. The resulting mean modal differences are comparable to their equivalents at the national and organizational levels. However, the wide distribution of survey mode effects across individuals is now clear. The minimum and maximum modal effects range between  $-8$  and  $8$  standard deviations, implying that some individuals might be particularly sensitive to the nature of measurement.<sup>20</sup>

Figure 19.6 displays the full distribution of survey mode effects. These are conditional on the matching process we undertook to generate paired observations, though our estimates are robust to including different sets of matching variables. A large fraction (12–15 percent) of paired individuals have a mode effect of at least two standard deviations for the management and motivation indexes.

Corresponding to the finding that particular organizations are more sensitive to survey mode effects, it would seem that the distribution of sensitivity across groups of individuals is also important in

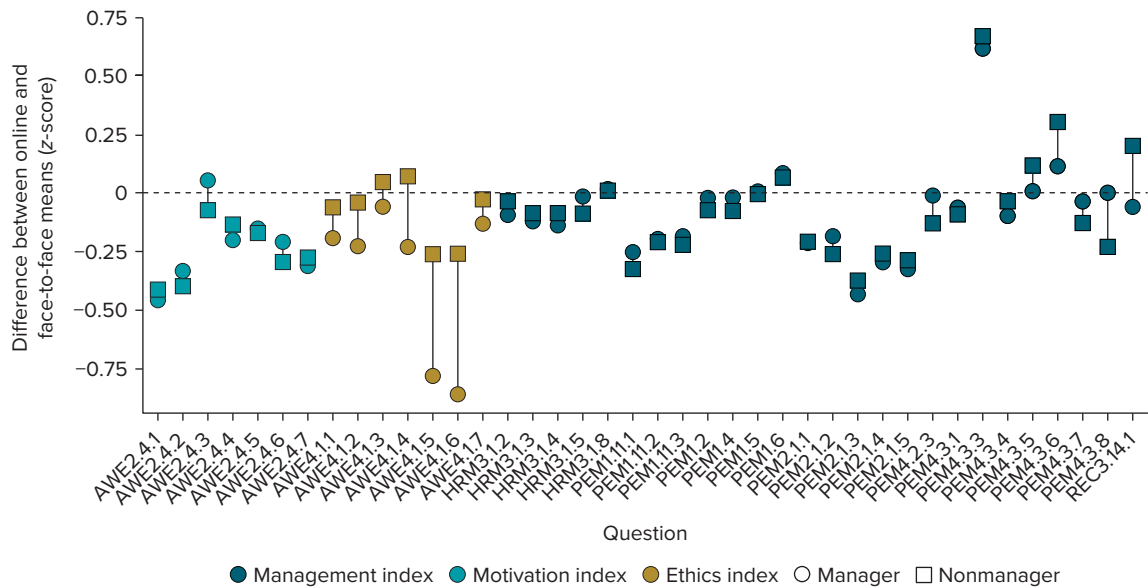
**FIGURE 19.6** Distribution of Survey Mode Differences across Matched Individuals



Source: Original figure for this publication.



**FIGURE 19.7** Average Item Difference, by Managerial Status



Source: Original figure for this publication. AWE = attitudes and work environment; HRM = human resources management; PEM = performance management; REC = recruitment.

understanding the wider nature of survey mode effects in survey design. We can explore how certain groups of public servants exhibit larger mode effects for certain topics. For instance, figure 19.7 shows survey mode effect differences separately for managers and nonmanagers for all individual questions included in each of our indexes (similar to figure 19.3 above). We might expect to see differences in sensitivity to the mode of survey enumeration between those two groups for multiple reasons. In a face-to-face interview with a human enumerator, managers might feel larger social pressure to keep up the good image of their work unit and therefore provide more-positive answers. Nonmanagers might feel less secure in their position, be warier of potential repercussions for answering truthfully, and, therefore, provide less-negative answers in a face-to-face setting, which is perceived as providing less anonymity. As the figure shows, the mean mode effects indeed vary between managers and nonmanagers by as much as 0.5 standard deviations. The differential sensitivity of these two groups to survey mode is particularly visible for some questions composing the ethics index, with the skew toward more-positive answers in the face-to-face mode being noticeably more pronounced for managers than for nonmanagers.

In a similar vein, we can analyze sensitivity to survey mode effects in other demographic groups. The OLS models in table 19.3 examine the relationship between the aggregate values of the three indexes, survey mode, and key individual characteristics, such as age, education level, gender, and tenure. They provide further evidence of the role of the survey mode for outcome measurement, which does not disappear after controlling for other respondent characteristics. For all three indexes, the dummy for the online mode is negative and statistically significant at 1 percent. These coefficients are also very similar in size to the coefficients in table 19.2, and they indicate that online respondents provide responses that are between 0.22 and 0.34 standard deviations more negative than face-to-face respondents.

The role of demographic controls is less consistent. Age and tenure stand out as highly significant for both the management and motivation indexes—with *older* respondents and those with *fewer* years of on-the-job experience providing more-positive answers. Table 19.3 and the further robustness checks discussed below suggest that there is little we can conclude about the independent role of measured demographic variables on our survey indexes. Across cultures, surveys, and agencies, the specific impacts of individual

**TABLE 19.3 Ordinary Least Squares Results: Individual Characteristics and Mean Survey Differences**

	Dependent variable		
	Management index (1)	Motivation index (2)	Ethics index (3)
Survey mode: Online	−0.244*** (0.029)	−0.341*** (0.029)	−0.222*** (0.032)
Age	0.009*** (0.002)	0.011*** (0.002)	0.002 (0.002)
Gender: Male	−0.013 (0.032)	−0.111*** (0.032)	−0.068* (0.035)
Education: Undergraduate	0.053 (0.073)	−0.097 (0.074)	−0.078 (0.083)
Education: Master's	0.045 (0.074)	−0.067 (0.075)	−0.172** (0.084)
Education: PhD	−0.112 (0.102)	−0.006 (0.103)	−0.123 (0.116)
Status: Civil servant	−0.109* (0.062)	−0.004 (0.062)	0.053 (0.067)
Pay grade	−0.020*** (0.006)	−0.006 (0.006)	−0.015** (0.007)
Managerial status: Manager	−0.470*** (0.049)	0.092* (0.049)	−0.052 (0.053)
Tenure	−0.011*** (0.003)	−0.008*** (0.003)	−0.001 (0.003)
Organizational tenure	0.009*** (0.003)	0.004 (0.003)	−0.006* (0.003)
Public administration tenure	−0.002 (0.003)	−0.003 (0.003)	−0.001 (0.003)
Constant	−0.028 (0.144)	−0.123 (0.144)	0.298* (0.158)
Observations	4,787	4,734	3,991
$R^2$	0.040	0.043	0.019
Adjusted $R^2$	0.038	0.040	0.016

Source: Original table for this publication.

Note: \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

characteristics on the size of mode effects will vary. This analysis has showcased a potential route for survey analysts to investigate these issues in their own data.

Our results suggest that the high degree of uncertainty around the impact of survey modes on the responses of different organizations and employee groups is an open area for research—both as an academic concern and for the improvement of specific public service surveys. Identifying those individuals sensitive to measurement will require experimentation in the modes to which specific individuals are subject. Identifying those characteristics of public servants that predict sensitivity will make the validity of inferences about differences between individual public servants across key organizational measures significantly more robust.

## THE IMPACT OF COMMON CORRECTIONS

Given the near-universal use of online surveys and the concerns that have motivated this chapter, many public servant surveys expend significant resources increasing response rates and analytical effort weighting their responses to correct for sample selection. Our experiment allows us to better understand the impacts of these efforts and their effects on the robustness of the quantities produced by analysis.

### How Does the Response Rate Mediate Survey Mode Effects?

A substantial criticism of online surveys—of all types—is that they achieve generally low and varying response rates across organizations relative to face-to-face surveys. Low response rates are typically interpreted as making surveys vulnerable to systematic differences in the sample of individuals who respond and their associated responses to questions. We have seen from the Romania experiment analyzed in this chapter that online surveys do have a lower response rate overall, that it varies more dramatically than the face-to-face survey response rate across organizations, and that respondents differ from a representative sample. However, the question remains whether this leads to differential inference.

As shown in figure 19.8, survey mode effects do not appear to be significantly correlated with survey response rates. In other words, mean modal differences at the organizational level do not differ systematically between organizations with low response rates to online surveys and organizations with high response rates to online surveys (relative to face-to-face surveys with consistently high response rates). Whether response rates are particularly high or low does not seem to explain the variation we see in the robustness of online surveys to replicating the responses generated by face-to-face surveys. This suggests that aggregate responses to online surveys may be compared across organizations even when response rates between these organizations vary widely, as was the case in our survey.

These results also imply that survey mode effects are driven by selection into response and by respondents' interaction with the survey mode rather than simply differing response rates. Given that even high online response rates still exhibit large mode effects, it must be some combination of these effects that drives the wider results of this paper rather than selection alone. Thus, we cannot ultimately conclude that either mode is more accurate, but we note that respondents do seem to respond differently to different approaches to enumeration under certain conditions.

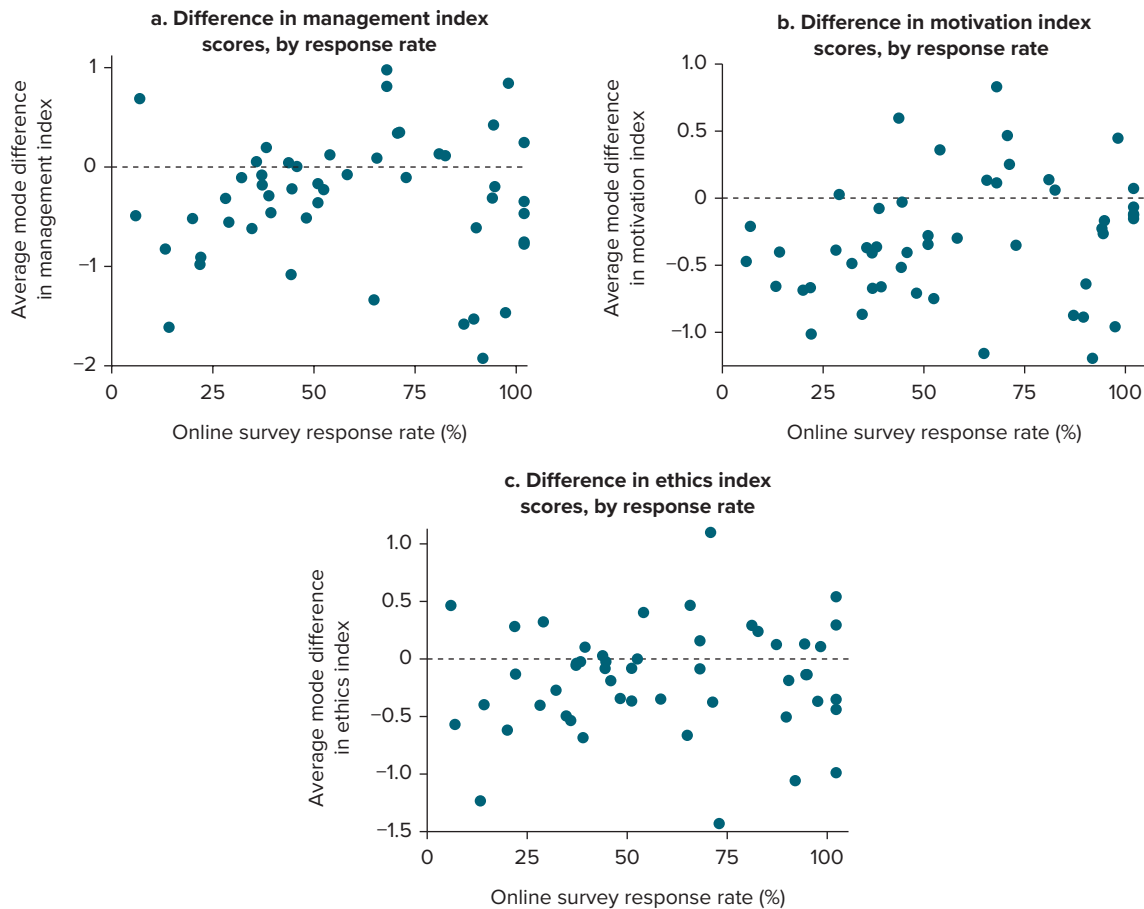
### What Is the Impact of Corrections Using Survey Weights?

Many public servant surveys use weighting schemes to upweight the responses of types of officials under-represented in the survey. To reflect these efforts, we estimate the mean modal difference, using a range of econometric weighting approaches to understand whether they impact the robustness of the corresponding estimates.

We recalculate the unweighted mean differences shown in table 19.4 using a sample weighted by a raking weight based on gender, age, and an inverse online survey response rate to adjust for differences between survey respondents and nonrespondents along these dimensions (Tourangeau, Conrad, and Couper 2013).<sup>21</sup> Finally, we calculate the mean modal difference using inverse probability weighting (IPW), which increases the weight of an official exactly inverse to its survey response rate. In doing so, we give responses from organizations with low response rates a larger weight.

As shown in table 19.4, the modal differences are relatively robust across the unweighted survey sample, a sample that is weighted using the raking method, and a sample that is weighted by the inverse of the organizational survey response rate. Figure 19.9 summarizes the average modal difference across all survey items at the aggregate level across three samples: one that is unweighted, one that is weighted using the raking method, and one that is weighted using IPW. The presence of mode effects is largely unchanged by either

**FIGURE 19.8** Difference in Scores, by Response Rate



Source: Original figure for this publication.

weighting method. Reweighting does little to improve the robustness of the estimates and, in several cases, actually increases the magnitude of the mode effects we observe.

This suggests that the application of weights, a statistical process undertaken by many major public administration surveys, including the FEVS, may not be effective in mitigating the biases introduced by their specific measurement approaches (for a full summary of the weighting methods undertaken by major public administration surveys, see chapter 18). These results are consistent with our preceding findings that the response rate and observable characteristics of individual public servants are not key determinants of the survey mode effects we find.

## DISCUSSION

Given the challenges of measuring critical aspects of public service life outside of surveys of public servants, survey design features will continue to be a critical input into our understanding of the state. Perhaps the most significant decision for a survey enumerator interviewing public officials is whether the survey should be administered in person or online. This chapter has reviewed the limited existing information on this question for the public service and presented a novel experiment that sheds light on various aspects of the choice.

This chapter has provided a framework for survey analysts to conceptualize testing survey mode effects in their own surveys, as well as benchmark evidence with which to compare their results. Experimental analysis, as in this chapter, provides a rigorous platform for better understanding the nature of the measurement of the state.

We undertake a field experiment with 6,037 public servants in 81 government institutions in Romania, in which we randomly assign each official to complete either a face-to-face or online survey. In line with predictions of the literature (Heerwegh and Loosveldt 2008; Krosnick and Presser 2010), the online survey exhibits significantly higher levels of survey nonresponse, breakoff, and item non-response than the corresponding face-to-face survey. This does change the sample of respondents answering each survey question, pushing the online survey away from a “representative” set of officials. Insofar as missing values impact the overall quality and usability of survey data collected, we can thus conclude that face-to-face survey modes provide higher-quality survey data with fewer missing or nonmeaningful responses. Government-run public servant surveys are almost universally online.

To what extent do we find evidence that the above quality concerns are leading to deviations in results from corresponding face-to-face surveys? The evidence from the experiment we analyze indicates that

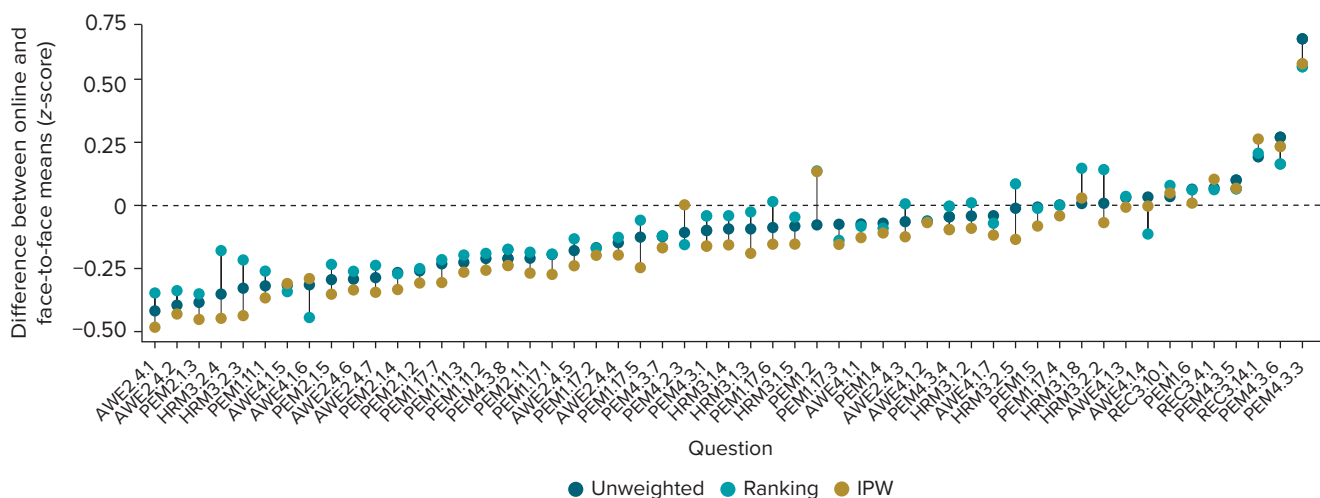
**TABLE 19.4 Mean Modal Differences at the National Level, by Weighting Approach**

(1) Unweighted	
Management index	−0.239
Motivation index	−0.350
Ethics index	−0.208
(2) Weighted (raking)	
Management index	−0.171
Motivation index	−0.263
Ethics index	−0.278
(3) Weighted (IPW)	
Management index	−0.331
Motivation index	−0.440
Ethics index	−0.248

Source: Original table for this publication.

Note: All values reflect the mean difference in the average index values between online and face-to-face samples ( $\hat{x}_{online} - \hat{x}_{f2f}$ ). Panel 1 shows unweighted means. Panel 2 shows the values for the sample weighted using the raking method, wherein weights are iteratively adjusted based on demographic characteristics for which the population distribution is known (in this case, age, gender, and the proportion of civil servants by employment status) until the weighted sample distribution aligns with the population distribution for those variables. Panel 3 weights the sample by inverse values of the organization-level response rate. IPW = inverse probability weights.

**FIGURE 19.9 Average Item Difference, by Sample**



Source: Original figure for this publication.

Note: IPW = inverse probability weighting. AWE = attitudes and work environment; HRM = human resources management; PEM = performance management; REC = recruitment.

the online treatment group provides more-negative evaluations across the topics of management, motivation, and ethics than the face-to-face group. This pattern also holds for the majority of individual survey questions, not only aggregate indexes. Though such a finding is consistent with the online mode's limiting social-desirability bias, the smallest mode effects are for the ethics index, where this bias should be the most pronounced.

Similar conclusions apply at the level of organizations and individuals. The majority of organizations record lower mean responses in surveys enumerated online. The magnitude of the difference at the national level is small, moving indexes 0.1 on a 1–5 scale. However, the difference for some organizations is above one standard deviation and is substantial enough to make rankings of organizations' scores very poorly correlated across the two survey modes. At the individual level, the magnitude of survey mode effects can be very large. Overall, we cannot make definitive statements about which survey mode is superior, but we note that measurement significantly mediates results at the organizational and individual levels. The burden of proof thus lies with survey analysts to show that results at these levels of aggregation are legitimate.

Based on a PSM analysis, we find that a number of public administrators are particularly sensitive to the survey enumeration approach. We present mode effects of considerable magnitude across matched individuals. These are not well predicted by standard observable characteristics, nor are they affected by common weighting schemes, suggesting that the survey mode is the factor responsible for the difference. Our findings hold with remarkable consistency for all three indexes. Interestingly, the mode effect is present across the whole distribution of response rates, implying that there is a limited correlation between the decision to participate and the deviation of the online survey results from a representative face-to-face survey.

These results suggest that the survey mode effects in public administration are substantial and, for some common survey conclusions to be valid, cannot be ignored. Though aggregates (say, at the national level) are least affected, the ranking of organizations, for example, can be substantially influenced by such effects. Therefore, this chapter proposes embedding an investigation of these issues into survey design generally. As particular national and service cultures mediate where mode effects are largest, corresponding survey analysts can refine their approach as each setting demands. For example, we find that certain groups of respondents and questions—like managers and ethical questions in our experiment—produce noticeably divergent results depending on the survey mode. Identifying the particular groups, questions, and circumstances that make the survey mode a more salient issue, as well as the mechanisms at work in those cases, will contribute to improving the way we measure public administration.

## NOTES

We gratefully acknowledge funding from the World Bank's i2i initiative, Equitable Growth, Finance, and Institutions Chief Economist's Office, and Governance Global Practice. We are grateful to Kerenssa Kay, Maria Ruth Jones, and Ravi Somani for helpful comments. We would like to thank Lior Kohanan, Robert Lipinski, Miguel Mangunpratomo, and Sean Tu for excellent research assistance; Kerenssa Kay and Anita Sobjak for guidance and advice; and seminar participants at the World Bank for their comments. Computational reproducibility was verified by DIME Analytics.

1. Though most online surveys follow a relatively standard form, there is potential to make online surveys more engaging for the respondent. For example, the gamification of surveys or the inclusion of short clips and other multimedia extensions may enable surveys to more effectively capture respondents' attention. These have generally not been taken up or experimented with in any setting, including in public administration. One notable exception is Haan et al. (2017), who examine whether adding a video of enumerators reading online survey questions increases engagement. The study finds a null effect and concludes that the interactive component of face-to-face surveys goes beyond a video recording of the enumerators.
2. To date, the existing literature has focused on the advantages and disadvantages of online versus face-to-face surveys in the general population (Couper et al. 2007; Daikeler, Bošnjak, and Lozar Manfreda 2020; Groves and Peytcheva 2008; Heerwegh and Loosveldt 2008; Krosnick and Presser 2010; Peytchev 2009). No studies, to our knowledge, focus on this debate in the context of public administration.
3. The value was calculated as a mean difference between the ratio of the number of respondents relative to the number of invited and eligible respondents in the web mode and the equivalent ratio for the other survey mode.



4. These large country differences and declining response rates are not unlike those observed in general public opinion surveys. For example, Beullens et al. (2018) find that response rates to the European Social Survey range from well below 50 percent in countries like the United Kingdom and Germany to above 70 percent in Cyprus, Bulgaria, and Israel, all while a double-digit decline in response rates is observable in many settings.
5. Implementation of the face-to-face surveys was successful, with 99 percent of face-to-face surveys rated as having gone well or very well.
6. Our concern is that in these institutions, the relevant survey links were not adequately distributed to targeted staff. To test the robustness of this decision to our results, we also use different cutoff points for online survey response rates of 3 percent and 7 percent, and our results are qualitatively the same.
7. A remaining concern is that an organization's response rate by mode may be high for distinct reasons, and these reasons may be correlated with the variables on which we collect data. However, given the low nonresponse rates in our matched sample, there is limited scope for endogenous selection to impact our estimates (Oster 2019).
8. This is contrary to some findings in the literature that online survey respondents tend to be male (Duffy et al. 2005). This difference may be due in part to the composition of the Romanian civil service, which is predominantly female across most organizations.
9. This difference is comparable in magnitude to other surveys in the literature, which find an average difference in educational attainment of approximately 6 percentage points (Braekman et al. 2020).
10. We also see a substantial number of individuals exiting where the demographic question block begins.
11. An additional 89 revisited the survey after previously completing it. These individuals are excluded from this analysis, as it is assumed that their returning to a survey they had already taken was inadvertent.
12. Though evidence on the impact of reminders in public servant surveys is scarce, data from the 2014 FEVS shows that the number of responses is at its peak in the first week of the survey, drops dramatically in subsequent weeks, and plateaus between weeks three and six (with a slight jump in the final week). This echoes our own experience and underlies the critical importance of the survey launch.
13. The full list of questions composing each index can be found in table G.8 in appendix G.
14. The z-scores are calculated over the full sample of individuals used for analysis.
15. For the list of questions and their phrasing, see table G.8 in appendix G.
16. In chapter 22, we specifically focus on how the complexity and sensitivity of each question influence response patterns. For that purpose, we develop a coding framework that assesses each question in the Romania questionnaire (among others) along various margins of complexity and sensitivity, like syntax, context familiarity, privacy, and the threat of disclosure.
17. More formal tests of the difference between mode effects at the organizational level are discussed in appendix G.
18. The correlation can be expected to be even lower for individual questions, which tend to exhibit greater variation.
19. As a reminder that these graphs are not an artifact of response bias arising from extreme response rates, note again that we restrict the sample of comparison here to only those organizations with an online response rate of at least 5 percent.
20. To assess the validity of our matched estimates, in table G.5 (see appendix G), we also present results obtained if PSM controls for a different set of demographic characteristics and also for organizational fixed effects only. We find that the estimates of mean differences are qualitatively similar across various PSM approaches.
21. Iterative proportional fitting, or raking, is among the most commonly used methods for weighting survey results. The method involves choosing a set of demographic variables where the population value is known and iteratively adjusting the weight for each case until the sample distribution aligns with the population distribution for those variables (Mercer, Lau, and Kennedy 2018).

## REFERENCES

- Anduiza, Eva, and Carol Galais. 2017. "Answering without Reading: IMCs and Strong Satisficing in Online Surveys." *International Journal of Public Opinion Research* 29 (3): 497–519. <https://doi.org/10.1093/ijpor/edw007>.
- Baumgartner, Hans, and Jan-Benedict E. M. Steenkamp. 2001. "Response Styles in Marketing Research: A Cross-National Investigation." *Journal of Marketing Research* 38 (2): 143–56. <https://doi.org/10.1509/jmkr.38.2.143.18840>.
- Beullens, Koen, Geert Loosveldt, Caroline Vandenplas, and Ineke Stoop. 2018. "Response Rates in the European Social Survey: Increasing, Decreasing, or a Matter of Fieldwork Efforts?" *Survey Methods: Insights from the Field*, April. <https://doi.org/10.13094/SMIF-2018-00003>.
- Biemer, Paul P., Joe Murphy, Stephanie Zimmer, Chip Berry, Grace Deng, and Katie Lewis. 2018. "Using Bonus Monetary Incentives to Encourage Web Response in Mixed-Mode Household Surveys." *Journal of Survey Statistics and Methodology* 6 (2): 240–61. <https://doi.org/10.1093/jssam/smx015>.
- Braekman, Elise, Rana Charafeddine, Stefaan Demarest, Sabine Drieskens, Finaba Berete, Lydia Gisle, Johan Van der Heyden, and Guido Van Hal. 2020. "Comparing Web-Based versus Face-to-Face and Paper-and-Pencil Questionnaire Data

- Collected through Two Belgian Health Surveys.” *International Journal of Public Health* 65 (1): 5–16. <https://doi.org/10.1007/s00038-019-01327-9>.
- Cornesse, Carina, and Michael Bošnjak. 2018. “Is There an Association between Survey Characteristics and Representativeness? A Meta-Analysis.” *Survey Research Methods* 12 (1): 1–13. <https://doi.org/10.18148/srm/2018.v12i1.7205>.
- Couper, Mick P., Arie Kapteyn, Matthias Schonlau, and Joachim Winter. 2007. “Noncoverage and Nonresponse in an Internet Survey.” *Social Science Research* 36 (1): 131–48. <https://doi.org/10.1016/j.ssresearch.2005.10.002>.
- Daikeler, Jessica, Michael Bošnjak, and Katja Lozar Manfreda. 2020. “Web versus Other Survey Modes: An Updated and Extended Meta-Analysis Comparing Response Rates.” *Journal of Survey Statistics and Methodology* 8 (3): 513–39. <https://doi.org/10.1093/jssam/smz008>.
- De la Rocha, Alexandra Maria. 2015. “The Relationship between Employee Engagement and Survey Response Rate with Union Membership as a Moderator.” Master’s thesis, San José State University. <https://doi.org/10.31979/etd.z4c6-uv9d>.
- Duffy, Bobby, Kate Smith, George Terhanian, and John Bremer. 2005. “Comparing Data from Online and Face-to-Face Surveys.” *International Journal of Market Research* 47 (6): 615–39. <https://doi.org/10.1177/147078530504700602>.
- Galesic, Mirta. 2006. “Dropouts on the Web: Effects of Interest and Burden Experienced During an Online Survey.” *Journal of Official Statistics* 22 (2): 313–28. <https://www.proquest.com/scholarly-journals/dropouts-on-web-effects-interest-burden/docview/1266792615/se-2>.
- Gnambs, Timo, and Kai Kaspar. 2015. “Disclosure of Sensitive Behaviors across Self-Administered Survey Modes: A Meta-Analysis.” *Behavior Research Methods* 47: 1237–59. <https://doi.org/10.3758/s13428-014-0533-4>.
- Groves, Robert M. 2006. “Nonresponse Rates and Nonresponse Bias in Household Surveys.” *Public Opinion Quarterly* 70 (5): 646–75. <https://doi.org/10.1093/poq/nfl033>.
- Groves, Robert M., and Emilia Peytcheva. 2008. “The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis.” *Public Opinion Quarterly* 72 (2): 167–89. <https://www.jstor.org/stable/25167621>.
- Haan, Marieke, Yfke P. Ongena, Jorre T. A. Vannieuwenhuyze, and Kees de Groot. 2017. “Response Behavior in a Video-Web Survey: A Mode Comparison Study.” *Journal of Survey Statistics and Methodology* 5 (1): 48–69. <https://doi.org/10.1093/jssam/smw023>.
- Heerwegh, Dirk. 2009. “Mode Differences between Face-to-Face and Web Surveys: An Experimental Investigation of Data Quality and Social Desirability Effects.” *International Journal of Public Opinion Research* 21 (1): 111–21. <https://doi.org/10.1093/ijpor/edn054>.
- Heerwegh, Dirk, and Geert Loosveldt. 2008. “Face-to-Face versus Web Surveying in a High-Internet-Coverage Population: Differences in Response Quality.” *Public Opinion Quarterly* 72 (5): 836–46. <https://doi.org/10.1093/poq/nfn045>.
- Jensen, Nathan M., Quan Li, and Aminur Rahman. 2010. “Understanding Corruption and Firm Responses in Cross-National Firm-Level Surveys.” *Journal of International Business Studies* 41 (9): 1481–504. <https://doi.org/10.1057/jibs.2010.8>.
- Kaminska, Olena, and Tom Foulsham. 2014. “Real-World Eye-Tracking in Face-to-Face and Web Modes.” *Journal of Survey Statistics and Methodology* 2 (3): 343–59. <https://doi.org/10.1093/jssam/smu010>.
- Kays, Kristina, Kathleen Gathercoal, and William Buhrow. 2012. “Does Survey Format Influence Self-Disclosure on Sensitive Question Items?” *Computers in Human Behavior* 28 (1): 251–56. <https://doi.org/10.1016/j.chb.2011.09.007>.
- Krosnick, Jon A. 1991. “Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys.” *Applied Cognitive Psychology* 5 (3): 213–36. <https://doi.org/10.1002/acp.2350050305>.
- Krosnick, Jon A., and Stanley Presser. 2010. “Question and Questionnaire Design.” In *Handbook of Survey Research*, edited by Peter V. Marsden and James D. Wright, 2nd ed., 263–314. Bingley: Emerald.
- Lee, Sunghee. 2006. “Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys.” *Journal of Official Statistics* 22 (2): 329–49. <https://www.researchgate.net/publication/259497319PropensityScoreAdjustmentasaWeightingSchemeForVolunteerPanelWebSurveys>.
- Lozar Manfreda, Katja, Michael Bošnjak, Jernej Berzelak, Iris Haas, and Vasja Vehovar. 2008. “Web Surveys versus Other Survey Modes: A Meta-Analysis Comparing Response Rates.” *International Journal of Market Research* 50 (1): 79–104. <https://doi.org/10.1177/147078530805000107>.
- Lozar Manfreda, Katja, and Vasja Vehovar. 2002. “Survey Design Features Influencing Response Rates in Web Surveys.” Paper delivered at the International Conference on Improving Surveys, Copenhagen, August 25–28, 2002. <http://www.websm.org/uploadi/editor/LozarVehovar2001Surveydesign.pdf>.
- Lugtig, Peter, Gerty J. L. M. Lensvelt-Mulders, Remco Frerichs, and Assyn Greven. 2011. “Estimating Nonresponse Bias and Mode Effects in a Mixed-Mode Survey.” *International Journal of Market Research* 53 (5): 669–86. <https://doi.org/10.2501/IJMR-53-5-669-686>.
- Mercer, Andrew, Arnold Lau, and Courtney Kennedy. 2018. “How Different Weighting Methods Work.” In *For Weighting Online Opt-In Samples, What Matters Most?*, 7–14. Washington, DC: Pew Research Center. <https://www.pewresearch.org/methods/2018/01/26/how-different-weighting-methods-work>.

- Moynihan, Donald P., and Sanjay K. Pandey. 2010. "The Big Question for Performance Management: Why Do Managers Use Performance Information?" *Journal of Public Administration Research and Theory* 20 (4): 849–66. <https://doi.org/10.1093/jopart/muq004>.
- Musch, Jochen, and Ulf-Dietrich Reips. 2000. "A Brief History of Web Experimenting." In *Psychological Experiments on the Internet*, edited by Michael H. Birnbaum, 61–87. Cambridge, MA: Academic Press. <https://doi.org/10.1016/B978-0-12-099980-4.X5000-X>.
- Newell, Carol E., Kimberly P. Whittam, Zannette A. Urielle, and Yeh-Chun (Anita) Kang. 2010. *Non-Response on U.S. Navy Quick Polls*. NPRST-TN-10-3. Millington, TN: Navy Personnel Research, Studies, and Technology, Bureau of Naval Personnel. <https://apps.dtic.mil/sti/citations/ADA516853>.
- Newman, Jessica Clark, Don C. Des Jarlais, Charles F. Turner, Jay Gribble, Phillip Cooley, and Denise Paone. 2002. "The Differential Effects of Face-to-Face and Computer Interview Modes." *American Journal of Public Health* 92 (2): 294–97. <https://doi.org/10.2105/ajph.92.2.294>.
- Oster, Emily. 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business and Economic Statistics* 37 (2): 187–204. <https://doi.org/10.1080/07350015.2016.1227711>.
- Peytchev, Andy. 2006. "A Framework for Survey Breakoffs." Paper presented at the 61st Annual Conference of the American Association for Public Opinion Research, Montréal, May 18–21, 2006. In JSM Proceedings, Survey Research Methods Section, 4205–12. Alexandria, VA: American Statistical Association. <http://www.asasrms.org/Proceedings/y2006/Files/JSM2006-000094.pdf>.
- Peytchev, Andy. 2009. "Survey Breakoff." *The Public Opinion Quarterly* 73 (1): 74–97. <https://www.jstor.org/stable/25548063>.
- Podsakoff, Philip M., Scott B. MacKenzie, Jeong-Yeon Lee, and Nathan P. Podsakoff. 2003. "Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies." *Journal of Applied Psychology* 88 (5): 879–903. <https://doi.org/10.1037/0021-9010.88.5.879>.
- Rich, Bruce Louis, Jeffrey A. Lepine, and Eean R. Crawford. 2010. "Job Engagement: Antecedents and Effects on Job Performance." *Academy of Management Journal* 53 (3): 617–35. <https://doi.org/10.5465/amj.2010.51468988>.
- Shih, Tse-Hua, and Xitao Fan. 2008. "Comparing Response Rates from Web and Mail Surveys: A Meta-Analysis." *Field Methods* 20 (3). <https://doi.org/10.1177/1525822X08317085>.
- Spitzmüller, Christiane, Dana M. Glenn, Christopher D. Barr, Steven G. Rogelberg, and Patrick Daniel. 2006. "If You Treat Me Right, I Reciprocate": Examining the Role of Exchange in Organizational Survey Response." *Journal of Organizational Behavior* 27 (1): 19–35. <https://doi.org/10.1002/job.363>.
- Steinbrecher, Markus, Joss Roßmann, and Jan Eric Blumenstiel. 2015. "Why Do Respondents Break Off Web Surveys and Does It Matter? Results from Four Follow-Up Surveys." *International Journal of Public Opinion Research* 27 (2): 289–302. <https://doi.org/10.1093/ijpor/edu025>.
- Tourangeau, Roger, Frederick G. Conrad, and Mick P. Couper. 2013. *The Science of Web Surveys*. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199747047.001.0001>.
- Tourangeau, Roger, and Ting Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133 (5): 859–83. <https://doi.org/10.1037/0033-2909.133.5.859>.
- Ye, Cong, Jenna Fulton, and Roger Tourangeau. 2011. "More Positive or More Extreme? A Meta-Analysis of Mode Differences in Response Choice." *Public Opinion Quarterly* 75 (2): 349–65. <https://doi.org/10.1093/poq/nfr009>.

## CHAPTER 20

# Determining Sample Sizes

## How Many Public Officials Should Be Surveyed?

*Robert Lipinski, Daniel Rogger, Christian Schuster,  
and Annabelle Wittels*

### SUMMARY

Determining the sample size of a public administration survey often entails a trade-off between the benefits of increasing the precision of survey estimates and the high costs of surveying a larger number of civil servants. Survey administrators ultimately have to decide on the sample size based on the types of inference they want the survey to yield. This chapter aims to quantify the sample sizes necessary to make a range of inferences that are commonly drawn from public administration surveys. It does so by employing Monte Carlo simulations and past survey results from Chile, Liberia, Romania, and the United States. The analyses show that civil service-wide estimates can be reliably derived using sample sizes considerably smaller than the ones currently used by these surveys. By contrast, comparison across demographic groups—gender and managerial status—and ranking individual public administration organizations both require large sample sizes, often substantially larger than those available to survey administrators. These results suggest that not all types of inference and comparison can be drawn from surveys of civil servants, which, instead, may need to be complemented by other research tools, like interviews or anthropological research. This chapter is also linked to an online toolkit that allows practitioners to estimate the optimal sample size for a survey given the types of inference expected to be drawn from it. Together, the chapter and the toolkit allow practitioners involved in survey design for the civil service to understand the trade-offs involved in sampling and what types of comparison can be reliably drawn from the data.

### ANALYTICS IN PRACTICE

- Sample size is one of the key factors affecting survey quality. An accurately selected sample of adequate size is indispensable to making survey results reliable and actionable. Choosing the number of respondents is, therefore, a crucial decision faced by any survey designer. This chapter details what factors should be considered to make an optimal choice in the context of sampling for civil servant surveys.

Robert Lipinski is a consultant and Daniel Rogger is a senior economist in the World Bank's Development Impact Evaluation (DIME) Department. Christian Schuster is a professor at University College London. Annabelle Wittels is an independent researcher.

- Efficient survey sampling strategies need to balance the precision of estimates against the costs of expanding the sample size. Sampling more people tends to improve the accuracy of survey results and the number of comparisons that can be reliably drawn from the responses. However, for logistical and financial reasons, it is not always possible to survey everyone. Thus, the benefits of increasing the sample size and the costs of running a survey need to be balanced against each other.
- The required survey sample size crucially depends on the types of comparison a researcher plans to make based on the data. Obtaining precise civil service–wide aggregates requires a considerably smaller sample than drawing comparisons between demographic groups of civil servants or institutions within public administration. Survey designers have to decide in advance what inferences they need to draw from their surveys and adjust the sample size accordingly.
- Civil servant surveys often oversample for the purpose of determining civil service–wide aggregate measures. On the basis of past civil servant surveys, we conclude that most common civil servant survey measures, like job satisfaction, work motivation, and merit-based recruitment, could be accurately estimated at the level of the civil service as a whole by surveying 50–70 percent of the current sample.
- Comparisons of survey responses between different demographic groups (such as male vs. female or manager vs. nonmanager) require sample sizes equivalent to or larger than those currently used. Decreasing current sample sizes would likely lead to incorrect comparisons between demographic groups—due to nonrepresentative samples—or prevent them altogether—due to insufficient responses from each group of interest to enable comparison. Although this topic is not covered here, the present analysis indicates that comparisons between more than two demographic groups, like civil servants of different education levels or ethnic backgrounds, would require sample sizes larger than the ones currently prevalent.
- Precise ranking of institutions within the civil service according to survey measures, like job satisfaction or motivation, requires larger sample sizes than currently prevalent. Given the standard sample sizes and the variation in estimates, survey questions are unlikely to determine an exact ranking of institutions within public administration. Institutions might not be sufficiently large for such comparisons, or samples of respondents drawn from them would need to become considerably larger than is currently the case. Rather than an exact ranking position, the quintile position of an institution (for example, if it is in the top 20 percent of institutions on a given measure) can be more reliably determined.

## INTRODUCTION

The usefulness of surveys as a research tool is determined by multiple factors, but one of the most crucial is sample size. The number of people who provide responses to a survey determines the confidence one can have in its results and the types of inference and comparison one can draw from it. In general, the more people are surveyed, the more reliable and actionable the results of a survey. To take the simplest example, a survey of 1,000 people in, say, a ministry of education is more likely to yield the true value of the quantity of interest, like the level of job satisfaction, than a survey of 10 people. It would also be more likely to allow for the comparison of job satisfaction levels between men and women, managers and nonmanagers, or different departments within the ministry.

However, surveying as many people as possible is not always a useful guideline for survey designers, especially in the context of public administration surveys. For one, many surveys in this context are administered face-to-face. This may be due to technical reasons (for example, low access to the internet) or methodological considerations (for example, face-to-face surveys tend to decrease item nonresponse; see chapter 19). Moreover, each additional person surveyed, regardless of the mode of survey delivery, increases



the direct and indirect costs associated with running a survey. The direct costs of survey administration are particularly pronounced in face-to-face surveys, in which travel time and enumerator staff costs increase for each extra person surveyed. Even in online surveys—in which survey administration costs are often fixed—indirect survey costs can be significant. For instance, completing surveys takes time. Each minute taken away from the workday of a public sector employee incurs a cost to the public purse. Half an hour of the time of the average public sector employee in the United States costs the taxpayer US\$19.81.<sup>1</sup> If the number of US civil servants surveyed in the annual Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS) were reduced by 10 percent, the opportunity cost of the survey would be reduced by US\$3 million. These costs cannot be eliminated, but they can be reduced by limiting survey time, which might limit the scope of inferences drawn from the survey, and—the focus of this chapter—by optimizing the number of people surveyed.

The goal of a public sector survey should be to sample efficiently in order to save resources on one survey and free up resources for other work tasks or for more frequent, targeted surveying, which can improve the quality and breadth of data available for decision-making. For instance, the United Kingdom's Office for National Statistics (ONS) publishes “experimental statistics.”<sup>2</sup> The ONS collects data on the UK labor market every three months but provides model estimates for single months and weeks. Their accuracy is repeatedly assessed to establish whether surveying on a three-month basis provides statistics that are accurate enough to make decisions about the performance of the labor market in a single month, or even in a single week. More frequent surveying of civil servants could be supported by creating surveys that sample a smaller pool of people and are thus quicker and less costly to administer. Slashing sample sizes, however, entails considerable risk: if sample sizes are too small, the error bounds around estimates become too large to reliably assess progress on key performance targets or to compare different groups of civil servants or individual organizations within public administration.

What, then, are the appropriate sample sizes for civil service surveys? And are existing approaches to civil service survey sampling efficient? To assess these conundrums, this chapter conducts Monte Carlo simulations with civil service survey data from Chile, Liberia, Romania, and the United States. Our results suggest that appropriate sample sizes depend on the inferences governments wish to make from the data. To estimate averages for countries or large organizations within public administration, sample sizes could often be reduced. This holds all the more for survey measures—such as measures of work motivation or job satisfaction—that vary only to a limited extent (cf. chapter 21). Where detailed comparisons among public sector organizations—ranking the organizations by the mean values of survey question responses—or groups of public servants—for example, by gender or managerial status—are sought, sample sizes are typically too small. This holds in particular for those survey measures with limited variation and high skew, such as work motivation, which require high levels of precision to enable comparisons that detect statistically significant differences between groups of public servants or organizations within public administration. Our chapter thus concludes that a detailed elaboration of the desired uses of the survey results should precede the determination of sample sizes. It also offers an online sampling toolkit for survey designers to estimate appropriate sample sizes depending on the intended uses of the survey data.<sup>3</sup>

## **SAMPLING BEST PRACTICES AND THE CIVIL SERVICE SURVEY CONTEXT**

Several governments regularly survey their employees, yet approaches to sampling vary. For instance, in Australia and the United Kingdom, all public sector employees are invited to take the survey (a census approach), whereas other countries employ a mix of random, ad hoc sampling, and census approaches.<sup>4</sup> For example, the FEVS uses stratified random sampling approaches in most years but conducts a census every few years (2012, 2018, and 2019) to update the sampling frames. Canada's Public Service Employee Survey recruits public sector organizations to reach out to their staff to complete the survey and also makes



the survey available online for anyone who decides they fit the eligibility criteria. In Colombia, the annual national public employee survey (the Survey of the Institutional Environment and Performance in the Public Sector) uses a mixed approach: for larger organizations, a stratified sampling approach is used, while for smaller organizations, a census is taken. This is similar to the approach that the United States uses during noncensus years. Countries that have run surveys for several years have the advantage of looking back at historical data to assess what future sample sizes would be adequate, given the distributions and variations of the indicators they use. However, in many countries, surveys are not yet routinized and survey questions or approaches have changed, so there is a dearth of data to make informed decisions. This chapter addresses this problem by illustrating how countries can determine what sample size is adequate for their needs.

Determining adequate sample sizes ideally requires information on the following factors:

- **The size and proportion of the units of comparison.** The ideal approach to sampling entails drawing up so-called sampling frames, which list all relevant persons to be surveyed. In countries that lack routinized surveys of the public sector, a common obstacle to efficient sampling for public sector surveys is that complete and up-to-date records of public sector staff are not centralized, not fully digitized, or generally contain gaps (Bertelli et al. 2020). The creation and maintenance of complete sampling frames is a first step toward improving the efficiency of sampling.
- **The types of comparison—between countries, organizations, subunits, key personnel groups, previous years, or industry benchmarks.** It is also important to consider what types of comparison governments want to make using survey results. In most cases, public sector organizations desire to provide feedback to the managers of organizational subunits. In these cases, sampling should be stratified at the subunit level to increase the chances of an adequate sample size at the subunit level. However, this is often not possible because staff lists at the subunit level are incomplete or not centralized. In such a case, a minimum number of observations per subunit should be used as a target. Another consideration is whether sampling approaches are adequate for the types of comparison that governments desire to make. For example, are organizations to be benchmarked against industry (public sector) averages? Should their performance be compared with the previous year? Are comparisons required between key employee groups, such as managers and nonmanagers? It might be the case that some comparisons are not possible in certain contexts. For example, if all subunits are composed of only a few civil servants, ranking them by average survey responses might not be possible even if all of them were surveyed. Therefore, the desired comparisons should account for all the external limitations present.
- **The distributions of key variables (for example, mean and variance).** Which sample sizes allow comparisons to be meaningful depends on the distribution of these indicators (and also, but to a lesser extent, the number of comparisons that are planned). If distributions are narrow (for example, for measures such as motivation; see chapter 21), then fewer respondents are needed to arrive at the true value of aggregate-level statistics, like the mean or median. However, such distributions make it difficult to discern differences between different groups or units within public administration.
- **The desired degree of precision for the estimates.** Pinpointing the exact value of the quantity of interest is almost never possible when sampling from a larger population. However, the sampling strategy depends on how wide of uncertainty survey designers are willing to tolerate. If the representativeness of the sample is maintained, having more respondents tends to mean a more precise estimate. However, survey designers have to decide what degree of precision is acceptable. For example, if a mean estimate within  $\pm 0.1$  points of the true value on a 1–5 Likert scale is sufficient, then it would be unnecessary to increase the sample size, and therefore the costs of running a survey, in order to narrow the precision even further.

Advice on sampling for surveys outside the public sector is available. Since Cochran (1977), conventions for how to sample have been well known. Textbooks, such as SAGE Publishing's "little green book" (Kalton 1983), an encyclopedia of common research methods, typically suggest the following approach to determining sample sizes for survey research: using simple random sampling, first determine the degree of precision

that is required from the estimates, add a design factor—a multiplier that inflates the sample size—if you use clustered sampling approaches, and adjust the sample for the expected level of nonresponse.

While this approach is sensible in many instances, simplification carries several dangers. As Fowler (2009) cautions, the size of the population from which a sample is drawn has little effect on the precision of the estimates, all sample size requirements need to be decided on a case-by-case basis, and increasing the sample size does not necessarily reduce the error of estimates.

The following illustrates Fowler's first point: although the population of the United States is 16 times larger than that of Romania, one would not need a sample size 16 times larger to make estimates about the public sector in the United States versus in Romania. Rather, the dispersion of scores matters. If civil servants in the United States answer more similarly to one another than those in Romania, it is possible that, despite the Romanian civil service being considerably smaller, a larger sample size would be needed for Romania than for the United States.

With regard to Fowler's second point, rules of thumb can be useful. For example, one rule that is often used is that one should have at least 30 observations in each subgroup in order to calculate nonparametric statistics, such as the chi-square statistic. However, without knowledge about the underlying distribution of metrics and likely error rates, rules of thumb can result in highly unsatisfactory sample sizes. What sample size is satisfactory is thus an empirical question. For instance, while many survey companies routinely use a target precision of  $\pm 3$  percentage points, one should ask how this compares to the dispersion of the underlying scores and whether it provides for meaningful differences. For instance, if one organization differs by 0.05 standard deviations from another in a given year, can this be considered a meaningful difference? If so, then the sample size should be large enough to detect such differences. If not, then the sample size should be revised to capture a difference that is meaningful to the question at hand.

Finally, error caused by insufficiently large sample sizes needs to be understood as a part of the total survey error. The total survey error refers to a compound measure of error. It includes, but is not limited to, error created by sampling; it includes error deriving from the choice of scale, the survey mode, and interview techniques. For instance, if more resources are deployed to sample more people, this might come at the expense of pretesting survey scales or training enumerators, which can inflate the variance of survey answers and thereby make estimates more imprecise.

What is more, algorithmic approaches to gauge sample size can lead to misleading conclusions when survey design and analysis approaches are more complex. For instance, one needs to assess whether clustered or stratified approaches were used.

## THEORETICAL BACKGROUND

Surveys can either be targeted at collecting information from the entire population or universe of interest (a census approach) or at collecting information from a fraction of the population—a sample. Typically, surveys are used to estimate means, medians, and modes for certain responses for the entire target sample and subgroups of interest. The desire is that these estimates are an accurate or accurate-enough representation of the measures of the population. This might be impossible because of errors introduced by sampling and such things as the interview process or the coding of data. Bjarkefur et al. (2021) provides more in-depth information on how to address issues related to nonsampling error. Sampling bias can occur because of issues related to who was targeted by the survey recruitment, self-selection into survey participation, and nonresponse bias (on this topic, see chapter 22). Finally, error can be introduced by sampling variance—the fact that measurements vary and that the sample technique and size need to be adequate for the underlying dispersion of responses targeted for estimation (on this topic, see chapter 21).

Why sample size matters can be demonstrated by looking at how the standard error of two group means is calculated. Typically, inferences from surveys will pertain to comparisons between groups of observations

(for example, between two agencies, between managers and nonmanagers, etc.). The standard error of the estimate of the difference in mean scores between the two groups is the square root of the sum of their individual squared standard errors:

$$SE(\mu_1 - \mu_2) = \sqrt{SE_1^2 + SE_2^2} = \sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}. \quad (20.1)$$

The standard error is mechanically smaller, the larger the respective sample sizes of each of the groups in the comparison are ( $n_1$  and  $n_2$ ). At the same time, it is positively correlated with the values of standard deviation.

## METHODOLOGY

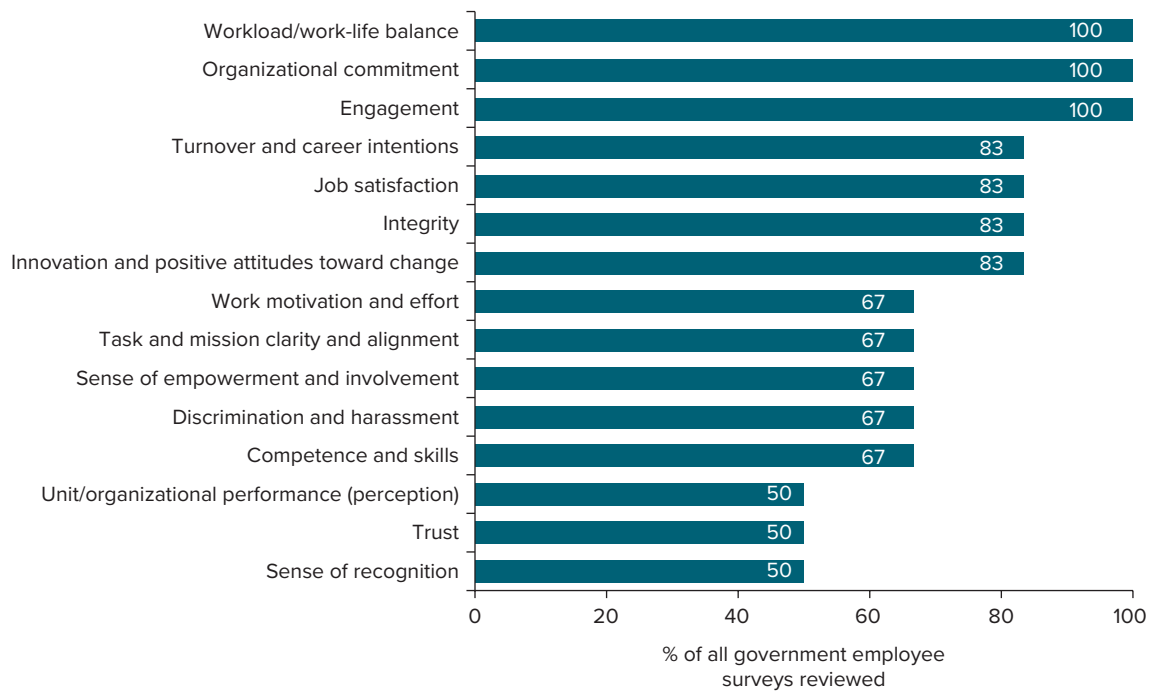
In this chapter, we illustrate what sampling error can be expected based on the variance observed in typical measures used in civil servant surveys and, consequently, what types of inference can be reliably drawn from them.

Since the approach taken in this chapter is to provide sampling guidelines for survey practitioners by extrapolating from existing civil service survey data and practice, we base the analyses upon the wealth of information provided by surveys of civil servants conducted in recent years by the World Bank, the Global Survey of Public Servants (GSPS) academic consortium, and national governments. Together, they allow us to present a wide range of statistical tests and a breadth of examples. The following surveys are included:

- A survey of civil servants in **Chile**, which takes a census approach, targeting all employees in a sample of 65 central government institutions. (The survey was part of the GSPS consortium's effort to collect more data on public administrations around the world.)
- A survey of civil servants in **Liberia**, which uses random sampling, stratified by institution. (The survey was conducted by the World Bank.)
- A survey of civil servants in **Romania**, which follows a stratified sampling approach, by which respondents are sampled in each department of a sample of organizations. (The survey was part of the GSPS consortium's effort to collect more data on public administrations around the world.)
- The Federal Employee Viewpoint Survey (FEVS), an annual survey administered by the **United States** Office of Personnel Management (OPM)—a federal agency—which first launched in 2002 under the name Federal Human Capital Survey. The FEVS aims to recruit a sample representative of the different types of US federal agencies. In 2012, 2018, and 2019, the FEVS took a census approach. In other years, the FEVS has used stratified random sampling, whereby the sample is stratified by work units within organizations. Work units smaller than 10 employees are merged. All senior executives are targeted by the survey, while lower-rank individuals are subject to random sampling within their strata. A target sample size for each organization is calculated. When this target rate amounts to 75 percent or more of an organization's entire staff, a census approach, whereby all employees are targeted, is employed instead. The FEVS has served as an important benchmark for multiple surveys of public administrators around the world.

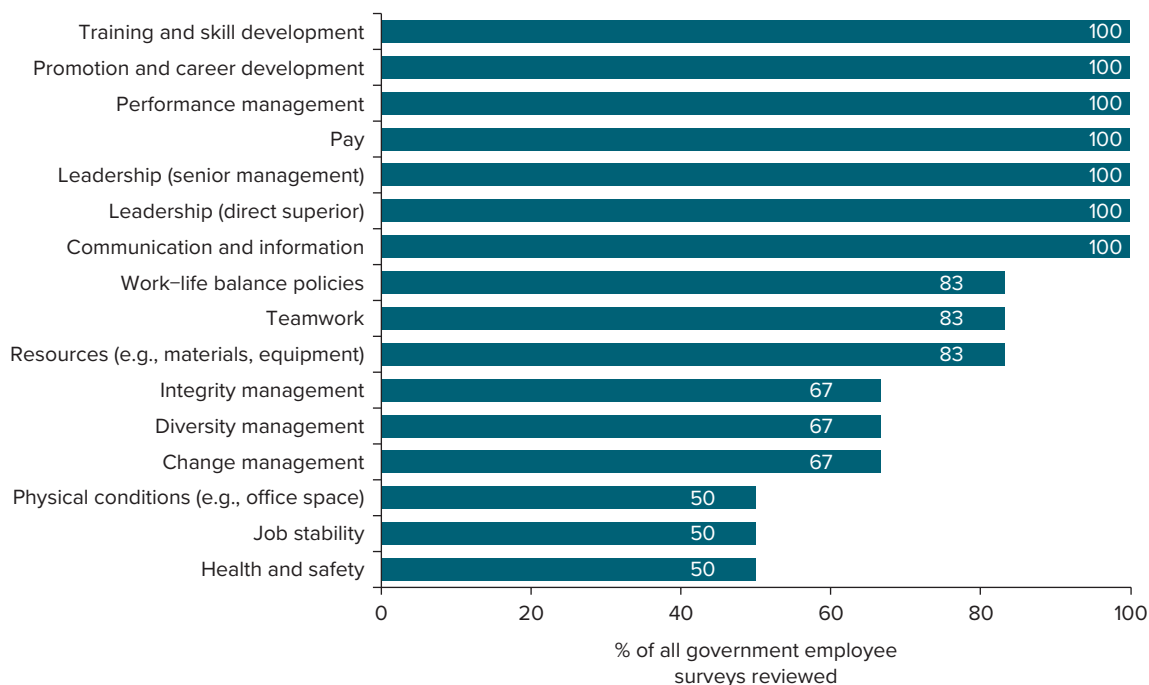
The selected surveys cover four continents and divergent socioeconomic contexts, as well as different sampling approaches and a range of widely used survey questions and indicators. Our analyses focus on a set of questions about job satisfaction, work motivation, performance review (evaluation), and merit-based recruitment. The chosen measures reflect some of the most commonly used indicators in surveys of public servants around the world, as a review by the GSPS has indicated (see figures 20.1 and 20.2). Most measures

**FIGURE 20.1 Most Commonly Used Survey Measures of Attitudes and Behaviors across Civil Servant Surveys**



Source: Meyer-Sahling et al. 2021.

**FIGURE 20.2 Most Commonly Used Survey Measures of Management Practices across Civil Servant Surveys**



Source: Meyer-Sahling et al. 2021.

are indicators, which are averages across several questions. We highlight where single questions, rather than indexes, are used for analysis.

As table 20.1 summarizes, the surveys selected for analysis in this chapter were all conducted between 2017 and 2019. Most surveys were conducted online using a structured format with closed-ended questions. The Liberia survey used a semi-structured format, akin to that used by the World Management Survey.<sup>5</sup> Trained enumerators asked open-ended questions and then selected a precoded answer option based on the responses that participants provided.

The Romania survey used two approaches: online and face-to-face. As chapter 19 on survey mode effects shows in more detail, surveys conducted via face-to-face enumeration tend to have higher response rates. For simplicity, in the simulations underpinning this chapter, we assume that these response rates remain the same.<sup>6</sup>

All surveys identify organizations within the sample. For each survey, the means for institutions were calculated in order to compare their performance. The number of organizations ranges from 30 to 65 per survey.

To foster comparability in our sampling simulations, we select survey questions that are similar, to the extent possible, across surveys. The exact wording can be found in table 20.2. To foster the generalizability of our findings to other surveys, the selected survey questions cover a range of core and frequently asked-about topics in civil service surveys—such as work motivation, job satisfaction, performance management, leadership, and the quality of management practices.

The distributions of each of the included variables in each of the countries and public sector organizations are visualized in figure 20.3.

## Monte Carlo Simulations

We show, based on these data, what sample sizes might be needed to draw the most common types of inference—defining country-level aggregates, comparing key demographic groups of civil servants (male vs. female and manager vs. nonmanager), and ranking organizations within public administration. Our hope is that these examples will help practitioners find examples that are similar to their own cases. This will provide

**TABLE 20.1** Characteristics of Surveys Included for Simulations

Country	Sampling strategy	Year	Key indicators	Key comparisons made	Mode	Sample size	No. of orgs.	Response rate
Chile	Simple random	2019	Motivation, leadership, performance, recruitment practices	Organization; unit	Online	23,636	65	44%
Liberia	Stratified random	2017	Management, recruitment practices	Organization; unit	Face-to-face	2,790	33	48%
Romania	Cluster random	2019	Motivation, leadership, performance, recruitment practices	Organization; unit	Face-to-face Online	2,721 3,721	30	92% 24%
United States	Cluster stratified random	2019	Engagement, satisfaction	Organization; previous years	Online	615,395	45	43%

Source: Original table for this publication.

**TABLE 20.2 Overview of Survey Questions for All Items Included in the Simulations, by Survey**

Survey	Indicator	Question	Original scale
Chile	Satisfaction question	I am satisfied with my job.	1 (strongly disagree) to 5 (strongly agree)
	Motivation question	I do my best to do my job, regardless of the difficulties.	1 (strongly disagree) to 5 (strongly agree)
	Performance review question	Did you have the opportunity to discuss the results of your last individual performance appraisal with your line manager?	0–1 dummy
	Merit-based recruitment question	Thinking about how you got your first job in the public sector—which of the following evaluations did you have to go through? (Written examination.)	0–1 dummy
	Motivation index	I am willing to start my workday earlier or stay after my hours of work to finish a pending job.	1 (strongly disagree) to 5 (strongly agree)
		I perform extra tasks at work, even if they are not really required.	1 (strongly disagree) to 5 (strongly agree)
		I put my best effort to perform my work, regardless of difficulties.	1 (strongly disagree) to 5 (strongly agree)
	Leadership index	My supervisor leads by setting a good example.	1 (strongly disagree) to 5 (strongly agree)
		My supervisor says things that make employees proud to be part of this institution.	1 (strongly disagree) to 5 (strongly agree)
		My supervisor communicates clear ethical standards to subordinates.	1 (strongly disagree) to 5 (strongly agree)
		My supervisor personally cares about me.	1 (strongly disagree) to 5 (strongly agree)
Liberia	Performance	My superior evaluates my performance in a just manner.	1 (strongly disagree) to 5 (strongly agree)
		The feedback that I receive about my work helps me to improve my performance.	1 (strongly disagree) to 5 (strongly agree)
		If I put more effort in my work, I will obtain a better evaluation of my performance.	1 (strongly disagree) to 5 (strongly agree)
		A positive evaluation of my performance could lead to an increase in my salary.	1 (strongly disagree) to 5 (strongly agree)
		A positive evaluation of my performance could help me in obtaining a promotion.	1 (strongly disagree) to 5 (strongly agree)
		A negative evaluation of my performance could be a reason for termination.	1 (strongly disagree) to 5 (strongly agree)
Liberia	Satisfaction question	To what extent would you say you are satisfied with your experience of the civil service?	1 (very dissatisfied) to 4 (very satisfied)
	Motivation question	How motivated are you to work as a civil servant today?	0 (not motivated at all) to 10 (extremely motivated)
	Management index	Does your unit have clearly defined targets?	Five descriptive answers progressively aligned from least to most positive description of the practices in question
		How are targets and performance measures communicated to staff in your unit?	Five descriptive answers progressively aligned from least to most positive description of the practices in question
Liberia	Management index	When arriving at work every day, do staff in the unit know what their individual roles and responsibilities are in achieving the unit's goals?	Five descriptive answers progressively aligned from least to most positive description of the practices in question
		Does your unit track its performance to deliver services?	0–1 dummy
		How does your unit track its performance to deliver services?	Five descriptive answers progressively aligned from least to most positive description of the practices in question

*(continues on next page)*



**TABLE 20.2 Overview of Survey Questions for All Items Included in the Simulations, by Survey (continued)**

Survey	Indicator	Question	Original scale
Liberia (continued)	Management index (continued)	<p>How much discretion do staff in your unit have when carrying out their assignments?</p> <p>Can most of the staff in your unit make substantive contributions to the policy formulation and implementation process?</p> <p>Is your unit's workload evenly distributed across its staff, or do some groups consistently shoulder a greater burden than others?</p> <p>Consider about the projects that your unit has worked on. Do the managers try to use the right staff for the right job?</p> <p>Does your unit try to adjust how it does its work based on the needs of the unit's clients/stakeholders who benefit from the work?</p> <p>How flexible is your unit in responding to new and improved work practices and reforms?</p> <p>How do problems in your unit get exposed and fixed?</p> <p>Consider if you and your colleagues agreed to an Action Plan at one of your meetings. What would happen if the plan was not being implemented or failed to meet the set deadlines?</p> <p>In your opinion, do the management of your unit think about attracting talented people to your unit and then do their best to keep them?</p> <p>If two senior-level staff joined your unit five (5) years ago and one performed better at their work than the other, would he/she be promoted through the service faster?</p>	<p>Five descriptive answers progressively aligned from least to most positive description of the practices in question</p> <p>Five descriptive answers progressively aligned from least to most positive description of the practices in question</p> <p>Five descriptive answers progressively aligned from least to most positive description of the practices in question</p> <p>Five descriptive answers progressively aligned from least to most positive description of the practices in question</p> <p>Five descriptive answers progressively aligned from least to most positive description of the practices in question</p> <p>Five descriptive answers progressively aligned from least to most positive description of the practices in question</p> <p>Five descriptive answers progressively aligned from least to most positive description of the practices in question</p> <p>Five descriptive answers progressively aligned from least to most positive description of the practices in question</p> <p>Five descriptive answers progressively aligned from least to most positive description of the practices in question</p> <p>Five descriptive answers progressively aligned from least to most positive description of the practices in question</p>
Romania	Satisfaction question	Overall, I am satisfied with my job.	1 (strongly disagree) to 5 (strongly agree)
	Motivation question	I put forth my best effort to get my job done regardless of any difficulties.	1 (strongly disagree) to 5 (strongly agree)
	Performance review question	Has your superior discussed the results of your last performance evaluation with you after filling in your performance evaluation report?	0–1 dummy
	Merit-based recruitment question	Have you ever participated in a recruitment competition in the public administration?	0–1 dummy
	Motivation index	<p>I am willing to do extra work for my job that isn't really expected of me.</p> <p>I put forth my best effort to get my job done regardless of any difficulties.</p> <p>I stay at work until the job is done.</p>	<p>1 (strongly disagree) to 5 (strongly agree)</p> <p>1 (strongly disagree) to 5 (strongly agree)</p> <p>1 (strongly disagree) to 5 (strongly agree)</p>

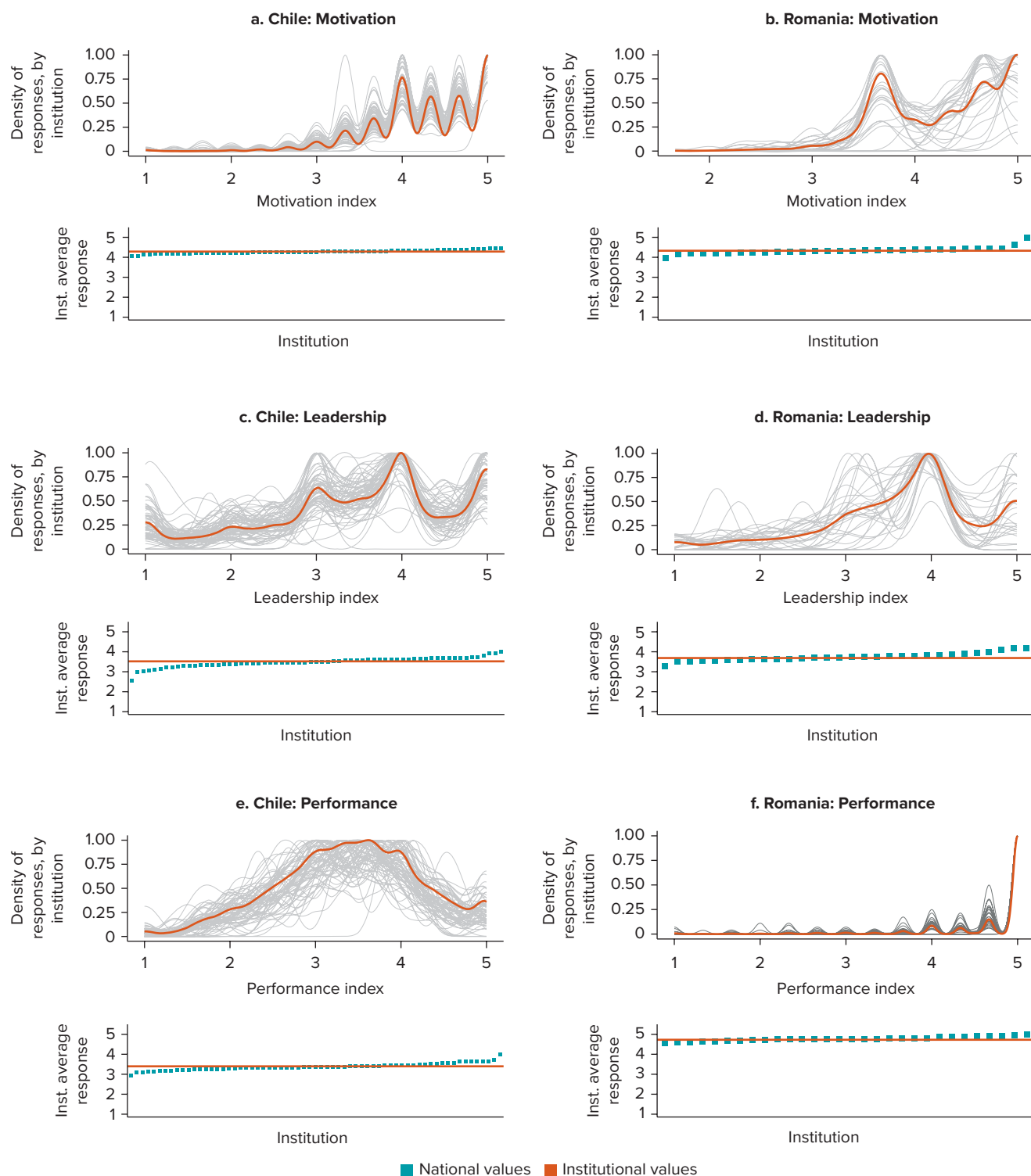
(continues on next page)

**TABLE 20.2 Overview of Survey Questions for All Items Included in the Simulations, by Survey (continued)**

Survey	Indicator	Question	Original scale
Romania (continued)	Leadership index	How frequently does your direct superior undertake the following actions? (Leads by setting a good example.)	1 (never) to 5 (always)
		How frequently does your direct superior undertake the following actions? (Says things that make employees proud to be part of this institution.)	1 (never) to 5 (always)
		How frequently does your direct superior undertake the following actions? (Communicates clear ethical standards to subordinates.)	1 (never) to 5 (always)
		How frequently does your direct superior undertake the following actions? (Personally cares about me.)	1 (never) to 5 (always)
	Performance	My performance indicators measure well the extent to which I contribute to my institution's success.	1 (strongly disagree) to 5 (strongly agree)
		My superior has enough information about my work performance to evaluate me.	1 (strongly disagree) to 5 (strongly agree)
		My superior evaluates my performance fairly.	1 (strongly disagree) to 5 (strongly agree)
United States	Satisfaction question	Considering everything, how satisfied are you with your job?	1 (strongly disagree) to 5 (strongly agree)
	Motivation question	I am willing to do extra work for my job that isn't really expected of me.	1 (strongly disagree) to 5 (strongly agree)
	Engagement index	In my organization, senior leaders generate high levels of motivation and commitment in the workforce.	1 (strongly disagree) to 5 (strongly agree)
		My organization's senior leaders maintain high standards of honesty and integrity.	1 (strongly disagree) to 5 (strongly agree)
		Managers communicate the goals of the organization.	1 (strongly disagree) to 5 (strongly agree)
		I have a high level of respect for my organization's senior leaders.	1 (strongly disagree) to 5 (strongly agree)
		Overall, how good a job do you feel is being done by the manager directly above your immediate supervisor?	1 (strongly disagree) to 5 (strongly agree)
		Supervisors in my work unit support employee development.	1 (strongly disagree) to 5 (strongly agree)
		My supervisor listens to what I have to say.	1 (strongly disagree) to 5 (strongly agree)
		My supervisor treats me with respect.	1 (strongly disagree) to 5 (strongly agree)
		I have trust and confidence in my supervisor.	1 (strongly disagree) to 5 (strongly agree)
		Overall, how good a job do you feel is being done by your immediate supervisor?	1 (strongly disagree) to 5 (strongly agree)
		I feel encouraged to come up with new and better ways of doing things.	1 (strongly disagree) to 5 (strongly agree)
		My work gives me a feeling of personal accomplishment.	1 (strongly disagree) to 5 (strongly agree)
		I know what is expected of me on the job.	1 (strongly disagree) to 5 (strongly agree)
		My talents are used well in the workplace.	1 (strongly disagree) to 5 (strongly agree)
		I know how my work relates to the agency's goals.	1 (strongly disagree) to 5 (strongly agree)

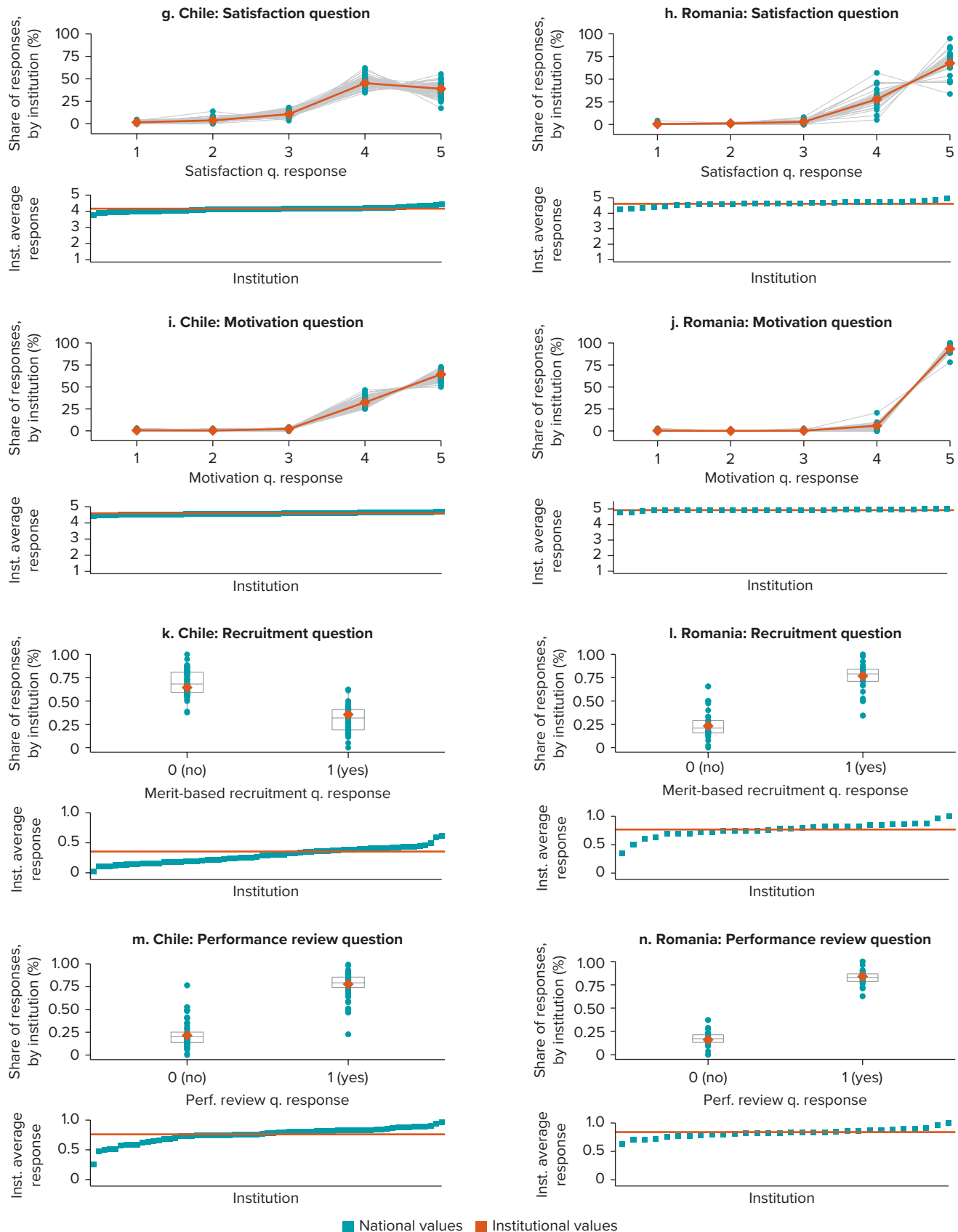
Source: Original table for this publication.

**FIGURE 20.3 Full-Sample Distribution of Key Indicators: National and Institutional Values across Surveys**



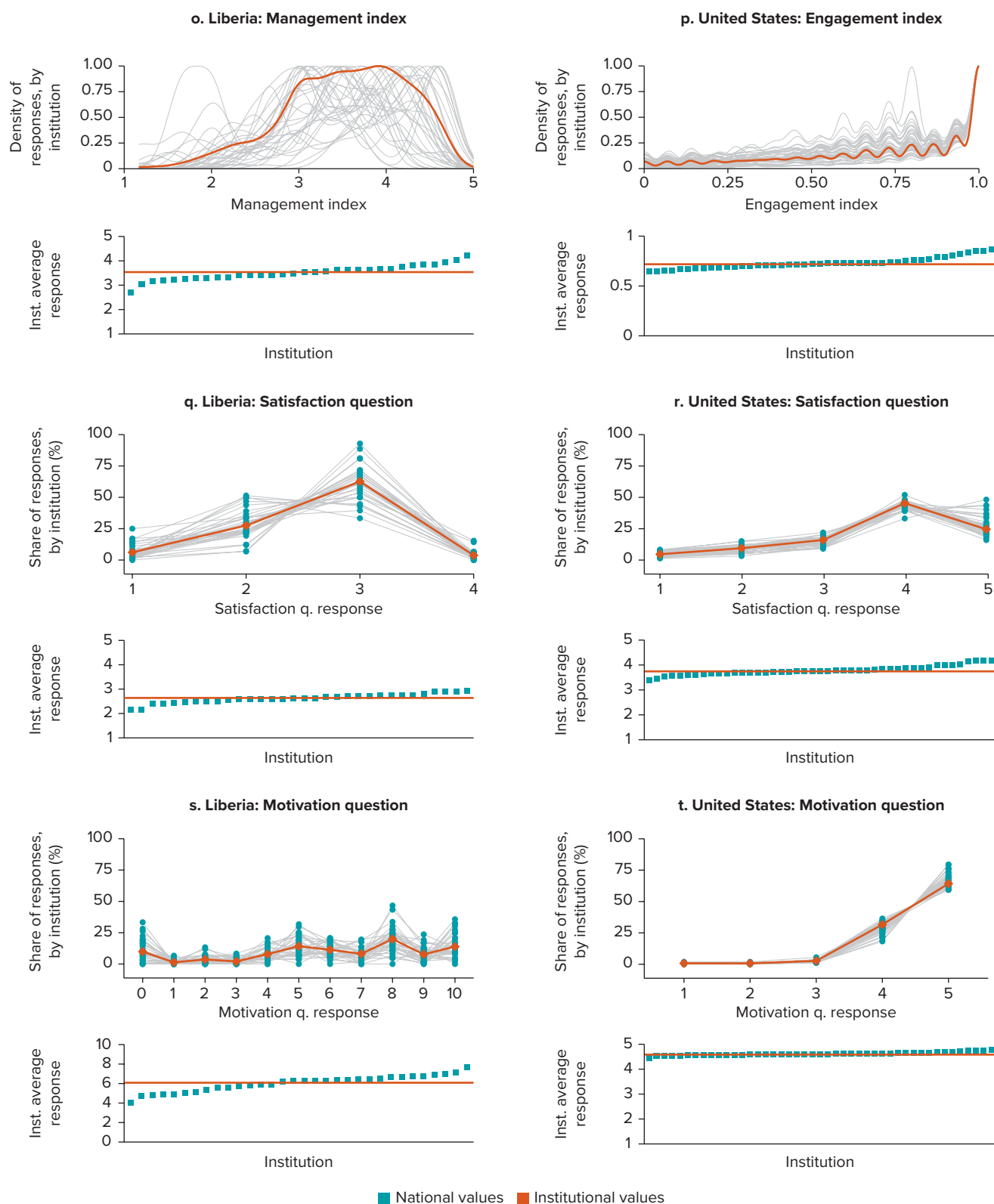
(continues on next page)

**FIGURE 20.3 Full-Sample Distribution of Key Indicators: National and Institutional Values across Surveys (continued)**



(continues on next page)

**FIGURE 20.3 Full-Sample Distribution of Key Indicators: National and Institutional Values across Surveys (continued)**



Source: Original figure for this publication.

Note: This figure shows the distribution of all key indicators analyzed in this chapter for the full sample of each of the surveys. Each subfigure refers to one indicator and survey and is divided into two panels. The top panel shows the distribution of responses, and the bottom panel shows ordered institution-level averages for a given indicator and survey. Red lines and points refer to aggregate values at the level of individual institutions within the civil service, whereas the blue ones refer to national-level values. Inst. = institution; perf. = performance; q. = question.

guidance on which sample sizes are more likely to yield satisfactory results—a goal that is further supported by the online sampling tool published alongside this chapter.

To do so, we use Monte Carlo simulation procedures to estimate:

- Sample means, standard deviations, and confidence intervals, to illustrate how sampling affects the statistical precision of estimates, and
- Differences in means between organizations and two groups of public servants that are often compared (managers and nonmanagers), to illustrate how sampling affects the possibility of statistically significant benchmarking between public sector institutions and groups of public servants—which is one primary use, in practice, of civil service survey data.

A random seed is set, from which a defined number of individual response IDs is randomly drawn, following the sampling strategy of the survey in question. This is repeated 1,000 times for each sampling proportion. All statistics presented here average across the number of simulations, providing an estimate for the average conclusions one would draw, given a certain number of individuals sampled, if the survey had been repeated 1,000 times. As a robustness check, we repeat each run of 1,000 simulations with a total of three different random seeds and record whether results deviate by more than 0.005 points on the answer scales. The results reported here have passed this robustness check.<sup>2</sup>

The results of the simulations are compared to means, standard deviations, confidence intervals, differences in means between manager and nonmanagers, and organizational rankings derived from the original surveys. In other words, we accept the statistics derived from these original surveys as the true sample statistics. We do not make statements of how these original means compare to “true” population means. We simply assume that the original sample sizes provide the best feasible estimates of underlying truths.

This approach has the advantage of not making assumptions about the population distribution beyond the information available to us. However, it is possible that the original sample sizes also over- or underestimated the true population parameters. If this is the case, results that indicate bias should be interpreted as lying even further away from the truth than when the original sample sizes were employed.

We evaluate the adequacy of the sample sizes using the following metrics:

- **The proportion of cases that fall within 95 percent of the confidence interval of the estimated means derived from the original samples.** Note that this metric is the inverse of what is typically used in statistics textbooks for the following reason: in our simulations, we sample smaller fractions of the original sample and see how well they perform in terms of recovering the original estimates. Mechanically, the confidence intervals for the estimates derived from samples with a small N will be larger than those derived from samples with a larger N. This means that it is more likely that a small sample includes the original mean, as it is wider. We instead want to know whether the estimated means of our new, smaller samples are close enough to the original mean (that is, within its confidence interval of 95 percent). For simplicity, we refer to estimates that fall within the 95 percent confidence interval of the original samples as estimates that have successfully been recovered.
- **The proportion of cases in which we find a significant difference between group means although there is none in the original data (type I error) and in which no significant difference is found although a difference between groups exists in the original data (type II error).** For the metrics presented here, we do not distinguish between the types of error that occur; we simply report the rate at which an error is made.
- **The proportion of cases in which an organization’s rank based on one of the metrics shifts into another performance quintile.** We use the proportion of shifts for ease of interpretation. For a more granular measure, we also calculate the Kendall’s tau rank correlation coefficient.<sup>8</sup>

The first metric illustrates the likelihood, given a sample size, that the means obtained are meaningfully different from those obtained from the original target sample size. The second metric illustrates the risk of drawing misleading inferences about differences in organizational subgroups. For smaller sample sizes, the



risk increases that one might wrongly conclude—for instance—that managers rate organizational characteristics differently than nonmanagers, when they do not, or conclude the opposite, when they indeed think differently. The third metric illustrates the extent to which the robustness of organizational rankings is affected by reductions in sample size. One frequent use of civil service surveys—and employee engagement surveys in the private sector (for example, Harter et al. 2020)—is the benchmarking of organizations and units—be that the benchmarking of different public sector organizations, units within public sector organizations, or organizations across the public and private sector, or benchmarking with other countries.

Benchmarking is often deemed crucial to understanding strengths and weaknesses by showcasing how well a unit or organization performs in comparison to other, similar organizations or units. Given the limited variation and skew of many variables typically included in civil service surveys (see chapter 21)—and, as a result, the small differences between organizations—the individual ranks of organizations are likely to be highly sensitive to sample composition changes. We thus instead assess whether changes in sample size can move an organization into an entirely different tranche of organizations in benchmarking. For instance, if a unit changes from ranking in the bottom 20 percent of performers to the midrange, this can have serious consequences for how problematic or nonproblematic its performance is perceived to be. We thus focus on quintile changes due to sample composition changes.

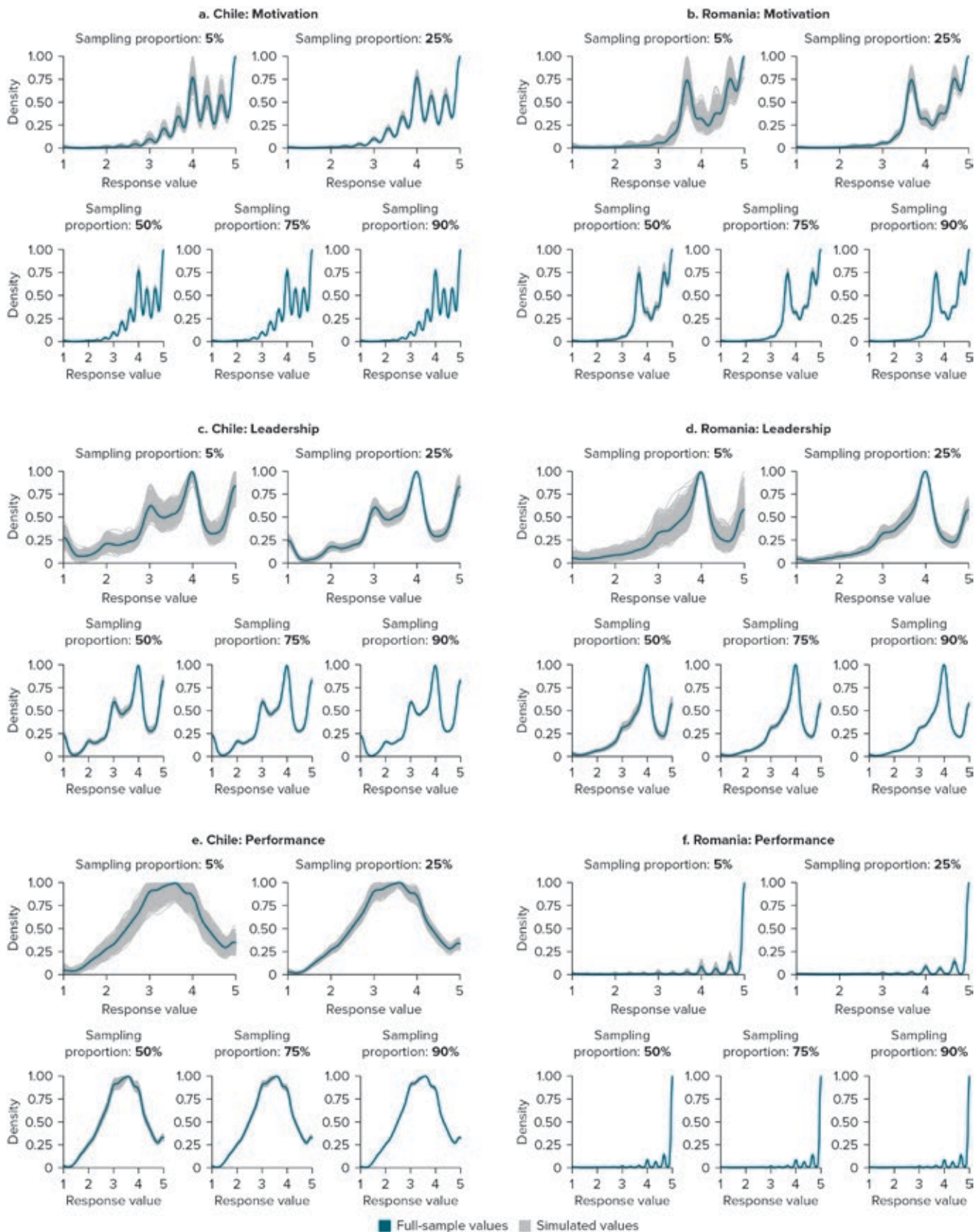
## EMPIRICAL FINDINGS ON SAMPLE SIZE REQUIREMENTS FOR PUBLIC ADMINISTRATION SURVEYS

Figures 20.4 and 20.5 present our results graphically. Figure 20.4 showcases how the distribution of results would change with distinct sample sizes relative to what was collected in the case of each survey. Figure 20.5 showcases how the accuracy of the results—benchmarked against the statistics derived from the full sample—varies with the proportion of sample that is used.

The results of our simulations against our three metrics underscore that appropriate sample sizes are largely a result of the intended use of the survey results. Assessing, first, statistical precision—our first metric—we find, for most metrics across all four surveys, 50–60 percent of the original sample size suffices to estimate means that fall within the 95 percent confidence interval of the original mean. In other words, if the objective of a civil service survey is to recover reasonably precise statistical estimates about civil servants at the country level, all four surveys currently oversample respondents. While single random surveys with a considerably smaller sample size can lead to substantial over- and underestimates of means, on average, differences are small. They range between 0.002 and 0.13 points on a five-point scale, or, expressed differently, 4 percent and 15 percent of the original standard deviation. This can be considered a very small difference. The extent of these deviations varies somewhat across questions and country. Most countries score very similarly on measures of motivation and job satisfaction. For such measures, smaller sample sizes suffice when the goal is to calculate simple country averages. As detailed in chapter 21, questions on management practices, by contrast, offer more variation. For instance, for countries like Chile, where there is considerable variation across organizations in terms of whether and how they conduct performance reviews, larger sample sizes are required to assess these indicators adequately.

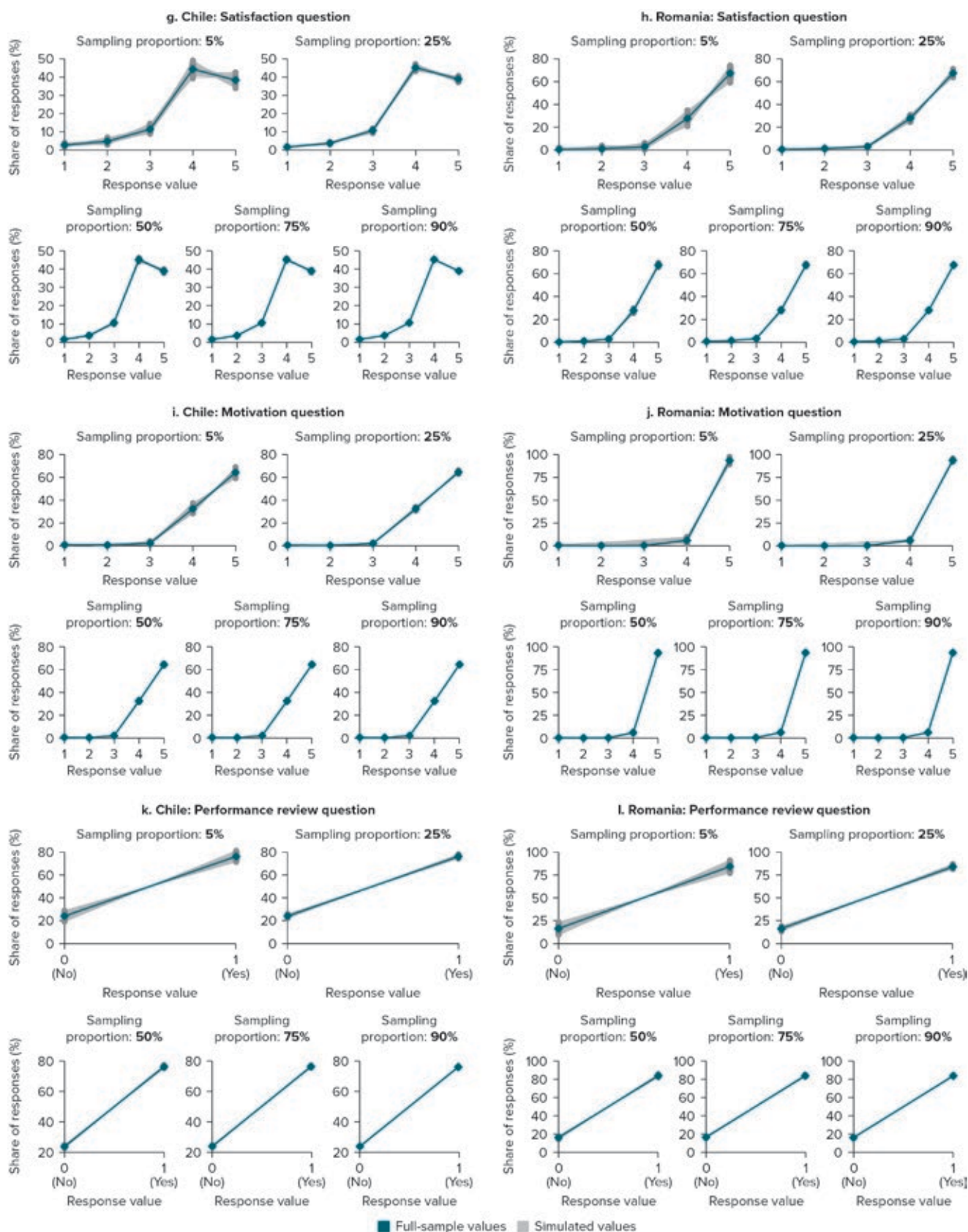
While our first metric suggests that countries oversample, our second and third metrics yield different conclusions. Consider, first, the results on the benchmarking of organizations. We find, as expected given the limited variation in many civil service survey indicators, that individual ranks are highly susceptible to changes in sample composition. In particular, if fewer than 80–90 percent of civil servants are sampled, conclusions about how institutions rank on key measures change significantly. For Romania, for instance, even when only 10 percent fewer civil servants are sampled, 50 percent of institutions change rank. At 90 percent sampled, most institutions get shuffled by one rank (there are 30 organizations in total in the sample). When only 60 percent are sampled, this increases to two to three ranks, and when only 40 percent are sampled, to

**FIGURE 20.4 Simulated Distribution of Key Indicators: National-Level Values across Surveys and Sampling Proportions**



(continues on next page)

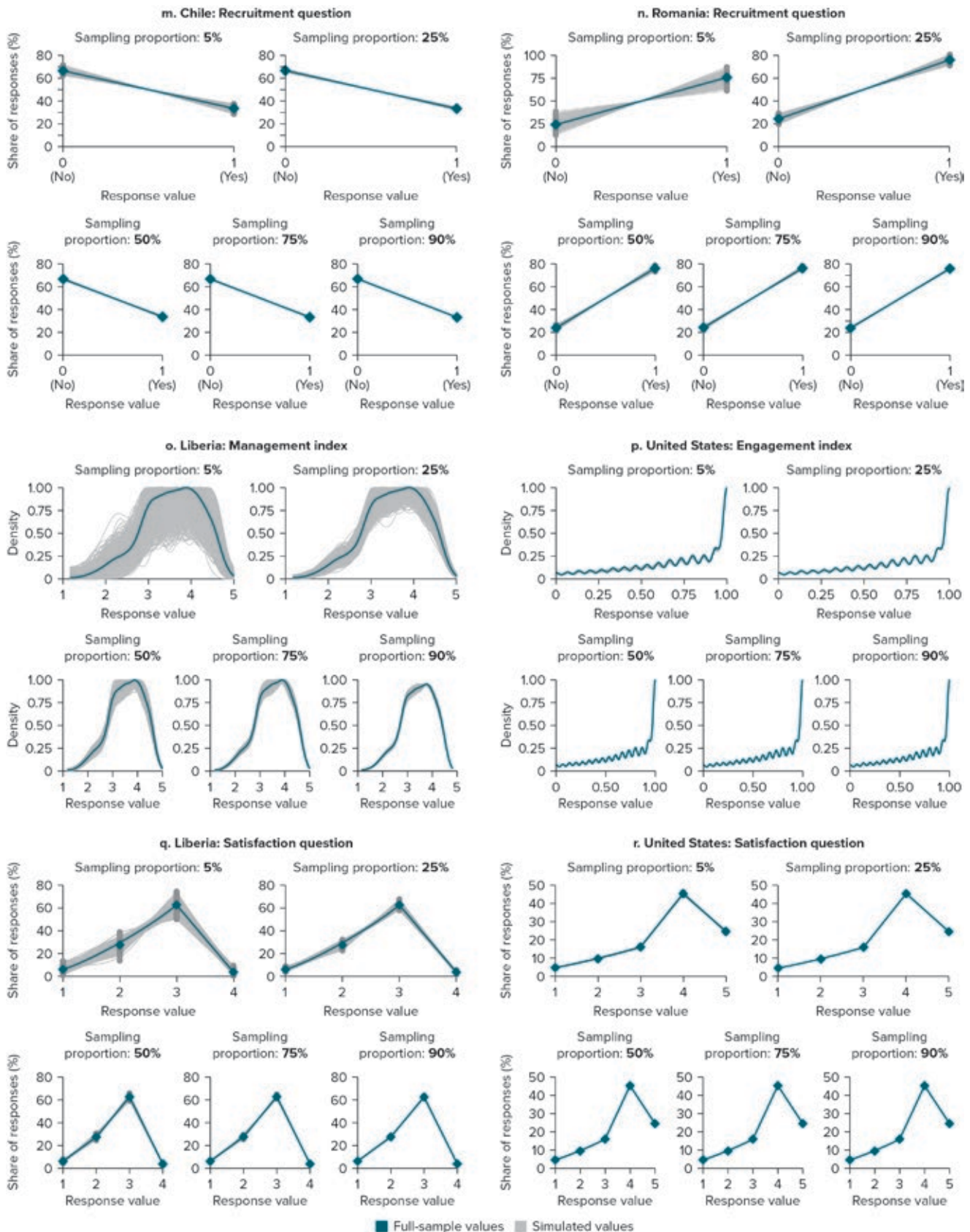
**FIGURE 20.4** Simulated Distribution of Key Indicators: National-Level Values across Surveys and Sampling Proportions (*continued*)



(continues on next page)

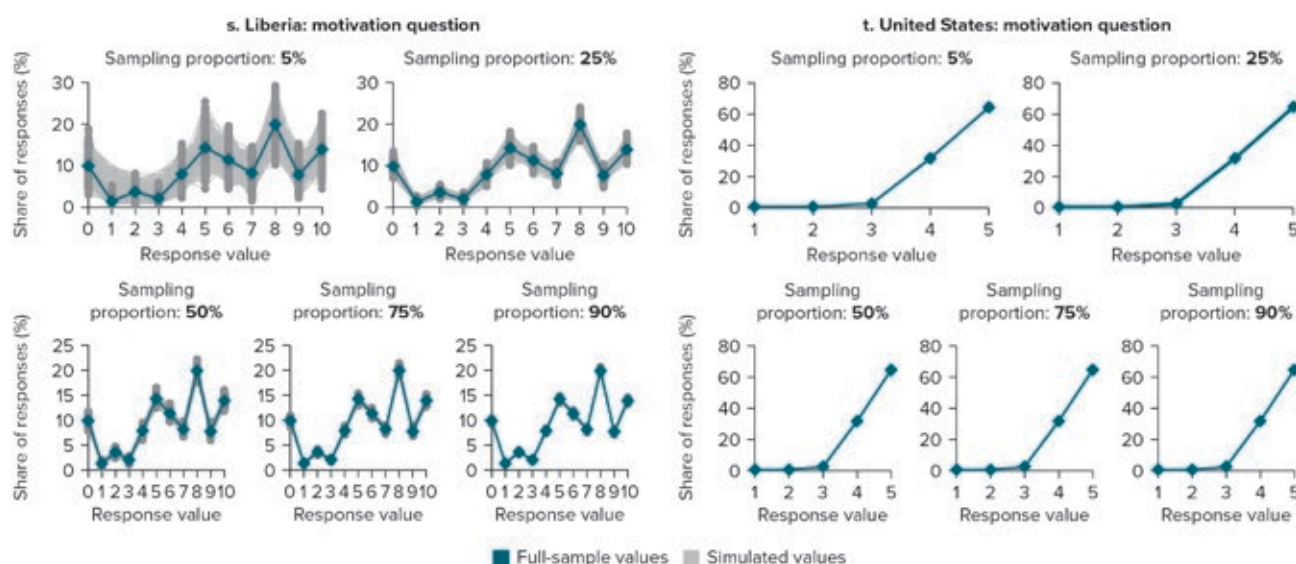


**FIGURE 20.4 Simulated Distribution of Key Indicators: National-Level Values across Surveys and Sampling Proportions (*continued*)**



(continues on next page)

**FIGURE 20.4 Simulated Distribution of Key Indicators: National-Level Values across Surveys and Sampling Proportions (*continued*)**



Source: Original figure for this publication.

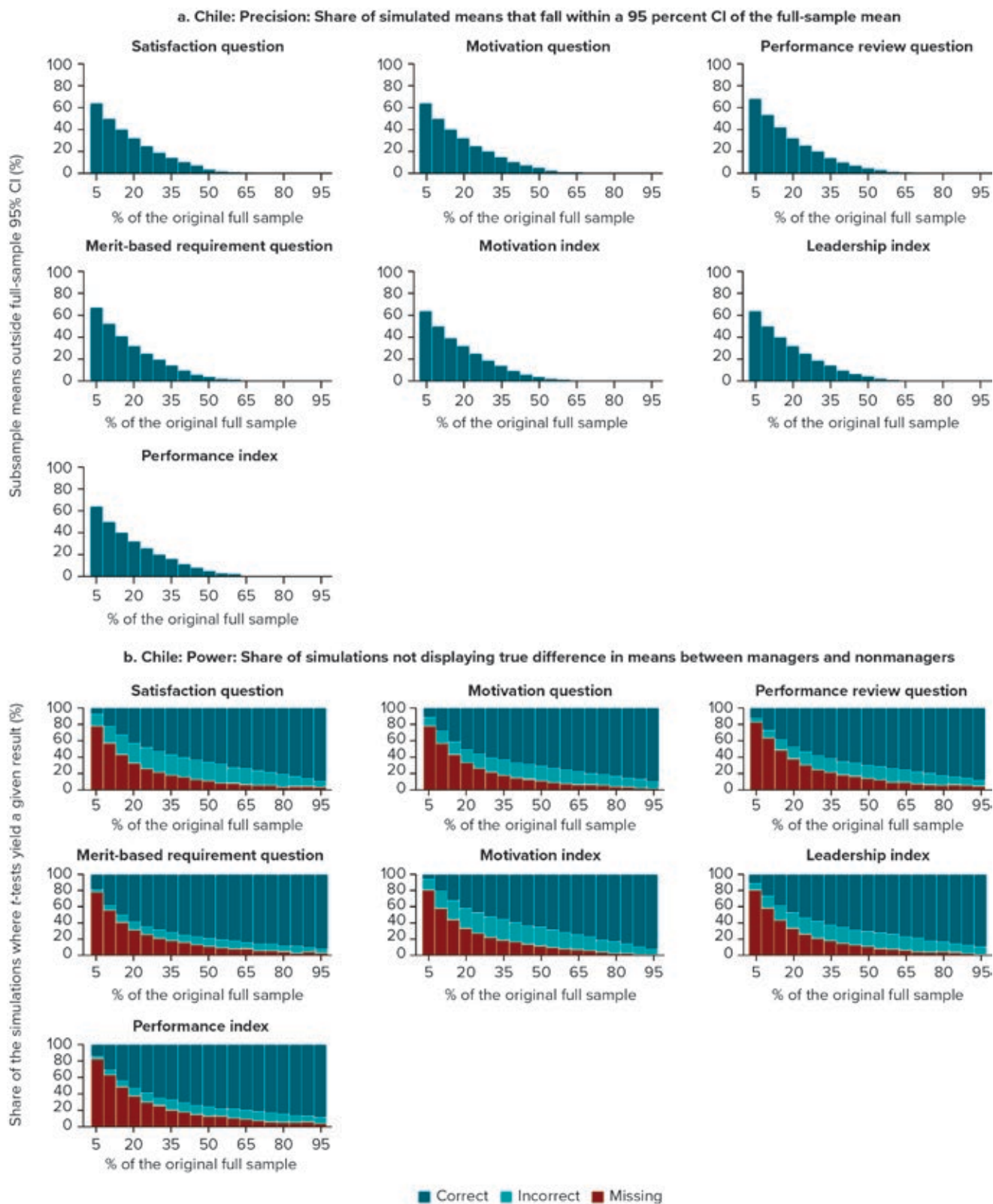
Note: This figure shows the distribution of all key indicators analyzed in this chapter across each of 1,000 simulations at different sampling proportions. The sampling proportion is specified in percentage terms on top of each line plot. Therefore, each line plot shows 1,000 simulated distributions of responses to a given question, which were obtained when a given percentage of respondents were randomly sampled from the original full-sample distribution. Gray lines and points refer to national-level distributions obtained from each simulation, whereas the blue ones refer to full-sample values obtained in the actual survey.

three to four ranks (see appendix H for the variation in institution-level values across simulations). We can also express this in terms of the Kendall's tau rank correlation coefficient, which indicates how well rankings obtained from the original data set correlate with those of the smaller samples. A rank correlation of one indicates a perfect match, and one of zero that no ranks matched. A correlation of 0.8 or more is considered desirable. This is only attainable when 80 percent or more of the original sample is surveyed for most measures. For measures with more condensed variances (motivation), sampling 60 percent or more of the original sample can achieve a similar result.

Looking at absolute shifts, however, might allow variability to appear disproportional. Often, governments, watchdogs, and international organizations group institutions into high and low performers. If we group institutions into quintiles, even at 80 percent sampled, 20–30 percent of them shift into another quintile. In other words, when 20 percent fewer civil servants are sampled, 20–30 percent of the institutions can end up being erroneously placed into the bottom 20 percent instead of the middle 20–40 percent of performers.

Another common type of analysis conducted on data derived from civil servant surveys is subgroup analysis. Statistics are typically broken down by characteristics such as job level, gender, or minority status. In our simulation example, we illustrate what the sample size requirements would be if one were to compare statistics for managers and nonmanagers. For simplicity, we report the rate of total errors committed in tests of independence. Across surveys, we find that error rates are high as soon as anything less than the original sample size is sampled. This is the case because initial differences on most indicators are very small. For example, in the original Chile survey, managers' and nonmanagers' assessments of leadership, motivation, and performance differ by less than 0.1 standard deviations (SD) for leadership and performance indicators and by about 0.2 SD for motivation. Differences in the original surveys conducted in Romania (0.1 SD), Liberia (0.2 SD), and the United States (0.1–0.2 SD) are similarly small.

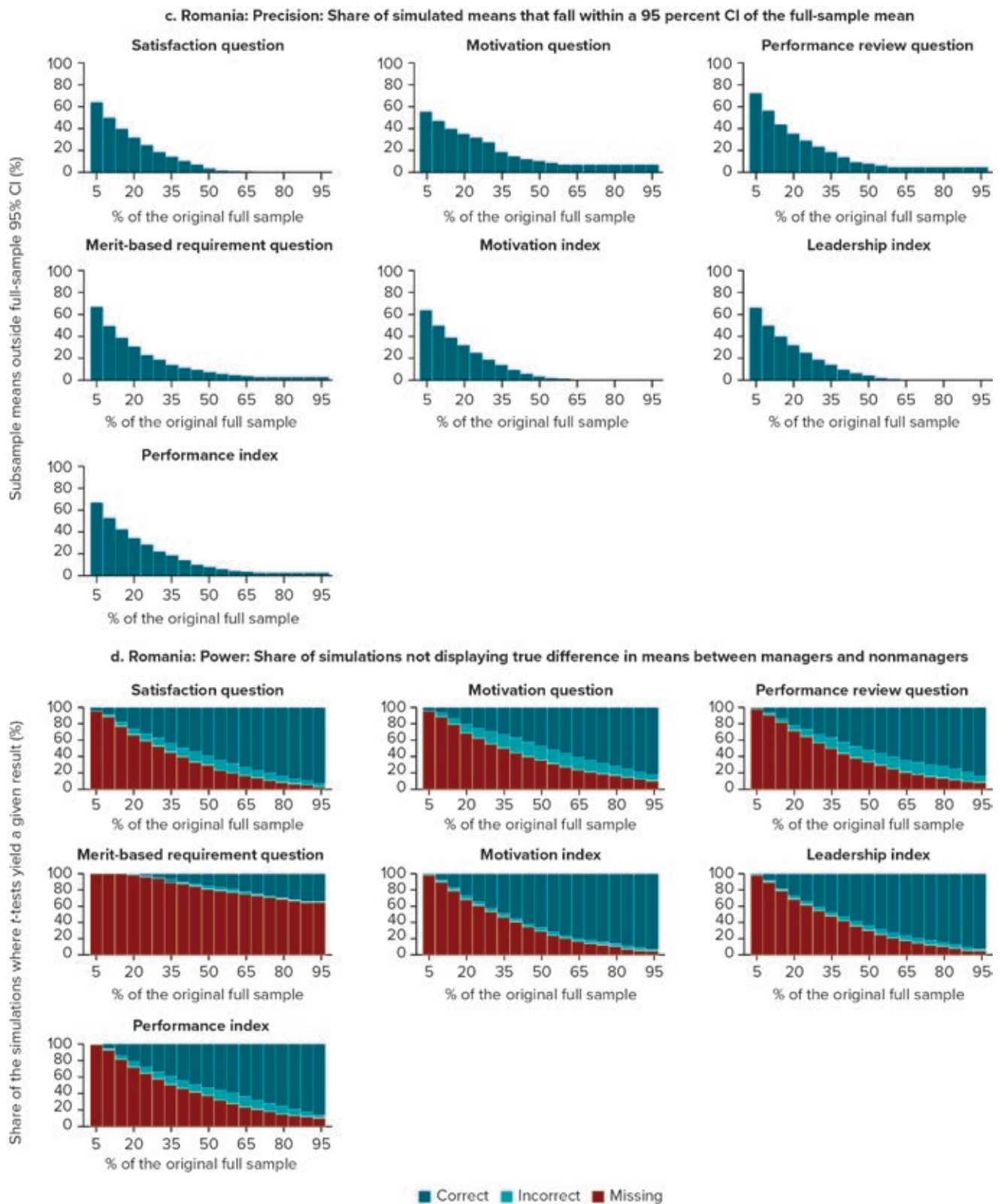
These small differences imply that the proportion of each of these subgroups needs to be rather large to be able to capture differences between the groups. Error rates for most indicators remain at around 20 percent with reduced sample sizes—considerably higher than the widely accepted 5–10 percent—until 90 percent or more of the original sample is recovered. For any sample sizes smaller than 50–60 percent of the original, indicators with an initially high variance, such as leadership in Chile, motivation in Romania,



(continues on next page)

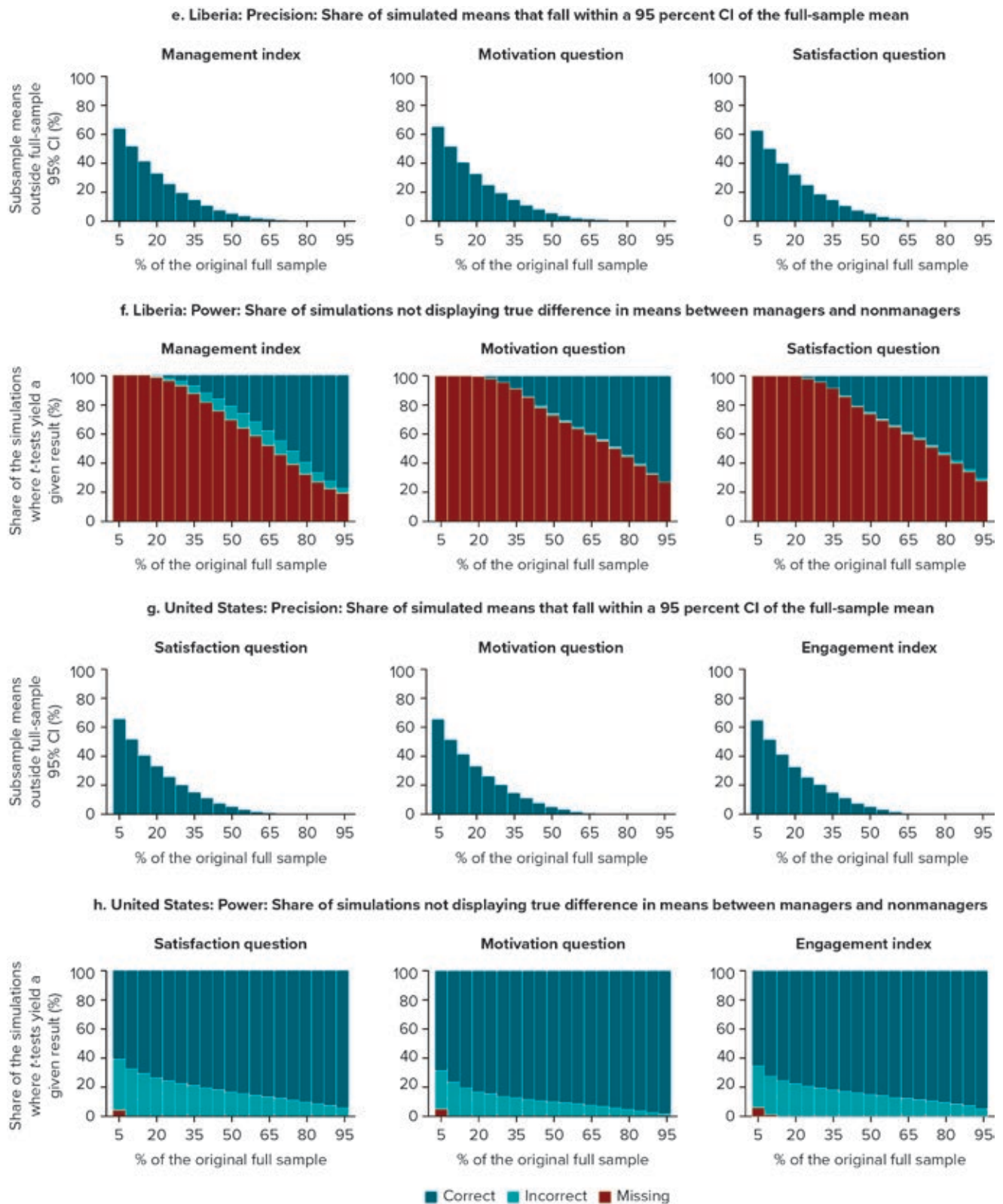


**FIGURE 20.5 Precision and Power of Simulated Distributions across Surveys (continued)**



(continues on next page)

**FIGURE 20.5 Precision and Power of Simulated Distributions across Surveys (continued)**



Source: Original figure for this publication.

Note: CI = confidence interval; perf. = performance; q. = question.

or management practices in Liberia, have very high error rates—most of the time, estimates fall far away from the true statistics.

The second challenge that conductors of civil service surveys should expect is that as sample sizes are reduced, it becomes more likely that statistics cannot be computed at all. For instance, simulations indicate that if one takes a rather conservative threshold of a minimum five observations per cell required to conduct comparisons, and if only 60–70 percent of the original sample is surveyed, in 20–30 percent of the cases, the statistic cannot be computed. The rate of failure quickly increases to 60–80 percent for questions that have a high rate of nonresponse (for example, the recruitment question in Romania).

## DISCUSSION AND CONCLUSION: IMPLICATIONS FOR CIVIL SURVEY SAMPLING

In sampling respondents, civil service survey designers face a trade-off between the costs of additional survey responses and the benefits of more precise survey estimates with greater sample sizes. What, then, are the appropriate sample sizes in civil service surveys? To assess this conundrum, this chapter has conducted Monte Carlo simulations with civil service survey data from the United States, Chile, Liberia, and Romania. Our results suggest that appropriate sample sizes depend, most of all, on the inferences governments wish to make from the data. Conclusions differ depending on which indicators are chosen and which comparisons are made. Assessing sample size requirements on a case-by-case basis, depending on government needs and survey topics, thus remains paramount.

With that said, some common patterns across civil service surveys do exist that can inform future sampling decisions. For one, on attitudinal measures—such as work motivation or job satisfaction—smaller sample sizes might be sufficient if the objective is relatively precise means (though no benchmarking). To estimate averages for countries or larger organizations, sample sizes could often be reduced. Where there are differences in practice that vary substantially by institution, however, the required sample sizes for the country increase.

At the same time, where detailed comparisons among public sector organizations—or individual rankings—are sought, sample sizes are typically too small, not least because many survey measures do not offer large variation between organizations and thus require high levels of precision to enable comparison.

However, in such instances, practitioners should first assess whether the magnitude of historical differences is likely to be sufficiently meaningful to increase sample sizes to obtain statistically significant differences. For instance, does it merit changing organizational strategies if nonmanagers are 0.05 standard deviations less satisfied? Or would the gap need to be closer to one full standard deviation (which suggests a sizeable gap) to be substantively meaningful? If the answer is the latter, then increasing sample sizes to obtain statistically significant differences on the former would not be meaningful.

This chapter thus concludes that a determination of the use for survey results should precede the determination of sample sizes. Once that discussion has been had, practitioners can turn to an online toolkit to estimate appropriate sample sizes depending on the intended uses of the survey data. We recommend that practitioners look for countries, survey measures, and comparisons or benchmarking similar to their own use case in the online tool for guidance on which sample sizes are likely required in their own surveys.

## NOTES

1. Our calculation is based on data from the CBO (2017).
2. More information about the experimental statistics program is available on the website of the ONS at <https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/guidetoexperimentalstatistics>.

3. Interested readers can access the toolkit at [https://encuesta-col.shinyapps.io/sampling\\_tool/](https://encuesta-col.shinyapps.io/sampling_tool/).
4. In this chapter, we assess sample sizes from the perspective of the types of inference that can be drawn from civil service survey data. For political reasons, governments may, of course, choose to undertake a census irrespective of whether this is necessary from a statistical perspective, to give every public employee the opportunity for *voice*—that is, the opportunity to give their feedback on matters of concern in the survey.
5. More information about the World Management Survey can be found on its website, <https://worldmanagementsurvey.org/>.
6. In practice, this would mean that if one were to sample again in the same country, using the same survey mode, response rates would look the same as for the last survey that was conducted.
7. Random seeds are used to enable replicable research. However, no computer-generated seed is truly random. Further, even if the starting seed is random, it is possible—although, by definition, very unlikely—that the random draws started from this seed end up being a very rare combination, leading to results not reflective of what most random draws would yield. Therefore, it is advisable to rerun all simulations with different seeds.
8. Kendall's tau is defined as: 
$$\tau = \frac{2}{n(n-1)} (n_{\text{concordant}} - n_{\text{discordant}})$$
.

## REFERENCES

- Bertelli, Anthony M., Mai Hassan, Dan Honig, Daniel Rogger, and Martin J. Williams. 2020. "An Agenda for the Study of Public Administration in Developing Countries." *Governance: An International Journal of Policy, Administration, and Institutions* 33 (4): 735–48. <https://doi.org/10.1111/gove.12520>.
- Bjarkefur, Kristoffer, Luiza Cardoso de Andrade, Benjamin Daniels, and Maria Ruth Jones. 2021. *Development Research in Practice: The DIME Analytics Data Handbook*. Washington, DC: World Bank. <http://hdl.handle.net/10986/35594>.
- CBO (Congressional Budget Office). 2017. *Comparing the Compensation of Federal and Private-Sector Employees, 2011 to 2015*. Washington, DC: CBO, Congress of the United States. <https://www.cbo.gov/publication/52637>.
- Cochran, William G. 1977. *Sampling Techniques*. 3rd ed. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.
- Fowler, Floyd. 2009. *Survey Research Methods*. 4th ed. Applied Social Research Methods. Thousand Oaks, CA: SAGE. <https://doi.org/10.4135/9781452230184>.
- Harter, James K., Frank L. Schmidt, Sangeeta Agrawal, Anthony Blue, Stephanie K. Plowman, Patrick Josh, and Jim Asplund. 2020. *The Relationship between Engagement at Work and Organizational Outcomes: 2020 Q12 Meta-Analysis*. 10th ed. Washington, DC: Gallup. <https://www.gallup.com/workplace/321725/gallup-q12-meta-analysisreport.aspx>.
- Kalton, Graham. 1983. *Introduction to Survey Sampling*. Quantitative Applications in the Social Sciences. Thousand Oaks, CA: SAGE. <https://doi.org/10.4135/9781412984683>.
- Meyer-Sahling, Jan, Dinsha Mistree, Kim Mikkelsen, Katherine Bersch, Francis Fukuyama, Kerenssa Kay, Christian Schuster, Zahid Hasnain, and Daniel Rogger. 2021. *The Global Survey of Public Servants: Approach and Conceptual Framework*. Global Survey of Public Servants. Last updated May 2021. <https://www.globalsurveyofpublicservants.org/about>.



## CHAPTER 21

# Designing Survey Questionnaires

## Which Survey Measures Vary and for Whom?

*Robert Lipinski, Daniel Rogger, Christian Schuster, and Annabelle Wittels*

### SUMMARY

Many aspects of public administration, such as employee satisfaction and engagement, are best measured using surveys of public servants. However, evaluating the extent to which survey measures are able to effectively capture underlying variation in these attributes can be challenging given the lack of objective benchmarks. At a minimum, such measures should provide a degree of discriminating variation across respondents to be useful. This chapter assesses variation in a set of typical indicators derived from data sets of public service surveys from administrations in Africa, Asia, Europe, and North and South America. It provides an overview of the most commonly used measures in public servant surveys and presents the variances and distributions of these measures. The chapter thus provides benchmarks against which analysts can compare their own surveys and an investigation of the determinants of variation in this field. Standard deviations of the measures we study range between 0.72 and 1.24 on a five-point scale. The determinants of variation are mediated by the focus of the variable, with country fixed effects the largest predictors for motivation and job satisfaction, and institutional structure key for leadership and goal clarity.

### ANALYTICS IN PRACTICE

- Effective questionnaire design and efficient sampling strategies both rely on an understanding of the performance of relevant survey measures. This chapter presents the variation in common measures used in public servant surveys from settings across the world (see table 21.2 later in the chapter for a full listing

---

Robert Lipinski is a consultant and Daniel Rogger is a senior economist in the World Bank's Development Impact Evaluation (DIME) Department. Annabelle Wittels is an independent researcher. Christian Schuster is a professor at University College London.



of summary statistics). Such statistics provide individual survey analysts a means of comparative analysis with their own results.

- In the sample of surveys we assess, measures related to personal characteristics, such as motivation, do not vary as substantially as those relating to institutional variables, such as leadership. The design of questions about motivation and satisfaction may therefore need to be reconceptualized to better discriminate between degrees of these categories with the highest response likelihood.
- Commonly used measures of quantities of interest in public administration show significant negative skew across most contexts and for most measures. This indicates that responses are located mostly on the more-positive (higher) end of the answer scale. This might reflect response bias or an underlying lack of latent variation. It also indicates a need for redesigning standard measures in public service surveys to better discriminate between values at the top end of indexes.
- Where analysts would like to test for positive-response (or social-desirability) bias, they can include scale items specifically designed to capture such bias and apply regression or weight adjustments to averages, include alternative methods of capturing more nuanced levels of variation, or apply transformation techniques to address extreme skew before data analysis.
- The determinants of the variation we observe are mediated by the nature of the variable. Demographics generally explain a very small proportion of variation across measures (less than 2 percent). Country fixed effects account for the highest degree of variation for measures of motivation and job satisfaction. Institutional divisions (unit and subunit identifiers) explain a greater proportion of the variation in measures related to the quality of leadership and the clarity of a respondent's goals and tasks. Thus, survey measures associated with organizational features of the public service are more likely to exhibit variation than those that probe aspects influenced by servicewide or national cultures.
- Since many countries use different survey approaches and questionnaires, it is difficult to establish to what extent these differences are artificially created by differences in measurement as opposed to differences in environment, institutions, and management practices. This underlines the necessity to further standardize a core of public servants' surveys in order to make cross-country comparisons meaningful and informative for public administration reform.

## INTRODUCTION

Surveys of attitudes, perceptions, and reported behaviors often try to assess two factors: the average or most common value of a concept of interest for respondents (or respondents in a particular subgroup) and variation in those responses. For example, one might want to find out the level of job satisfaction for public servants in an agency and the variation in satisfaction across that agency, across agencies, or across public servants of different managerial ranks.

Large-scale surveys are of particular relevance where a feature of the population being surveyed varies substantially. If satisfaction, or any other variable of interest, were known to be the same everywhere, analysts would only be required to carefully measure a single instance of the phenomenon. This would then be sufficient to know the value of the variable in the population at large. A practical example of this in the public service is the *de jure* nature of laws and regulations. Recording a single instance of a universal law is sufficient to understand its nature everywhere.

By contrast, once a phenomenon can vary across individuals, units, departments, agencies, time, and so on, surveys provide a tool to measure the underlying variation. Mapping this variation allows analysts to understand the average of the variable, its spread, where the feature takes extreme or unusual values,

and so on. Again taking satisfaction as an example, a central public service agency may look for agencies that have the lowest levels of staff satisfaction, those where satisfaction is falling fastest, or those in which there are the largest disparities. Or, taking the universal-law example, analysts may be interested in how the law is de facto implemented across agencies, which may differ significantly. Surveys provide a tool for mapping the corresponding variation.<sup>1</sup>

Features of public administration increase the likelihood that there will be variation in key elements of the work environment. Unlike the private sector, there are no market forces driving work units to specific standards. The diverse range of activities undertaken by the public sector and the myriad outputs it produces imply potentially very different approaches to production. The challenges to measuring many aspects of public administration—externalities created by both tasks and public outputs, for example—compound the challenges to creating a common approach to management.

Not all phenomena of interest in the public service vary. It is conceivable that in some settings, public officials are universally oriented toward public service, or the opposite, such that even the best measures will exhibit no variation. The tension at the heart of measurement in public administration, where so few benchmark measures exist of a wide range of phenomena of interest, is how to identify which elements of the public service truly do vary, and for whom.

We proxy this underlying variation through observed variation in survey measures. Yet it is variation in the underlying phenomenon that we are interested in, rather than a proxy measured by a survey. We want measures that reflect true levels of satisfaction and allow us to discriminate between levels that are meaningfully different. If variation in survey measures solely reflects biases induced by the way questions are formulated or measurement error, it does not provide valid information upon which to base decision-making.<sup>2</sup> The validity of the variation in surveys of public administration is thus of key concern to understanding the public service.

This chapter aims to investigate the validity of measures from public servant surveys. The challenge all such exercises face is that many important concepts in public administration—such as satisfaction, motivation, and quality of management—are inherently internal phenomena. Assessing different measures against objective benchmarks, such as measures of satisfaction against turnover data, presents many issues of comparison. Alternative indicators of the validity of survey data are that conceptually related items should covary and that the same measurement should attain comparable variation across measurement in time and across survey contexts. This chapter assesses validity by comparing common measures across settings. By benchmarking which measures consistently vary across settings, we identify those measures that consistently provide differentiating variation.

The assumption of this approach to assessing the validity of variation is that the biases and measurement error that may drive variation in one setting are distinct from those in another. Thus, where we observe a measure providing discriminating variation across settings, we can infer that it is providing valid data. The disadvantage of such an approach is revealed when this assumption fails and measurement is affected by common bias across settings. For our approach to be valid, we also require that there be variation in the underlying phenomena across settings.

The payoffs for undertaking a valid assessment of variation in public servant surveys are substantial. Knowing what type and shape distributions of measurements of concepts of interest take is important for picking appropriate survey designs. Sampling strategies (see more on this in chapter 20) and question design (see chapters 22 and 23) require an understanding of underlying variances to be fit for purpose. The validity of our measurements of public administration, and of the corresponding survey designs, is at the core of our ability to understand the functioning of the state.

This chapter's perspective is that despite concerns about comparing measures of public service across time and space, such assessments act as rare and therefore valuable benchmarks to a single survey's results. Knowing that a specific measure has limited variation in many other settings allows analysts to take a more informed perspective on the use of that measure in their own contexts.

The importance of variation and its relation to validity and reliability has been investigated widely in volumes on statistics and survey sciences (for example, Brandler et al. 2007; Fink and Litwin 1995;

Grosh and Glewwe 2000; Wright and Marsden 2010). However, there is little systematic evidence available on patterns of variation of the measurements typically used to assess concepts central to the analysis of public administration. A review of common source bias in civil servant surveys conducted by George and Pandey (2017) makes evident that public administration as a field suffers from a reliance on surveys to measure both independent and dependent variables.<sup>3</sup> This approach inflates correlations between different variables and can make it difficult to distinguish between effects driven by individual-level error, individual-level differences, and generalizable relationships between different factors.<sup>4</sup> Most other work relates to scale validation. (For example, for public service motivation, see Kim 2009; Mikkelsen, Schuster, and Meyer-Sahling 2021; Perry 1996; for job satisfaction, see Cantarelli, Belardinelli, and Bellé 2016; for policy alienation, see van Engen 2017; for public leadership, see Tummers and Knies 2016 and chapter 24 on invariance.)

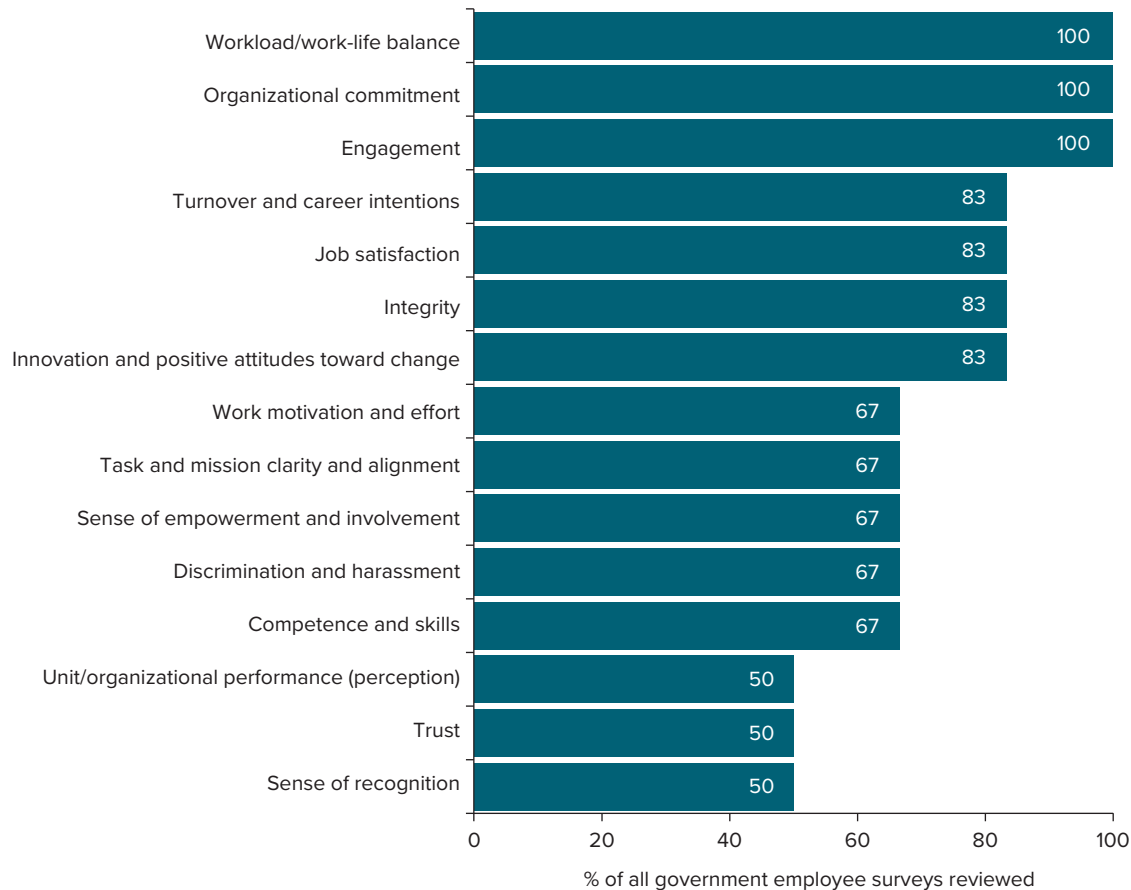
This chapter, therefore, assesses variation in a set of typical indicators derived from data sets of public service surveys to add to the existing literature and provide survey designers and analysts with benchmarks against which to assess their own efforts. The data span administrations in Africa, Asia, Europe, and North and South America. The chapter provides an overview of the most commonly used measures in public servant surveys and presents the variances and distributions of these measures. It then describes the extent to which observed variance can be explained by demographic and institutional characteristics typical of the analysis undertaken by analysts of public servant surveys.

## KEY CONCEPTS IN PUBLIC SERVANT SURVEYS AND THEIR MEASUREMENT

What are the phenomena that surveys of public servants typically seek to measure? Meyer-Sahling et al. (2021) undertake a review of major surveys of public servants. They find that government employee surveys almost universally ask questions about workload or work-life balance and organizational commitment and engagement, and they tend to ask about career intentions, job satisfaction, integrity, and attitudes toward organizational change. Figure 21.1 provides a breakdown of the proportion of surveys that attempt to measure each of these phenomena. The figure indicates that though the precise set of measures used in government employee surveys varies, the core concepts themselves overlap significantly. Why have these specific topics become the major areas of investigation in surveys of public servants? Some measures that are central to existing government employee surveys, such as engagement, do not emerge clearly from reviews of academic models of public service governance (Meyer-Sahling et al. 2021). This is, in large part, due to the fact that “models” of public service governance do not engage in depth with organizational psychology. Such considerations are critical to the actual management of the public service. Management practices, such as work-life balance policies or leadership to generate enthusiasm for the mission of a public sector organization, are important predictors of the attitudes and behaviors of public servants (see, for example, Esteve and Schuster 2019) and feature prominently in government employee surveys, yet models of public service governance are (with some exceptions) silent about them.<sup>5</sup> Major surveys of public servants thus reflect the priorities of those who manage them, typically central agencies of public service management.

For this chapter, we assess the topics within surveys for which we both have access to the underlying microdata and which contain required identifiers (such as organization). We focus on those measures dealt with in a comparable way across these surveys. As we describe in more detail later, these topics are job satisfaction, pay satisfaction, motivation, assessments of leadership’s trustworthiness and tendency to motivate, a measure of performance management related to promotion, and the clarity respondents have over goals and tasks. As can be seen from figure 21.1, these overlap closely with many of the standard topics in surveys of public servants. In this section, we review the existing evidence on the quality of measurement of these topics in surveys of public servants and their relationship to the effective functioning of public administration.

**FIGURE 21.1** Topics Measured in Government Employee Surveys



Source: Meyer-Sahling et al. 2021.

Note: Meyer-Sahling et al. (2021) review the Federal Employee Viewpoint Survey (FEVS) in the United States, Canada's Public Service Employee Survey, the United Kingdom's Civil Service People Survey, the Australian Public Service Employee Census, Colombia's Survey of the Institutional Environment and Performance in the Public Sector, and Ireland's Civil Service Employee Engagement Survey. Questionnaires were reviewed for the last survey year prior to the COVID-19 pandemic. Only concepts covered in at least half of the surveys are shown.

## Measuring Job Satisfaction

A review conducted by Cantarelli, Belardinelli, and Bellé (2016) finds that about a quarter of studies in public administration use a single item to measure job satisfaction. Three-quarters use an index based on several question items. They all use Likert-type response scales. Some measure overall feeling ("How satisfied are you?") while others measure specific aspects of job satisfaction, such as pay, career prospects, and work-life balance. Several surveys, such as the United States Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS), use a mixture.

This diversity in measurement approach can partly be explained by the lack of a coherent theory as to how job satisfaction factors into public sector performance. Myriad authors have postulated links between satisfaction and public service performance, as well as corresponding interpretations of satisfaction (Hameduddin and Engbers 2022; Mehra and Joshi 2010; Potts and Kastle 2010).<sup>6</sup>

Cantarelli, Belardinelli, and Bellé's (2016) meta-analysis reports significant correlations between the various measures of job satisfaction used in the literature and organizational commitment, feelings of inclusion, justice, goal clarity, turnover intentions, leadership perceptions, and positive perceptions of promotion systems and monetary incentives. Job satisfaction measures have also been shown to be strongly correlated with measures of employee mental health and burnout (Faragher, Cass, and Cooper 2013; Scanlan and Still 2019). These within-survey correlations are taken as indicators of the validity of measures of satisfaction.

Returning to the discussion on objective benchmarks of satisfaction, studies typically use turnover as an independent measure of satisfaction. However, longitudinal studies suggest that the relationship between job satisfaction and staff turnover might be temporal or spurious (Cramer 1996; Sousa-Poza and Sousa-Poza 2007). It is simply unclear to what extent turnover is a good marker of organizational productivity and good management in the context of a noncompetitive job market. Staff might simply stay because they have no attractive exit options.

## Measuring Motivation

Motivation is most commonly measured via a scale developed by Perry (1996) or its adaptation by Kim (2009). Perry's scale consists of 24 items and has six dimensions: attraction to public policy making, commitment to the public interest, civic duty, social justice, self-sacrifice, and compassion. Each of the dimensions is measured by four questions with a Likert-type response scale. The measure developed by Kim consists of 12 items and has four dimensions: attraction to policy making (APM), commitment to the public interest (CPI), compassion (COM), and self-sacrifice (SS).

Mikkelsen, Schuster, and Meyer-Sahling (2021) have found that Kim's scale can be used to compare relationships between motivation and other variables across cultural contexts but that means cannot be meaningfully compared across country contexts. In terms of concept validity, a large body of correlation-based studies shows robust correlations between measures of public service motivation and effort (Camilleri and Van Der Heijden 2007; Moynihan and Pandey 2007; Naff and Crum 1999). However, as the vast majority of these studies rely on correlations between self-assessments completed in a survey, they all suffer from the threat of common source bias (Favero and Bullock 2015; George and Pandey 2017; Meier and O'Toole 2010): the risk that correlations are an artifact of individuals' providing multiple ratings about themselves at the same point in time, without any external validation.

Causal studies are limited because intrinsic motivation cannot be manipulated directly. However, several experiments have shown that when extrinsic, nonfinancial rewards are increased, the performance of public sector employees improves (Ashraf et al. 2020; Bellé 2014, 2015; Bellé and Cantarelli 2015).

## Measuring Leadership

Leadership in public administration has typically been measured with scales based in or borrowed from management science and psychology (Tummers and Knies 2016). Scales commonly measure specific types of leadership. Two types studied extensively are transformational and transactional leadership (for example, Hameduddin and Engbers 2022; Kroll and Vogel 2014; Pandey et al. 2016; Vigoda-Gadot 2007). More recently, there has been a movement to develop scales that are particular to the leadership challenges faced by public sector managers, such as working with a large and diverse network of stakeholders and responding to the demands of political principals, all while remaining accountable to a broad public (Tummers and Knies 2016; Vogel, Reuber, and Vogel 2020).

In terms of the validity of leadership scales, Hameduddin and Engbers (2022) show in a meta-analysis of studies on public service motivation that there is considerable evidence that assessments of leadership predict motivation levels in staff. The relationship seems to be consistent across country contexts. Problematically, as with motivation, the majority of studies measure both motivation and leadership with a single survey. They are thus subject to common source bias, and there is a risk that relationships are spurious (see George and Pandey 2017). Evidence derived through other methods increases confidence, however, that this is not the case. For example, a field experiment conducted by Bellé (2014) finds that transformational leadership interventions can increase motivation as measured by output quantity. Other examples are 360-degree assessments of leadership, whereby assessments are collected from managers themselves and staff (Vogel and Kroll 2019), and varying whether questions are asked at the organizational or individual level (see chapter 23). Both approaches have found significant relationships between assessments of leadership and motivation.

## Measuring Performance Management

In the academic literature, the measurement of public performance is often equated with the management approaches in place (Bouckaert 2021). A lot of measurement on public sector productivity now is concerned with integrating the measurement of inputs with administrative data on outputs (for recent reviews and discussions, see Heinrich 2002; Somani 2021). Questions about performance management in public service surveys typically ask staff to report what management approaches are used in their organization. Survey items include questions on the frequency and adequacy of performance reviews; goal setting and goal clarity; and the recognition of good performance, promotion, and financial incentives. However, little consensus exists on the appropriate approach to the measurement of these concepts in public service surveys.

To the knowledge of the authors, no comprehensive reviews of the validity or reliability of such question items and scales have been published. However, government agencies in charge of running public service surveys report routinely undertaking reviews of the relevance and internal validity of such measures.<sup>2</sup>

## Measuring Goal and Task Clarity

The concept of *goal clarity* was first developed under the umbrella of organizational psychology. Latham and Locke (1991) define it as a spectrum varying from “vague (‘work on this task’) to specific (‘try for a score of 62 correct on this task within the next 30 minutes’).” Greater goal clarity is theorized to improve performance because resources can be targeted at goals and there is less waste. It is also theorized to have a motivational role because it becomes clear to individuals what they need to do in order to do well (Latham and Locke 1991). Goal clarity and its opposite—goal ambiguity—have been regarded as particularly important in the public sector because the mission statements of government organizations are often vast, and outcomes are hard to measure (Jung 2014). Jung (2012) distinguishes between clarity relating to “target, time limit, and external evaluation.”

Goal clarity is typically measured via self-ratings with questions such as “The work I do is meaningful to me,” “I understand my agency’s mission,” and “I understand how I contribute to my agency’s mission” (see, for example, Hoek, Groeneveld, and Kuipers 2018). Survey research and laboratory studies have suggested that such self-ratings are positively associated with perceived performance (Hoek, Groeneveld, and Kuipers 2018) and performance as measured by quantity and quality of output (Anderson and Stritch 2016).

Rasul, Rogger, and Williams (2021) demonstrate with observational data that expert ratings of *task clarity* (“How precise, specific, and measurable is what the division actually achieved?” and “How precise, specific, and measurable is the target?”) are strongly associated with differences in tasks completed by public sector workers. Importantly, they find that for tasks rated as high in ambiguity, greater control over workers backfires (a reduction of 14 percentage points in completion rates in response to a standard deviation increase in their corresponding measure of management), while giving workers greater autonomy over how they manage their work increases task completion rates (a corresponding increase of 21 percentage points).

In sum, the public administration literature has established the relevance of the latent concepts we focus on in this chapter to key concerns of management: performance, motivation, and turnover. However, most studies have focused on correlations—and, to a lesser extent, causal relationships—without providing an overview of the expected distribution of these variables. It remains unclear to what extent practitioners and scholars should expect substantial variation from these variables, how they are typically distributed across civil servant populations, and to what extent such distributions should be expected to be uniform across employee groups. Answers to these questions are crucial for picking appropriate research designs—including questionnaire design, sampling strategies, and analytic approaches—and for spurring the improvement of existing survey measures.



## SELECTION OF SURVEYS AND MEASURES

Assessing variation requires access to micro-level public service survey data across countries. To maximize microdata coverage, we combine data collected in public service surveys conducted by the Global Survey of Public Servants (GSPS) consortium (which was cofounded by two of the authors) with micro-level public service survey data published by the US federal government. To undertake the required analysis, we also require surveys that can identify the public sector organization (unit) of respondents, can identify the department (subunit) within the organization within which the sampled public administrator works, and that measure the topics most commonly covered in public administration surveys.

This process leads us to focus our analysis on 10 surveys from across Africa (Ethiopia, Ghana, and Liberia), Asia (China and the Philippines), Europe (Romania), North America (the FEVS in the United States), and South America (Chile, Colombia, and Guatemala). All surveys except the FEVS were undertaken by members of the GSPS. They were conducted between 2014 and 2020 and include a mix of online and face-to-face efforts. Each survey featured in this analysis targets core administrative entities—ministries, all nationwide (or federal) agencies, and, where applicable, local governments.

Although we select surveys to maximize comparability, we are not able to measure all concepts consistently across all settings. We therefore focus our analyses on concepts that can be compared across the majority of surveys and have been identified as concepts of interest for the public administration literature, as described above. The resulting set of questions pertains to job satisfaction, pay satisfaction, motivation, assessments of leadership's trustworthiness and tendency to motivate, a measure of performance management related to promotion, and the clarity respondents have over goals and tasks. Comparison with the topics in figure 21.1 indicates that the topics we focus on are key elements of major surveys of public servants.

Table I.1 in appendix I provides further details on the survey questions used from each of the surveys, their original and transformed scales, and the extent of missing observations in the underlying data. Across surveys, question items related to job and pay satisfaction, leadership, performance incentives, and goal and task clarity are nearly identical except for some small adjustments implemented in response to testing in the local context.<sup>8</sup>

All the measures are based on single items. Though this deviates from some common practices, such as the use of Perry scales to measure public service motivation, it reduces the dimensionality of comparison in our setting, where few surveys used similar indexes. Most survey outcomes use a response scale ranging from 1 to 5, where 1 is the most negative and 5 is the most positive response.<sup>2</sup> For surveys where this was not the case, we rescale outcomes to fit on a 1–5 scale. Where a midpoint is missing, scores are split and rounded up to the next full point on a 1–5 scale.

The resulting data set combines multiple surveys of public servants in as coherent a way as possible given the differences in the underlying questions. Given the paucity of published public servant survey data, this provides a relatively unique opportunity to understand the spread of responses to the typical measures contained in such surveys.

To augment the analysis of variance, we consolidate a set of explanatory variables from the surveys that are frequently used for subgroup analysis in reporting on public servant surveys. One of the two most common ways public administration statistics are investigated is by demographic categories. Breaking down statistics by employee demographics can be valid in its function to provide accountability to different interest groups (for example, ethnic minorities and women in the workforce).

It is unclear to what extent demographic characteristics have explanatory value. Parola et al. (2019) find in a meta-analysis that age and gender are significantly related to different levels of public sector motivation. However, the confidence intervals are large and span zero for gender in many specifications. The analysis does not control for other individual and job characteristics, such as time in the job and job type. Cantarelli, Belardinelli, and Bellé (2016) find no significant association between gender and age and leadership assessments. The literature on correlates between demographic variables and other measurements, such as job satisfaction, is largely lacking or based on studies with ad hoc and very small samples. For our analysis, we

**TABLE 21.1 Surveys Used in the Analysis**

Survey country	Year	Unit	N	Subunit	N
Chile	2019	Organization	31	Subunit within the government organization	417
China	2015	City administration	4	Subunit defined by nature of sector and associated task	28
Colombia	2020	Central government/local government	84	Ministry within central government/local government organization	488
Ethiopia	2016	Central government/local government	53	Ministry within central government/sectoral office within local government	198
Ghana	2018	Central government organization	40	Subunit within the government organization	196
Guatemala	2019	Organization	15	Region	176
Liberia	2016	Central government organization	30	Subunit within the government organization	104
Philippines	2014	Central government agency	6	Locality within central agency	18
Romania	2019	Central government/regional government/local government	13	Ministry within central government/sectoral office within regional and local government	54
United States	2019	Agency	30	Level one units (one level below agency)	222

Source: Original table for this publication.

thus refer to the following set of variables as *demographics*: gender, age, tenure in public service, and managerial level. Where demographic characteristics are missing, we impute the median response for continuous and ordinal variables and the mode for categorical variables.<sup>10</sup>

The second type of explanatory variable typically used in studies of public administrations is institutional markers (for example, local or regional governments, organizations, agencies, and teams). Governments naturally want to understand how different government organizations and subunits compare in order to develop targeted strategies to improve performance. Once again, whether institutional divisions are strong predictors of variation in public service surveys is unclear. In their review of studies on motivation and leadership, Hameduddin and Engbers (2022) find no significant differences in the relationship between the two variables by the level of government. Table 21.1 provides details of the hierarchical level we use in each country to approximate organization (unit) and department (subunit).

## WHICH PUBLIC SERVICE SURVEY MEASURES VARY?

Table 21.2 presents descriptive statistics for the surveys we assess, and figure 21.2 presents corresponding standardized distributions of the variables across surveys. In general, pay satisfaction is low while motivation and, to some extent, job satisfaction is high, in line with theories of the public service that see pay satisfaction as a limited component of public sector motivation. Assessments of leadership and meritocratic promotion receive some of the lowest scores across countries.

There is a degree of variation in all measures and in all countries. As shown in table 21.2, standard deviations in the aggregate panel (the means of the statistics in the rest of the table) range between 0.72 and 1.24 on a five-point scale. As a benchmark, if responses are uniformly distributed over a five-point scale, the standard deviation is 1.15.

A number of features stand out. First, there is a distinct negative skew to the variables, with modal responses of 4 or 5. This interpretation is summarized by the motivation scales' highly negative skew (−2.44 in the aggregate panel), indicating that most people report high levels of motivation.<sup>11</sup> Assessments of task and goal clarity and job satisfaction also show considerable—albeit more positive—skew, followed by those

**TABLE 21.2 Descriptive Statistics for Surveys of Public Servants**

Country	Variable	Mean	Median	SD	Skew	Shannon's entropy	N
Aggregate	Job satisfaction	3.88	4.29	0.92	-1.16	1.05	7
	Pay satisfaction	2.8	2.88	1.12	0.21	1.21	8
	Motivation	4.42	4.67	0.72	-2.44	0.85	6
	Leadership trust	3.84	4.2	1.13	-1.01	1.27	5
	Leadership motivates	3.66	4	1.12	-0.88	1.36	5
	Meritocratic promotion	3.54	3.75	1.24	-0.83	1.3	8
	Goal clarity	4.01	4.25	0.89	-1.31	1.13	8
	Task clarity	4.15	4.38	0.8	-1.5	1.02	8
Chile	Job satisfaction	4.16	4	0.87	-1.23	1.15	10,926
	Pay satisfaction	2.82	3	1.23	0.11	1.53	11,082
	Motivation	4.6	5	0.61	-1.99	0.78	10,955
	Leadership trust	3.75	4	1.18	-0.85	1.43	10,605
	Leadership motivates	3.52	4	1.22	-0.57	1.51	10,675
	Meritocratic promotion	2.7	3	1.4	0.24	1.58	9,303
	Goal clarity	4.42	5	0.75	-1.66	0.97	10,973
	Task clarity	4.46	5	0.73	-1.67	0.94	10,978
China	Job satisfaction	3.85	4	0.68	-1	0.96	2,477
	Pay satisfaction	—	—	—	—	—	—
	Motivation	—	—	—	—	—	—
	Leadership trust	—	—	—	—	—	—
	Leadership motivates	—	—	—	—	—	—
	Meritocratic promotion	3.62	4	0.93	-0.64	1.3	2,473
	Goal clarity	—	—	—	—	—	—
	Task clarity	—	—	—	—	—	—
Colombia	Job satisfaction	4.43	5	0.76	-1.76	0.96	9,693
	Pay satisfaction	—	—	—	—	—	—
	Motivation	4.57	5	0.59	-1.7	0.77	9,710
	Leadership trust	—	—	—	—	—	—
	Leadership motivates	—	—	—	—	—	—
	Meritocratic promotion	—	—	—	—	—	—
	Goal clarity	—	—	—	—	—	—
	Task clarity	4.27	4	0.84	-1.51	1.08	17,595
Ethiopia	Job satisfaction	3.04	4	1.31	-0.3	1.2	1,117
	Pay satisfaction	2.12	2	1.17	0.81	1.13	1,125
	Motivation	—	—	—	—	—	—
	Leadership trust	—	—	—	—	—	—
	Leadership motivates	—	—	—	—	—	—
	Meritocratic promotion	2.91	3	1.54	-0.01	1.56	1,121
	Goal clarity	3.13	3	0.85	0.03	1.25	368
	Task clarity	2.93	3	0.81	0.41	1.17	368

*(continues on next page)*

**TABLE 21.2 Descriptive Statistics for Surveys of Public Servants (continued)**

Country	Variable	Mean	Median	SD	Skew	Shannon's entropy	N
Ghana	Job satisfaction	—	—	—	—	—	—
	Pay satisfaction	1.33	1	0.87	2.78	0.65	2,632
	Motivation	4.49	5	0.81	−2.18	0.93	1,103
	Leadership trust	—	—	—	—	—	—
	Leadership motivates	4.25	5	1.09	−1.61	1.15	1,384
	Meritocratic promotion	4.6	5	1.05	−2.73	0.66	1,276
	Goal clarity	4.32	5	0.95	−1.37	1.12	1,503
	Task clarity	4.44	5	0.82	−1.51	1.01	1,510
Guatemala	Job satisfaction	—	—	—	—	—	—
	Pay satisfaction	3.18	4	1.07	−0.3	1.2	1,138
	Motivation	—	—	—	—	—	—
	Leadership trust	3.59	4	1.05	−1.11	1.26	579
	Leadership motivates	3.47	4	1.06	−0.93	1.32	585
	Meritocratic promotion	3.02	3	1.22	−0.18	1.54	574
	Goal clarity	4.08	4	1.01	−0.9	1.27	748
	Task clarity	4.28	5	0.88	−1.01	1.13	747
Liberia	Job satisfaction	3.31	4	1.09	−0.74	0.94	2,651
	Pay satisfaction	2.33	2	1.13	0.61	1.09	2,670
	Motivation	3.33	3	1.31	−0.43	1.55	2,687
	Leadership trust	3.89	5	1.36	−0.77	1.13	839
	Leadership motivates	—	—	—	—	—	—
	Meritocratic promotion	4.39	5	1.03	−1.91	0.94	486
	Goal clarity	3.84	4	1.04	−0.71	1.35	1,407
	Task clarity	3.82	4	1	−0.4	1.34	1,410
Philippines	Job satisfaction	—	—	—	—	—	—
	Pay satisfaction	2.9	3	1.14	−0.08	1.42	1,768
	Motivation	—	—	—	—	—	—
	Leadership trust	—	—	—	—	—	—
	Leadership motivates	—	—	—	—	—	—
	Meritocratic promotion	—	—	—	—	—	—
	Goal clarity	3.78	4	0.93	−1.04	1.2	1,766
	Task clarity	—	—	—	—	—	—
Romania	Job satisfaction	4.62	5	0.65	−2.18	0.79	2,716
	Pay satisfaction	4.09	4	1.17	−1.5	1.2	2,690
	Motivation	4.92	5	0.35	−6.29	0.27	2,726
	Leadership trust	4.01	4	0.88	−1.26	1.16	1,624
	Leadership motivates	3.88	4	0.96	−1.02	1.26	1,667
	Meritocratic promotion	3.82	4	1.5	−0.96	1.33	612
	Goal clarity	4.83	5	0.52	−3.97	0.48	2,707
	Task clarity	4.88	5	0.45	−4.94	0.38	2,723

*(continues on next page)*

**TABLE 21.2 Descriptive Statistics for Surveys of Public Servants (continued)**

Country	Variable	Mean	Median	SD	Skew	Shannon's entropy	N
United States	Job satisfaction	3.75	4	1.07	−0.88	1.36	573,255
	Pay satisfaction	3.59	4	1.15	−0.73	1.43	572,853
	Motivation	4.58	5	0.65	−2.04	0.82	601,274
	Leadership trust	3.96	4	1.18	−1.08	1.36	582,758
	Leadership motivates	3.17	3	1.25	−0.29	1.55	565,650
	Meritocratic promotion	3.25	3	1.23	−0.43	1.52	558,198
	Goal clarity	3.64	4	1.1	−0.84	1.39	569,466
	Task clarity	4.13	4	0.87	−1.38	1.12	598,601

Source: Original table for this publication.

Note: The table shows the mean, median, standard deviation (SD), skew, and Shannon's entropy for each variable in each survey data set we analyze. Skew indicates the extent to which observed values diverge from the balance of observations on each side of the scale characteristic of normal distributions. Shannon's entropy is a measure of variation for categorical variables. It describes the log likelihood of a category's being observed. If the measure equals zero, then all observations are of the same category. If the measure equals one, then the number of observations per category is equal (or near equal). For the aggregate panel, the numbers presented are a simple average of those presented for individual surveys in the rest of the table. — = not measured.

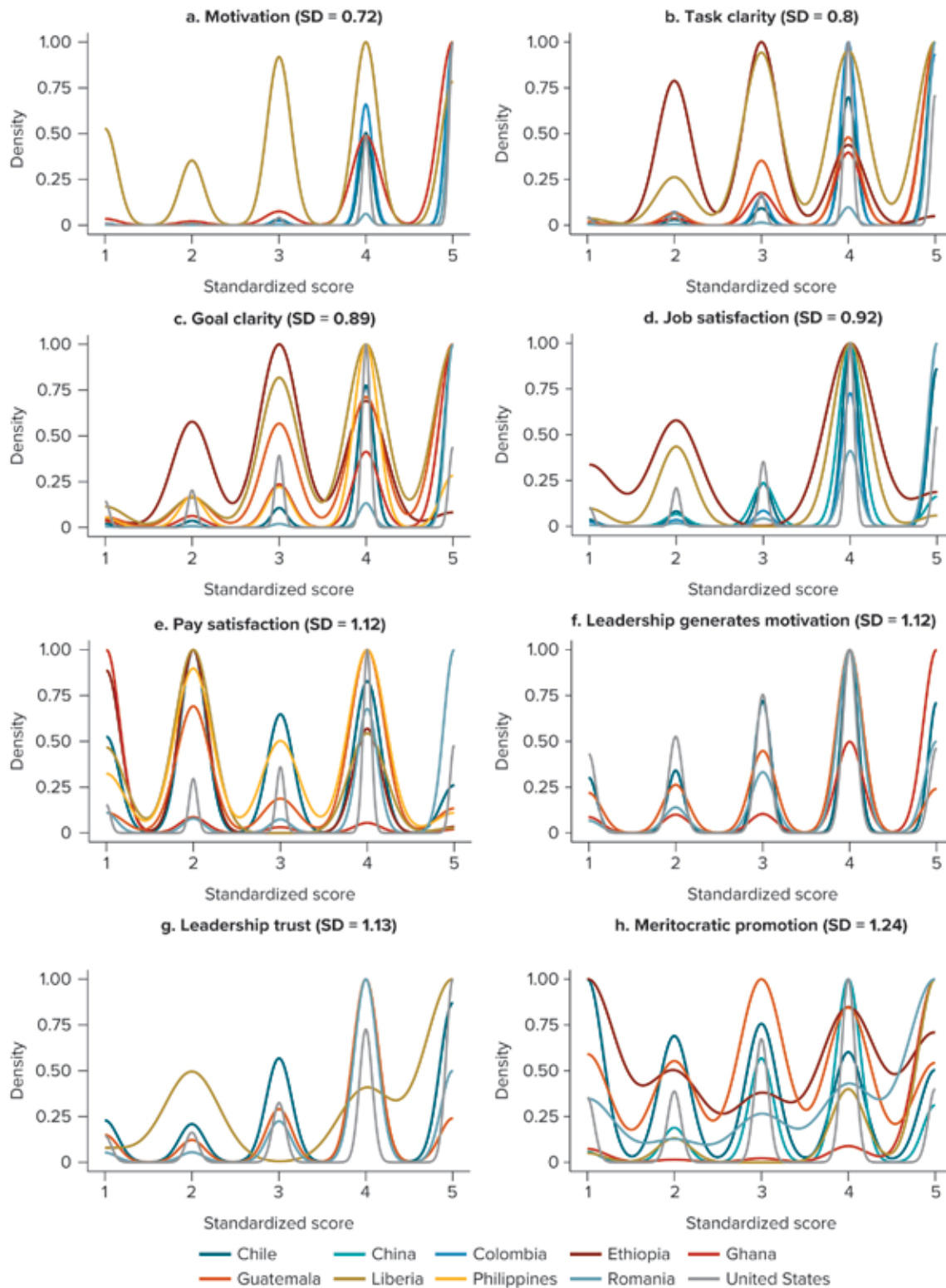
on leadership and promotion. Thus, current practice in measure design compacts a significant proportion of the variation into a minority of the response categories. This limits discriminating variation and the effectiveness of the measures.

Second, we find that responses vary more for measures that assess institutional characteristics, such as leadership, and the extent to which respondents perceive promotion to be linked to performance, relative to questions about individual characteristics. As the aggregate panel of table 21.2 shows, measures related to these topics all have standard deviations above one, while variation in measures of job satisfaction and motivation are lower. As can be seen in figure 21.2, responses about the respondent's characteristics tend to be rated toward the top of the scale for most individuals. This pattern tends to hold in all of the countries we assess.

A different way to see the relative variation in variables related to organizational features is presented in figure 21.3. The figure presents the standard deviation of individual topics across our surveys. The first three topics relate to measures of the self (how motivated a person is), the second group to interactions between the individual and the organization (the extent to which an individual understands the organization's goals), and the third group to perceptions of organizational characteristics (whether leadership is trusted). Figure 21.3 makes clear that across surveys and countries, we see a surprisingly large degree of commonality in which topics vary more than others and in the extent of variation for a single topic. This implies that there are commonalities in the nature of survey responses across settings. Comparing across topic groups, we see that concepts related to a general assessment of the organization exhibit greater variation than those more focused on the self.

Since the data are rather skewed, relying on standard deviation as a summary statistic has its drawbacks. Patterns gleaned from looking at the standard deviation and skew of all measures are reflected in Shannon's entropy index. The index equals zero when all observations assume one value and increases as observations tend to assume different values in equal proportions. Using Shannon's entropy index, the variation in the aggregate panel ranges from 0.85 to 1.36. As a benchmark, if responses are uniformly distributed over a five-point scale, Shannon's entropy index equals one. Similar to values of standard deviations, the highest diversity according to Shannon's entropy index lies in leadership and performance questions, while motivation has the lowest index value (0.85), meaning the least diversity. When looking at measures within countries, rankings of diversity hold up in most cases where comparisons can be made, suggesting that patterns of diversity in responses apply across country contexts.

**FIGURE 21.2** Distributions of Standardized Scores

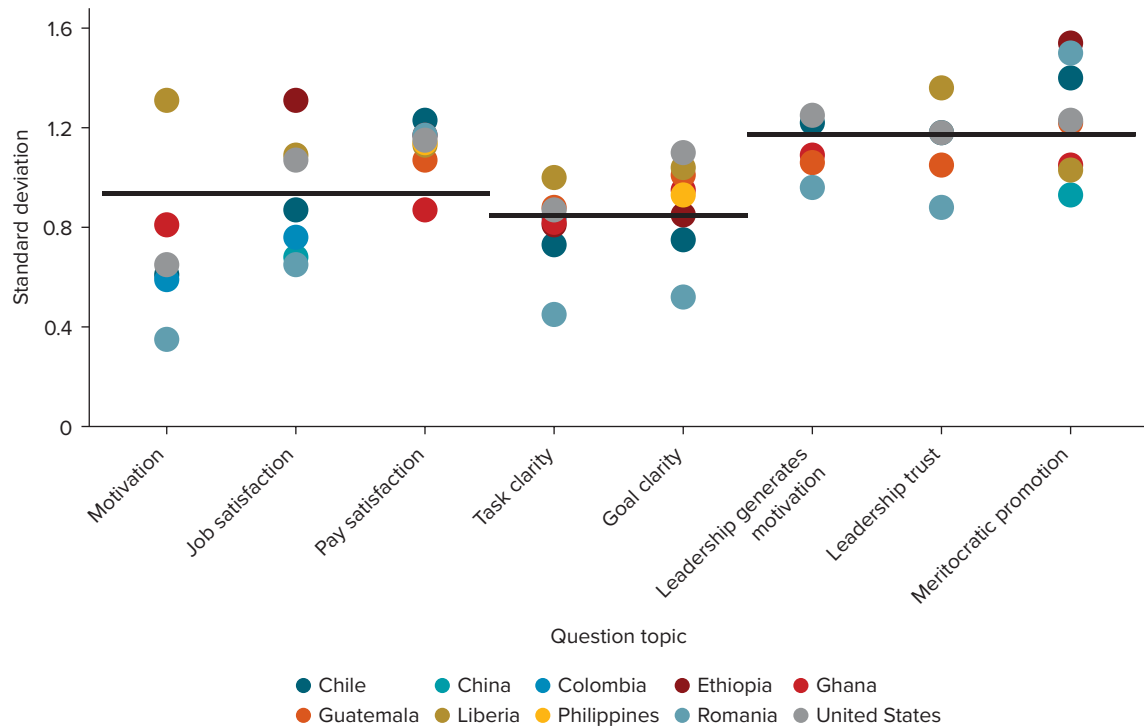


Source: Original figure for this publication.

Note: The panels (corresponding to distinct topics) are ordered in ascending order of the standard deviation (SD) of the measure across surveys. The standard deviation shown above each panel is an average of those in the underlying surveys.



**FIGURE 21.3** Average Response Variance across Surveys and Question Topics



Source: Original figure for this publication.

Note: Horizontal lines illustrate the average standard deviation (y-axis variable) across a given group of questions.

Fourth, the results in table 21.2 suggest that sample size is not a central mediating factor in the extent of variation. The degree of variation in surveys with a thousand or so respondents is not dissimilar to those with tens or hundreds of thousands. One interpretation of this fact is that the underlying distributions across public service entities are relatively stable and are not simply artifacts of measurement error. This could be taken as validation of our approach.

## FOR WHOM DO PUBLIC SERVICE SURVEY MEASURES VARY?

To what extent do the survey measures documented in the last section vary substantively by country, organization, unit within the organization, and demographics? To investigate this question, we fit fixed-effect models to our focal measures and compare the explanatory power of these features.<sup>12</sup> For all surveys, we exclude subunits that have fewer than two respondents (2.2 percent of all respondents with nonmissing values of unit and subunit variables).

The results of a series of analysis of variance (ANOVA) exercises are displayed in table 21.3. Here, the dependent variable is the measure of the topic of interest, and each row of a panel is a distinct regression including variables outlined in the “Models” column. Demographic models include the variables listed above—that is, gender and tenure in public service, as well as age (for all except the United States) and managerial status (for Chile, Colombia, Ghana, Guatemala, and the United States). “Country,” “unit,” and “subunit” indicate the inclusion of fixed effects at the corresponding level, with unit and subunit defined along the lines outlined in table 21.1.

A broad assessment of the measures we include, which are typical factors drawn on in reports on public service surveys, indicates that they all explain a significant portion of the variation we seek to explore. Of the 24 *F*-tests we undertake, all are significant at the 5 percent level.

**TABLE 21.3 Compare Models: ANOVAs, Nested**

Variable	Model	Residual df	RSS	df	Sum of squares	F-stat	Pr	R <sup>2</sup>	Adjusted R <sup>2</sup>
Job satisfaction	Demographics	616,400	694,485					0.01	0.01
	Demographics + country	616,394	685,684	6	8,801.12	1,354	0.00	0.02	0.02
	Demographics + country + unit	616,091	675,248	303	10,436.06	32	0.00	0.04	0.04
	Demographics + country + unit + subunit	614,047	665,135	2,044	10,113.33	4.57	0.00	0.05	0.05
	Nested RE model							0.24	
Pay satisfaction	Demographics	609,406	826,499					0.01	0.01
	Demographics + country	609,399	795,521	7	30,977.73	3,586	0.00	0.05	0.05
	Demographics + country + unit	609,130	782,511	269	13,010.01	39.2	0.00	0.07	0.06
	Demographics + country + unit + subunit	607,351	749,507	1,779	33,004.34	15.03	0.00	0.10	0.10
	Nested RE model							0.40	
Motivation	Demographics	642,380	277,121					0.01	0.01
	Demographics + country	642,375	272,429	5	4,692.82	2,243	0.00	0.02	0.02
	Demographics + country + unit	642,093	270,640	282	1,788.93	15.16	0.00	0.03	0.03
	Demographics + country + unit + subunit	640,062	267,870	2,031	2,769.50	3.26	0.00	0.04	0.04
	Nested RE model							0.44	
Leader: Trust	Demographics	608,171	839,113					0.01	0.01
	Demographics + country	608,167	838,490	4	622.87	115.30	0.00	0.01	0.01
	Demographics + country + unit	608,033	830,463	134	8,027.20	44.36	0.00	0.02	0.02
	Demographics + country + unit + subunit	607,044	819,823	989	10,640.20	7.97	0.00	0.03	0.03
	Nested RE model							0.05	
Leader: Motivation	Demographics	591,842	928,523					0.01	0.01
	Demographics + country	591,838	924,733	4	3,789.80	638.46	0.00	0.01	0.01
	Demographics + country + unit	591,679	89,700	159	27,730.55	117.5	0.00	0.04	0.04
	Demographics + country + unit + subunit	590,411	876,150	1,268	20,852.75	11.08	0.00	0.06	0.06
	Nested RE model							0.14	
Meritocratic promotion	Demographics	585,446	868,379					0.02	0.02
	Demographics + country	585,439	861,128	7	7,250.71	733.26	0.00	0.03	0.03
	Demographics + country + unit	585,171	842,951	268	18,177.21	48.01	0.00	0.05	0.05
	Demographics + country + unit + subunit	583,544	824,321	1,627	18,629.89	8.11	0.00	0.07	0.07
	Nested RE model							0.27	
Goal clarity	Demographics	601,764	723,096					0.01	0.01
	Demographics + country	601,757	712,487	7	10,608.59	1,327	0.00	0.03	0.03
	Demographics + country + unit	601,488	698,168	269	14,319.39	46.60	0.00	0.05	0.04
	Demographics + country + unit + subunit	599,824	685,129	1,664	13,039.43	6.86	0.00	0.06	0.06
	Nested RE model							0.23	
Task clarity	Demographics	649,539	486,185					0.01	0.01
	Demographics + country	649,532	482,436	7	3,748.90	735.4	0.00	0.02	0.02
	Demographics + country + unit	649,155	477,083	377	5,353.09	19.50	0.00	0.03	0.03
	Demographics + country + unit + subunit	646,343	470,732	2,812	6,351.41	3.10	0.00	0.04	0.04
	Nested RE model							0.34	

Source: Original table for this publication.

Note: The first four lines for each variable summarize test statistics for analyses of variance and how the model fit compares to the next more complex model. The first row refers to a model that only includes demographic predictor variables. These include the respondent's gender and tenure in public service, as well as age (present in all surveys except for the United States) and managerial status (for Chile, Colombia, Ghana, Guatemala, and the United States). Individual missing values for age and tenure are imputed using the median and mean values, respectively. Missing values for the gender and managerial status variables are assigned to the "missing" category. Rows two through four progressively add country, unit, and subunit level dummies to the model. The *F*-test for each model indicates whether it has a better fit than the simpler model specified above. Models with lower residual sums of squares (RSS) and a higher (adjusted) *R*-squared explain a larger proportion of the variance. The last, fifth, line for each variable reports the model fit for a nested model that nests subunits into units and units into countries. If the *R*-squared of the nested model is larger than the values in the lines above it, the nested model is a better fit. ANOVAs = analyses of variance; df = degrees of freedom; Pr = probability associated with the *F*-statistic; RE = residual error.

Our analysis begins with an assessment of the extent to which basic demographic characteristics of respondents are predictive of their answers. Demographics explain between 0 and 2 percent of the variation across measures, with no clear pattern across different measures. Of the demographic variables, managerial position tends to explain the largest portion of variation, followed by age, gender, and tenure. Thus, public service measures vary most for managers compared with nonmanagers in the data sets we study.

Our ANOVA results suggest that the determinants of variation we observe are mediated by the nature of the variable. Country effects are significant throughout the analysis, but these may pick up both national commonalities in responses as well as differences in survey wording, enumeration, and so on. They are particularly important for respondents' assessments of their own characteristics, such as motivation, job satisfaction, and pay satisfaction. Thus, though more intimate features of self-identity vary the least, they are the most likely to be predicted by demographic features or national boundaries.

In rows three and four of each panel in table 21.3, we add measures of institutional structures indicating the unit and subunit the respondent works in. Focusing on the sum of squares each set of variables explains, we see that relative to country fixed effects, the institutional features explain a small proportion of the variance in job satisfaction, pay satisfaction, and motivation, in comparison to their much more significant role in assessments of leadership and organizational features (such as the extent to which promotions are generally meritocratic and how individual respondents understand organizational goals and tasks and their relationships to them). Institutional variables therefore appear to have more predictive power for those variables more closely aligned to hierarchy.

Intuitively, institutional structures are more predictive of those features of public service life generated by those structures. This implies that elements of public service defined most fully within the individual respondent, such as motivation, are in fact relatively stable across institutional settings. The core motivation of public servants seems relatively robust to their office, while perceptions of the quality of leadership are highly dependent on the unit and subunit in which an official works.

We perform a series of robustness exercises. Since three countries use different scales for three measures, we perform a robustness check whereby we rerun the main models excluding these countries. The results are presented in table I.2 in appendix I. We also rerun all analyses on data without imputation, using a listwise deletion instead. The results are presented in table I.3 in appendix I. Regression diagnostics indicate that none of the variables of interest has normally distributed error terms (see table I.4 in appendix I). Therefore, we rerun all models with the outcome variables transformed using Box-Cox transformations (see figure I.1 and table I.3 in appendix I for details). The robustness checks broadly support our core results.

Finally, we also fit mixed models in row five of each panel of table 21.3. Fixed-effect models do not account for the nested structure of data—public administrators who are located within subunits are nested in units that belong to organizations.<sup>13</sup> The mixed models have fixed effects for demographics, country, and unit and random effects for subunits nested within these. We do not fit random slopes as our main set of predictor variables is categorical and we have no clear hypotheses of interactions between predictor variables. Taking into account the nested structure of the explanatory variables does not significantly alter the interpretation.

## DISCUSSION

There is little systematic evidence available on variation in the measurements typically used to assess the nature of public administration. In this chapter, we have provided descriptive statistics for, and assessed variation in, a range of the most common indicators of public administration. We have done so based on a unique data set of public service surveys conducted in 10 countries in Africa, Asia, Europe, and North and South America. The statistics presented in table 21.2 provide benchmarks for other analysts to use in assessing variation in their own surveys of public servants. They answer the question “Which public service survey measures vary?” The analysis in table 21.3 provides evidence of which features of public

administration are predictive of these measures, and thus answers “For whom do public service survey measures vary?”

Our results point to less variation in measures related to personal characteristics, such as motivation, than in institutional variables, such as assessments of leadership. Personal characteristics are predicted more strongly by demographics and country fixed effects than institutional features, which are more strongly predicted by the units and subunits in which respondents work. The most substantial variation in surveys of public servants is in organizational characteristics, and these are determined by the office a respondent works in.

Our findings may reflect both the design of questions common in public servant surveys as well as a skew in the latent features of public service on which we have focused. For example, it may be that motivation is very high across all the public service entities we study, and our measures accurately reflect this. However, the negative skew we observe may be an indication that survey questions could be better designed and analyzed to explore the variation at the top of affected measures. Given the ambition of this chapter to inform the design of public servant surveys, we conclude with a discussion of avenues for responding to this finding.

## Developing More-Discriminatory Measures

### Validity of Scales and Skew

The first question raised by the compressed variance and extreme skew in most of these measures is whether these are artifacts of the survey measures employed, or whether they capture the realities of public administrations. As the introduction of this chapter summarized, for some measures, particularly motivation and leadership scales, an extant body of research on their validation reinforces our findings as reflections of reality. However, this does not preclude the possibility that current measures do not adequately capture distributions of concepts in real populations.

Observed patterns of skew could be driven by several factors related to measurement: social-desirability bias (see Kim and Kim [2016] for a discussion related to public service motivation), cognitive biases related to the choice of reference category, and extreme response bias (Tourangeau 2003). Public administrators may feel pressured to indicate high levels of motivation, for instance, in case their responses are ever disclosed (even if such disclosure never occurs in practice). Alternatively, there may be no desirability bias at play, but skew and kurtosis may simply be driven by cognitive biases. For instance, the *medium fallacy* is a common psychological bias that makes people believe they are better than the average person (which, statistically, cannot be true for everyone). Extreme response bias may also explain some of the observed patterns. It has been shown that some individuals have a greater tendency to pick extreme points on scales than others (Hibbing et al. 2019). One approach to these concerns is to tweak questions so that their scales have a greater range of options to discriminate between higher values of response. Another is to provide anchors to which respondents can relate their experiences.

### Analysis Strategies and Skew

If measures are valid, the second concern raised by our observations of extreme skew in the data pertains to analysis methods. How can an analyst approach highly skewed data? There are several strategies that can be pursued to help address them.

The first is to include other questions in surveys that allow analysts to quantify the potential drivers of skew. For example, surveys could include social-desirability scales, which could then be used in regression analysis to (partially) control for bias introduced via this avenue.

A second strategy is to reweight data points by using transformations such as the log or Box-Cox transformations, as used in this chapter. Such an approach can “smooth out” the distribution of a skewed variable, conditional on a reinterpretation of the corresponding results.

Another strategy is to approach skewed responses differently than other points in the data. Several sophisticated strategies have also been developed to deal with extreme response bias. For example, item response tree (IRTree) models adjust for extreme responses by modeling a two-stage decision-making process. The multidimensional nominal response model (MNRM) recodes extreme responses as a separate dimension and includes them as dummies in regressions. Partial credit models use random effects to control for biases introduced by extreme responses (see Falk and Ju [2020] for a recent evaluation of their comparative performance).<sup>14</sup>

As this chapter has illustrated, there is a danger that the error terms of skewed variables are not normally distributed (and are potentially also heteroskedastic). Analysts can employ regression diagnostics, as used in table I.4 in appendix I, to assess the nature of their data more thoroughly. In response to the nonnormality of measures, they might consider employing bootstrapping methods in their analysis (see Afifi et al. 2007).

### **Building the Evidence Base Further**

The specific culture of public service will determine the challenges to survey measurement that analysts will face. Though international comparisons are useful, particularly given the commonalities we have observed across surveys in this chapter, generating evidence on survey design is best done at the survey level. The analysis undertaken in this chapter could be repeated for multiple rounds of the same survey or for distinct departments or geographical regions covered by a survey. Such work builds a picture of which measures of public administration provide discriminating variation and which do not.

It has been difficult to assess the predictive validity of measures standard in public servant surveys. Assessments of discriminant validity are more common, but they could be expanded to address the theoretical overlap and imprecision of many concepts utilized in public administration research (see chapter 24 on discriminant validity for a recent evaluation). One key problem is that the vast majority of research in public administration, and the validation relating to the measurement used, is reliant on surveys (see Strauss and Smith [2009] for a discussion of developments in the philosophy of science on construct validity). Using the same methodology to test a measure can severely inflate its construct validity. Future research thus faces a pressing need to link survey and self-reported data to other ways of measuring the same concepts, such as administrative and behavioral data (for example, turnover, sick leave, performance ratings, output efficiency, and career progression). None of these measures is superior on its own to survey measurement. However, using Campbell and Fiske's (1959) multitrait-multimethod matrix methodology, the robustness of validity assessments of key concepts in public administration research can be improved: if the measured concepts are universal, they should manifest in different contexts and be detectable with a variety of methods. Their quantities should not change substantially as a function of method. Where adequate quantitative data are missing, qualitative methods could help to assess the validity of survey measures (see chapter 4 for a discussion of the problems with monolithic approaches to methodology).

Where experimentation is feasible, analysts may build evidence as to what is driving the (skewed) variation in responses. Cognitive biases could be addressed by using randomized controlled trials to systematically evaluate which features of a survey might cause greater skew in response. By combining this evidence with objective measures, where available, analysts can answer the questions posed in the title of this chapter with increasingly granular detail for their survey(s) of interest.

## **NOTES**

1. Given that so many features of public administration may vary across units of observation, and the challenge of measuring these features, the use of surveys seems a natural response. An alternative approach would be to use administrative data to measure variation—for example, by using the extent to which officials leave a department (turnover data) as a measure of

satisfaction. But such a measure is very crude, only measuring satisfaction once it is at its lowest level and an official leaves the department, and has a range of other issues. Survey variation helps us understand the extent to which respondents perceive or experience things differently or similarly across the full distribution of values by asking the party of interest directly.

2. A survey measure is valid when it appears to measure the concept of interest (*face validity*) and covers relevant dimensions of the concept of interest (*content validity*), as well as to the extent that the measurement correlates with those that theoretically should be correlated (*criterion validity*) and captures variation not already captured by other variables (*discriminant validity*) (see chapter 24).
3. This reliance is very much based in the difficulty of accessing alternative data sources and the latent nature of most concepts of interest.
4. Guajardo (1996) looks into variation in demographic variables used as a proxy for diversity and representation in the public sector but restricts attention to studies with this common source bias.
5. This is not to say, of course, that there is not significant scholarship on these organizational psychology concepts in the public service (Esteve and Schuster 2019). Dozens of studies have, for instance, focused on leadership in the public sector. These organizational psychology concepts have, however, not been aggregated into a separate model of civil service governance, akin to Weberianism or new public management.
6. Measures of pay satisfaction are intimately linked to job satisfaction. They are commonly measured as a part of job satisfaction or separately, via a single item.
7. As evidenced by private correspondence between the authors and the Australian, Canadian, Irish, UK, and US governments, which forms the basis of the more in-depth case studies of measurement featured in chapters 25 and 26.
8. For example, in Ethiopia, the question on pay satisfaction was phrased “To what extent would you say you are satisfied with your salary?”; in Liberia, it was “How satisfied are you with your total income?”; and in Ghana, it was “My salary is very satisfactory.”
9. The exception is work motivation, which is measured the same for all but three surveys. In Ethiopia, Liberia, and the Philippines, civil servants were asked to compare their motivation today to when they started. In Ethiopia and the Philippines, they were provided with an answer scale ranging from 0 to 100, while in Ethiopia, a 0–10 scale was used. In Ghana, civil servants were asked to rate the extent to which they “would feel an obligation to take time from my personal schedule to generate ideas/solutions for the organization if it is needed.”
10. All surveys have data available on the respondent’s gender and tenure in public service. The age variable is missing for the United States, and the managerial level is missing for China, Ethiopia, Liberia, the Philippines, and Romania.
11. Regression diagnostics indicate that none of the variables of interest has normally distributed error terms (see appendix I).
12. Our core approach uses *F*-tests to test for statistical significance, but we also run Wald tests using robust standard errors, and results do not differ.
13. Note that the unit-inside-organization classifier is not homogeneous across countries. For instance, in some cases, units are ministries, while in others, they are local governments, while subunits may refer to teams inside ministries or regional offices of ministries, for instance.
14. All such models can easily be implemented in standard statistical software; for example, in R, using packages such as *mirt* and *eRm*.

## REFERENCES

- Afifi, Abdelmonem A., Jenny B. Kotlerman, Susan L. Ettner, and Marie Cowan. 2007. “Methods for Improving Regression Analysis for Skewed Continuous or Counted Responses.” *Annual Review of Public Health* 28: 95–111. <https://doi.org/10.1146/annurev.publhealth.28.082206.094100>.
- Anderson, Derrick M., and Justin M. Stritch. 2016. “Goal Clarity, Task Significance, and Performance: Evidence from a Laboratory Experiment.” *Journal of Public Administration Research and Theory* 26 (2): 211–25. <https://doi.org/10.1093/jopart/muv019>.
- Ashraf, Nava, Oriana Bandiera, Edward Davenport, and Scott Lee. 2020. “Losing Prosociality in the Quest for Talent? Sorting, Selection, and Productivity in the Delivery of Public Services.” *American Economic Review* 110 (5): 1355–94.
- Bellé, Nicola. 2014. “Leading to Make a Difference: A Field Experiment on the Performance Effects of Transformational Leadership, Perceived Social Impact, and Public Service Motivation.” *Journal of Public Administration Research and Theory* 24 (1): 109–36. <https://doi.org/10.1093/jopart/mut033>.
- Bellé, Nicola. 2015. “Performance-Related Pay and the Crowding Out of Motivation in the Public Sector: A Randomized Field Experiment.” *Public Administration Review* 75 (2): 230–41. <https://doi.org/10.1111/puar.12313>.



- Bellé, Nicola, and Paola Cantarelli. 2015. "Monetary Incentives, Motivation, and Job Effort in the Public Sector: An Experimental Study with Italian Government Executives." *Review of Public Personnel Administration* 35 (2): 99–123. <https://doi.org/10.1177/0734371X13520460>.
- Bouckaert, Geert. 2021. "Public Performance: Some Reflections and Lessons Learned." In *The Public Productivity and Performance Handbook*, 3rd ed., edited by Marc Holzer and Andrew Ballard, 68–73. New York: Routledge.
- Brandler, Sondra, Camille P. Roman, Gerald J. Miller, and Kaifeng Yang. 2007. *Handbook of Research Methods in Public Administration*. Boca Raton, FL: CRC Press.
- Camilleri, Emanuel, and Beatrice I. J. M. Van Der Heijden. 2007. "Organizational Commitment, Public Service Motivation, and Performance within the Public Sector." *Public Performance and Management Review* 31 (2): 241–74. <https://doi.org/10.2753/PMR1530-9576310205>.
- Campbell, Donald T., and Donald W. Fiske. 1959. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin* 56 (2). <https://doi.org/10.1037/h0046016>.
- Cantarelli, Paola, Paolo Belardinelli, and Nicola Bellé. 2016. "A Meta-analysis of Job Satisfaction Correlates in the Public Administration Literature." *Review of Public Personnel Administration* 36 (2): 115–44. <https://doi.org/10.1177/0734371X15578534>.
- Cramer, Duncan. 1996. "Job Satisfaction and Organizational Continuance Commitment: A Two-Wave Panel Study." *Journal of Organizational Behavior* 17 (4): 389–400. <https://www.jstor.org/stable/2488549>.
- Esteve, Marc, and Christian Schuster. 2019. *Motivating Public Employees*. Elements in Public and Nonprofit Administration. Cambridge, UK: Cambridge University Press. <https://doi.org/10.1017/9781108559720>.
- Falk, Carl F., and Unhee Ju. 2020. "Estimation of Response Styles Using the Multidimensional Nominal Response Model: A Tutorial and Comparison with Sum Scores." *Frontiers in Psychology* 11: 72. <https://doi.org/10.3389/fpsyg.2020.00072>.
- Faragher, E. Brian, Monica Cass, and Cary L. Cooper. 2013. "The Relationship between Job Satisfaction and Health: A Meta-analysis." In *From Stress to Wellbeing: The Theory and Research on Occupational Stress and Wellbeing*, edited by Cary L. Cooper, 254–71. London: Palgrave.
- Favero, Nathan, and Justin B. Bullock. 2015. "How (Not) to Solve the Problem: An Evaluation of Scholarly Responses to Common Source Bias." *Journal of Public Administration Research and Theory* 25 (1): 285–308. <https://doi.org/10.1093/jopart/muu020>.
- Fink, Arlene, and Mark S. Litwin. 1995. *How to Measure Survey Reliability and Validity*. Thousand Oaks, CA: SAGE Publications. <https://doi.org/10.4135/9781483348957>.
- George, Bert, and Sanjay K. Pandey. 2017. "We Know the Yin—But Where Is the Yang? Toward a Balanced Approach on Common Source Bias in Public Administration Scholarship." *Review of Public Personnel Administration* 37 (2): 245–70. <https://doi.org/10.1177/0734371X17698189>.
- Grosh, Margaret, and Paul Glewwe, eds. 2000. *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study*. 3 vols. Washington, DC: World Bank. <https://www.worldbank.org/en/programs/lsm/publication/designing-household-survey-questionnaires-for-developing-countries>.
- Guajardo, Salomon A. 1996. "Representative Bureaucracy: An Estimation of the Reliability and Validity of the Nachmias-Rosenbloom MV Index." *Public Administration Review* 56 (5): 467–77. <https://doi.org/10.2307/977046>.
- Hameduddin, Taha, and Trent Engbers. 2022. "Leadership and Public Service Motivation: A Systematic Synthesis." *International Public Management Journal* 25 (1): 86–119. <https://doi.org/10.1080/10967494.2021.1884150>.
- Heinrich, Carolyn J. 2002. "Outcomes-Based Performance Management in the Public Sector: Implications for Government Accountability and Effectiveness." *Public Administration Review* 62 (6): 712–25. <https://doi.org/10.1111/1540-6210.00253>.
- Hibbing, Matthew V., Matthew Cawvey, Raman Deol, Andrew J. Bloeser, and Jeffery J. Mondak. 2019. "The Relationship between Personality and Response Patterns on Public Opinion Surveys: The Big Five, Extreme Response Style, and Acquiescence Response Style." *International Journal of Public Opinion Research* 31 (1): 161–77. <https://doi.org/10.1093/ijpor/edx005>.
- Hoek, Marieke van der, Sandra Groeneveld, and Ben Kuipers. 2018. "Goal Setting in Teams: Goal Clarity and Team Performance in the Public Sector." *Review of Public Personnel Administration* 38 (4): 472–93. <https://doi.org/10.1177/0734371X16682815>.
- Jung, Chan Su. 2012. "Developing and Validating New Concepts and Measures of Program Goal Ambiguity in the US Federal Government." *Administration and Society* 44 (6): 675–701. <https://doi.org/10.1177/0095399711413730>.
- Jung, Chan Su. 2014. "Why Are Goals Important in the Public Sector? Exploring the Benefits of Goal Clarity for Reducing Turnover Intention." *Journal of Public Administration Research and Theory* 24 (1): 209–34. <https://doi.org/10.1093/jopart/mus058>.
- Kim, Sangmook. 2009. "Revising Perry's Measurement Scale of Public Service Motivation." *American Review of Public Administration* 39 (2): 149–63. <https://doi.org/10.1177/0275074008317681>.
- Kim, Seung Hyun, and Sangmook Kim. 2016. "Social Desirability Bias in Measuring Public Service Motivation." *International Public Management Journal* 19 (3): 293–319. <https://doi.org/10.1080/10967494.2015.1021497>.

- Kroll, Alexander, and Dominik Vogel. 2014. "The PSM-Leadership Fit: A Model of Performance Information Use." *Public Administration* 92 (4): 974–91. <https://doi.org/10.1111/padm.12014>.
- Latham, Gary P., and Edwin A. Locke. 1991. "Self-Regulation through Goal Setting." *Organizational Behavior and Human Decision Processes* 50 (2): 212–47. [https://doi.org/10.1016/0749-5978\(91\)90021-K](https://doi.org/10.1016/0749-5978(91)90021-K).
- Mehra, Kavita, and Kirti Joshi. 2010. "The Enabling Role of the Public Sector in Innovation: A Case Study of Drug Development in India." *Innovation* 12 (2): 227–37. <https://doi.org/10.5172/impp.12.2.227>.
- Meier, Kenneth J., and Laurence J. O'Toole. 2010. "Organizational Performance: Measurement Theory and an Application: Or, Common Source Bias, the Achilles Heel of Public Management Research." Paper presented at the Annual Meeting of the American Political Science Association, September 1–5, 2010, Washington, DC.
- Meyer-Sahling, Jan, Dinsha Mistree, Kim Mikkelsen, Katherine Bersch, Francis Fukuyama, Kerenssa Kay, Christian Schuster, Zahid Hasnain, and Daniel Rogger. 2021. *The Global Survey of Public Servants: Approach and Conceptual Framework*. Global Survey of Public Servants. Last updated May 2021. <https://www.globalsurveyofpublicservants.org/about>.
- Mikkelsen, Kim Sass, Christian Schuster, and Jan-Hinrik Meyer-Sahling. 2021. "A Cross-Cultural Basis for Public Service? Public Service Motivation Measurement Invariance in an Original Survey of 23,000 Public Servants in Ten Countries and Four World Regions." *International Public Management Journal* 24 (6): 739–61. <https://doi.org/10.1080/10967494.2020.1809580>.
- Moynihan, Donald P., and Sanjay K. Pandey. 2007. "The Role of Organizations in Fostering Public Service Motivation." *Public Administration Review* 67 (1): 40–53. <https://doi.org/10.1111/j.1540-6210.2006.00695.x>.
- Naff, Katherine C., and John Crum. 1999. "Working for America: Does Public Service Motivation Make a Difference?" *Review of Public Personnel Administration* 19 (4): 5–16. <https://doi.org/10.1177/0734371X9901900402>.
- Pandey, Sanjay K., Randall S. Davis, Sheela Pandey, and Shuyang Peng. 2016. "Transformational Leadership and the Use of Normative Public Values: Can Employees Be Inspired to Serve Larger Public Purposes?" *Public Administration* 94 (1): 204–22. <https://doi.org/10.1111/padm.12214>.
- Parola, Heather R., Michael B. Harari, David E. L. Herst, and Palina Prysmakova. 2019. "Demographic Determinants of Public Service Motivation: A Meta-analysis of PSM-Age and -Gender Relationships." *Public Management Review* 21 (10): 1397–1419. <https://doi.org/10.1080/14719037.2018.1550108>.
- Perry, James L. 1996. "Measuring Public Service Motivation: An Assessment of Construct Reliability and Validity." *Journal of Public Administration Research and Theory* 6 (1): 5–22. <https://doi.org/10.1093/oxfordjournals.jpart.a024303>.
- Potts, Jason, and Tim Kastle. 2010. "Public Sector Innovation Research: What's Next?" *Innovation* 12 (2): 122–37. <https://doi.org/10.5172/impp.12.2.122>.
- Rasul, Imran, Daniel Rogger, and Martin J. Williams. 2021. "Management, Organizational Performance, and Task Clarity: Evidence from Ghana's Civil Service." *Journal of Public Administration Research and Theory* 31 (2): 259–77. <https://doi.org/10.1093/jopart/muaa034>.
- Scanlan, Justin Newton, and Megan Still. 2019. "Relationships between Burnout, Turnover Intention, Job Satisfaction, Job Demands and Job Resources for Mental Health Personnel in an Australian Mental Health Service." *BMC Health Services Research* 19 (1): 1–11. <https://doi.org/10.1186/s12913-018-3841-z>.
- Somani, Ravi. 2021. *Public-Sector Productivity (Part 1): Why Is It Important and How Can We Measure It?* Equitable Growth, Finance and Institutions Insight. Washington, DC: World Bank. <http://hdl.handle.net/10986/35165>.
- Sousa-Poza, Alfonso, and Andrés A. Sousa-Poza. 2007. "The Effect of Job Satisfaction on Labor Turnover by Gender: An Analysis for Switzerland." *Journal of Socio-Economics* 36 (6): 895–913. <https://doi.org/10.1016/j.socsec.2007.01.022>.
- Strauss, Milton E., and Gregory T. Smith. 2009. "Construct Validity: Advances in Theory and Methodology." *Annual Review of Clinical Psychology* 5: 1–25. <https://doi.org/10.1146/annurev.clinpsy.032408.153639>.
- Tourangeau, Roger. 2003. "Cognitive Aspects of Survey Measurement and Mismeasurement." *International Journal of Public Opinion Research* 15 (1): 3–7. <https://doi.org/10.1093/ijpor/15.1.3>.
- Tummers, Lars, and Eva Knies. 2016. "Measuring Public Leadership: Developing Scales for Four Key Public Leadership Roles." *Public Administration* 94 (2): 433–51. <https://doi.org/10.1111/padm.12224>.
- van Engen, Nadine A. M. 2017. "A Short Measure of General Policy Alienation: Scale Development Using a 10-Step Procedure." *Public Administration* 95 (2): 512–26. <https://doi.org/10.1111/padm.12318>.
- Vigoda-Gadot, Eran. 2007. "Leadership Style, Organizational Politics, and Employees' Performance: An Empirical Examination of Two Competing Models." *Personnel Review* 36 (5): 661–83. <https://doi.org/10.1108/00483480710773981>.
- Vogel, Dominik, and Alexander Kroll. 2019. "Agreeing to Disagree? Explaining Self–Other Disagreement on Leadership Behaviour." *Public Management Review* 21 (12): 1867–92. <https://doi.org/10.1080/14719037.2019.1577910>.
- Vogel, Dominik, Artur Reuber, and Rick Vogel. 2020. "Developing a Short Scale to Assess Public Leadership." *Public Administration* 98 (4): 958–73. <https://doi.org/10.1111/padm.12665>.
- Wright, James D., and Peter V. Marsden. 2010. "Survey Research and Social Science: History, Current Practice, and Future Prospects." In *Handbook of Survey Research*, 2nd ed., edited by Peter V. Marsden and James D. Wright, 3–26. Bingley, UK: Emerald.



## CHAPTER 22

# Designing Survey Questionnaires

## To What Types of Survey Questions Do Public Servants Not Respond?

*Robert Lipinski, Daniel Rogger, and Christian Schuster*

### SUMMARY

Surveys of public servants differ sharply in the extent of item nonresponse: respondents' skipping or refusing to respond to questions. Item nonresponse can affect the legitimacy and quality of public servant survey data. Survey results may be biased, for instance, if those least satisfied with their jobs are also most prone to skipping survey questions. Understanding why public servants respond to some survey questions but not others is thus important. This chapter offers a conceptual framework and empirical evidence to further this understanding. Drawing on the existing literature on survey nonresponse, the chapter theorizes that public servants are less likely to respond to questions that are complex (because they are unable to) or sensitive (because they are unwilling to). This argument is assessed using a newly developed coding framework for survey question complexity and sensitivity, which is applied to public service surveys in Guatemala, Romania, and the United States. The results imply that one indicator of complexity—the unfamiliarity of respondents with the subject question—to be the most robust predictor of item nonresponse across countries. By contrast, other indicators in the framework or machine-coded algorithms of textual complexity do not predict item nonresponse. The findings point to the importance of avoiding questions that require public servants to speculate about topics with which they are less familiar.

---

Robert Lipinski is a consultant and Daniel Rogger is a senior economist in the World Bank's Development Impact Evaluation (DIME) Department. Christian Schuster is a professor at University College London.

## ANALYTICS IN PRACTICE

- Surveys of public servants typically rely on voluntary responses from public servants. For this reason, they may suffer not only from unit nonresponse—that is, public servants’ not responding to surveys at all—but also item nonresponse—that is, public servants’ not responding to particular survey questions.
- Assessments of three public servant surveys spanning three continents imply that item nonresponse is a significant concern in the public sector. In some survey modules, nonresponse can be as high as 30 percent.
- Public servants are typically more educated than the average survey respondent, and their daily duties are closely aligned with the task of filling in a questionnaire. As such, the determinants of nonresponse in surveys of public servants may be distinct from those identified in the existing literature.
- This chapter presents a coding framework that allows survey analysts to measure the complexity and sensitivity of different questions in a public service questionnaire. Such assessments provide an important exercise in assessing survey quality.
- The analysis finds one indicator of complexity—the unfamiliarity of respondents with the subject question—to be the most robust predictor of item nonresponse across countries. Surveys of public servants should carefully consider the need for questions that require public servants to speculate about topics they are less familiar with, as they are associated with greater item nonresponse.
- In contrast, no other margin of complexity or sensitivity is a particularly acute source of nonresponse. At least in terms of missing data, the current analysis implies that public officials can handle many aspects of complex and sensitive topics.
- The manual coding approach is compared to common machine-coded assessments of complexity and find that a manually coded assessment of unfamiliarity outperforms machine-coded variables.

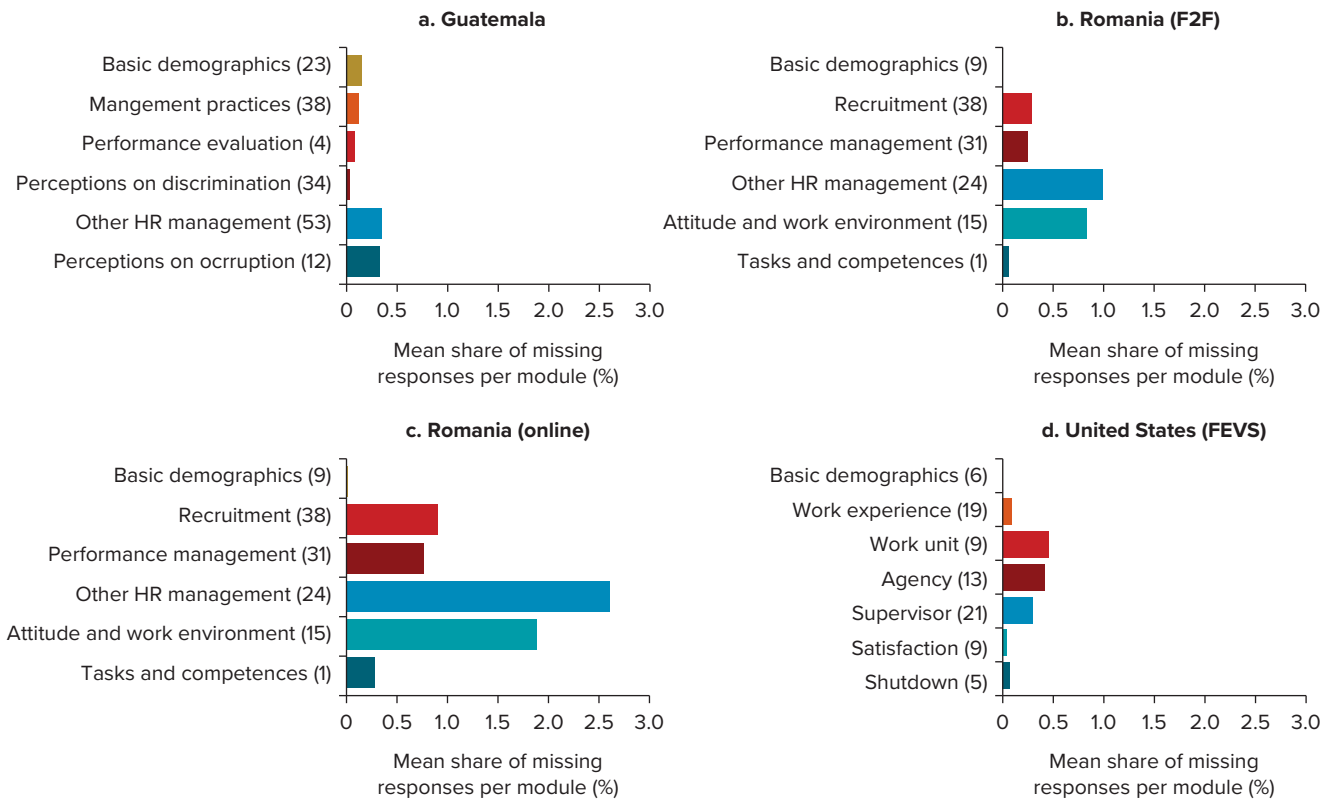
## INTRODUCTION

Surveys of public servants typically rely on voluntary responses from public servants. For this reason, they may suffer not only from unit nonresponse—that is, public servants’ not responding to surveys at all or dropping out of the survey (see chapter 19)—but also item nonresponse: public servants’ not responding to particular survey questions. They may, for instance, skip survey questions in online surveys, refuse to answer questions in face-to-face surveys, or simply indicate “I don’t know” in response to questions.

Item nonresponse is a challenge for both the quality and legitimacy of public service survey data. Item nonresponse may undermine the quality of public service survey data because having fewer responses enhances the variance of items. From a legitimacy perspective, high item nonresponse undermines potential uses of the data, as skeptics can critique the inferences drawn from items with high nonresponse as not representative of the survey population. If nonrespondents differ in a systematic way from respondents, questions can produce biased point estimates (Haziza and Kuromi 2007). This is not inconceivable: survey results may be biased, for instance, if those least satisfied with their jobs or those with reason to hide their behavior are also most prone to skipping survey questions.

Understanding what types of questions public servants tend to respond to and what types of questions prompt item nonresponse is thus important for survey designers. It provides a basis for designing questions that reduce item nonresponse and thus for enhancing public service survey quality and legitimacy. This is

**FIGURE 22.1** Share of Missing Responses, by Survey Module



Source: Original figure for this publication.

Note: The labels on the y axis in each graph contain numbers in parentheses indicating the number of questions in each module. FEVS = Federal Employee Viewpoint Survey; F2F = face-to-face; HR = human resources.

important not least because surveying public servants about certain topics—such as satisfaction, motivation, or assessments of leadership—is often the only means to obtain data on these topics. Given the absence of other data sources to measure them, improved questionnaire design is the only alternative for valid data collection.

To date, public service surveys have varied in the extent to which their questions yield nonresponse. As illustrated in figure 22.1—which draws on data from public service surveys in Guatemala, Romania, and the United States (and which will be used throughout this chapter)—item nonresponse varies across survey modules from almost 0 percent to almost 30 percent in some settings (and up to 60 percent for certain individual questions). These figures imply that for certain topics, nonresponse is a substantive concern in public service surveys. The variation observed across questions also implies that question characteristics determine the likelihood that a question will be answered.

Why do public servants respond to some survey questions but not to others? This chapter offers a conceptual framework and empirical evidence to better understand this question. Conceptually, we build on the survey methodology literature, which has broadly argued for two causes of item nonresponse: question complexity and question sensitivity. Question complexity leads to item nonresponse when respondents are unable to answer a question, even if they are willing. This is due to an excessive cognitive burden on one or more steps in the mental process of answering a question: (1) comprehension of the question, (2) information retrieval from memory, (3) information integration, and (4) translation to the correct response option (Tourangeau 1984; Tourangeau and Rasinski 1988). As detailed below, this burden might arise because a question is formulated using complicated or vague language, because a question asks for information that is not readily accessible in the respondent's memory, because a question asks for a simultaneous evaluation of



several factors, making it more difficult to render a judgment, or because a respondent's judgment does not correspond to the available answer categories. This burden might also be larger for certain groups of respondents—for example, the elderly.

Question sensitivity, by contrast, leads to item nonresponse when respondents are not willing to answer a question, even if they are able to. A sensitive question might infringe on respondents' privacy or make them reluctant to answer due to a fear of social or legal repercussions should the answer become known to third parties (Tourangeau, Rips, and Rasinski 2000; Tourangeau and Smith 1996).

While the survey methodology literature on question complexity and sensitivity is substantial, it is typically based on assessments of citizen or household surveys. It is unclear whether its findings are applicable to surveys of public servants. Public servants typically respond to employee surveys as part of their work duties, thus potentially enhancing their willingness to invest cognitive effort into question understanding. Moreover, public servants are usually relatively educated and accustomed to bureaucratic language, which is often highly technical and more complex than the language used in regular conversations.<sup>1</sup> Therefore, public officials should find it easier to interpret complex syntax and vague terms, and their education should enable them to integrate varied information and perform required calculations or information retrieval from memory more easily. At the same time, questions in public employee surveys often ask for more complex inferences than household surveys—for instance, about employees' perception of the organization or senior management practices. These diverging characteristics of public officials, the environment in which they respond to surveys, and the content of surveys put a premium on empirically assessing item nonresponse in public employee surveys, rather than simply extrapolating findings about item nonresponse from household surveys.

This chapter does this by analyzing missing response patterns in three public administration surveys—the United States Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS) and two World Bank surveys of public officials in Guatemala and Romania.<sup>2</sup> The analysis is based on the creation of a coding framework to assess different elements of question complexity and sensitivity, the application of this framework to code each of the questions in the aforementioned surveys in terms of complexity and sensitivity, and, finally, regressions to assess which of the elements of complexity and sensitivity predict item nonresponse.

We find, contrary to literature findings in other contexts, that public officials do not appear to shy away from answering questions that are longer, that are characterized by more complex syntax, or that require more cognitive effort to answer. We also find only limited evidence that question sensitivity is associated with greater item nonresponse. By contrast, we find robust evidence that one subindicator of complexity—the unfamiliarity of topics in questions—is associated with item nonresponse across all countries. Public officials prefer to not answer questions about topics outside of their immediate experience—for instance, about practices in their units and organization at large—a feature we term *unfamiliarity*. In sum, it appears that relatively highly educated public officials do not struggle with terminologically complex questions but are more unwilling or unable to answer questions about their broader working environment or to integrate different aspects of the functioning of their organization into one response option.

Given the manual nature of the approach to coding the complexity and sensitivity of survey questions, one natural criticism is that machine-coded measures may perform as effectively in determining problem questions but at a lower cost. We therefore perform a comparison of the core results to the predictive ability of machine-coded measures. We find that the unfamiliarity index continues to be the most effective approach to identifying questions that suffer from nonresponse in public servant surveys.

The chapter is organized as follows. Section 2 presents an overview of past work on survey complexity and sensitivity. Section 3 shows how the coding framework was constructed and how it relates to the past research, as well as the present research design. Section 4 details the results, which are followed by a discussion in section 5. The final section concludes and outlines avenues for future research.

## UNDERSTANDING ITEM NONRESPONSE: LESSONS FROM THE SURVEY METHODOLOGY LITERATURE

In essence, the survey methodology literature posits two broad underlying causes of item nonresponse: respondents are either unable to answer survey questions (due to different dimensions of question complexity) or are unwilling to answer survey questions (due to different dimensions of question sensitivity) (Rässler and Riphahn 2006). We follow the literature in assessing these two central causes of potential item nonresponse. To build the coding framework, we discuss the literature on complexity and then on sensitivity.

### Complexity

Questions assessing the same underlying concept can be expressed in more or less complex ways. “What is your age?” is an extremely common survey question. It is also a question that virtually everyone can understand and answer. “How many orbital periods have passed on the third planet from the sun since your hour of birth?” asks for the same information but could leave respondents confused about what the question is actually asking for. Although this example is needlessly complicated, some survey questions are longer and more convoluted or otherwise hinder respondents who are willing to respond from providing answers. The literature typically refers to this quality as *question complexity* (Knäuper et al. 1997; Yan and Tourangeau 2008).

Complexity is a multidimensional concept. While its definition is contested, it can perhaps best be conceptualized as a set of hurdles that respondents can encounter on their mental pathway from the moment they are presented with a question to providing an answer (Tourangeau and Rasinski 1988). Or as Knäuper et al. (1997, 181) phrase it, “Question answering involves a series of cognitive tasks that respondents have to resolve to provide high-quality data.” These tasks may be objectively more or less difficult, but the effort they require may also depend on respondents’ characteristics. To go back to the example used at the beginning of this section, the more complex version of the age question would likely pose relatively less trouble to a native English-speaking astrophysicist than to someone for whom English is a second language and who has never learned about physics.

The literature on cognitive psychology commonly refers to four steps in the question-answering process, as outlined by Tourangeau (1984), Tourangeau and Rasinski (1988), and Tourangeau, Rips, and Rasinski (2000). These steps are as follows: (1) question comprehension, (2) the retrieval of necessary information from memory, (3) the integration of the retrieved information into a judgment or estimate, and (4) the translation of the judgment into an appropriate response. Depending on the type and format of a question, these steps might vary in length and cognitive difficulty. For example, for a question about age, information is easily retrieved from memory but might require some mapping process if the response is not numerical but rather matched to predefined age bands.

In the first step, respondents have to comprehend the language used in a question and its intent (Holbrook, Cho, and Johnson 2006). Faaß, Kaczmarek, and Lenzner (2008, 2) write that “comprehending a question involves two processes which cannot be separated: decoding semantic meaning and inferring pragmatic meaning.” Therefore, a question with more elaborate syntax and sentence construction, as well as technical or unfamiliar words, requires more cognitive effort to be understood by respondents (Knäuper et al. 1997)—an effort they may or may not be able or willing to perform.

It is less obvious whether questions that are longer have a positive or negative impact on comprehension. On the one hand, a question might be longer because it explains its purpose and content in more detail, thus reducing the cognitive effort required on the part of respondents. On the other hand, a long question may simply be convoluted, touch on too many topics, or be difficult to remember in full when providing the answer, thus increasing difficulties for respondents (Holbrook, Cho, and Johnson 2006; Knäuper et al. 1997).

Other features of a question, like the number of propositions and logical operators (for example, *or* and *not*), dense nouns (accompanied by many adjectives or adverbs), or left-embedded syntax, can interact with the above to complicate even relatively short words and sentences (Faaß, Kaczmirek, and Lenzner 2008). Cognitive difficulties in comprehension might also depend on individual working memory capacity. Research by Just and Carpenter (1992) shows that working memory is a key element of both information storage and the computations necessary for language comprehension.

Once respondents have comprehended what information is required, they have to search their memories to retrieve it. This task is more difficult when the required information refers to the more distant past (Krosnick 1991). It is clear that recalling what one had for breakfast this morning, for example, is easier than recalling the same information from a week ago. In psychology, this is the well-known phenomenon of *attitude (or information) accessibility* (Fazio 1986). More-accessible attitudes are retrieved from the memory more easily and quickly, or, in other words, with lower cognitive effort. The more recently an individual has thought about a particular matter, the more accessible this and related considerations are when answering a survey (Zaller 1992). Zaller (1992) terms the predominant use of easily retrievable information the “accessibility axiom.”

Apart from the temporal reference frame, attitudes that refer to direct, more recent, or recurrent experiences tend to be more accessible (Berger and Mitchell 1989; Fazio 1989; Fazio and Roskos-Ewoldsen 2005). Memories of events that were emotional, unique, or drawn out are more likely to be accessible from memory, possibly biasing survey responses in favor of such events (Tourangeau 1984). Finally, it is less burdensome to retrieve information related to one item or topic rather than two or more, and surveys should therefore avoid what are called *double-barreled* questions (Krosnick 1991).

The information retrieved then needs to be integrated into a judgment. Depending on the question, the difficulty of this process can range from null to very high. Information about one’s gender or age and other factual questions about oneself require little integration. By contrast, in other cases, the format in which questions are asked can shape the difficulty of integration. Consider the following example of three different question formats to measure the role of personal connections in public sector recruitment:

**1) Were personal connections (friends and family in the institution) important to get your first public sector job?**

*1 - Yes; 2 - No; 3 - Don’t know*

**2) How important were personal connections (friends and family in the institution) to getting your first public sector job?**

*1 - Very unimportant; 2 - Somewhat unimportant; 3 - Neither important nor unimportant; 4 - Somewhat important; 5 - Very important; 6 - Don’t know*

**3) Please rank the following criteria in order of the importance they had for obtaining your first public sector job:**

*1 - Personal connections (friends and family in the institution); 2 - Political connections; 3 - Educational background; 4 - Previous work experience; 5 - Work-related skills*

The first version of the question only requires respondents to make a binary choice about the importance of personal connections. The second version requires a more fine-grain evaluation—not only about whether personal connections were important but also how important. In the third version, respondents have not only to judge the importance of personal connections but also of four other considerations and to evaluate them against each other. Clearly, this last approach requires the greatest cognitive effort from respondents.

Much work in psychology has been conducted to determine how people formulate judgments from available information. According to Anderson’s (1971) information integration theory, when people formulate a judgment, they gather all available pieces of information, assigning value and weight to each of them, before summing them up to form a final judgment. Another view, developed mainly in the work of Tversky and Kahneman, is that people tend to use a range of heuristic methods to arrive at judgments, like using only readily available instances and examples, using resemblance to a prototype, or anchoring based on

initial information (Tourangeau 1984; Tversky and Kahneman 1974). A combination of these views has been adopted by Zaller (1992) in his “response axiom,” which argues that individuals answer survey questions by averaging different considerations, but only those that are immediately salient or accessible to them.

The final stage of question answering is mapping the answer onto the available response options (Tourangeau and Rasinski 1988). Holbrook, Cho, and Johnson (2006) mention two possible difficulties at this stage. One is the problem of mental multitasking, which occurs because respondents have to simultaneously remember the question and the answer options and to map their formed judgments onto them. This might be an issue, particularly for individuals who have problems with remembering information—for example, the elderly. It might also be overly taxing if response options are descriptive rather than articulated on a frequency or Likert-like scale, or if they contain vague words and complex phrases. Second, response formats that are hard to understand or that have an ambiguous set of possible responses might compound mapping difficulties. Whereas multitasking as an obstacle depends mainly on the respondent, problems with the response format are usually due to faulty questionnaire design. To ease the process of translating a formed judgment into a response, it is particularly important to ensure that the set of responses to each question is both exhaustive and mutually exclusive (Krosnick and Presser 2009).

Across each of these stages, survey question complexity can have multiple effects. Some are less consequential for survey data quality—such as longer response times (Faaß, Kaczmirek, and Lenzner 2008; Yan and Tourangeau 2008) or respondents’ asking the interviewer for clarification (Holbrook, Cho, and Johnson 2006). Some effects of question complexity, however, are more consequential. In particular, complexity can invite *acquiescence bias* or *satisficing*, in which respondents tend to agree with a complex statement, regardless of their true position, in order to avoid cognitive overload (Knäuper et al. 1997; Krosnick 1991; Lenski and Leggett 1960). Apart from agreeing with a statement, respondents might ease the cognitive burden by selecting the first available response option, choosing randomly, skipping the question, or selecting the “I don’t know” option. This last option is an example of strong satisficing because it requires no cognitive effort whatsoever.<sup>3</sup>

In short, the survey methodology literature suggests that complex survey questions heighten the cognitive effort required along the mental process of answering a question and may thus lead to satisficing, including item nonresponse. The empirical literature that complements the theoretical considerations outlined here finds supporting evidence that each of these answering stages can increase nonresponse. For example, Knäuper et al. (1997) find that respondents answer “I don’t know” more often to questions that, among other things, contain ambiguous terms or require retrospective or quantity reports. Including these more complex question characteristics raised item nonresponse in their study by between 0.5 and 7.7 percentage points (and, as expected, more so for individuals with lower cognitive ability). This is substantial, considering that in most subgroups, the total share of “I don’t know” responses stayed well below 10 percent.

## Sensitivity

Irrespective of how complicated a question is, the extent to which it requests personally sensitive information may also impact nonresponse. “How many bribes have you accepted in the last month?” has simple syntax, uses precise terms, and has a clearly defined, short, and direct reference frame. It is not a complex question. However, the question is sensitive—it asks about behavior that is typically both morally wrong and illegal—which is a second source of concern for survey designers.

Unlike complex questions, when people are asked sensitive questions, they usually know the correct or true response but are unwilling to provide it. Or, in other words, “data quality does not only depend on the accurate recall of facts but also depends on the degree of peoples’ self-disclosure” (Gnambs and Kaspar 2015, 1238). Sensitivity is unavoidable in some surveys. In fact, the whole purpose of a survey might be to elicit information that cannot be obtained from other data sources because people conceal it and avoid discussing it in public (Lensvelt-Mulders 2008). Typical topics of concern include drug use, sexuality, and gambling. In the context of public administration, the issue of sensitivity may arise with topics such as corruption and integrity, discrimination inside the public service, or the sexual harassment of employees.

The most commonly used classification of sources of sensitivity was developed by Roger Tourangeau, along with several coauthors (Tourangeau, Rips, and Rasinski 2000; Tourangeau and Yan 2007). According to them, sensitivity derives from three primary sources—a question may touch on a taboo subject, a truthful answer may violate social norms, or a truthful answer may lead to negative formal consequences.<sup>4</sup>

In the first instance, respondents might feel that the topic of a question is not supposed to be discussed in public but rather kept private. In other words, it is considered a *taboo* subject (Tourangeau and Yan 2007). This may be a concern for various topics, from sexual orientation to salary level. McNeeley (2012) describes how talking about such topics may lead to distress and uneasiness for respondents (and, in some cases, enumerators as well) and, for demographic items, also lead to the threat of identification.<sup>5</sup> Unlike social-desirability bias and other sources of sensitivity discussed below, these topics are not problematic because revealing the requested information could lead to some type of sanctions. Instead, these topics are perceived as sensitive regardless of a respondent's true position (Krumpal 2013; Tourangeau and Yan 2007) because it is not common to discuss them in public or with strangers, like an enumerator. Therefore, these questions often lead to item nonresponse rather than misreporting (Höglinger, Jann, and Diekmann 2016) because respondents simply do not want to discuss the topics at all.

Second, but arguably most commonly, the wariness to truthfully answer sensitive questions is explained with reference to *social-desirability bias*. This refers to an inner desire to conform to established social norms in a given circle, be it a workplace, a family, or society at large. Admitting that one has committed an action that violates a common norm, either by doing something considered “wrong” (for example, taking a bribe) or failing to do “good” (for example, not helping a colleague in need), is undesirable (Tourangeau and Yan 2007) because, if someone found out, the violator could be frowned upon, criticized, or shunned. The impact social-desirability bias has on responses further depends on the specific social norms respondents identify with and how concerned they are about not violating them. For example, Kim and Kim (2016) find that national culture significantly moderates the degree and pattern of social-desirability bias in public service motivation surveys.

Apart from this *extrinsic* threat, answering sensitive questions also poses *intrinsic* threats to the self-image of respondents (Lensvelt-Mulders 2008, 462). Touching on sensitive topics may raise feelings of guilt, embarrassment, or shame in respondents for having done (or for failing to do) something, or they may be stressful to discuss for respondents in general. Therefore, to avoid negative consequences from others as well as one's own conscience, respondents may prefer not to answer a sensitive question or to answer it in a socially “expected” way. In face-to-face surveys, even respondents who believe in the full confidentiality of their responses may want to create a positive image of themselves or earn social approval from the enumerator (Krumpal 2013) and thus may succumb to social-desirability bias.

Psychologists have long debated the precise causes of social-desirability bias. Paulhus (1984) suggests it has two parts. One is impression management—that is, a desire to present oneself in a positive light in front of others to avoid negative feedback from them. Another is self-deception, which means holding favorably biased views about oneself while honestly believing them to be true.

Third, a related but distinct source of sensitivity comes from questions that ask about actions that are formally (rather than socially or informally) prohibited. For example, hiring one's family members and friends might be a widespread and socially accepted practice. However, if nepotism is formally prohibited, then admitting it in a survey might lead to legal sanctions, like a fine, a disciplinary note, or being fired. Or, as Tourangeau and Yan (2007, 859) note, “possessing cocaine is not just socially undesirable; it is illegal, and people may misreport in a drug survey to avoid legal consequences rather than merely to avoid creating an unfavorable impression.”

Informal and formal sources of sensitivity might also interact with each other. Research by Galletly and Pinkerton (2006) suggests that there is an interaction between social stigma and formal sanctions in the case of HIV disclosure laws in US states. The introduction of legal sanctions for some actions may add to the already existing social stigma around them. Alternatively, the threat of social disapproval may be a more undesirable consequence than a formal sanction that is small or unlikely to follow. Likewise, if social and legal norms are not perfectly matched, admitting to an illegal but socially acceptable practice might be



less difficult for respondents. For example, if the law prohibits hiring one's family members and friends but society generally accepts this practice, then admitting to some degree of nepotism might come more easily to a survey respondent than if this practice were socially unacceptable.

As with question complexity, item nonresponse is one of several possible behavioral responses to sensitivity (Krumpal 2013; Lensvelt-Mulders 2008; McNeeley 2012; Tourangeau and Smith 1996; Tourangeau and Yan 2007). Respondents, when aware of survey topics, may decline to participate altogether (McNeeley 2012). Moreover, respondents may believe that not answering sensitive questions is “revealing” in itself (Tourangeau and Yan 2007, 877). Instead, respondents may choose simply to answer in an expected way that is certain not to result in any negative consequences (Bradburn et al. 1978; Krumpal 2013; McNeeley 2012). For example, refusing to answer a question about bribe-taking might seem suspicious in itself, so bribe-takers may avoid any suspicion or feeling of shame by simply saying that they have never taken bribes rather than refusing to answer.

In sum, item nonresponse may increase as a result of increased question sensitivity—though, compared with complexity, this effect may be diluted by respondents who answer sensitive items in a socially desirable way rather than not answering at all (Sakshaug, Yan, and Tourangeau 2010). Tourangeau and Yan (2007), for example, report that item nonresponse in the National Survey of Family Growth (NSFG) Cycle 6 female questionnaire tends to rise by fewer than 3 percentage points when comparing questions with very low sensitivity (for example, education [0.04 percent nonresponse rate] and age [0.39 percent]) with high-sensitivity items (for example, the number of times the respondent had sex in the past four weeks [1.37 percent] and their number of sexual partners [3.05 percent]). Only the income question has more noticeable nonresponse, at 8.15 percent. And whereas experimental methods that aim to reduce question sensitivity, such as the unmatched count technique, do significantly affect the mean estimates obtained, they have a far smaller effect on item nonresponse (Coutts and Jann 2011), suggesting that biasing rather than avoiding an answer is a more prevalent response for people presented with sensitive questions.<sup>6</sup> Comparing the effects of unit (although not item) nonresponse and measurement error in reports of voting behavior, Tourangeau, Groves, and Redline (2010) suggest that the latter is around two times larger and can elevate the reported prevalence of voting from the true value of 47.6 percent to 69.4 percent.

## METHODOLOGY

### Case Selection

We evaluate question complexity and sensitivity and their relationship to item nonresponse in three 2019 governmentwide public administration surveys in Guatemala, Romania, and the United States.

The surveys in Guatemala and Romania were nationally representative surveys of public officials conducted by the World Bank in 2019 and 2020. The survey in Guatemala was a face-to-face survey conducted from November to December 2019. It covered 14 central government and four decentralized institutions. A sample of 205 respondents was selected from each institution (of which one-quarter were supervisors and three-quarters were subordinates). In total, 3,465 public officials provided answers, resulting in a response rate of 96 percent (World Bank 2020a). All respondents were surveyed in person by trained enumerators.

The survey in Romania used a mixed-mode delivery, with a randomly chosen set of officials answering the survey online and another set answering it in face-to-face (F2F) interviews with enumerators. The face-to-face questionnaire was longer than the online one, and, therefore, only the questions overlapping between the two versions are used in the analyses below. The Romanian data were collected from June 2019 to January 2020 across 81 institutions that agreed to participate (out of 103 invited). The targeted sample of respondents was drawn from the institutional census of employees. In total, 2,721 public officials answered the online questionnaire (for a response rate of 24 percent), and 3,316 answered the face-to-face one (for a response rate of 92 percent; for details see World Bank [2020b]).



Responding to a survey online may increase respondents' sense of comfort and privacy, thus reducing the perceived threat posed by sensitive questions (McNeeley 2012). On the other hand, online surveys lack an enumerator, who can clarify complex questions or encourage respondents to answer (De Leeuw 1992). We thus estimate the effects of complexity and sensitivity for Romania separately for online and face-to-face respondents.

The FEVS has been fielded by the US OPM biannually since 2004 and annually since 2010 (see chapter 26). It covers all types of employees across federal government departments and agencies that choose to participate. It is delivered in an online, self-administered form. In the latest available iteration, from 2019, which is used here, it was conducted as “a census administration that included all eligible employees from 36 departments and large agencies as well as 47 small and independent agencies” (OPM 2019). In total, over 615,000 government employees responded to the survey, for a response rate of 42.6 percent.

The case selection enables us to understand item nonresponse in public service surveys of countries from across diverse cultures, regions, and levels of development and education. Findings about item nonresponse that travel across all three contexts are plausibly generalizable to other surveys of public administrators.

## Coding Framework

Understanding item nonresponse—and whether different dimensions of complexity and sensitivity shape item nonresponse in public service surveys—requires measuring complexity and sensitivity consistently across and within surveys. To do so, coding framework is developed that allows us to assign a numerical value reflecting the degree of complexity and sensitivity of every survey question. The approach builds on the existing literature summarized above and resembles research by Bais et al. (2019), who similarly integrate several aspects of complexity and sensitivity into a manual coding framework.<sup>2</sup>

The complexity and sensitivity indexes comprise several subdimensions, as described in tables 22.1 and 22.2. The complexity index is composed of 10 subdimensions, which are conceptually based on the four-stage mental process of answering a question (see Tourangeau and Rasinski 1988; Tourangeau, Rips, and Rasinski 2000), and synthesizes the measures proposed by, among others, Belson (1981); Holbrook, Cho, and Johnson (2006); and Knäuper et al. (1997). The subdimensions include the complexity of the syntax, the number of subquestions, the presence of a reference frame, and the unfamiliarity of the subject.

The sensitivity index is constructed using four subdimensions suggested by the literature: invasion of privacy, the social-emotional threat of disclosure, the threat of formal sanctions (Tourangeau, Rips, and Rasinski 2000; Tourangeau and Yan 2007), and the interaction between informal and formal sanctions (see, for example, Galletly and Pinkerton 2006).

We evaluate each question in the three surveys studied along the dimensions outlined in tables 22.1 and 22.2 and score it a value of 0, 1, or 2. The value of 0 is given to questions that do not present a particular subdimension of complexity or sensitivity at all. The value of 1 refers to cases in which questions potentially create problems for respondents in a given subdimension, whereas 2 is used for cases where such problems are clearly substantive. The full coding framework is presented in appendix J.

Three research assistants applied the coding framework to assess the complexity and sensitivity of each of the questions in the three surveys. (For examples of this process, see box 22.1.) Each research assistant first coded the values independently, and then their scores were compared. In 86.8 percent of cases, all coders working on a given question agreed on the score. In the instances where they were not in agreement, differences were discussed and resolved with a view to maximizing consistency in coding across survey questions. The values of both indexes are calculated as arithmetic means of the scores across their respective subdimensions.

**TABLE 22.1 Complexity Coding Framework**

Subdimension	Description	Guatemala	Romania	United States (FEVS)	Aggregate
<i>Comprehension</i>					
Complex syntax	This component assesses the length of a question, which is measured by the number of characters ( <i>n</i> ), and the complexity of the syntax or the grammatical arrangements of words and phrases, which is determined by the sentence structure. The term <i>simple syntax</i> indicates simple sentence(s) with three parts of speech, <i>moderately difficult syntax</i> indicates simple sentence(s) with more than three parts of speech, and <i>complicated syntax</i> indicates complex or complex-compound sentences.	0.85 (0.65)	1.09 (0.61)	1.2 (0.53)	1 (0.63)
Vagueness	This component assesses the extent to which the language used in a question is vague, unclear, imprecise, ambiguous (Edwards et al. 1997), or open to interpretation. Common terms such as “good” are predetermined in a list of vague words.	0.34 (0.48)	0.64 (0.5)	0.54 (0.55)	0.48 (0.52)
Reference category	This component assesses the extent to which the necessary frame(s) of reference are available in a question so that respondents understand the question in the way intended.	0.17 (0.47)	0.26 (0.51)	0.16 (0.48)	0.2 (0.48)
Number of questions	This component measures the number of subquestions embedded in the question block to which a question belongs. A subquestion must only ask for one issue, so a compound subquestion is not counted as one subquestion.	0.3 (0.55)	0.11 (0.33)	0.21 (0.46)	0.22 (0.48)
<i>Information retrieval</i>					
Unfamiliarity	This component assesses the extent to which respondents are knowledgeable on the subject of a question. The coding presumes that respondents are more familiar with subjects they have a closer knowledge of (for instance, their own experience versus their perceptions of the experiences of other employees in the organization).	0.93 (0.79)	0.34 (0.53)	0.35 (0.51)	0.62 (0.72)
Recalling	This component assesses the extent to which respondents are required to remember information based on the question's level of specificity and time frame of interest (past/present).	0.91 (0.4)	1 (0.44)	1.04 (0.4)	0.96 (0.42)
<i>Information integration</i>					
Computational intensity	This component assesses the extent to which basic arithmetic computations (addition, subtraction, multiplication, and division) are required to reach an answer.	0.07 (0.29)	0.07 (0.26)	0.02 (0.16)	0.06 (0.26)
Scope of information	This component assesses the extent to which answers are derived from information beyond the personal experience of respondents.	0.38 (0.52)	0.46 (0.55)	0.32 (0.47)	0.39 (0.52)
<i>Translation to answer</i>					
Category mismatch	This component assesses the extent to which the available answer options match the true answer to the question.	0.06 (0.28)	0.09 (0.41)	0.04 (0.25)	0.06 (0.32)
Number of responses	This component assesses the extent to which respondents are required to pick more than one answer to the question.	0.06 (0.28)	0.01 (0.09)	0 (0)	0.03 (0.2)

Source: Original table for this publication.

Note: The final four columns show the mean and standard deviation (in parentheses) of scores for each subdimension and survey. FEVS = Federal Employee Viewpoint Survey.

**TABLE 22.2 Sensitivity Coding Framework**

Subdimension	Description	Guatemala	Romania	United States (FEVS)	Aggregate
<i>Privacy</i>					
Invasion of privacy	This subindicator measures the extent to which respondents are asked to discuss taboo or private topics that may be inappropriate in everyday conversation. Questions related to a respondent's income or religion may fall into this category.	0.08 (0.27)	0.31 (0.6)	0.04 (0.19)	0.14 (0.41)
<i>Informal sensitivity</i>					
Social-emotional threat of disclosure	This subindicator measures the degree to which respondents may be concerned with the social or emotional consequences of a truthful answer, should the information become known to a third party. In the case of informal sensitivities, this type of question is only considered sensitive if the respondent's truthful answer departs from socially desirable behaviors or social norms.	0.55 (0.51)	0.7 (0.65)	0.79 (0.56)	0.65 (0.58)
<i>Formal sensitivity</i>					
Threat of formal sanctions	This subindicator measures the degree to which respondents may be concerned with the legal and/or formal consequences of a truthful answer, should the information become known to a third party. This type of question is only sensitive if the respondent's truthful answer departs from legal behaviors defined by formal institutions and legal regulations.	0.26 (0.6)	0.15 (0.46)	0.15 (0.5)	0.2 (0.54)
<i>Interaction</i>					
Relationship between informal and formal sensitivity	This subindicator measures the likelihood that a behavior or attitude may cause a threat of both social-emotional disclosure and formal sanctions. This type of question is logically more sensitive than ones that violate one type of institution while conforming to another. A behavior may be frowned upon in one's social circle—for example, reporting colleagues taking bribes might be considered “snitching”—but it may also be a legal obligation. In such instances, asking about it should be less sensitive compared to a situation where both informal and formal norms were violated. Galletly and Pinkerton (2006) suggest such an interaction between social stigma and formal sanctions (in the case of HIV disclosure laws).	0.2 (0.41)	0.2 (0.48)	0.13 (0.49)	0.19 (0.45)

Source: Original table for this publication.

Note: The final four columns show the mean and standard deviation (in parentheses) of scores for each subdimension and survey. FEVS = Federal Employee Viewpoint Survey.

## Analysis

To investigate nonresponse in a public administration setting, we assess the impact of the complexity and sensitivity measures outlined above on responsiveness in the three surveys under study. The regressions take the respondent-question as the unit of observation, meaning that each row corresponds to a particular respondent's answer to a given question. We define item nonresponse as an “I don't know” answer, a refusal to answer, or skipping the question.

We control for individual-level characteristics that might affect nonresponse, including age and education (both of which are correlated with respondents' cognitive abilities to deal with complexity; Holbrook, Cho, and Johnson [2006]; Yan and Tourangeau [2008]), gender (which can shape item nonresponse for

## BOX 22.1 Applying the Coding Framework: Illustrative Examples from Romania

**Complexity—information retrieval (recalling):** “Which of the following factors were important for getting your current job in the public administration?” This question pertains to the past, asks for a specific level of information, and requires the respondent to consider the importance of many factors: academic qualifications, job-specific skills, knowing someone with political links, having personal connections, and so on. Therefore, this question is coded as 2.

**Complexity—information integration (computational intensity):** “How many years have you been in your current institution?” This question requires respondents to calculate their length of service in their current institution by subtracting their starting year from the current year. This is not a complicated calculation, but it still is not likely to be performed often and may require some mental effort if respondents joined a long time ago, are confused about whether periods like initial internships should be included, and so on. Therefore, this question is coded as 1.

**Sensitivity—threat of formal sanctions:** “How frequently do employees in your institution undertake the following actions? Accepting gifts or money from citizens.” This question is coded as 2 because respondents may feel that there are social consequences for disclosing this information or even formal ones if they did not inform relevant authorities about the bribe-taking behavior.

sensitive questions—for instance, on harassment), tenure in the organization, and managerial status (more-experienced workers and managers might have more work-related knowledge and a different cost-benefit calculus when deciding whether to answer a survey), as well as job satisfaction as a proxy measure for willingness to respond (with more-satisfied respondents potentially more willing to respond to employee surveys or, alternatively, dissatisfied workers more eager to respond to report reasons for their dissatisfaction).<sup>8</sup>

We further control for the overall response rate in the government organization or agency to which a respondent belongs. A lower response rate might reflect unobservable characteristics of the organization or its employees that shape item nonresponse. We also control for the position of a question within a questionnaire (coded as integer variables starting from one). This is to take into account the fact that respondents might skip more questions or become less willing to cognitively engage with questions as the survey progresses and fatigue or dullness sets in (Krosnick 1991). Our data thus take a “long” format, with each row corresponding to a particular respondent’s answer to a question, accompanied by the respondent’s individual characteristics and the variables pertaining to his or her organization; the question’s complexity, sensitivity, and position in the questionnaire; and, finally, whether the respondent answered a given question (1) or not (0). In general, it is found that men tend to have lower item nonresponse and that nonmanagers and less-satisfied employees skip questions more often, although the pattern doesn’t hold in all settings and regression specifications. Questions appearing later in the questionnaire are also omitted more often, as hypothesized.

We first look at simple correlations between the key variables of interest and then go on to regress the item nonresponse variable on the indexes of complexity and sensitivity, as well as their various subdimensions in ordinary least squares (OLS) regressions. In order to account for the possible correlation of residual errors in the data set, we use multiway clustering on the individual and question levels, which allows us to correctly estimate standard errors and corresponding significance levels.

## RESULTS

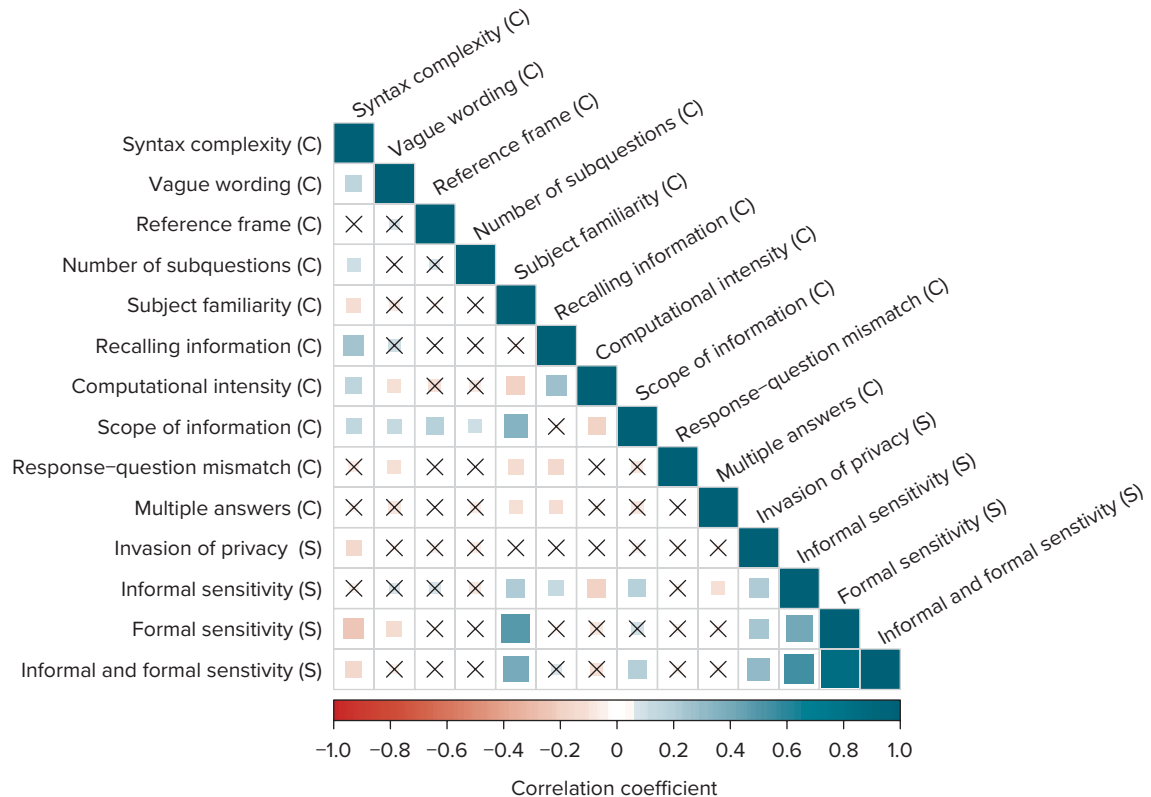
Descriptively, our independent variables vary across the components of complexity and sensitivity we code. As detailed in tables 22.1 and 22.2, among the components of complexity, complexity of syntax and unfamiliarity are the variables with the highest variance. On the other end of the scale, components related to translation to answer seem the least variable. Among the sensitivity components, the social-emotional threat of disclosure records both the highest mean score and the greatest variation. Invasion of privacy scores the lowest in mean and standard deviation.

To ensure the coding framework meaningfully captures distinct subdimensions or components of complexity and sensitivity, we assess correlations between different components or subdimensions of complexity and sensitivity. Figure 22.2 shows that for complexity, most of the correlations are not significant, suggesting, as theorized, that different components relate to different mental processes and aspects of a question. Where there is some conceptual overlap, however, we do see significant correlations, such as between syntax complexity and vague wording or between the scope of information and subject unfamiliarity.

In the case of sensitivity, all correlations are significant and strong. This is conceptually plausible. Informal and formal sensitivity most often occur simultaneously, while questions about illegal or socially disapproved behaviors are plausibly also often too private or embarrassing to discuss in public. In sum, the observed correlations yield a degree of credibility to the coding framework and its application.

Next, as presented below, we can observe that item nonresponse is a challenge across the three surveys, though to a varying extent. As illustrated in figure 22.3, in the FEVS online survey, questions have an average item nonresponse of 2.4 percent. This number increases to 2.6 percent in the face-to-face public service

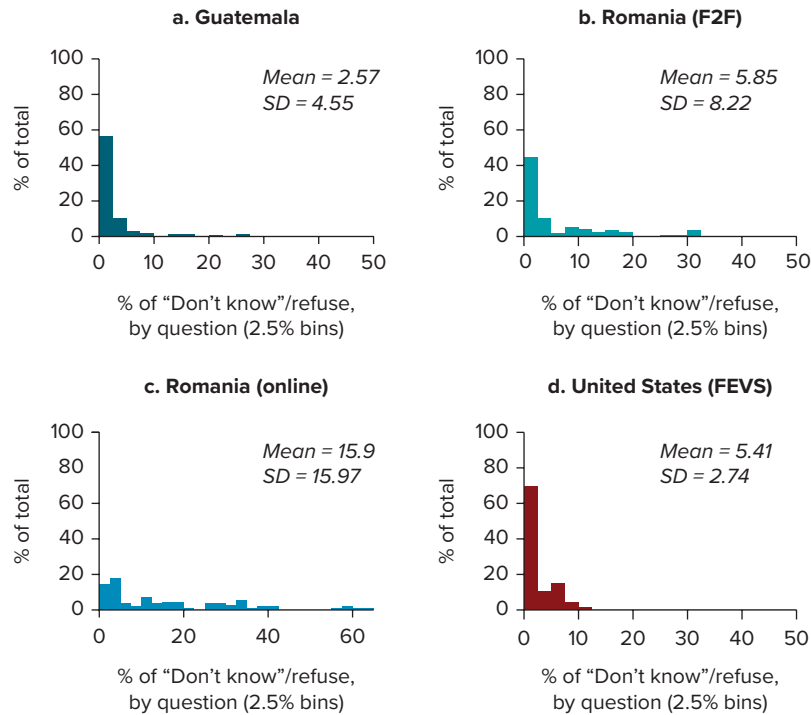
**FIGURE 22.2** Correlation between Subdimensions of Complexity and Sensitivity



Source: Original figure for this publication.

Note: Correlations are obtained by pooling questions across all three surveys. Crosses mark correlations that are insignificant at the 5 percent level. C = complexity; S = sensitivity.

**FIGURE 22.3** Share of Missing Responses



Source: Original figure for this publication.

Note: FEVS = Federal Employee Viewpoint Survey; F2F = face-to-face; SD = standard deviation.

survey in Guatemala and 5.9 percent in the face-to-face public service survey in Romania. In the online version of the survey in Romania, in turn, average item nonresponse increases to 15.9 percent.<sup>2</sup> To what extent do complexity and sensitivity predict item nonresponse?

Looking first at correlations, we find that there is a positive association between complexity and item nonresponse. Figure 22.4 presents the correlations separately for each survey. Correlation coefficients range from 0.066 to 0.404 and are significant at the 5 percent level, except for Guatemala. We observe a similar pattern, although slightly weaker in Romania, for correlation between item nonresponse and sensitivity.

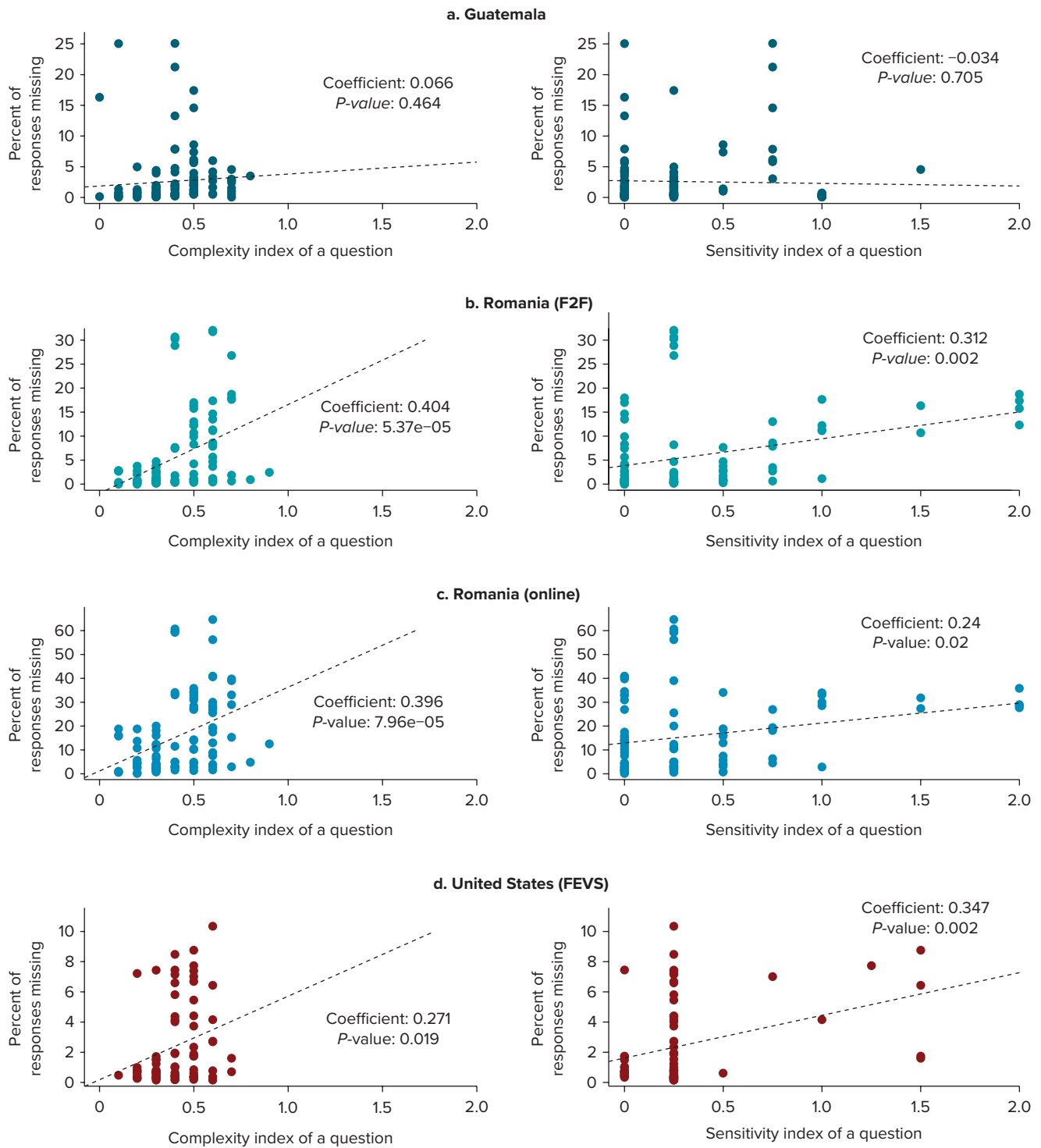
Table 22.3 presents regressions of item nonresponse on standardized sensitivity and complexity (both separately and jointly) with and without the aforementioned set of controls. We observe evidence from the United States and Romania that both complexity and sensitivity increase the probability of survey nonresponse in surveys of public officials. The results in Guatemala are not significant at the 10 percent level. The effect sizes are relatively small, with a standard deviation increase in the indexes having a 1 percentage point increase in nonresponse in the United States. In Romania, a one standard deviation increase in complexity is associated with an at most 6 percentage point increase in nonresponse, depending on the specification and mode of enumeration.

On average, the indexes of complexity and sensitivity thus predict item nonresponse in some but not all cases. Of course, however, it could be that our indexes—which simply average out different potentially relevant subcomponents of complexity and sensitivity—are not appropriately aggregated. The various subcomponents of complexity and sensitivity may not, as theorized, measure a single underlying dimension. To assess this, exploratory factor analysis (EFA) is performed across all 14 subdimensions pooled together. Indeed, instead of finding that two factors are sufficient to describe the data (as would be expected if the subdimensions measured only two dimensions: complexity and sensitivity), we find that at least four factors are needed to properly describe the data in each survey.<sup>10</sup>

The results of the EFA with four factors are presented in table 22.4. The results suggest that across countries, sensitivity subdimensions load onto a single factor (first factor). While the scores for the second



**FIGURE 22.4 Relationship between Complexity and Sensitivity Indexes and the Share of Missing Responses**



Source: Original figure for this publication.

Note: FEVS = Federal Employee Viewpoint Survey; F2F = face-to-face.

**TABLE 22.3 The Impacts of Complexity and Sensitivity on Item Nonresponse**

	OLS estimates					
	(1)	(2)	(3)	(4)	(5)	(6)
<i>Guatemala</i>						
Sensitivity	–0.002 (0.004)		–0.002 (0.004)	0.002 (0.004)		–0.002 (0.004)
Complexity		0.003 (0.006)	0.003 (0.006)		0.003 (0.005)	0.002 (0.006)
<i>Romania (F2F)</i>						
Sensitivity	0.025*** (0.004)		0.020*** (0.004)	0.020*** (0.006)		0.015** (0.006)
Complexity		0.035*** (0.009)	0.030*** (0.009)		0.033*** (0.009)	0.031*** (0.009)
<i>Romania (online)</i>						
Sensitivity	0.039*** (0.007)		0.030*** (0.009)	0.027** (0.010)		0.017 (0.010)
Complexity		0.062*** (0.015)	0.056*** (0.015)		0.060*** (0.015)	0.057*** (0.015)
<i>United States (FEVS)</i>						
Sensitivity	0.009** (0.004)		0.008* (0.004)	0.009** (0.003)		0.008* (0.004)
Complexity		0.007* (0.003)	0.004 (0.003)		0.008* (0.003)	0.005 (0.003)
Controls	No	No	No	Yes	Yes	Yes

Source: Original table for this publication.

Note: Standard errors are in parentheses and are clustered using multiway clustering at the level of an individual and a question throughout. All columns report OLS (ordinary least squares) estimates. The dependent variable in all columns is a dummy indicating whether a respondent answered “I don’t know,” refused to answer, or skipped a particular question. Sensitivity and complexity scores are calculated as z-scores estimated across questions in a given survey. Controls at the individual level include gender (0 = female; 1 = male), education (0 = below university level; 1 = university level), occupational status (0 = nonmanager; 1 = manager), experience in public administration (0 = more than 10 years; 1 = fewer than 10 years), and job satisfaction (0 = respondent not satisfied with his or her job; 1 = respondent satisfied with his or her job; in Guatemala, a pay-satisfaction variable is used instead). The control at the organizational level is response rate (0–1 scale) and, at the question level, the position of a question within a questionnaire. The controls are described in detail in the analysis subsection of the methodology section. FEVS = Federal Employee Viewpoint Survey; F2F = face-to-face.

Significance level: \* = 10 percent, \*\* = 5 percent, \*\*\* = 1 percent.

**TABLE 22.4 Exploratory Factor Analysis**

	First factor	Second factor	Third factor	Fourth factor
<b>Guatemala</b>				
<i>Complexity</i>				
Comprehension: complex syntax	–0.304	0.219		0.476
Comprehension: vagueness				0.508
Comprehension: reference category				
Comprehension: number of questions				0.348
Information retrieval: unfamiliarity	0.349	0.798	–0.355	–0.333
Information retrieval: recalling		0.445	0.263	
Information integration: computational intensity	–0.237		0.952	
Information integration: scope of information		0.329	–0.215	0.416
Translation to answer: categories mismatch	0.290	–0.342		
Translation to answer: number of responses		–0.281		

(continues on next page)

**TABLE 22.4 Exploratory Factor Analysis (continued)**

	First factor	Second factor	Third factor	Fourth factor
<i>Sensitivity</i>				
Invasion of privacy	0.259			–0.325
Social-emotional threat	0.484			
Formal threat of sanctions	0.776	0.241		–0.404
Informal-formal threat interaction	0.964			
<b>Romania</b>				
<i>Complexity</i>				
Comprehension: complex syntax			0.247	0.566
Comprehension: vagueness				–0.236
Comprehension: reference category		0.37		–0.447
Comprehension: number of questions				
Information retrieval: unfamiliarity		0.715		
Information retrieval: recalling			0.967	0.206
Information integration: computational intensity		–0.235		0.376
Information integration: scope of information		0.989		
Translation to answer: categories mismatch				
Translation to answer: number of responses			–0.252	
<i>Sensitivity</i>				
Invasion of privacy	0.532			
Social-emotional threat	0.65			
Formal threat of sanctions	0.806	0.314		
Informal-formal threat interaction	0.938	0.321		
<b>United States (FEVS)</b>				
<i>Complexity</i>				
Comprehension: complex syntax			0.982	
Comprehension: vagueness				–0.378
Comprehension: reference category		–0.356		–0.24
Comprehension: number of questions	0.403			
Information retrieval: unfamiliarity	0.207	0.544		
Information retrieval: recalling				
Information integration: computational intensity				0.597
Information integration: scope of information	0.23	0.795		
Translation to answer: category mismatch				
Translation to answer: number of responses				
<i>Sensitivity</i>				
Invasion of privacy				0.547
Social-emotional threat	0.489	0.438		–0.259
Formal threat of sanctions	0.989			
Informal-formal threat interaction	0.909			

Source: Original table for this publication.

Note: Only loadings with absolute values higher than 0.2 are shown. FEVS = Federal Employee Viewpoint Survey.

factor exhibit more variation, two subdimensions consistently score highly across countries: *unfamiliarity* and *scope of information*. Both these factors measure whether a question asks about the personal or at least proximate experiences of a respondent rather than the broader working environment (for example, the behavior of employees in the organization as a whole). Both thus relate closely to the unfamiliarity (of a topic). The remaining two factors vary, in terms of significant subdimensions, across countries and thus do not offer a clear conceptual interpretation.

We next assess whether the four factors from the EFA models—and, in particular, the sensitivity factor (first factor) and the unfamiliarity factor (second factor)—predict item nonresponse (table 22.5). We find that the first factor (sensitivity) does not predict item nonresponse. By contrast, the second factor (unfamiliarity) does predict item nonresponse in two of the three countries (Romania and the United States) (the third and fourth factors do not display clear patterns).<sup>11</sup>

As the EFA pointed to the sensitivity index as meaningfully reflecting the empirical structure of the subdimensions, while the complexity index consists of unfamiliarity and other complexity items, we next regress item nonresponse on unfamiliarity, sensitivity, and complexity without unfamiliarity (table 22.6). We find that unfamiliarity significantly predicts item nonresponse in Romania and the United States. It is also associated with greater item nonresponse in Guatemala, though this relationship is not significant at the standard significance levels.

The coefficients on the unfamiliarity index are larger than those on the basic indexes in table 22.3. A standard deviation increase in the unfamiliarity index (implying that the questions are *less* familiar) increases nonresponse by 3 percentage points in the United States and by almost 20 percentage points in the online survey in Romania. Relative to the baseline levels of nonresponse of 2.4 percent in the US FEVS and 5.9 percent and 15.9 percent in Romania's face-to-face and online surveys, respectively, these are large effects. By contrast, within this framework, the sensitivity index and complexity without unfamiliarity do not have significant effects. The evidence we present points to unfamiliarity, in the sense we have coded it, as the key driver of nonresponse.

**TABLE 22.5 Factor Analysis Regression**

	Guatemala	Romania (F2F)	Romania (online)	United States (FEVS)
First factor	0.005 (0.005)	0.006 (0.007)	0.0001 (0.010)	0.006 (0.004)
Second factor	−0.002 (0.005)	0.048*** (0.007)	0.095*** (0.013)	0.018*** (0.003)
Third factor	−0.002 (0.003)	−0.009* (0.003)	−0.011 (0.006)	0.003 (0.003)
Fourth factor	0.011* (0.005)	0.016* (0.008)	0.029 (0.015)	−0.002 (0.002)
Controls	Yes	Yes	Yes	Yes
N	378,472	181,614	161,793	667,425
Adjusted R <sup>2</sup>	0.005	0.057	0.094	0.015

Source: Original table for this publication.

Note: Standard errors are in parentheses and are clustered using multiway clustering at the level of an individual and a question throughout. All columns report OLS (ordinary least squares) estimates. The dependent variable in all columns is a dummy indicating whether a respondent answered “I don’t know,” refused to answer, or skipped a particular question. Factor scores are obtained from exploratory factor analysis models with four factors, as presented in table 22.4. Controls at the individual level include gender (0 = female; 1 = male), education (0 = below university level; 1 = university level), occupational status (0 = nonmanager; 1 = manager), experience in public administration (0 = more than 10 years; 1 = fewer than 10 years), and job satisfaction (0 = respondent not satisfied with his or her job; 1 = respondent satisfied with his or her job; in Guatemala, a pay-satisfaction variable is used instead). The control at the organizational level is the response rate (0–1 scale) and, at the question level, the position of a question within a questionnaire. FEVS = Federal Employee Viewpoint Survey; F2F = face-to-face.

Significance level: \* = 10 percent, \*\* = 5 percent, \*\*\* = 1 percent.

**TABLE 22.6** Impact of Sensitivity, Complexity, and Unfamiliarity on Nonresponse Rate

	Guatemala	Romania (F2F)	Romania (online)	United States (FEVS)
Sensitivity	–0.005 (0.005)	0.004 (0.007)	–0.005 (0.011)	0.004 (0.004)
Complexity (without unfamiliarity subdimensions)	–0.004 (0.005)	–0.011 (0.008)	–0.027 (0.015)	–0.002 (0.003)
Unfamiliarity	0.007 (0.007)	0.097*** (0.017)	0.196*** (0.028)	0.030*** (0.007)
Controls	Yes	Yes	Yes	Yes
<i>N</i>	378,472	181,614	161,793	667,425
Adjusted <i>R</i> <sup>2</sup>	0.001	0.061	0.100	0.012

Source: Original table for this publication.

Note: Standard errors are in parentheses and are clustered using multiway clustering at the level of an individual and a question throughout. All columns report OLS (ordinary least squares) estimates. The dependent variable in all columns is a dummy indicating whether a respondent answered “I don’t know,” refused to answer, or skipped a particular question. Sensitivity and complexity scores are calculated as z-scores estimated across questions in a given survey. Unfamiliarity is calculated as a mean value of the “information retrieval: unfamiliarity” and “information integration: scope of information” subdimensions. Controls at the individual level include gender (0 = female; 1 = male), education (0 = below university level; 1 = university level), occupational status (0 = nonmanager; 1 = manager), experience in public administration (0 = more than 10 years; 1 = fewer than 10 years), and job satisfaction (0 = respondent not satisfied with his or her job; 1 = respondent satisfied with his or her job; in Guatemala, a pay-satisfaction variable is used instead). The control at the organizational level is the response rate (0–1 scale) and, at the question level, the position of a question within a questionnaire. FEVS = Federal Employee Viewpoint Survey; F2F = face-to-face. Significance level: \* = 10 percent, \*\* = 5 percent, \*\*\* = 1 percent.

### The Performance of Manual versus Machine-Coded Complexity

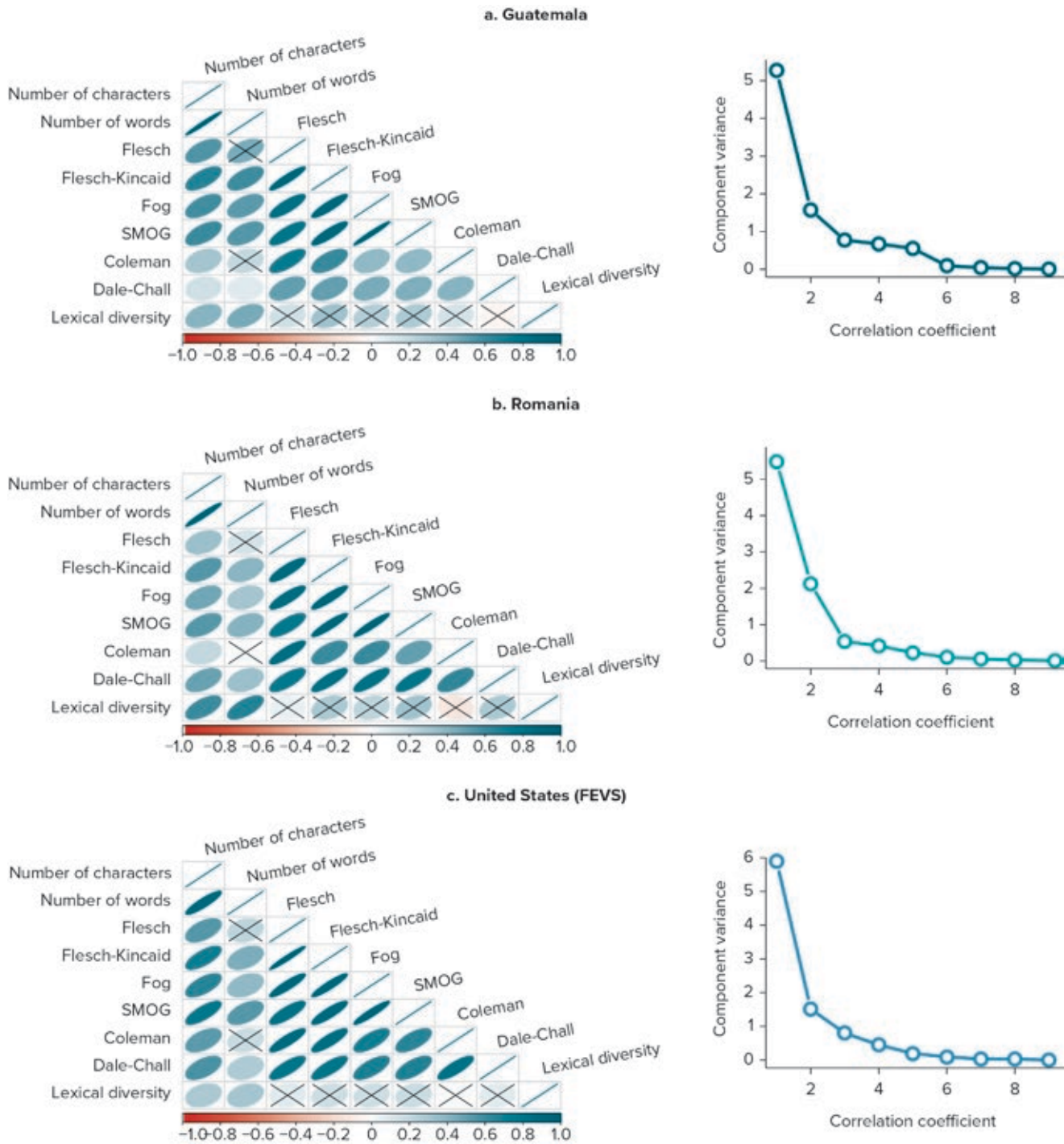
A potential criticism of our approach is that automated measures of complexity may be as effective a means of identifying potential problem questions—or more—while requiring far less investment. We therefore turn to assessing the relative effectiveness of the approach with respect to machine-coded complexity.

Machine coding has the disadvantage that it must rely on relatively basic indicators of syntax. Computer-based complexity indicators are usually based on mathematical formulas that score the complexity (or, as it is described more commonly, the *readability*) of a text based on purely textual features, like the number of characters, syllables, words, and sentences. Some measures also check the text against predefined lists of words regarded as easy or difficult. As there are dozens of such indexes, with no agreement on which one is optimal, we select nine that are commonly used and calculate their values for each survey question. Correlations among their final scores can be seen in the first column of figure 22.5. The correlation between models (shown by the intensity of shading) varies, though is understandably relatively high across the comparisons made. Given a very high degree of correlation, instead of using all nine scores in a regression, a principal component analysis is performed across them and extract the first principal component (which explains between 59 and 66 percent of the overall variance—see the second column in figure 22.5) to serve as a predictor in the regressions.

Tables 22.7 and 22.8 evaluate the predictive power of machine-driven complexity scores. When not controlling for the manually coded indexes (sensitivity, unfamiliarity, and complexity without unfamiliarity), we find some evidence for an effect of machine-coded complexity, with significant positive effects in the United States only.

Once we condition on the indexes of complexity (excluding complexity of syntax to avoid multicollinearity with the machine-coded measure) and sensitivity, we no longer find any evidence that the machine-coded complexity measure is predictive of greater item nonresponse. Table 22.8 presents the full regressions. The measure of how unfamiliar questions are is a significant and positive predictor of item nonresponse for the United States and both modes of the Romania survey. In Guatemala, the coefficient on unfamiliar is positive,

**FIGURE 22.5** Relationship between Machine-Coded Complexity Scores: Correlograms and Scree Plots from Principal Component Analysis



Source: Original figure for this publication.

Note: Crosses mark correlations that are insignificant at the 5 percent level. FEVS = Federal Employee Viewpoint Survey; SMOG = simple measure of gobbledygook.



**TABLE 22.7 Impact of Machine-Coded Complexity on Nonresponse Rate**

	Guatemala	Romania (F2F)	Romania (online)	United States (FEVS)
Machine-coded complexity	−0.008 (0.006)	0.009 (0.010)	0.022 (0.015)	0.010*** (0.003)
Controls	Yes	Yes	Yes	Yes
<i>N</i>	378,472	181,614	161,793	667,425
Adjusted <i>R</i> <sup>2</sup>	0.002	0.014	0.027	0.007

Source: Original table for this publication.

Note: Standard errors are in parentheses and are clustered using multiway clustering at the level of an individual and a question throughout. All columns report OLS (ordinary least squares) estimates. The dependent variable in all columns is a dummy indicating whether a respondent answered “I don’t know,” refused to answer, or skipped a particular question. Machine-coded complexity is calculated as the first principal component across nine different machine-coded complexity scores (number of characters, number of words, Flesch’s Reading Ease Score, Flesch-Kincaid Readability Score, Gunning’s Fog Index, SMOG Index, Coleman’s Readability Formula, Dale-Chall Readability Formula, and lexical diversity), as described in detail in appendix J. Controls at the individual level include gender (0 = female; 1 = male), education (0 = below university level; 1 = university level), occupational status (0 = nonmanager; 1 = manager), experience in public administration (0 = more than 10 years; 1 = fewer than 10 years), and job satisfaction (0 = respondent not satisfied with his or her job; 1 = respondent satisfied with his or her job; in Guatemala, a pay-satisfaction variable is used instead). The control at the organizational level is the response rate (0–1 scale) and, at the question level, the position of a question within a questionnaire. FEVS = Federal Employee Viewpoint Survey; F2F = face-to-face.

Significance level: \* = 10 percent, \*\* = 5 percent, \*\*\* = 1 percent.

**TABLE 22.8 Full Model: Impact of Sensitivity, Complexity, Unfamiliarity, and Machine-Coded Complexity**

	Guatemala	Romania (F2F)	Romania (online)	United States (FEVS)
Sensitivity	−0.002 (0.007)	0.002 (0.006)	−0.008 (0.010)	0.003 (0.003)
Complexity	−0.003 (0.005)	−0.011 (0.008)	−0.026 (0.015)	−0.002 (0.003)
Unfamiliarity	0.010 (0.006)	0.109*** (0.016)	0.217*** (0.026)	0.028*** (0.008)
Machine-coded complexity	−0.009 (0.009)	−0.017* (0.008)	−0.029 (0.015)	0.003 (0.003)
Controls	Yes	Yes	Yes	Yes
<i>N</i>	378,472	181,614	161,793	667,425
Adjusted <i>R</i> <sup>2</sup>	0.003	0.064	0.103	0.013

Source: Original table for this publication.

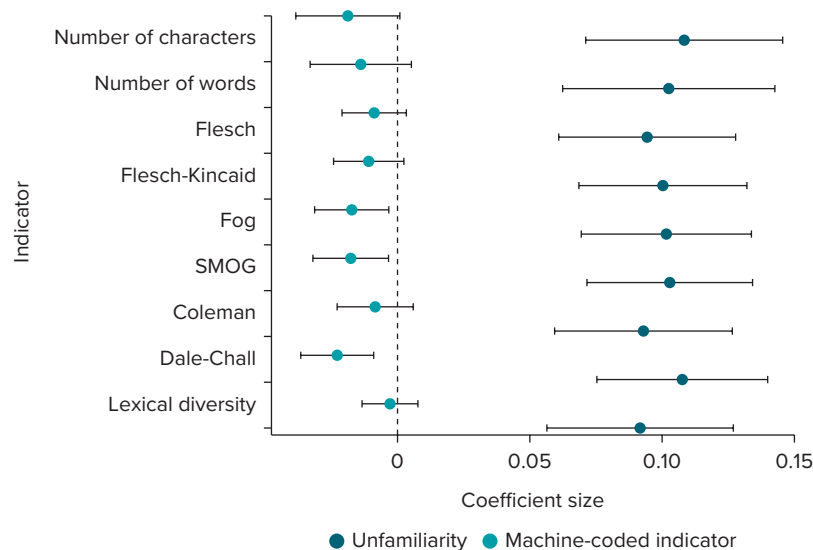
Note: Standard errors are in parentheses and are clustered using multiway clustering at the level of an individual and a question throughout. All columns report OLS (ordinary least squares) estimates. The dependent variable in all columns is a dummy indicating whether a respondent answered “I don’t know,” refused to answer, or skipped a particular question. Sensitivity and complexity scores are calculated as z-scores estimated across questions in a given survey. Unfamiliarity is calculated as a mean value of the “information retrieval: unfamiliarity” and “information integration: scope of information” subdimensions. Machine-coded complexity is calculated as the first principal component across nine different machine-coded complexity scores (number of characters, number of words, Flesch’s Reading Ease Score, Flesch-Kincaid Readability Score, Gunning’s Fog Index, SMOG Index, Coleman’s Readability Formula, Dale-Chall Readability Formula, and lexical diversity), as described in detail in appendix J. Controls at the individual level include gender (0 = female; 1 = male), education (0 = below university level; 1 = university level), occupational status (0 = nonmanager; 1 = manager), experience in public administration (0 = more than 10 years; 1 = fewer than 10 years), and job satisfaction (0 = respondent not satisfied with his or her job; 1 = respondent satisfied with his or her job; in Guatemala, a pay-satisfaction variable is used instead). The control at the organizational level is the response rate (0–1 scale) and, at the question level, the position of a question within a questionnaire. FEVS = Federal Employee Viewpoint Survey; F2F = face-to-face.

Significance level: \* = 10 percent, \*\* = 5 percent, \*\*\* = 1 percent.

although smaller in size, and just misses the 10 percent threshold of statistical significance. The coefficients vary in size from 0.010 in Guatemala to 0.217 in the online mode of the Romania survey. The coefficients for sensitivity and the restricted measure of hand-coded complexity are insignificant and small across all the models.

To illustrate the relative predictive power of the framework relative to machine-coded methods, figure 22.6 also presents the coefficient sizes of each individual measure of machine-coded readability

**FIGURE 22.6 Machine-Coded Complexity Indicators, Romania (F2F)**



Source: Original figure for this publication.

Note: Values show the size of the coefficients of machine-coded complexity indicators when they are entered individually into a regression model with the standard dependent and control variables. The size of the coefficient for unfamiliarity entered into the same regression is shown for comparison. Error bars indicate 95 percent confidence intervals. F2F = face-to-face; SMOG = simple measure of gobbledygook.

and the unfamiliarity index (see appendix J to get details on each of the individual readability scores presented). For ease of presentation, we present coefficients from the Romanian face-to-face survey only, but the patterns are similar throughout. Measuring lack of familiarity clearly has a far greater predictive ability than any of the syntax-based, machine-coded measures.

## CONCLUSION

The loss of precision and potential biases introduced by item nonresponse can hinder valid inference from surveys of public servants. Why do public servants respond to some questions but not others? The importance of this question stems from the proliferation of such surveys and their use for management reforms in government. Yet, to our knowledge, prior studies have not assessed item nonresponse in surveys of public officials.

This chapter contributes to addressing this gap. Building on the survey methodology literature, we design a unique coding framework to coherently assess the roles of question complexity and sensitivity in nonresponse in surveys of public servants. We apply this framework to governmentwide surveys of public officials in Guatemala, Romania, and the United States. As in the existing literature, we find that complexity matters for item nonresponse. Contrary to much prior work on item nonresponse, however, public servants do not seem to shy away from questions that are complex due to, for instance, syntax, computational intensity, or the number of response options (see, for example, Knäuper et al. 1997). As we argued in the introduction, this may be because public officials tend to be more educated and more accustomed to complex technical language in their day-to-day bureaucratic work. As such, they may be better able to cope with these dimensions of complexity. We find that asking public officials about issues with which they have lower familiarity is the feature of question design that is most robustly associated with item nonresponse. Questions that ask for assessments of public sector organizations as a whole or departments within them, for instance, lead to greater item nonresponse than questions about public servants themselves. By contrast, the findings provide little evidence that public officials shy away from answering sensitive questions. This does not, however, imply that responses to sensitive questions are not biased.

The implication for survey designers is clear: asking about topics public officials are less familiar with—such as their organizations or departments, rather than their immediate work environment—is associated with greater item nonresponse, with concomitant concerns about greater variance and potential biases in estimates. Where data aim to assess practices in larger units or institutions, it would thus be preferable, from a nonresponse perspective, to ask respondents about their individual-level experiences with organizational practices and aggregate these.

We have also compared the predictive ability of the findings to models that include machine-coded measures of complexity. The findings underscore the importance of manual assessments by survey designers to assess question complexity. While machine-coded estimates have some predictive power, this was eclipsed by the manual coding approach, once it was added to the models. Algorithms themselves appear to be an imprecise guide when assessing question complexity in surveys of public servants.

Future research could, in the first place, use the coding framework to understand whether our findings travel beyond Guatemala, Romania, and the United States. The diverse case contexts give us confidence that the findings are generalizable. Probing generalizability should not only extend to testing different country settings but also different survey administrators. The noticeably lower item nonresponse in the FEVS compared to the World Bank–administered surveys may reflect differential levels of trust in the survey administrator itself, for instance. The framework could equally be employed in employee surveys in private sector companies. One worthwhile area of investigation is to understand whether the findings are unique to public officials or would apply similarly to (educated) private sector administrators in a workplace survey.

Survey designers in the public service can utilize the coding framework to adjust survey questions in terms of their complexity and sensitivity. They can randomly roll out survey variations with different levels of these concepts—in particular, unfamiliarity—and assess experimentally whether this leads to improvements in item response rates in their setting.

The limitations of the findings should be kept in mind. In the first place, we only assess item nonresponse. Other threats to validity—such as overall survey nonresponse or response bias—may be of equal or greater concern. Sensitivity, for instance, was not robustly associated with greater item nonresponse across all of the surveys but may well lead to significant response bias.

Moreover, the inferences are necessarily limited by the number of surveys (three countries) and the types of questions included. In particular, the surveys contained relatively few highly sensitive questions (see figure 22.4), which might partially explain the null results obtained. It is possible that more discernible patterns in item nonresponse could be observed in surveys focused more squarely on sensitive topics—say, for instance, a corruption survey.

Overall, we present an analytically coherent approach to assessing survey item nonresponse that highlights a particular aspect of complexity—unfamiliarity—as the fundamental driver of nonresponse.

## NOTES

The authors would like to thank Lior Kohanan, Miguel Mangunpratomo, and Sean Tu for excellent research assistance, Kerenssa Kay for guidance and advice, and seminar participants at the World Bank for their comments.

1. According to the Worldwide Bureaucracy Indicators published by the World Bank, the share of publicly paid employees with tertiary education across the world is 54.2 percent, whereas in the private sector, it is around half of that: 26.9 percent (average over 2010–18).
2. More information on the surveys used and the reason for their selection is presented in the methodology section of this chapter.
3. Apart from the degree of complexity, which is a stable feature of a question, the likelihood of engaging in satisficing also depends on respondents' characteristics that might increase or decrease their cognitive capacity (for example, age, education, or tiredness) and on their willingness to answer the question. Contextual variables, like the pace at which the interviewer conducts an interview or time pressure (for example, having only a 20-minute slot to take a survey), can also impact the degree to which respondents are willing and able to engage in high cognitive effort (Fazio and Roskos-Ewoldsen 2005; Lessler, Tourangeau, and Salter 1989).

4. For a more in-depth discussion, see, for example, Paulhus (2002).
5. This might be a particular concern in restricted-sample settings, like the ones in which public administration surveys are usually conducted. This is because “individuals who complete surveys may worry that their unique responses to demographic questions could allow researchers to identify them, especially if they are part of a known sample, such as a survey conducted within one’s workplace” (McNeeley 2012, 4380).
6. Some other methods, like the randomized response technique, actually lead to higher item nonresponse, but due to the convoluted instructions they entail rather than the nature of the question itself.
7. Bais et al. (2019) do not disaggregate sensitivity to the same extent we do and, more importantly, do not assess whether their measures of complexity and sensitivity predict item nonresponse.
8. The exclusion of this variable does not change the substantive conclusions.
9. This is consistent with prior work that associates online surveys with higher nonresponse than face-to-face surveys (Heerwegh and Loosveldt 2008).
10. The decision was made based on the  $p$ -values of the EFA models. In Guatemala, the model was significant at the 5 percent level only when five factors were used, but for cross-country consistency, we employ a four-factor model throughout the analyses.
11. In the following regression tables, we present results only with the standard set of controls. The results, however, hold in unconditional regressions as well.

## REFERENCES

- Anderson, N. 1971. “Integration Theory and Attitude Change.” *Psychological Review* 78 (3): 171–206.
- Bais, F., B. Schouten, P. Lugtig, V. Toepoel, J. Arends-Töth, S. Douhou, N. Kieruj, M. Morren, and C. Vis. 2019. “Can Survey Item Characteristics Relevant to Measurement Error Be Coded Reliably? A Case Study on 11 Dutch General Population Surveys.” *Sociological Research Methods and Research* 48 (2): 263–95.
- Belson, W. A. 1981. *The Design and Understanding of Survey Questions*. Aldershot, UK: Gower.
- Berger, I., and A. Mitchell. 1989. “The Effect of Advertising on Attitude Accessibility, Attitude Confidence, and the Attitude-Behavior Relationship.” *Journal of Consumer Research* 16 (3): 269–79.
- Bradburn, N., S. Sudman, E. Blair, and C. Stocking. 1978. “Question Threat and Response Bias.” *Public Opinion Quarterly* 42 (2): 221–34.
- Coutts, E., and B. Jann. 2011. “Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT).” *Sociological Methods and Research* 40 (1): 169–93.
- De Leeuw, E. 1992. *Data Quality in Mail, Telephone and Face to Face Surveys*. Amsterdam: Netherlands Organization for Scientific Research.
- Edwards, J., M. Thomas, P. Rosenfeld, and S. Booth-Kewley. 1997. *How to Conduct Organizational Surveys: A Step-by-Step Guide*. Thousand Oaks, CA: SAGE Publications.
- Faaß, T., L. Kaczmirek, and A. Lenzner. 2008. “Psycholinguistic Determinants of Question Difficulty: A Web Experiment.” Paper presented at the seventh International Conference on Social Science Methodology.
- Fazio, R. 1986. “How Do Attitudes Guide Behavior?” Chap. 8 in *Handbook of Motivation and Cognition: Foundations of Social Behaviour*, edited by R. Sorrentonio and E. Higgins, 204–43. New York: Guilford Press.
- Fazio, R. 1989. “Attitude Accessibility, Attitude-Behavior Consistency, and the Strength of the Object-Evaluation Association.” *Journal of Experimental Social Psychology* 18 (4): 339–57.
- Fazio, R., and D. Roskos-Ewoldsen. 2005. “Acting as We Feel: When and How Attitudes Guide Behavior.” In *Persuasion: Psychological Insights and Perspectives*, edited by T. Brock and M. Green, 41–62. Thousand Oaks, CA: SAGE Publications.
- Galletly, C., and S. Pinkerton. 2006. “Conflicting Messages: How Criminal HIV Disclosure Laws Undermine Public Health Efforts to Control the Spread of HIV.” *AIDS and Behaviour* 10: 451–61.
- Gnambs, T., and K. Kaspar. 2015. “Disclosure of Sensitive Behaviors across Self-Administered Survey Modes: A Meta-analysis.” *Behaviour Research Methods* 47: 1237–59.
- Haziza, D., and G. Kuromi. 2007. “Handling Item Nonresponse in Surveys.” *Journal of Case Studies in Business, Industry and Government Statistics* 1 (2): 102–18.
- Heerwegh, D., and G. Loosveldt. 2008. “Face-to-Face versus Web Surveying in a High-Internet-Coverage Population: Differences in Response Quality.” *Public Opinion Quarterly* 72 (5): 836–46.
- Höglinger, M., B. Jann, and A. Diekmann. 2016. “Sensitive Questions in Online Surveys: An Experimental Evaluation of Different Implementations of the Randomized Response Technique and the Crosswise Model.” *Survey Research Methods* 10 (3): 171–87.

- Holbrook, A., Y. Cho, and T. Johnson. 2006. "The Impact of Question and Respondent Characteristics on Comprehension and Mapping Difficulties." *Public Opinion Quarterly* 70 (4): 565–95.
- Just, M., and P. Carpenter. 1992. "A Capacity Theory of Comprehension: Individual Differences in Working Memory." *Psychological Review* 99 (1): 122–49.
- Kim, S., and S. Kim. 2016. "Social Desirability Bias in Measuring Public Service Motivation." *International Public Management Journal* 19 (3): 293–319.
- Knäuper, B., R. Belli, D. Hill, and R. Herzog. 1997. "Question Difficulty and Respondents' Cognitive Ability: The Effect on Data Quality." *Journal of Official Statistics* 13 (2): 181–99.
- Krosnick, J. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5 (3): 213–36.
- Krosnick, J., and S. Presser. 2009. "Question and Questionnaire Design." In *Handbook of Survey Research*, 2nd edition, edited by J. Wright and P. Marsden. San Diego, CA: Elsevier.
- Krumpal, I. 2013. "Determinants of Social Desirability Bias in Sensitive Surveys: A Literature Review." *Quality and Quantity* 47: 2025–47.
- Lenski, G., and J. Leggett. 1960. "Caste, Class, and Deference in the Research Interview." *American Journal of Sociology* 65 (5): 463–67.
- Lensvelt-Mulders, G. 2008. "Surveying Sensitive Topics." In *International Handbook of Survey Methodology*, edited by E. de Leeuw, J. Hox, and D. Dillman, 461–578. New York: Routledge.
- Lessler, J., R. Tourangeau, and W. Salter. 1989. *Questionnaire Design in the Cognitive Research Laboratory*. Vital and Health Statistics 6, Cognitive and Survey Measurement 1. DHHS Publication No. (PHS) 89-1076. Hyattsville, MD: US Department of Health and Human Services.
- McNeeley, S. 2012. "Sensitive Issues in Surveys: Reducing Refusals While Increasing Reliability and Quality of Responses to Sensitive Survey Items." In *Handbook of Survey Methodology for the Social Sciences*, edited by L. Gideon, 4377–96. New York: Springer.
- OPM (Office of Personnel Management). 2019. *2019 Office of Personnel Management Federal Employee Viewpoint Survey: Technical Report*. Washington, DC: OPM.
- Paulhus, D. 1984. "Two-Component Models of Socially Desirable Responding." *Journal of Personality and Social Psychology* 46 (3): 598–609.
- Paulhus, D. 2002. "Socially Desirable Responding: The Evolution of a Construct." In *The Role of Constructs in Psychological and Educational Measurement*, edited by H. Braun, D. Jackson, and D. Wiley, 49–69. Mahwah, NJ: Erlbaum.
- Rässler, S., and R. T. Riphahn. 2006. "Survey Item Nonresponse and Its Treatment." *Allgemeines Statistisches Arch* 90: 217–32.
- Sakshaug, J., T. Yan, and R. Tourangeau. 2010. "Nonresponse Error, Measurement Error, and Mode of Data Collection: Tradeoffs in a Multimode Survey of Sensitive and Non-sensitive Items." *Public Opinion Quarterly* 74 (5): 907–33.
- Tourangeau, R. 1984. "Cognitive Sciences and Survey Methods." In *Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines*, edited by T. Jabine, M. Straf, J. Tanur, and R. Tourangeau, 73–100. Washington, DC: National Academy Press.
- Tourangeau, R., R. Groves, and C. Redline. 2010. "Sensitive Topics and Reluctant Respondents: Demonstrating a Link between Nonresponse Bias and Measurement Error." *Public Opinion Quarterly* 74 (3): 413–32.
- Tourangeau, R., and K. A. Rasinski. 1988. "Cognitive Processes Underlying Context Effects in Attitude Measurement." *Psychological Bulletin* 103 (3): 299–314.
- Tourangeau, R., L. Rips, and K. Rasinski. 2000. *The Psychology of Survey Response*. Cambridge, UK: Cambridge University Press.
- Tourangeau, R., and T. Smith. 1996. "Asking Sensitive Questions: The Impact of Data Collection Mode, Question Format, and Question Context." *Public Opinion Quarterly* 60 (2): 275–304.
- Tourangeau, R., and T. Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133 (5): 859–83.
- Tversky, A., and D. Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science* 185 (4157): 1124–31.
- World Bank. 2020a. *Final Field Report: Encuesta General de Servidores Públicos y Contratistas del Organismo Ejecutivo y Entidades Descentralizadas 2019*. Washington, DC: World Bank.
- World Bank. 2020b. *Selecting the Right Staff and Keeping Them Motivated for a High-Performing Public Administration in Romania: Key Findings from a Public Administration Employee Survey*. Washington, DC: World Bank.
- Yan, T., and R. Tourangeau. 2008. "Fast Times and Easy Questions: The Effects of Age, Experience and Question Complexity on Web Survey Response Times." *Applied Cognitive Psychology* 22 (1): 51–68.
- Zaller, J. 1992. *The Nature and Origins of Mass Opinion*. Cambridge, UK: Cambridge University Press.

## CHAPTER 23

# Designing Survey Questionnaires

## Should Surveys Ask about Public Servants' Perceptions of Their Organization or Their Individual Experience?

*Kim Sass Mikkelsen and Camille Mercedes Parker*

### SUMMARY

Civil service surveys are often interested in organizational aggregates and comparisons across organizations. Therefore, the choice of question referent is important in questionnaire design. Should survey questions refer to individual employees or to employees' assessments of their organizations? This chapter provides tools for thinking through this choice. Moreover, experimental evidence from representative public service surveys in Romania and Guatemala shows that the choice of referent matters to how employees respond. Finally, the chapter provides evidence that organizational referents can help reduce socially desirable responding, particularly for highly sensitive questions, and that referent effects may be larger for attitudes and behaviors that are uncommon, but that the size of referent effects beyond this is difficult to predict.

### ANALYTICS IN PRACTICE

- Many civil service surveys are centrally interested in organizational aggregates. Therefore, the choice of question referent is important in questionnaire design. Should questions refer to individual employees or to employees' assessments of their organizations?
- Inside organizations, perceptions of management practices are often only weakly correlated across respondents, suggesting that they are not organizational *constructs*. Organizational referents can—but often do not—better enable survey questions to reflect organizational constructs.

---

Kim Sass Mikkelsen is an associate professor of politics and public administration at Roskilde University. Camille Mercedes Parker is an economist at the United States Agency for International Development.



- Experimental evidence from representative public service surveys in Romania and Guatemala shows that the choice of referent matters to how employees respond.
- We provide evidence that organizational referents may help reduce socially desirable responding, particularly for highly sensitive questions.
- We examine, but uncover little systematic evidence for, a set of other factors that could conceivably influence question-referent effects. We conclude that organizational referents may be less useful in situations where attitudes and behaviors are uncommon because respondents may not have the needed information to answer them. Beyond this, however, the size of referent effects is difficult to predict.

## INTRODUCTION

Many civil service surveys are centrally interested in organizational aggregates. Which surveyed organization has the highest level of job satisfaction among its employees? Which organizations need additional ethics training to keep up with the ethical awareness of employees in other organizations? Questions such as these are core both to internal government benchmarking and, since aggregates attached to recognizable labels (like organization names) are simple to interpret, to government communication of data from civil service surveys.

The focus on organizational aggregates has an intuitive implication for how questions should be asked in civil service surveys: ask civil servants to evaluate their organizations. Indeed, practitioners and academics alike routinely ask civil servants for such evaluations. For example, the United States Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS) asks its respondents to evaluate the extent to which “employees are protected from health and safety hazards on the job” as a measure of workplace safety (OPM 2018). This practice is sensible. If the target of evaluations is the organization, it seems reasonable to align the *level of measurement*—the level in reference to which respondents are asked to provide answers—with the level at which claims are made (Klein, Dansereau, and Hall 1994). *Referents*, the entities to which a survey question refers, can be sensibly chosen to reflect the entities researchers wish to learn about. The question cited above is an example of the use of *organizational referents* in civil service surveys.

It is not always clear, however, that the organization is the most appropriate or most useful level of measurement. While recognizable labels make organizational comparisons simple and appealing, using organizational referents implies one of two claims: that the question measures the respondent’s *perceptions* of his or her organizational surroundings, or that the subject of the question is an *organization-level phenomenon* (Klein, Dansereau, and Hall 1994; Klein and Kozlowski 2000). In the first instance, *top-down* claims can be made about how respondents react to their perceptions of organizational characteristics, management practices, leadership, culture, and so on. In the second instance, *bottom-up* claims can be made about organization-level phenomena principally detached from any individual public servant’s experiences and beliefs. Both of these claims may be true, but they are infrequently stated explicitly.

Levels of measurement have been subject to contention in leadership research (for example, Schriesheim, Wu, and Scandura 2009), organizational research (for example, Baltes, Zhdanova, and Parker 2009; Chan 1998; Klein and Kozlowski 2000), and survey methods research (for example, Blair, Menon, and Bickart 2004). However, the issue is rarely discussed in the inherently multilevel field focused on civil service surveys. Is the common practice of asking civil service survey questions at the organizational level sensible? Or is the use of *individual referents*, asking respondents to provide information about themselves, a more appropriate strategy? And does the choice matter for survey results?

What is at issue is not whether the level of analysis should match the level at which claims and comparisons are made. There is already good evidence that these levels should match and that the consequence of

mismatches is potentially biased results (for example, Gingerich 2013). Instead, we examine the advantages and disadvantages of using individual versus organizational referents in civil surveys. Should the level of *measurement* match the level at which claims and comparisons are made as well? In which situations should the level of measurement be the individual respondent, and in which should it be individuals' assessments of their employing organizations?

Intuitively, the answer parallels the match between claims and levels of analysis. If one is interested in individual public servants, individual referents should be used. If, by contrast, one is interested in organizations, organizational referents should be used. However, this answer is too simple. It underestimates the complexity of the consequences of choosing question referents. Our chapter describes some of this complexity and provides guidelines for understanding what is at stake in choosing question referents and when to choose which referent.

We are—as far as we know—the first to assess the issue of referent choice in civil service survey design. Yet the public organizational setting likely matters. Organizational referents require information from respondents that civil servants may not possess to the same degree private sector employees do, for instance. Public organizations are frequently very large in terms of both personnel and budget and are often hierarchically organized into relatively segmented and informationally insular parts (for example, Eggers 2007). This can make organizational-referent questions difficult for a public official in one part of an organization to answer due to a simple lack of knowledge about other parts of that same organization (cf. Homburg et al. 2012).

For organizations like ministries, this problem may even grow with managerial reforms that further segment and fragment the ministerial hierarchy into deliberately insular agencies (Dunleavy et al. 2006). In a sense, organizational referents in civil service surveys may have to grapple with issues similar to those that whole-of-government approaches to public sector organizations were intended to solve: information and knowledge can have a hard time traversing the organizations that respondents are asked to evaluate (for example, Christensen and Lægreid 2007).

Our advice to civil service survey designers is not to abandon one question referent in favor of another. Instead, we provide a set of important considerations that designers can use when choosing referents. In particular, designers should consider:

- Whether what they are measuring is, conceptually, an organizational phenomenon. Does it make sense to think of all respondents within an organization rating the same entity when responding? If designers are not measuring an organizational phenomenon, organizational referents are less attractive.
- How sensitive their measures are. Respondents tend to respond as they believe is socially desirable when questions are sensitive, and this effect is more pronounced for questions about them as individuals. If questions are very sensitive, organizational referents may be more attractive.
- How easily accessible to respondents the information required for the measure is. Respondents often have better access to their own experiences, beliefs, and attitudes than those of their colleagues. If questions require information that is not readily available to respondents, organizational referents are less attractive.

These conclusions are based partly on a conceptual discussion and partly on empirical evidence. Empirically, we use experiments embedded in two civil service surveys. We embedded experiments in a survey of more than 6,000 civil servants in Romania's central government, randomly assigning respondents to answer questions about human resource management practices—specifically, recruitment, promotion, dismissal, and turnover intent—using individual or organizational referents. We embedded a similar experiment in a survey of more than 3,000 civil servants in Guatemala's central government.

The basic thrust of the experiment is that, if referent choice matters, otherwise similar questions using different referents will result in different average responses. If referents do not matter, the average employee evaluation of the organization will correspond to the average of the employees' evaluation of themselves. Thus, the experiment can provide evidence that referents matter, the core interest of this chapter. The drawback is that the sources of divergences are harder to determine. We do conduct a series of tests attempting

to determine sources, but the question we can answer most clearly is whether referents matter. Despite its simplicity, a strong answer to this question is useful to survey designers, many of whom do not seem to know whether referents matter or how they matter to the responses they get.

Beyond the questions it can answer, this experimental approach is valuable for the strength of our conclusions. And it sets our study apart from previous examinations of the use of referents in organizational surveys (for example, Baltes, Zhdanova, and Parker 2009; Klein et al. 2001). Prior examinations of referent issues have asked the same respondents to provide information both about themselves (using individual referents) *and* about their organizations (using organizational referents). This is needed, of course, to show that each of the two referents contributes separate information (Klein et al. 2001). However, it creates the risk that respondents anchor their responses to one set of questions to their answers to the other set of questions in order to appear consistent, or that they respond to both sets of questions relative to one another, either to maintain that they are “above average” on relevant metrics (Guenther and Alicke 2010) or because they form their answers relative to comparisons with significant colleagues (Baltes, Zhdanova, and Parker 2009). Thus, responses to questions with individual referents can affect subsequent responses to question sets with organizational referents and vice versa. Our experimental design avoids this issue, permitting a causal assessment of the relative differences between responses stemming from the two referents.<sup>1</sup>

We proceed in four steps. In section 2, we discuss what difference organizational as opposed to individual referents might make theoretically. We focus particularly on concept levels, socially desirable responding, and information availability. In section 3, we describe our survey experiments. Section 4 presents our results. Section 5 contains our discussion of these results for the design of civil service surveys.

## WHAT IS AT STAKE?

In this section, we provide a more detailed account of the already-noted reasons why the choice of question referent might matter. This takes us into the psychology of survey response and questions about levels of theory and measurement from organizational studies. But the point is not the theory. Rather, we aim to provide readers with a rough and simple understanding of the stakes in choosing between individual and organizational referents. Table 23.1 provides an overview of the arguments we discuss. These fall along three main lines: the match between the measure and the target entity of interest, socially desirable responding, and the informational requirements placed on respondents.

### Do Analyses Concern Individuals or Organizations?

At base, the choice of referent should reflect the interest of subsequent analyses. If the interest is in measuring, comparing, or benchmarking organizations, organizational referents appear to be the obvious choice because they create a clear match between the entities in subsequent analyses (the target) and the measure. However, this is not as obvious as it would at first appear. Table 23.2 provides an overview of the discussion.

Table 23.2 distinguishes between referents, the entities referred to in survey questions, and target entities, the entities the survey aims to learn something about. The intuition is that referents should be chosen to match the downward diagonal of the table. Inquiries with an individual focus should ask individual-referent questions, while organizational inquiries should use organizational referents.

The first half of this intuition holds. Inquiries with an individual focus should likely ask about individuals. But the second half of the intuition is more complicated. There are three ways of thinking about settings where questions either use organizational referents or aim to learn about organizations: the *top-down* perspective, which asks about organizations to learn about individual employees, the *bottom-up* perspective, which asks about organizations to learn about organizations (when possible) (Klein and Kozlowski 2000),

**TABLE 23.1 Advantages and Disadvantages of Organizational and Individual Referents When Used for Calculating and Analyzing Organizational Aggregates**

Type of cost or benefit	Organizational referents	Individual referents
Conceptual	+ Match between target and measure – Disagreement	+ No agreement requirement – Possible mismatch between target and measure
Measurement	+ Decreased social-desirability bias – Informational requirements	+ Fewer informational requirements – Social-desirability bias

Source: Original table for this publication.

Note: This table shows a summary of the discussion in the three following subsections. Columns represent question referents (organizational vs. individual). Rows are divided into conceptual concerns (discussed in the first subsection) and measurement concerns (discussed in the second and third subsections). Plus signs indicate competitive advantages relative to the referent in the other column; minus signs indicate competitive disadvantages. Advantages and disadvantages are relative to data used for calculating organizational aggregates. Some points are not relevant in other contexts (for example, “match between target and measure” is not a competitive advantage for organizational-referent questions if an individual’s beliefs are the target, as in the top-down perspective).

**TABLE 23.2 Question Referents and Target Entities**

		Target entity	
		Organization	Individual employee
Question referent	Organizational	Bottom-up	Top-down
	Individual	Summary bottom-up	Individual focus

Source: Original table for this publication.

and finally, what we call the *summary bottom-up* perspective, which asks about individuals to learn about organizations through data summaries.

The *top-down* perspective interprets organizational-referent questions as asking about respondents’ perceptions of their working environment. Even questions that appear to be intrinsically at the organizational level may be best thought of at the individual level in terms of definitions, causal efficacy, or both. For instance, Parker et al. (2003, 390) define the *psychological climate* in organizations—a term that, intuitively, has a clear organizational focus, though it does not have this connotation in the relevant literature—as “an individual’s psychologically meaningful representations of proximal organizational structures, processes, and events.” Such representations—including perceptions of management practices and attributions related to those perceptions—are often proposed as causally efficacious for important employee outcomes (for example, Nishii, Lepak, and Schneider 2008). They are related to organizational practices, but they are not themselves organization-level phenomena. Rather, it is employee perceptions or experiences that matter for outcomes. From this perspective, organizational-referent questions are not asking respondents to rate the same entity—indeed, they are, in a sense, not organizational at all. Instead, they are asking about individuals’ representations, beliefs, or experiences. From this perspective, answers to the FEVS question about whether “employees are protected from health and safety hazards on the job” can be interpreted as reflecting individual respondents’ beliefs about health and safety in their workplaces—which can be relevant to understanding their commitment to their workplaces, their job satisfaction, or their turnover intent—but not, strictly, as offering descriptions of their workplaces as they are.

The *bottom-up* perspective is more complicated. It involves interpreting respondents’ evaluations as reflecting genuine organizational constructs—that is, features of the organization—over and above the perspective of the individual respondent. It is not perceptions but features of the organization that are the target of organizational-referent questions, from this perspective. Respondents within an organization are all seen as rating the same entity with the same characteristics.<sup>2</sup> The bottom-up perspective on organizational referents assumes that the characteristic of concern in a question is a characteristic of the organization,

not of the respondent. From this perspective, answers to the question about whether “employees are protected from health and safety hazards on the job” are ratings of the organization; they ask the responding employee to evaluate the organization (principally) as a whole. Consequently, since respondents within an organization are rating the same entity, the bottom-up perspective assumes a substantial level of agreement among respondents in the same organization.

Based on the assumptions behind the bottom-up perspective, it seems reasonable to believe that using organizational referents furthers agreement on responses within organizations because individual respondents are essentially instructed to disregard their personal experiences and report using a *referent shift*. From this perspective, there may be reason to prefer organizational referents because they may further the agreement necessary for the desired bottom-up interpretation of organizational aggregates as reliable descriptions of the organization as one entity evaluated by multiple raters.

But what if respondents within organizations do not agree? The answer can be stated, likely too succinctly: then the measures do not appropriately measure an organization-level characteristic but a construct at a lower level (such as an employee perception). This brings us to the *summary bottom-up* perspective, which construes descriptions of organizations using survey data as summaries of individual perspectives. Employee responses to organizational referents can be thought of as such summaries, but they do not have the advantage of capturing the organization above individual perceptions and experiences. This is because the perspective does not consider employees as rating the organization but as providing their own views.

Uneven implementation within organizations is often proposed as a vehicle for intraorganizational differentiation in civil service management practices when these are measured using organizational referents (Bezes and Jeannot 2018; Meyer-Sahling and Mikkelsen 2016, 2020). From this argument, questions with organizational referents do not necessarily result in organization-level assessments by respondents but rather elicit the experience of respondents in their immediate working environments. The disadvantage is uncertainty about the width of the assessments provided by individual respondents if these are not at the organizational level to which survey items refer. If organizational referents do not prompt consensus on ratings of the same entity, it is not clear what level the questions measure. Instead of capturing their organizational target, organizational aggregates are reduced to *summaries* of features of lower levels, be these sections, teams, or individuals.

Indeed, when organizational aggregates of responses are seen as summaries of individual perspectives, organizations are arguably better described using individual-referent questions: the width of the assessment is determined by the question, and the result is still a useful organizational summary. The cost of this view is that organizational characteristics are redefined to nothing more than aggregates of individual answers. Organizational workplace safety, for example, becomes the proportion of employees who think their work is safe.

In sum, if civil service survey designers are primarily interested in organizational aggregates, should they ask questions with organizational referents to ensure correspondence between levels of measurement and levels of theory? It depends. If respondents’ within-organization responses are strongly correlated, individual and organizational referents are both useful measures of organizational characteristics. While they entail different perspectives on interpreting answers, and the bottom-up perspective has a more intuitive appeal, both kinds of referent can be used.

However, if responses do not strongly correlate within organizations, this indicates that the use of individual referents is preferable on a conceptual basis. The bottom-up perspective, in this situation, does not lend as much analytical leverage as the summary bottom-up perspective because responses do not reflect an organization-level construct; instead, organizational aggregates are more readily understood as summaries of employee information.

In sum, if employees’ beliefs and perceptions are of central interest—as in the right column of table 23.2—the choice of referent is conceptual, not statistical. In that case, organizational referents should be used if respondents’ beliefs about the organization are of central interest, and individual questions should be used if respondents’ own experiences and behaviors are of interest. However, if the organization is the target, the preferable choice of referent is, in part, statistical because organizational-referent



questions impose the requirement of *interrater agreement* among employees of the same organization, while individual-referent questions do not. Even such statistically based choices have conceptual consequences, however, since the bottom-up and summary bottom-up perspectives use different ideas about the composition of individual responses and hence capture somewhat different ideas about what organizational aggregates are (Chan 1998).

### Are Questions Sensitive?

It is often less embarrassing and feels less threatening to respond to a question in a socially undesirable way if the question is not about oneself. “Do you ever steal stationery from work?” is a much more sensitive question on its face than “Do colleagues in your organization ever steal stationery from work?” Consequently, many researchers utilize organizational referents not on conceptual grounds but to limit socially desirable responses. Organizational referents are used to make sensitive questions less sensitive to respondents, on the assumption that they will provide more truthful answers and avoid social-desirability bias (SDB) due to question sensitivity (for example, Graaf, Huberts, and Strüwer 2018; Meyer-Sahling and Mikkelsen 2016).

This assumption is plausible and has been indirectly tested in other fields under labels such as “proxy questioning” (Blair, Menon, and Bickart 2004) and “structured projective questioning” (Fisher 1993). For instance, in marketing, Fisher (1993) studies whether questions that ask for the opinion of others rather than the respondent’s own opinion can reduce SDB. Fisher’s finding accords with the assumptions made in analyses of civil service survey data: indirect questions reduce SDB on questions subject to social influence. Thematically closer to our purpose, Bardasi et al. (2011) find that reported male labor market participation rates dropped substantially when others provided proxy answers, rather than the men themselves. Like these approaches, the use of organizational referents is sometimes interpreted as an indirect question technique because respondents provide information about others, not about themselves.

Questions engender SDB through several channels. Questions can be intrusive, threatening, or socially undesirable (Tourangeau and Yan 2007). Intrusive questions can be seen as offensive, nosy, or taboo. Threatening questions make respondents worry about the disclosure of their responses and the negative consequences that may ensue. Finally, socially undesirable questions are questions for which certain answers violate social norms.

Disclosure threats and socially undesirable answers are particularly relevant to our discussion. In organizational settings, the disclosure of attitudes and behaviors to which colleagues, management, political superiors, the media, or the public will react negatively is a real concern. This is true of questions for which admitting to behaviors can have negative career consequences—such as admitting to kickbacks (Meyer-Sahling and Mikkelsen 2016). And it is true of questions for which agreeing or disagreeing can be seen as negative by colleagues or management and have negative consequences in terms of careers or ostracization at work. *Sensitive* questions—for example, questions about corruption or absenteeism—thus engender one form of SDB, but not the only form. Socially desirable responding can also occur for questions in which anything but a strong endorsement of the question’s content can be seen as undesirable—such as questions about helping colleagues or working hard.

If SDB were all about threats of disclosure, however, anonymity safeguards for individual responses should help the problem. Unfortunately, SDB persists—albeit to varying degrees—even when anonymity is guaranteed (Kreuter, Presser, and Tourangeau 2008).<sup>3</sup> This is why many contemporary studies of very sensitive topics, such as corruption, employ indirect questioning techniques, such as the randomized response method (for example, Gingerich 2013) or conjoint experiments (for example, Schuster, Meyer-Sahling, and Mikkelsen 2020) to protect respondents’ answers. When such techniques are too cumbersome or are not available, the use of organizational referents may be an attractive way to combat residual socially desirable responding by asking respondents about sensitive topics less directly. The cost of doing so, as we discuss below, is that organizational-referent questions on sensitive topics often place strong demands on respondents for information that may not be accessible to them.



In situations where SDB is severe and information is at least somewhat readily available to organizational outsiders, organizational aggregates may even be obtained from raters external to the organization. Such individuals will likely be less affected by SDB, although they may have other interests at stake in responding. However, using their answers comes at the cost of losing access to information from inside organizational boundaries, which may make their assessments noisy or inaccurate (for example, Razafindrakoto and Roubaud 2010). And, of course, this problem is likely to be particularly pernicious for sensitive questions, in which information is likely to be deliberately concealed from external assessment.

In sum, question sensitivity is a common reason for the use of organizational referents. There are good reasons to think this is an effective strategy, but, as far as we know, it has not been empirically examined in the context of civil service surveys. We do so below.

### Is Organizational Information Available to Respondents?

The third topic we cover concerns information. Specifically, in some circumstances, it may be difficult for respondents to have the information that organizational referents ask them to provide. When asked a question with an individual referent, respondents work to retrieve or recall information about the question (Tourangeau, Rips, and Rasinski 2000). For past behaviors, recall involves respondents' remembering what they have previously done. For beliefs or attitudes, following Zaller (1992), we can think of recall as respondents' process of deciding what beliefs or attitudes they hold, which can be either remembered or formed on the spot based on available information.

Recall and introspection are not perfectly reliable, and respondents tend to "fill in" information they are unsure about or do not recall accurately (Tourangeau, Rips, and Rasinski 2000). Yet the difficulties can multiply when questions are posed using organizational referents. Organizational referents impose an additional challenge for respondents. If organizational referents work as intended, respondents rely on different sources of information when answering questions about themselves or about others (cf. Blair, Menon, and Bickart 2004). It is reasonable to believe that information about aggregates, such as organizations, will often be harder for respondents to access, and perhaps harder to recall, than information obtained by introspection (that is, information about themselves). Consequently, respondents' beliefs about their organizational surroundings may be mistaken or biased, which may influence their responses.

When Meyer-Sahling and Mikkelsen (2016), for instance, ask respondents whether "political parties place their supporters in the ministerial structure" as a measure of personnel politicization, they are asking respondents for an evaluation they may not have sufficient information to provide accurately. Did new recruits get their positions due to political influence? Politicization may be hidden, particularly where—as in Central and Eastern Europe, where the authors collect their data—political influence over recruitment often extends to positions formally codified as career posts (for example, Meyer-Sahling 2011).

Due to these difficulties, respondents who are asked questions using organizational referents may get their answers wrong, with consequences for measurement. The literature on *establishment surveys*—instruments in which one respondent replies on behalf of an organization—assumes that respondents use records from the establishment to counteract these difficulties (Edwards and Cantor 2004). However, it is certainly optimistic to expect respondents in civil service surveys to do the same. Even if they could and were willing, many of the topics of central interest to civil service surveys—like politicization—are often not formally recorded. As such, errors rooted in mistaken beliefs are likely to persist.

Moreover, when respondents lack necessary information, they may default to public sector stereotypes or other heuristic shortcuts to construct an answer. If public servants hold views similar to the general public, for instance, they may default to considering their colleagues as stereotypically caring or dedicated (Willems 2020), irrespective of their own concrete knowledge about the caring or dedication of organizational members beyond their immediate coworkers. Or respondents may extrapolate from anecdotes or stories to a systemic evaluation, particularly if they are asked to evaluate questions on topics they view as threatening or emotionally engaging (Freling et al. 2020).

From the perspective of the response process, the built-in assumption behind the use of organizational referents can easily come to seem somewhat heroic in large and complex organizations. Findings from previous studies do not help. Baltes, Zhdanova, and Parker (2009) propose that respondents may rely on “better-off” or “worse-off” colleagues when responding to questions with organizational referents. This may bias estimates of organizational aggregates because the implicit referents that are actually used are no longer representative of the organization.<sup>4</sup> Similarly, Shah (1998) finds that job-related information is often obtained from people in similar positions, whereas organization-related information is obtained from friends within the organization. This means organizational-referent questions are answered using networked information rather than representative information or simple ratings of features of the organization.

However, organizational information may not be equally difficult to obtain in all organizations or by all public servants. When answering questions about others, respondents may start with themselves and subsequently take in the stories and observed behavior of others (Hoch 1987). This information may be sourced from networks, but there are predictable situations in which it is more likely to accurately represent the organization. In those situations, question referents are likely to matter less for responses, and hence concerns with the information requirements of organizational referents may not matter in practice.

First, drawing information from unrepresentative colleagues, stories, and observed behaviors should matter less when questions concern attitudes or behaviors that are either very rare or very common. In these situations, most colleagues, stories, and observed behaviors will provide the same information: that the attitude or behavior is very rare or very common. This means that while respondents may not, in fact, know the answer to a question using an organizational referent, their assessment is likely to be less affected by how they arrive at it. A similar point holds when most members of an organization hold roughly similar views because networked information in this situation is also more likely to be representative of the common view in the organization.

Learning is another factor that may limit how much questions using organizational referents elicit biased assessments. For instance, years of employment in an organization may improve the accuracy of reports about it (cf. Blair, Menon, and Bickart 2004). That is, respondents may learn to answer questions using organizational referents more accurately after years in an organization because they acquire more information over time.<sup>5</sup>

In sum, questions using organizational referents ask a lot of respondents informationally. Employees are asked to assess the characteristics of large and diverse organizations based on information they may not have. This is concerning because responses may come to rely on unrepresentative information, stories, observed behaviors, networks of colleagues, and social comparisons within public organizations rather than the real features of these organizations. This makes such questions less attractive where information is hard to obtain. The more we know about which respondents in which organizations are most likely to have the necessary information, however, the more we can counteract this disadvantage of organizational referents. In our analysis below, we seek to provide such knowledge, but we find that patterns are difficult to uncover.

To summarize, we arrive at the advantages and disadvantages outlined in table 23.1. Organizational referents have the advantage of matching target to measure when an inquiry is interested in describing organizations. This is the promise of the bottom-up perspective on organizational measurement. The disadvantage is that the perspective underpinning them requires substantial agreement in answers between employees within the same organization. This may not obtain. When agreement does not obtain, as our discussion of the top-down and summary bottom-up perspectives reveals, the conceptual advantage of organizational-referent questions for inquiries interested in organizations diminishes.

Moreover, asking about organizations may decrease SDB but may do so at the cost of placing large informational requirements on responding employees. Conversely, questions about respondents themselves require less external information and no within-organization agreement in responses. But this comes at the cost of greater SDB and of presenting organizational summaries rather than describing organizational features beyond individual respondents’ aggregated perspectives.

## DATA

We rely on two survey experiments to examine the questions we have raised in the previous section. We first describe the surveys in which these experiments were embedded, then the experiments themselves and how they help us gain strong leverage on question referents.

### Surveys

Our experiments were embedded in two surveys of central government public servants. We implemented the first survey in Romania between June 2019 and January 2020. Respondents were randomly assigned to face-to-face or online survey formats and partook in our experiment as part of a longer survey on civil service management practices. In all, we interviewed 3,316 respondents face-to-face (for a response rate of 92 percent) and 2,721 respondents online through Qualtrics (for a 24 percent response rate). The representativeness of our samples and the extent to which it differs according to the survey mode is covered in detail elsewhere in *The Government Analytics Handbook* (chapter 19).

We fielded the second survey in 18 Guatemalan government institutions between October and December 2019. Our experiments were embedded in a longer civil service survey. Respondents were sampled through the sample frame used for the Human Resources National Census, comprising staff lists of 14 central and four decentralized government institutions, and were asked to participate in face-to-face interviews. In all, we interviewed 3,465 respondents (for a 96 percent response rate).<sup>6</sup>

Though both surveys included responses concerning a range of civil service management practices of potential interest for questions surrounding the use of referents, we focus our attention on the analysis of the question-referent experiments. This is, as we explain next, where we get the strongest leverage on question-referent issues.

### Experiments

Our experiments all share the same essential strategy. Each survey respondent was randomly assigned to one of two survey flows. In one flow, the respondent was asked a set of questions (see below) that use organizational referents. In the other flow, the respondent was asked a set of questions differing from the first questions only in their use of individual rather than organizational referents.

Assignment to each survey flow was random for reasons of causal identification: random assignment ensures that the respondents who answered questions using individual referents and those who answered otherwise equivalent questions using organizational referents are identical, on average, on all observed and unobserved characteristics. As a result, any difference between average responses in the two flows must be due to the difference between them: whether question referents are individual or organizational. This ensures that we can causally identify the difference referents make to respondents' answers. It is the experimental setup that enables us to say with confidence that referents matter, how much, and for which organizations or groups of people.

The gist of our argument is this: if we ask some respondents a question on, say, salary satisfaction with reference to themselves and other respondents a salary-satisfaction question with reference to their organization, the average response from all respondents in an organization to each question should be the same if the question referent does not matter. The respondents who answered questions with individual referents are a random sample of all respondents and, thus, representative of them. The respondents who answered questions with organizational referents are also a random sample and, thus, representative in their views of their organization.<sup>7</sup> Therefore, any average difference between respondents assigned to different question referents must be due to the question referents.

In both surveys we fielded, we manipulated different sets of questions in this manner. In the survey in Romania, we assigned respondents to individual- or organizational-referent versions of questions surrounding recruitment (two questions), promotions (two questions), turnover (five questions), and dismissals (two questions). All questions were in five-point Likert scales. Additional follow-up questions on the use of various sources of information in recruitment and the questions asked at recruitment interviews were similarly randomized. We include these only in some of our analyses as they are scaled differently than the questions listed above. Questions were assigned to respondents in groups such that respondents either got all questions using individual referents or all questions using organizational referents. For instance, the group of respondents who received organizational-referent questions was asked the question “Please indicate the extent to which you agree or disagree with the following statements: The promotion process in my institution is fair.” By contrast, the other group of respondents, who received individual-referent questions, was asked the question “Please indicate the extent to which you agree or disagree with the following statements: The promotion process I have to go through in my institution is fair.” Appendix K.1 shows the full lists of questions in both versions.

Similarly, in the survey in Guatemala, we assigned respondents to individual- or organizational-referent versions of questions surrounding promotion confidence (one question), promotion fairness (one question), turnover (three questions), dismissals (two questions), and leadership (nine questions). Appendix K.1 shows the full lists of questions in both versions. Even where themes overlap, questions were formulated somewhat differently in Romania and Guatemala. Consequently, a comparison of results between the two countries should be made with caution.

From a design perspective, the experiments illuminate question-referent effects, but they do share a common drawback: we lack an objective benchmark for the phenomena, behaviors, or attitudes they measure. This means that while we are willing to interpret average higher scores on sensitive questions as diminishing SDB, we are often not strictly able to say whether individual or organizational referents caused the stronger method effect grounded solely in the way the question was posed. This is a weakness shared by most nonlaboratory experiments of this type, but we are still able to examine differences between individual- and organizational-referent questions, which are often informative. With this caveat noted, we proceed to our results.

## RESULTS

There is much we can examine within our framework using our data. Within the confines of this chapter, we cannot address every possible question. Instead, we opt to answer four questions directly related to the issues of substantive interest, information availability, and social desirability that we have outlined. Each subsection poses a question, which is immediately answered before detailed results are provided.

### Do Organizational-Referent Questions Reliably Reflect Organizational Characteristics?

Organizational-referent questions do not generally reflect organizational characteristics, though they often reflect them better than individual-referent questions do. For this reason, individual-referent questions may be preferred on conceptual grounds in many instances, given that organizational referents—while often resulting in increased agreement—are by no means guaranteed to ensure that questions result in clear ratings of organizational characteristics rather than summaries of individual perspectives.

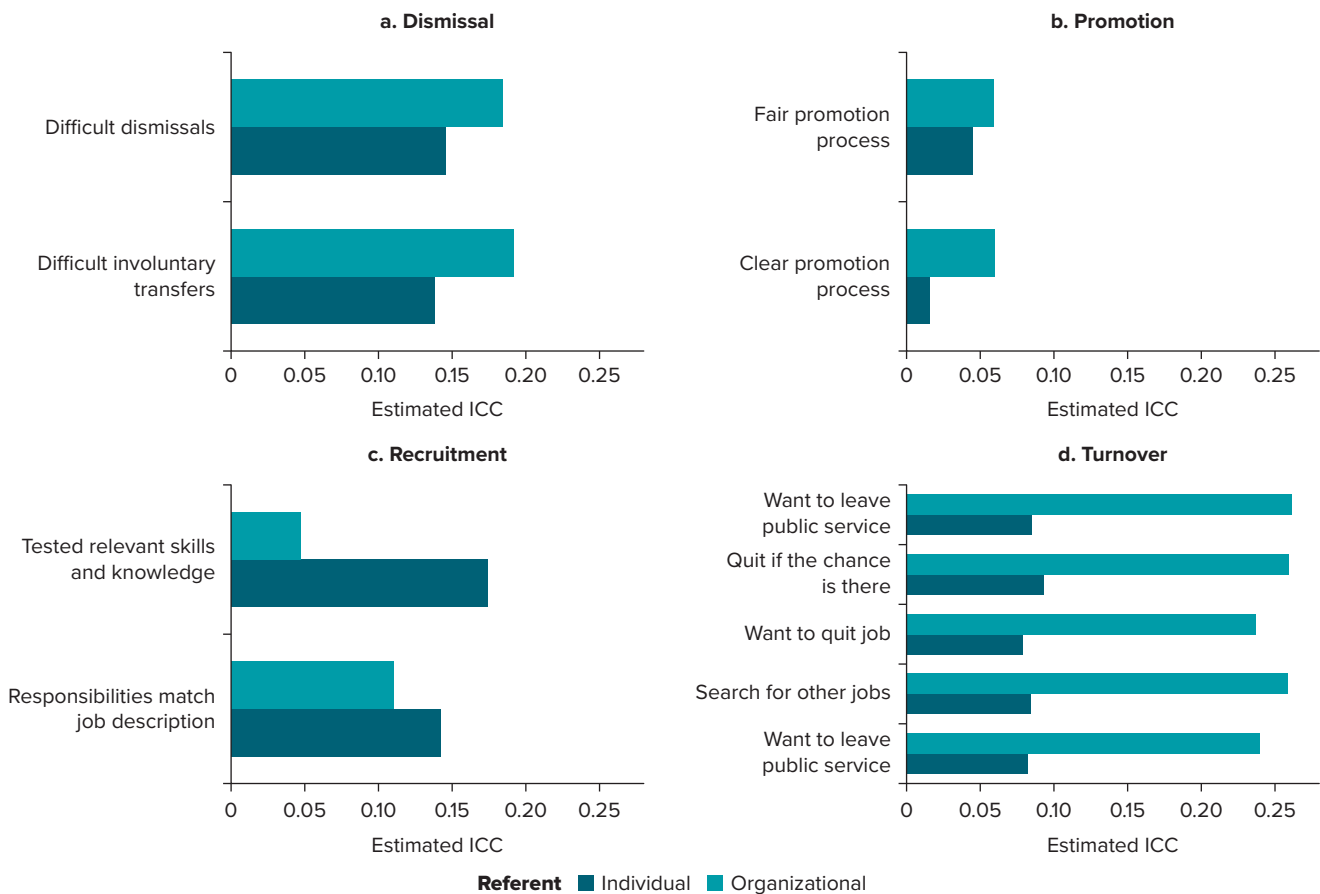
As noted, one central question for the bottom-up perspective on the utility of organizational and individual referents in civil service surveys concerns agreement within organizations. If respondents within an organization tend to agree in their responses to questions about their organization, we can more plausibly claim that their responses evaluate the same organizational phenomenon. If respondents rate the same entity in the world, they should agree in their ratings.

There are many measures of within-group agreement on survey measures. Here we opt for a common and simple measure, intraclass correlation (ICC). ICC is a measure of how much responses to questions rely on respondents' organizational setting. It can be interpreted as the percentage of variation in responses accounted for by organizational level. The higher the ICC, the more responses correlate within organizations—that is, the more respondents within organizations agree on their answers—and the more we can think of measures as reflecting objective organizational characteristics, which are simply observed and reported by respondents.

Our data permit the examination of two questions regarding ICC. First, are responses within organizations correlated to a high enough degree that we can think of the concepts they measure as genuine organization-level constructs? Second, is the correlation affected by the use of organizational or individual referents? If it is, this could indicate that organizational-referent questions can help survey designers elicit answers that characterize organizations from the appealing bottom-up perspective. Other things being equal, responses to questions about organizations *should* correlate more within organizations than responses to individual-referent questions.

Figures 23.1 and 23.2 examine these questions using the surveys from Romania (figure 23.1) and Guatemala (figure 23.2). Analysis of the Romanian data reveals that there is non-negligible agreement on responses within organizations for several questions but not for others. For some questions, the ICC is low enough that we might ask whether questions using either of the two referents elicit responses that refer to the same underlying phenomenon (rather than reporting two different perspectives).<sup>8</sup>

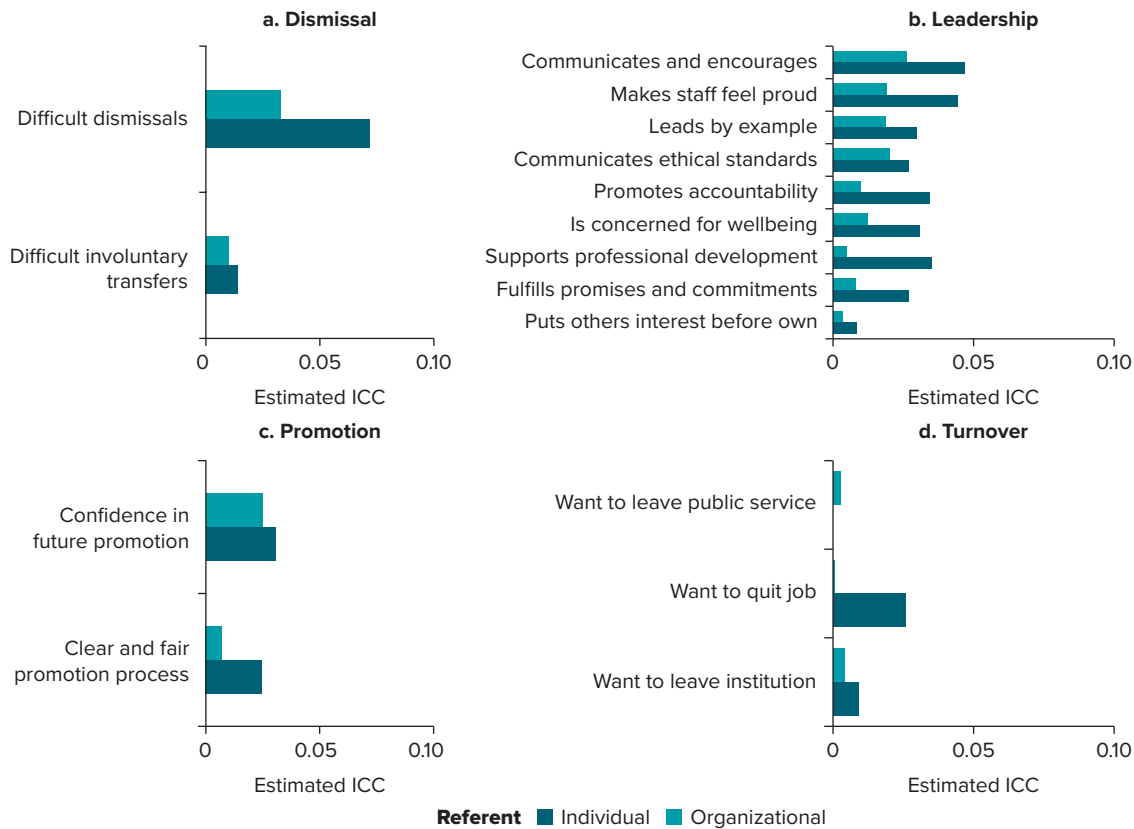
**FIGURE 23.1 Intraclass Correlations for the Romanian Data**



Source: Original figure for this publication.

Note: Bars show the calculated organizational ICC for each variable in the survey experiment in Romania, divided by HR area and treatment status. Positive differences between organizational (light blue) and individual (dark blue) referent questions indicate stronger agreement for the former than for the latter. See appendix K.1 for full items and question labels. ICC = intraclass correlation.

**FIGURE 23.2** Intraclass Correlations for the Guatemalan Data



Source: Original figure for this publication.

Note: Bars show the calculated organizational ICC for each variable in the survey experiment in Guatemala, divided by HR area and treatment status. Positive differences between organizational (light blue) and individual (dark blue) referent questions indicate stronger agreement for the former than for the latter. The horizontal axis is kept on the same scale as in figure 23.1 for ease of comparison. See appendix K.1 for full items and question labels. ICC = intraclass correlation.

Equally important for our purposes, these data show that organizational-referent questions *do* generally correlate more strongly within organizations than questions with individual referents. The expected agreement effect from organizational referents does emerge for some questions. The ICC for questions using organizational referents is higher for all but two recruitment questions in the Romanian data, though some differences are slight.

Question-referent effects are particularly pronounced for the turnover and recruitment questions. For turnover questions, within-organization agreement climbs by a factor of four. One possible explanation for this is social desirability. If respondents differ in their propensity to provide socially desirable answers more than they differ in their views on turnover intention among their colleagues, we could arrive at the pattern we observe. For now, however, this has to be considered speculative.

Somewhat puzzlingly, referent effects on recruitment items are reversed relative to what we would expect. Respondents to individual-referent questions agree more within organizations than respondents to organizational-referent questions. One possible explanation for this is that the questions using individual referent ask about recruitment processes that may have occurred years ago. This could lead to larger differences within organizations that have changed practices over time. However, our data do not reveal substantial differences in estimated ICCs if we split them along years of service.

Another possible explanation is that respondents' beliefs about public sector recruitment generally lead to the underestimation of differences between organizations, which drives down the ICC for questions using organizational referents, while individual-referent questions capture the diversity in recruitment practices.<sup>2</sup>



This is consistent with the fact that the between-organization variance of organizational-referent questions for recruitment is among the lowest in our data (alongside variables related to career advancement).

Analysis of the Guatemalan data reveals a similar pattern, although with a lower ICC across the board (figure 23.2). This offers two important lessons. First, many of the questions we examine do not appear to be statistically sound measures of bottom-up, organization-level constructs in Guatemala. The lower ICCs are due in part to the larger size of Guatemalan institutions, which leads to more variation within them. But this is precisely the point: respondents in these large organizations may be rating effectively different entities. It seems responses in our Guatemalan data are often better seen as employee perspectives, from the summary bottom-up perspective. Second, organizational referents do sometimes, as expected, help consolidate responses around agreeing ratings of organizational constructs, particularly for leadership and turnover.

What does this mean? From the bottom-up perspective of using survey responses to describe organizational characteristics, these analyses are not generally good news. Instead, they indicate that many organizational aggregates are perhaps better thought of, from the summary bottom-up perspective, as data summaries, particularly in Guatemala. That is, the summary bottom-up perspective appears to have more traction here than the pure bottom-up perspective. As noted, there may be good structural reasons for this. Public organizations are large, segmented, and complex entities in which management practices can vary by team, division, or section—particularly where management and human resources tasks are decentralized to line managers. Expecting consistent organizational characteristics to emerge under these conditions is, perhaps, expecting too much. The use of organizational referents does seem to consolidate a unified description by respondents, but to a limited degree, leaving plenty of disagreement behind.

Conceptually, then, while civil service surveys may benefit from the use of organizational referents, the big prize—the reliable description of organizational phenomena as rated by organizational members above and beyond their individual perspectives—appears elusive in our data. Given this conclusion, the question arises whether the use of organizational or individual referents matters to the data summaries both questions can provide.

### Does the Choice of Referent Matter for Responses?

Yes, in most instances, the choice of referent matters for responses, although it matters more for average responses than for relationships between response variables or for the tendency to respond at all. Average responses are sometimes higher and sometimes lower for organizational-referent questions, depending on the question. Similarly, nonresponse is sometimes more common and sometimes less common for organizational-referent questions, depending on the question. There is little systematic evidence that question referents matter to associations between different measures and less evidence still that associations are systematically stronger or weaker.

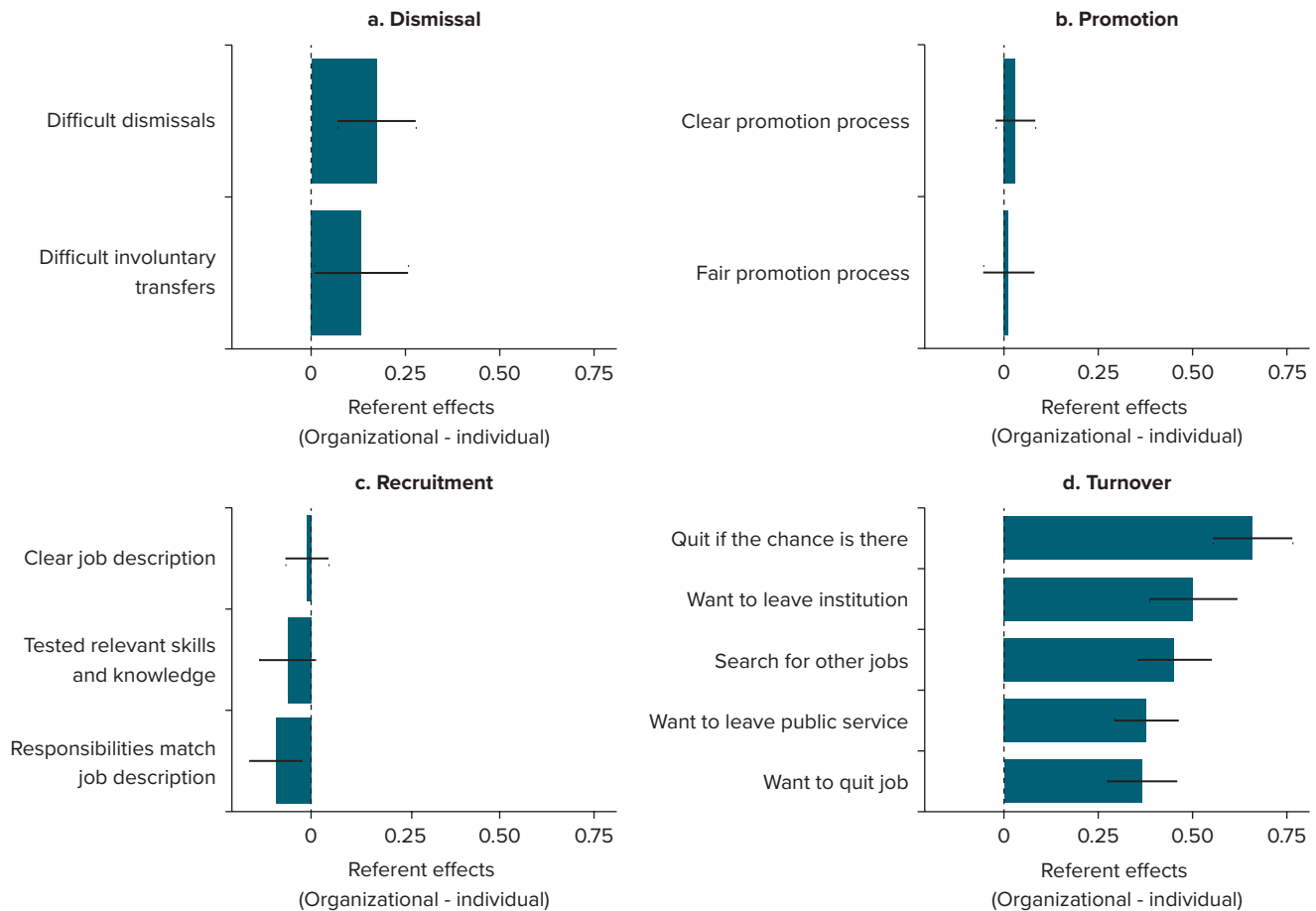
In figure 23.3, we show, using the Romanian data, the differences in average responses to questions on recruitment, turnover, dismissals, and promotion, varying only the use of organizational versus individual referents. As the figure shows, respondents who were asked about themselves rather than their colleagues are, on average,

- More convinced that they are difficult to dismiss or transfer,
- Less convinced that their responsibilities match their job descriptions, and, most markedly,
- Less willing to quit their jobs, organizations, or the public service.

Notably, two recruitment questions and both promotion questions do not show clear evidence that referents matter to responses.

These results provide the minimally expected result that different question referents result in different responses. Moreover, they are our first indication that the use of organizational referents really does make respondents more willing to admit to sensitive attitudes and behaviors, such as turnover intentions, as well as

**FIGURE 23.3 Organizational and Individual Referents in the Romanian Data**



Source: Original figure for this publication.

Note: Bars show estimated differences between organizational- and individual-referent questions in the survey experiment in Romania with 95 percent confidence intervals based on cluster-robust standard errors. Bars left of zero on the horizontal axis indicate higher scores on the individual-referent version of the question, whereas bars right of zero indicate higher scores on the organizational-referent version. All variables are scaled on the same 1–5 Likert scale. See appendix K.1 for full items and question labels.

slightly less prone to exaggerate their views on dismissals and their job descriptions. We return to this issue in more detail below.

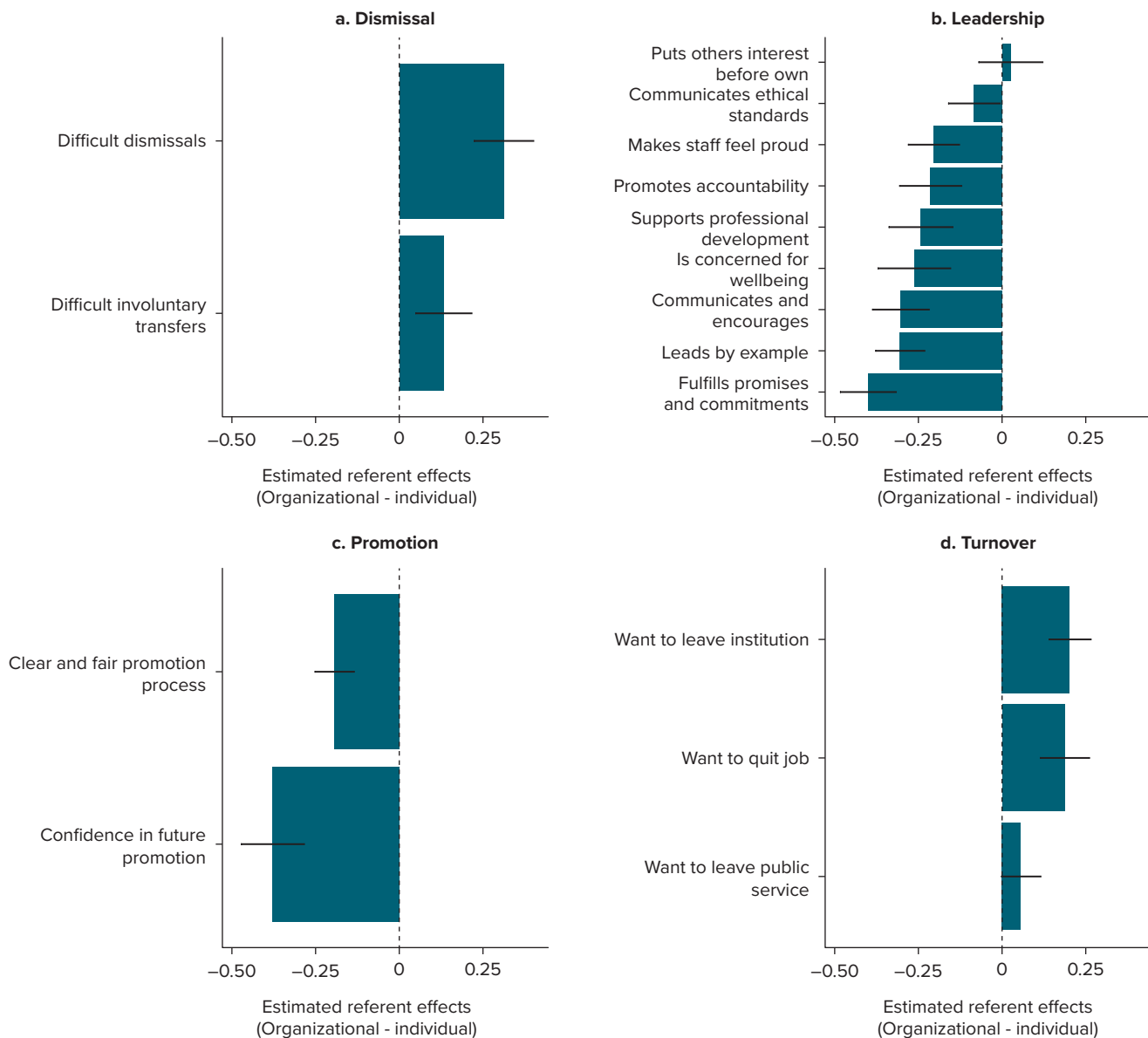
Figure 23.4 shows the results of a similar analysis using the Guatemalan data. In these data, the use of organizational versus individual referents matters more to average responses than in the Romanian data. Individual referents make respondents

- More likely to report that their direct managers are more transformational and ethical in their leadership styles on nearly any measure,
- Less prone to report turnover intentions,
- Less concerned about involuntary dismissals and transfers, and
- More convinced that promotions are within reach and that the process for achieving promotion is fair.

We can conclude at this stage that the choice of referent often matters to average responses—sometimes not a lot, but substantially for some questions. We return to plausible determinants of when referent choice matters below. Qualitatively speaking, however, we can already establish that referents do matter.

The average responses provided to survey questions matter a great deal, not least because they feed the organizational descriptive statistics commonly used in benchmarking organizations (about which, more shortly).

**FIGURE 23.4 Organizational and Individual Referents in the Guatemalan Data**



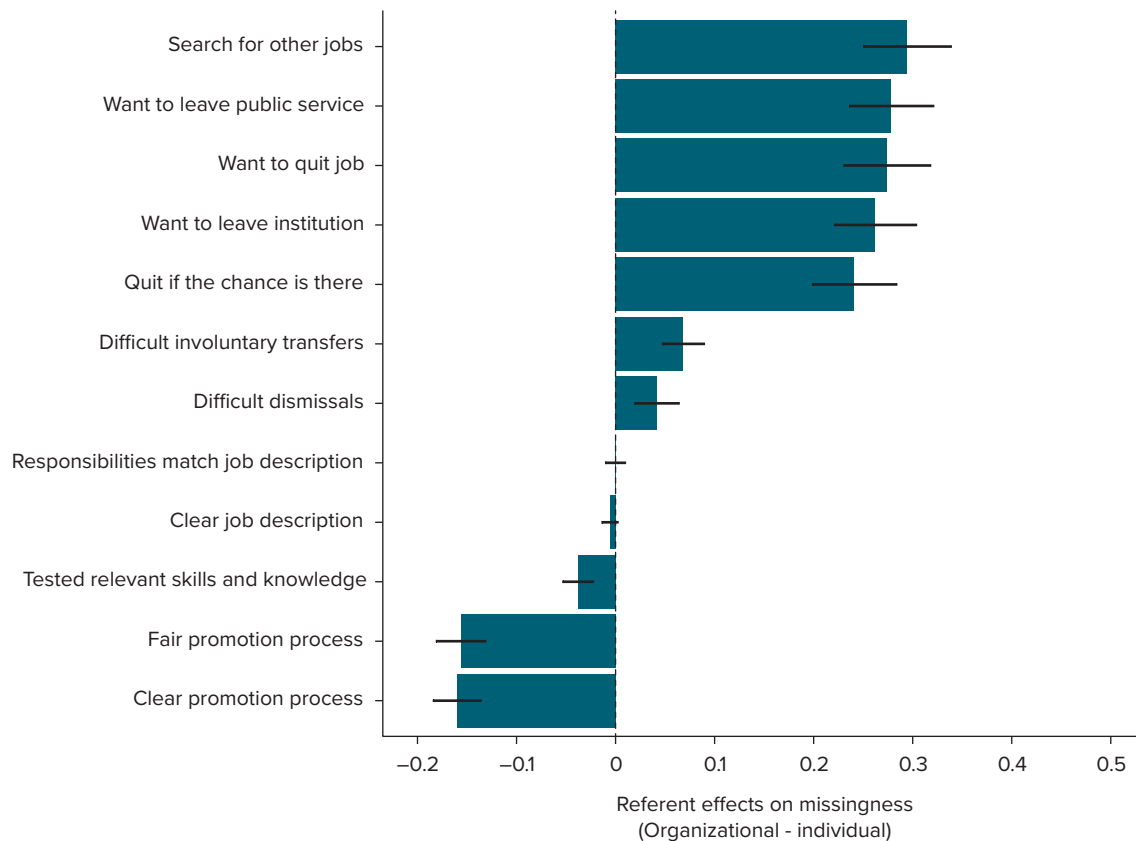
Source: Original figure for this publication.

Note: Bars show estimated differences between organizational- and individual-referent questions in the survey experiment in Guatemala with 95 percent confidence intervals based on cluster-robust standard errors. Bars left of zero on the horizontal axis indicate higher scores on the individual-referent version of the question, whereas bars right of zero indicate higher scores on the organizational-referent version. All variables are scaled on the same 1–5 Likert scale. See appendix K.1 for full items and question labels.

Yet average responses are not the only quantity that question referents may affect. It is possible, for instance, that organizational-referent questions are harder for respondents to understand, prompting item nonresponse—that is, respondents' not responding to individual items (see chapter 22).

Figure 23.5 examines this question using our Romanian data. Using a set of linear probability models with institution fixed effects, we find evidence of substantial nonresponse effects, particularly for questions relating to turnover. For each turnover question, the estimated probability of respondents not responding to individual-referent questions is increased by more than 20 percent relative to otherwise identical organizational-referent questions. This effect is substantial and worth considering. It is also worth noting, however, that less-sensitive questions on dismissal show much smaller effects, and questions on recruitment show no clear evidence of an effect at all. Moreover, individual referents substantially *reduce* nonresponse

**FIGURE 23.5** Estimates of Referent Effects on the Likelihood of Item Nonresponse



Source: Original figure for this publication.

Note: Bars show linear probability estimates of the differences between organizational- and individual-referent questions in the survey experiment in Romania with 95 percent confidence intervals based on cluster-robust standard errors. Bars left of zero on the horizontal axis indicate a higher probability of missingness on the individual-referent version of the question, whereas bars right of zero indicate a higher probability of missingness on the organizational-referent version. All variables are scaled on the same 0–1 scale, where 1 indicates “missing.” See appendix K.1 for full items and question labels.

relative to organizational referents for questions relating to promotion. One explanation for this finding may be that questions surrounding promotion processes are difficult to answer on behalf of the organization as a whole, leading respondents to nonresponse as a way of indicating they do not know the answer (see chapter 22). We return to the consequences of these findings below.

A final question we can examine is whether there are referent effects not on responses to individual survey variables but on relationships between survey variables. It is possible, for instance, that respondents fall back on their general opinions about the organization when asked for specific information about it, forming their attitudes as they go. This could result in increased statistical relationships between variables because they all tap into the same overarching attitude.

Table 23.3 examines this question using the leadership questions from the Guatemalan data. The table shows differences in statistical association between respondents who answered individual-referent questions and respondents who answered organizational-referent questions. Positive values indicate that organizational-referent questions correlate more strongly than similar individual-referent questions.

Table 23.3 does give some indication that variables covary differently when using organizational- rather than individual-referent questions. The effects we find indicate that the relevant relationships are generally—though not always—stronger when organizational referents are used. The differences vary in size, and not all are substantial. However, qualitative conclusions about the relationships between factors do sometimes hinge on the choice of referent. For example, when using our leadership, recruitment, and promotion variables to

**TABLE 23.3** Estimated Differences in Relationships between Leadership Variables for Different Referents, Guatemala (Organizational—Individual)

	Communicates and encourages	Communicates ethical standards	Fulfills promises and commitments	Is concerned for wellbeing	Leads by example	Makes staff feel proud	Promotes accountability	Puts others interest before own	Supports professional development
Communicates and encourages		−0.064* (0.028)	−0.046 (0.033)	−0.027 (0.022)	−0.085*** (0.019)	−0.037 (0.022)	−0.069* (0.031)	0.095* (0.036)	−0.042‡ (0.021)
Communicates ethical standards	−0.092** (0.026)		−0.122** (0.034)	−0.130*** (0.028)	−0.125*** (0.030)	−0.108*** (0.021)	−0.138*** (0.031)	0.059 (0.034)	−0.128*** (0.029)
Fulfills promises and commitments	0.012 (0.040)	−0.027 (0.043)		0.020 (0.041)	−0.055 (0.032)	−0.017 (0.033)	−0.007 (0.044)	0.179*** (0.037)	0.020 (0.026)
Is concerned for wellbeing	−0.033 (0.025)	−0.117** (0.036)	−0.035 (0.026)		−0.086* (0.034)	−0.021 (0.021)	−0.056‡ (0.031)	0.161*** (0.031)	−0.038* (0.017)
Leads by example	−0.037 (0.036)	−0.044 (0.037)	−0.073* (0.034)	−0.032 (0.028)		0.000 (0.025)	−0.046 (0.036)	0.154** (0.042)	−0.028 (0.027)
Makes staff feel proud	−0.038 (0.031)	−0.078* (0.028)	−0.086* (0.033)	−0.022 (0.026)	−0.056* (0.022)		−0.075** (0.022)	0.126*** (0.033)	−0.035 (0.032)
Promotes accountability	−0.013 (0.028)	−0.032 (0.027)	−0.009 (0.038)	−0.003 (0.025)	−0.035 (0.030)	−0.007 (0.023)		0.144*** (0.036)	−0.003 (0.028)
Puts others interest before own	0.137** (0.047)	0.117* (0.048)	0.208** (0.056)	0.191*** (0.044)	0.194** (0.056)	0.162*** (0.038)	0.176** (0.049)		0.156** (0.042)
Supports professional development	−0.019 (0.030)	−0.089* (0.038)	−0.009 (0.021)	−0.004 (0.027)	−0.049 (0.028)	−0.006 (0.025)	−0.032 (0.031)	0.150*** (0.037)	

Source: Original table for this publication.

Note: Results from ordinary least squares models with institution fixed effects and standard errors clustered by institution. Each cell in the table is the estimated interaction between our experimental treatment and the question in the cell's row in a model predicting the question in the cell's column. All variables are scaled on the same 1–5 Likert scale. See appendix K.1 for full items and question labels. *p*-values: ‡ *p* < 0.100, \* *p* < 0.050, \*\* *p* < 0.010, \*\*\* *p* < 0.001.

predict turnover variables in the Guatemalan data, 13 percent of estimated associations have different signs depending on the referent used.<sup>10</sup>

In sum, the choice of question referent matters. We find often small but sometimes substantial referent effects on the average responses to most questions we examine. Given our experimental setup, these differences must be due to the way we pose our questions. Hence, average differences are, in most cases, plausibly interpreted as being due to the question referent. We also find substantial referent effects on nonresponse patterns, but without a single direction of the effect. Whether referents make people respond more or less often appears to hinge on the question, its sensitivity, and how difficult it is to respond to. Finally, we find referent effects on relationships between some variables, but not in any clear direction.

### Can Organizational Referents Limit Social-Desirability Bias?

Yes, organizational referents limit SDB, but mostly for strongly sensitive items. We find evidence that more-sensitive questions show larger differences between individual- and organizational-referent questions in our experiment. This likely indicates that organizational referents can help limit SDB in civil service surveys. We find indications that this effect may be particularly pronounced for very sensitive questions.

As noted above, combatting SDB is a sensible reason for the use of organizational referents. To examine this question in more depth, we coded our individual questions in the Romania and Guatemala experiments for their sensitivity (see chapter 22 for details on the procedure). For the sake of statistical power in the analyses that follow, we now include the follow-up questions on the use of various sources of information in recruitment and the questions asked at recruitment interviews from the Romania questionnaire we have excluded from our analysis up to this point.

We regress this measure on the absolute difference between average responses to questions using individual and organizational referents (the referent effect), which we standardize to make our different response scales comparable. We run regressions with two sets of observations: one (model 1 in table 23.4) in which each observation is a question—from either survey—with its associated referent effect and sensitivity score, and one (model 2) in which each observation is an organizational aggregate for a question.

If organizational referents guard against SDB, we would expect a positive association between the sensitivity of questions and referent effect sizes because a reduction in SDB for sensitive questions would increase the difference between responses using organizational and individual referents. In our analysis, we find evidence for this assertion. In model 1, the expected positive association is significant only at the 10 percent level due to a low number of observations. In the more well-powered model 2, the expected association is highly significant. As expected, sensitive questions see larger question-referent effects, indicating that organizational referents may diminish socially desirable responding. It is worth noting, however, that this analysis

**TABLE 23.4** Standardized Question-Referent Effects, by Sensitivity

	Model 1 (Questions as observations)	Model 2 (Institution aggregates as observations)
Sensitivity	0.107 <sup>‡</sup> (0.056)	0.079*** (0.016)
(Intercept)	0.201*** (0.043)	0.328*** (0.013)
<i>N</i>	40	2,664
<i>R</i> -squared adjusted	0.065	0.008

Source: Original table for this publication.

Note: Results from ordinary least squares models. Each observation in model 1 is a question; each observation in model 2 is a question aggregate from an institution. The dependent variable is the absolute referent effect—the absolute difference in average responses between individual- and organizational-referent questions—standardized to account for the different scales of the included variables. See appendix K.2 for model results using other measures.

<sup>‡</sup>  $p < 0.100$ ; \*  $p < 0.050$ ; \*\*  $p < 0.010$ ; \*\*\*  $p < 0.001$ .



cannot leverage randomization to the same extent that our previous analyses do and, consequently, that it cannot be conclusively established whether the associations we document are due to sensitivity.

However, this analysis masks an additional finding: some very sensitive questions do appear to display larger differences than less sensitive questions. To see this, consider the violin plot in figure 23.6, showing the distribution of standardized referent effects for nonsensitive and sensitive questions (the thicker the “violin” at a certain height, the more questions have referent effects at the corresponding value on the second axis).

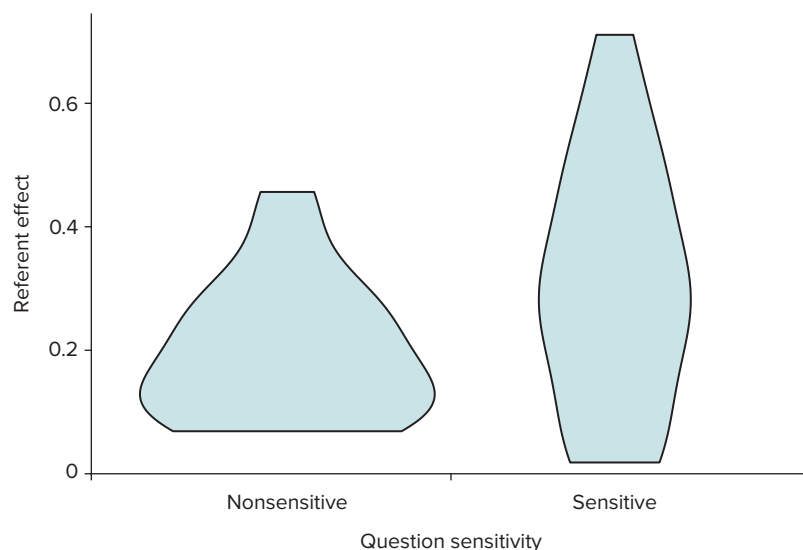
As the figure shows, the top of the referent effects distribution is far above the sensitivity effect observed in the table above, indicating that some sensitive questions have larger-than-predicted referent effects. On a qualitative inspection, these turn out to be very sensitive questions—particularly concerning turnover. This is a valuable conclusion. When examining sensitive issues—particularly highly sensitive issues such as corruption, politicization, or absenteeism—organizational referents appear to be able to combat SDB. For nonsensitive issues, the difference organizational referents make is more limited. The implication is that if the use of individual referents is preferred on other grounds, shifting to organizational referents may be justified on the grounds of SDB if questions are highly sensitive.

### Does Information Availability Matter?

Yes, information availability matters, but not in all the ways one might think. We find evidence that referent effects are smaller for very common attitudes and behaviors. However, we find no statistically clear evidence that respondents who have served longer in their organizations are less prone to referent effects.

As discussed, the availability of information may determine how much question referents matter if they are partly rooted in information availability. In these instances, we would expect smaller referent effects for questions about attitudes or behaviors that are either very common or very uncommon in respondents’ surroundings. Respondents are less (more) likely to report rare (common) behaviors about themselves by definition, but they are also less (more) likely to report rare (common) behaviors about their organizations because they encounter them rarely (commonly). By contrast, attitudes and behaviors that some hold but others do not can give rise to substantial referent effects, particularly if they are unevenly distributed within organizations.

**FIGURE 23.6** Distributions of Referent Effects for Sensitive and Nonsensitive Questions



Source: Original figure for this publication.

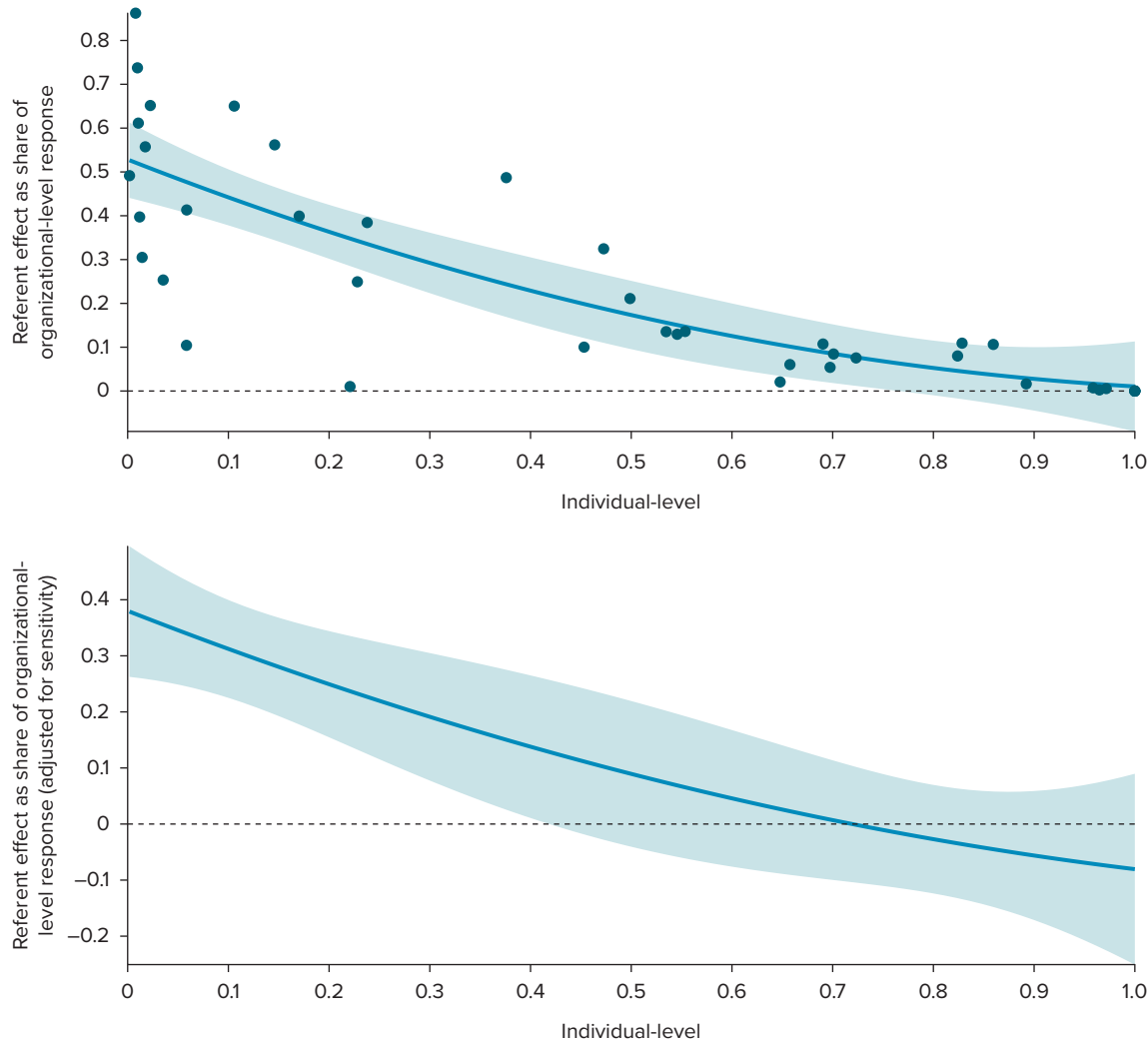
Note: The figure shows the distributions of referent effects split by question sensitivity. The width of the “violins” indicates the number of referent effects at or around the size indicated on the vertical axis. Thus, sensitive questions have a smaller range of referent effects, with the largest and smallest referent effects larger and smaller than for nonsensitive questions in our sample.

We examine this prediction by looking at patterns in referent effects. If very common or very rare attitudes and behaviors give rise to smaller referent effects, we should expect the referent effects, relative to reported commonality on organizational- (individual-) referent measures, to depend on how commonly the attitude or behavior in question is reported by respondents who are asked individual- (organizational-) referent questions. Specifically, we would expect an inverted-U relationship, in which referent effects are smaller for very rare or very common attitudes or behaviors.

Figure 23.7 speaks to this prediction. The figure plots the organizational proportion of affirmative responses to each question in our Romanian experiment (using individual referents) against the absolute difference between individual- and organizational-referent questions as a proportion of responses to the organizational-referent question.<sup>11</sup> Affirmative responses are interpreted as responses scoring on the upper two quintiles of the possible answers for scale questions (for example, “Strongly agree” and “Agree” on a Likert scale) and affirmative answers to follow-up questions, where respondents could indicate “yes” or “no.”

As the top panel in the figure shows, differences are not generally smaller for questions where scores are generally very low or very high, behaviors or practices are very rare or very common, and information is

**FIGURE 23.7** Response Score and Question-Referent Effects in the Romanian Data



Source: Original figure for this publication.

Note: The figure shows the average referent effect—here defined as the absolute difference between organizational- and individual-referent versions of each question in the experiment as a proportion of the score of the organizational-referent version—as a polynomial regression function of responses to the individual-referent version of the questions. The top panel shows the raw association, with individual questions plotted as points. The lower panel shows the association adjusted for question sensitivity.

more readily available. Instead, it shows referent effects declining as a function of commonality. One interpretation of this aligns, albeit asymmetrically, with information availability: it is easier for respondents to provide information about their organizations if they experience the relevant attitudes or behaviors around them, and this renders relative referent effects smaller for questions about common attitudes and behaviors than for questions where attitudes and behaviors are less common.

One obvious objection to this finding is that sensitive questions often result in indications that behaviors are rare, either because the behaviors in question *are* rare or because of SDB. As a result, the association depicted in the top panel of figure 23.7 could reflect sensitivity rather than information availability. To examine this issue, the lower panel in figure 23.7 shows the same association adjusted for question sensitivity. Indeed, the identified referent effects are weaker, but the pattern holds: referent effects appear to be smaller for questions targeting attitudes and behaviors that are common. Of course, one cannot definitively conclude from this simple analysis that greater information availability to respondents either will or will not result in smaller question referent effects. But the analysis does suggest that the use of organization-level referents may require caution when targeting rare behaviors or attitudes.

In our data, at least, we are not able to further pin down plausible determinants of information availability that give rise to the predicted changes in referent effects. To exemplify, further analysis of our two data sets (not shown) shows that organization size does not appear to matter for respondents' reactions to organizational versus individual referents, although one might expect smaller organizations to be easier to rate for respondents who use the information available to them, all else being equal. Moreover, as shown in table 23.5, the effect of using organizational referents in our Romanian sample does not generally vary with years of service. The exception is recruitment, where a negative referent effect grows with years of service (contrary to the idea that organizational experience would facilitate learning and diminish information-based referent effects). This effect could reflect changing recruitment practices over time, which would be consistent with the finding not being recovered when limiting the sample to relatively recently recruited public servants (model 6 in table 23.5).

However, some organizational characteristics do matter for referent effect sizes. If split by organization, the average referent effect size in the Romanian data is 0.15 (standardized across all experimental questions), but effect sizes range widely from one organization to another, from 0.02 to 0.31, the latter being a moderately sized effect, whereas the former is negligible.

The conclusion, then, is that if information availability matters in our data, we are not able to get very far in pinpointing its determinants. We can offer two suggestive conclusions, however. First, referent effects appear smaller for questions targeting attitudes that are very common. Second, arguing when information is available is no simple matter and is not a function of simple structural characteristics, such as organization size, or respondent characteristics, such as years of service.

**TABLE 23.5 Question-Referent Effects, by Years of Service, Romania**

	Model 3 (Dismissals)	Model 4 (Recruitment)	Model 5 (Turnover)	Model 6 (Recruitment, <5 years)
Organizational level	0.148 (0.091)	0.005 (0.041)	0.453*** (0.062)	-0.009 (0.073)
Years of service	-0.000 (0.004)	0.001 (0.002)	-0.003 (0.002)	-0.002 (0.023)
Organizational level × Years of service	-0.001 (0.005)	-0.005* (0.002)	0.002 (0.004)	-0.011 (0.021)
<i>N</i>	3,016	3,298	2,898	656
<i>R</i> -squared adjusted	0.137	0.088	0.216	0.212

Source: Original table for this publication.

Note: Results from ordinary least squares models with cluster-robust standard errors by institution. Each observation is a respondent in the Romania data set. The dependent variable is the indexes for our experimental measures of dismissals (model 3), recruitment (models 4 and 6), and turnover (model 5). All are kept on the same 1–5 scale as their items. Years of service is a single-item measure of how long respondents have served in public administration (measured in years). See appendix K.2 for model results using other measures.

‡  $p < 0.100$ ; \*  $p < 0.050$ ; \*\*  $p < 0.010$ ; \*\*\*  $p < 0.001$ .

## DISCUSSION AND CONCLUSION

Where do our experiments leave us? What do we learn from them? While they do give valuable insights on the effects of individual versus organizational referents in civil service surveys, they also raise new and interesting questions to which we do not yet have the answers.

The primary lesson is that the choice of referent matters. Using organizational referents often leads to more agreement between respondents in the same organization—a finding consistent with Glick (1985) and Klein et al. (2001). Yet in our measures, this agreement is often too low for responses to reliably track bottom-up, organization-level features above and beyond the perspective of respondents. The use of organizational referents can, however, provide summaries of respondents' perspectives and experiences, which are also, per the summary bottom-up perspective, valuable organizational metrics.

Moreover, average responses to survey questions often change when question referents change. For some questions, respondents report stronger agreement when asked about their organizations than when asked about themselves. For other questions, the pattern is reversed. In general, these effects are of modest size, but for some questions, they are substantial—and predicting for which questions referents will matter the most is not straightforward. Similarly, we find substantial question-referent effects on nonresponse but without uncovering one clear direction. For some questions, organizational referents substantially reduce nonresponse; for others, they exacerbate it. We also find some evidence that relationships between variables are affected by referents. But not all associations between variables are clearly impacted by the choice of referent, and we cannot propose a general direction of effects when they are.

We have examined the determinants of referent effect sizes: when does the choice between individual and organizational referents matter the most? From our analyses, we can draw only a few lessons about the question of referent effect size. First, referent effects seem to be larger for (highly) sensitive questions. This is consistent with organizational referents' ability to mitigate SDB for sensitive questions. Second, referent effects seem to be larger for attitudes, behaviors, and practices that are not common among respondents. This is consistent with the view that organization-level questions can pose higher informational demands than respondents can meet. It is also notable that question-referent effects are stronger in some organizations than others, but it is not clear which organizational characteristics drive these differences. And question-referent effects are not negatively associated with experience in the organization, suggesting that learning may have limited consequences for their size.

What does all this mean for civil service survey designers? It means they must be aware of the referents used in the questions they include in their surveys. Using organizational referents, as is common practice today, is not uniformly preferable on conceptual grounds—since responses often track but do not directly reflect organizational characteristics over and above respondents' perspectives. However, using individual referents is not uniformly preferable either. Particularly on measurement grounds, there is evidence that individual referents may suffer from SDB both for sensitive questions and for questions for which respondents wish to positively manage impressions.

Beyond awareness, we can make a few recommendations for more specific situations. First, a survey designer including very sensitive questions in a survey should consider posing these questions using organizational referents to combat SDB. It is important to recognize the limitations of this advice, however. Our analysis shows that more sensitive behavior is reported when using organizational referents. Yet this does not mean organizational referents provide an accurate estimate of how frequently the sensitive behavior or attitude occurs.

Moreover, using organizational referents comes at a heavy conceptual cost if the survey is interested in anything more than organizational aggregates. Predicting individual behaviors and attitudes with individual responses to sensitive organizational-referent questions implies a shift in what is studied (Klein and Kozlowski 2000). There is a difference between saying that a respondent's manager is abusive and that managers in the organization generally are abusive. Predicting sensitive organizational-referent questions with individual attitudes and experiences is often problematic because it tends to operationally conflate beliefs

about the organizational collective with individual attitudes and behaviors. If survey designers want to know why individual public servants behave and think as they do, the conceptual cost of organizational referents may be higher than the measurement gain, even for sensitive topics.

Second, survey designers should consider how the information needed to answer a question will be acquired by respondents. If using an organizational referent, can individual respondents reasonably be expected to know the answer? Individual referents are preferable if introspection provides more or more-reliable information than beliefs and available information about the organization. Our findings indicate few systematic patterns in which questions are most affected by this or in which respondents are most prone to provide the needed information accurately, rather than information infused with impressions, rumors, and beliefs. However, this does not mean that information availability can be glossed over by survey designers. Instead, it highlights the need for more measurement studies specifically targeting information availability and its determinants.

Third, our results may help survey designers think about utilizing other levels of measurement than individual or organizational. Of course, this implication is somewhat speculative, and more data are needed. Consider the conceptual issue with an organizational-referent question that elicits low levels of intraorganizational agreement in response. This means that respondents perceive their organization differently even though they all work within it. As noted, the usual interpretation of this occurrence is that organizational practices differ, that implementation of policies and procedures is uneven, and that management and leadership matter to how organizational practices are felt by public servants. There is nothing intrinsically wrong with this interpretation—but it is uncertain. After all, respondents were asked about their organization, not their section, team, manager, or other lower-level entities. It is not clear from our responses which level respondents draw on the most for information. This is an important weakness of organizational-referent questions in such a situation.

The interpretation gives rise to a question we cannot examine in detail using our data. Would it be a better strategy to use team referents or section referents rather than organizational or individual ones? Is it possible that using team referents would combat socially desirable responding without posing too high of informational demands on respondents? Our results cannot speak directly to this question. They do suggest that the answer is likely contingent on the type of question. Teams are often psychologically closer to people than whole organizations (Riketta and Dick 2005), which might mean that for some questions, team referents will do little to combat SDB. Similarly, some information can be difficult or impossible to access even within teams. Yet is likely to be more easily accessible within teams than for the entire organization. As such, team referent measures may be preferred to organizational-referent questions on measurement grounds if questions are not too sensitive. On the other hand, civil service survey designers may be less interested in reporting team aggregates to decision-makers or other audiences. And aggregating team aggregates to the organizational level is not likely to resolve the issues we discuss in this chapter.

Let us end with a few open questions for which both research and practice would benefit from systematic answers. We know much, both in conceptual and measurement terms, about multilevel theory, measurement, and modeling (Humphrey and LeBreton 2019; Klein and Kozlowski 2000). However, the literature on referent choice is limited, seemingly on the assumption that matching to the level of stated claims is all there is to it. This is sensible enough if one requires organizational measures to reflect organizational characteristics over and above respondents' perspectives in order to be useful. Yet such a perspective is overly limiting, not least for the practice of civil service survey design. For many variables, including management practices, perspectives on leadership, human resources functions, and more, data summaries of employee views and perspectives—interpretable from what we have referred to as the summary bottom-up perspective—can be valuable forms of decision support.

If we accept that organizational—or other higher-level—measurement referents can be useful even if respondents do not strongly align in response to them, our analyses point to a series of underexamined questions. First, which questions are particularly exposed to referent effects? We have found very sensitive questions to be affected, but much more knowledge is needed to reliably provide the type of advice survey designers want. Second, we have scarcely any evidence on whether the choice of referent affects different

survey respondents differently. We have not found any such effects in a few exploratory analyses, but this does not mean they do not exist. Third, it appears in our data that organizations affect the size of referent effects. We note that organization size does not appear to matter systematically, but we can see in our data that *something* about organizations does. Yet again, much more knowledge is needed on this issue.

The fact that our findings are not straightforward should highlight for both interested academics and survey designers that the choice of levels of measurement is a complicated issue, and, as we have shown, it is a choice that matters more than current practice seems to be aware.

## NOTES

1. An alternative design could randomly assign respondents a question order, with one group being asked individual-referent questions before organizational-referent questions and another group being asked organizational-referent questions before individual-referent questions. This would permit estimation of the average anchoring effect. However, as there is likely to be substantial heterogeneity in this effect, adjusting for the effect can become challenging. For this reason, we opt to ask each respondent only one set of questions.
2. For attitudinal variables, the equivalent of this perspective is that survey aggregates capture shared attitudes in the organization (Chan 1998).
3. This is true, in part, because impression management—wanting to control how one is viewed normatively—concerns both others (impression management proper) and oneself (self-deception) (for example, Millham and Kellogg 1980; Paulhus 1986).
4. The findings of Baltes, Zhdanova, and Parker (2009) suggest that organizational aggregates may depend on the unknown mixture of respondents using “upward” and “downward” comparisons to arrive at their answers. (Their findings also suggest that downward comparison is more common in their sample, but they are unable to assess the specific mixture.)
5. Respondents who have served longer in organizations have been shown in previous studies to be less prone to using heuristics in their decision-making because they can substitute their experience (cf. Pedersen, Stritch, and Thuesen 2018). Translated into the survey-response setting, more experienced personnel may not need to rely on stories and other heuristic devices when assessing their organizations.
6. Some respondents were interviewed even though they were not included on the original staff lists, meaning this number is somewhat inflated relative to those staff lists.
7. Note the assumption behind this null hypothesis is that respondents aggregate information in a way that approximates averaging when responding with reference to their organization. If this assumption does not hold, it poses an additional problem for organizational-referent questions because the aggregation used by respondents is then both unknown and does not approximate common-sense (though not the only sensible) aggregation procedures. Theoretically, this simply adds complexity to the information-processing discussion already noted.
8. This is because the organizational construct assessment of the ICC treats it as a measure of reliability. One way to think of this is to consider each respondent a rater of his or her organization. From this perspective, if at most 15 percent of variance is accounted for by organizations, for an ICC of 0.15, and at least 85 percent is accounted for by the raters, this does not indicate a reliable assessment of organizational characteristics. Raters affect responses too much.
9. We are grateful to an external reviewer for pointing us to this possibility and regret we have no better options available for examining it.
10. This figure includes only associations where effects in at least one direction are statistically significant at the 5 percent level. In none of the included cases are effects in both directions both statistically different from zero.
11. We thank a reviewer for pointing us in this direction. We originally considered simply presenting the absolute differences between answers to questions using different referents, but this created downward trends on the extremes of figure 23.7, consistent both with the prediction and a methodological artifact related only to question scaling.

## REFERENCES

- Baltes, Boris B., Ludmila S. Zhdanova, and Christopher P. Parker. 2009. “Psychological Climate: A Comparison of Organizational and Individual Level Referents.” *Human Relations* 62 (5): 669–700. <https://doi.org/10.1177/0018726709103454>.



- Bardasi, Elena, Kathleen Beegle, Andrew Dillon, and Pieter Serneels. 2011. "Do Labor Statistics Depend on How and to Whom the Questions Are Asked? Results from a Survey Experiment in Tanzania." *The World Bank Economic Review* 25 (3): 418–47. <https://doi.org/10.1093/wber/lhr022>.
- Bezes, Philippe, and Gilles Jeannot. 2018. "Autonomy and Managerial Reforms in Europe: Let or Make Public Managers Manage?" *Public Administration* 96 (1): 3–22. <https://doi.org/10.1111/padm.12361>.
- Blair, Johnny, Geeta Menon, and Barbara Bickart. 2004. "Measurement Effects in Self vs. Proxy Response to Survey Questions: An Information-Processing Perspective." In *Measurement Errors in Surveys*, edited by Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman, 145–66. Hoboken, NJ: Wiley. <https://doi.org/10.1002/9781118150382.ch9>.
- Chan, David. 1998. "Functional Relations among Constructs in the Same Content Domain at Different Levels of Analysis: A Typology of Composition Models." *Journal of Applied Psychology* 83 (2): 234–46. <https://doi.org/10.1037/0021-9010.83.2.234>.
- Christensen, Tom, and Per Lægrend. 2007. "The Whole-of-Government Approach to Public Sector Reform." *Public Administration Review* 67 (6): 1059–66. <https://doi.org/10.1111/j.1540-6210.2007.00797.x>.
- Dunleavy, Patrick, Helen Margetts, Simon Bastow, and Jane Tinkler. 2006. "New Public Management Is Dead—Long Live Digital-Era Governance." *Journal of Public Administration Research and Theory* 16 (3): 467–94. <https://doi.org/10.1093/jopart/mui057>.
- Edwards, W. Sherman, and David Cantor. 2004. "Toward a Response Model in Establishment Surveys." In *Measurement Errors in Surveys*, edited by Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman, 211–33. Hoboken, NJ: Wiley. <https://doi.org/10.1002/9781118150382.ch12>.
- Eggers, William D. 2007. *Government 2.0: Using Technology to Improve Education, Cut Red Tape, Reduce Gridlock, and Enhance Democracy*. Lanham, MD: Rowman & Littlefield.
- Fisher, Robert J. 1993. "Social Desirability Bias and the Validity of Indirect Questioning." *Journal of Consumer Research* 20 (2): 303–15. <https://doi.org/10.1086/209351>.
- Freling, Traci H., Zhiyong Yang, Ritesh Saini, Omar S. Itani, and Ryan Rashad Abualsamh. 2020. "When Poignant Stories Outweigh Cold Hard Facts: A Meta-Analysis of the Anecdotal Bias." *Organizational Behavior and Human Decision Processes* 160: 51–67. <https://doi.org/10.1016/j.obhdp.2020.01.006>.
- Gingerich, Daniel W. 2013. "Governance Indicators and the Level of Analysis Problem: Empirical Findings from South America." *British Journal of Political Science* 43 (3): 505–40. <https://doi.org/10.1017/S0007123412000403>.
- Glick, William H. 1985. "Conceptualizing and Measuring Organizational and Psychological Climate: Pitfalls in Multilevel Research." *Academy of Management Review* 10 (3): 601–16. <https://doi.org/10.2307/258140>.
- Graaf, Gjalt de, Leo Huberts, and Tebbine Strüwer. 2018. "Integrity Violations and Corruption in Western Public Governance: Empirical Evidence and Reflection from the Netherlands." *Public Integrity* 20 (2): 131–49. <https://doi.org/10.1080/10999922.2017.1350796>.
- Guenther, Corey L., and Mark D. Alicke. 2010. "Deconstructing the Better-Than-Average Effect." *Journal of Personality and Social Psychology* 99 (5): 755–70. <https://doi.org/10.1037/a0020959>.
- Hoch, Stephen J. 1987. "Perceived Consensus and Predictive Accuracy: The Pros and Cons of Projection." *Journal of Personality and Social Psychology* 53 (2): 221–34. <https://doi.org/10.1037/0022-3514.53.2.221>.
- Homburg, Christian, Martin Klarmann, Martin Reimann, and Oliver Schilke. 2012. "What Drives Key Informant Accuracy?" *Journal of Marketing Research* 49 (4): 594–608. <https://doi.org/10.1509/jmr.09.0174>.
- Humphrey, Stephen E., and James M. LeBreton, eds. 2019. *The Handbook of Multilevel Theory, Measurement, and Analysis*. Washington, DC: American Psychological Association.
- Klein, Katherine J., Amy Buhl Conn, D. Brent Smith, and Joann Speer Sorra. 2001. "Is Everyone in Agreement? An Exploration of Within-Group Agreement in Employee Perceptions of the Work Environment." *Journal of Applied Psychology* 86 (1): 3–16. <https://doi.org/10.1037/0021-9010.86.1.3>.
- Klein, Katherine J., Fred Dansereau, and Rosalie J. Hall. 1994. "Levels Issues in Theory Development, Data Collection, and Analysis." *Academy of Management Review* 19 (2): 195–229. <https://doi.org/10.5465/amr.1994.9410210745>.
- Klein, Katherine J., and Steve W. J. Kozlowski, eds. 2000. *Multilevel Theory, Research, and Methods in Organizations: Foundations, Extensions, and New Directions*. San Francisco: Jossey-Bass.
- Kreuter, Frauke, Stanley Presser, and Roger Tourangeau. 2008. "Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity." *Public Opinion Quarterly* 72 (5): 847–65. <https://doi.org/10.1093/poq/nfn063>.
- Meyer-Sahling, Jan-Hinrik. 2011. "The Durability of EU Civil Service Policy in Central and Eastern Europe after Accession." *Governance: An International Journal of Policy, Administration, and Institutions* 24 (2): 231–60. <https://doi.org/10.1111/j.1468-0491.2011.01523.x>.
- Meyer-Sahling, Jan-Hinrik, and Kim Sass Mikkelsen. 2016. "Civil Service Laws, Merit, Politicization, and Corruption: The Perspective of Public Officials from Five East European Countries." *Public Administration* 94 (4): 1105–23. <https://doi.org/10.1111/padm.12276>.

- Meyer-Sahling, Jan-Hinrik, and Kim Sass Mikkelsen. 2020. "Codes of Ethics, Disciplinary Codes, and the Effectiveness of Anti-Corruption Frameworks: Evidence from a Survey of Civil Servants in Poland." *Review of Public Personnel Administration* 42 (1): 142–64. <https://doi.org/10.1177/0734371X20949420>.
- Millham, Jim, and Richard W. Kellogg. 1980. "Need for Social Approval: Impression Management or Self-Deception?" *Journal of Research in Personality* 14 (4): 445–57. [https://doi.org/10.1016/0092-6566\(80\)90003-3](https://doi.org/10.1016/0092-6566(80)90003-3).
- Nishii, Lisa H., David P. Lepak, and Benjamin Schneider. 2008. "Employee Attributions of the 'Why' of HR Practices: Their Effects on Employee Attitudes and Behaviors, and Customer Satisfaction." *Personnel Psychology* 61 (3): 503–45. <https://doi.org/10.1111/j.1744-6570.2008.00121.x>.
- OPM (Office of Personnel Management). 2018. *Governmentwide Management Report: Results from the 2018 Federal Employee Viewpoint Survey*. Washington, DC: US Office of Personnel Management, US Government. <https://www.opm.gov/fevs/reports/governmentwide-reports/governmentwide-reports/governmentwide-management-report/2018/2018-governmentwide-management-report.pdf>.
- Parker, Christopher P., Boris B. Baltes, Scott A. Young, Joseph W. Huff, Robert A. Altmann, Heather A. Lacost, and Joanne E. Roberts. 2003. "Relationships between Psychological Climate Perceptions and Work Outcomes: A Meta-Analytic Review." *Journal of Organizational Behavior* 24 (4): 389–416. <https://doi.org/10.1002/job.198>.
- Paulhus, Delroy L. 1986. "Self-Deception and Impression Management in Test Responses." In *Personality Assessment via Questionnaires*, edited by Alois Angleitner and Jerry S. Wiggins, 143–65. Berlin: Springer-Verlag. [https://doi.org/10.1007/978-3-642-70751-3\\_8](https://doi.org/10.1007/978-3-642-70751-3_8).
- Pedersen, Mogens Jin, Justin M. Stritch, and Frederik Thuesen. 2018. "Punishment on the Frontlines of Public Service Delivery: Client Ethnicity and Caseworker Sanctioning Decisions in a Scandinavian Welfare State." *Journal of Public Administration Research and Theory* 28 (3): 339–54. <https://doi.org/10.1093/jopart/muy018>.
- Razafindrakoto, Mireille, and François Roubaud. 2010. "Are International Databases on Corruption Reliable? A Comparison of Expert Opinion Surveys and Household Surveys in Sub-Saharan Africa." *World Development* 38 (8): 1057–69.
- Riketta, Michael, and Rolf van Dick. 2005. "Foci of Attachment in Organizations: A Meta-Analytic Comparison of the Strength and Correlates of Workgroup versus Organizational Identification and Commitment." *Journal of Vocational Behavior* 67 (3): 490–510. <https://doi.org/10.1016/j.jvb.2004.06.001>.
- Schriesheim, Chester A., Joshua B. Wu, and Terri A. Scandura. 2009. "A Meso Measure? Examination of the Levels of Analysis of the Multifactor Leadership Questionnaire (MLQ)." *The Leadership Quarterly* 20 (4): 604–16. <https://doi.org/10.1016/j.leaqua.2009.04.005>.
- Schuster, Christian, Jan-Hinrik Meyer-Sahling, and Kim Sass Mikkelsen. 2020. "(Un)principled Principals, (Un)principled Agents: The Differential Effects of Managerial Civil Service Reforms on Corruption in Developing and OECD Countries." *Governance: An International Journal of Policy, Administration, and Institutions* 33 (4): 829–48. <https://doi.org/10.1111/gove.12461>.
- Shah, Priti Pradhan. 1998. "Who Are Employees' Social Referents? Using a Network Perspective to Determine Referent Others." *Academy of Management Journal* 41 (3): 249–68. <https://doi.org/10.2307/256906>.
- Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. Cambridge, UK: Cambridge University Press.
- Tourangeau, Roger, and Ting Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133 (5): 859–83. <https://doi.org/10.1037/0033-2909.133.5.859>.
- Willems, Jurgen. 2020. "Public Servant Stereotypes: It Is Not (At) All about Being Lazy, Greedy and Corrupt." *Public Administration* 98 (4): 807–23. <https://doi.org/10.1111/padm.12686>.
- Zaller, John R. 1992. *The Nature and Origins of Mass Opinion*. Cambridge, UK: Cambridge University Press.



## CHAPTER 24

# Interpreting Survey Findings

## Can Survey Results Be Compared across Organizations and Countries?

*Robert Lipinski, Jan-Hinrik Meyer-Sahling, Kim Sass Mikkelsen, and Christian Schuster*

### SUMMARY

With the rise in worldwide efforts to understand public administration by surveying civil servants, issues of survey question comparability become paramount. Surveys can rarely be understood in a void but rather require benchmarks and points of reference. However, it is not clear whether survey questions, even when phrased and structured in the same manner, measure the same concepts in the same way and, therefore, can be compared. For multiple reasons, including work environment, adaptive expectations, and cultural factors, different people might understand the same question in different ways and adjust their answers accordingly. This might make survey results incomparable, not only across countries but also across different groups of civil servants within a national public administration. This chapter uses results from seven public service surveys from across Europe, Latin America, and South Asia to investigate the extent to which the same survey questions measure the same concepts similarly—that is, are measurement invariant—using as an example questions related to *transformational leadership*. To ascertain measurement invariance, models of a hypothesized relationship between questions measuring transformational leadership are compared across countries, as well as along gender, educational, and organizational lines within countries. Solid evidence of metric invariance and tentative evidence of scalar invariance is found in cross-country comparisons. Moreover, factor loadings can be judged equal (*metric invariance*) across gender, education level, and organization in most countries, as can latent factor means (*scalar invariance*). Our results suggest that groups of public servants within countries—delineated, for instance, by gender, education, or organization—can typically be benchmarked without invariance concerns. Across countries, evidence for valid benchmarking—that is, scalar invariance—is strongest for countries in similar regions and at similar income levels. It is weaker—though still suggestive—when comparing all countries in the sample. Our chapter concludes that less culturally contingent concepts may be plausibly benchmarked with care across countries.

Robert Lipinski is a consultant in the World Bank's Development Impact Evaluation (DIME) Department. Jan-Hinrik Meyer-Sahling is a professor at the University of Nottingham. Kim Sass Mikkelsen is an associate professor at Roskilde University. Christian Schuster is a professor at University College London.

## ANALYTICS IN PRACTICE

- Many theoretical insights and practical lessons from surveys of civil servants depend on the ability to draw comparisons between countries and demographic groups. This chapter focuses on the comparability of the concept of *transformational leadership* across different contexts and groups—a premise known as *measurement invariance*. Transformational leadership measures the extent to which managers lead by setting a good example, making employees proud, and generating enthusiasm about an organization's mission.
- Equality in the understanding of a single overarching concept, such as transformational leadership, across different countries and groups can be conceptualized in three ways. The same concept can be measured by the same set of questions (*configural invariance*), those questions can have the same strength of a relationship with the underlying concept (*metric invariance*), and the concept can have the latent mean structure (*scalar invariance*).
- When comparing survey measures and concepts across countries, practitioners should consider the extent to which different cultural interpretations of a concept (such as leadership), different social-desirability biases, different pressures in the work environment, and even differences in language may lead to differences in survey means across countries that do not reflect substantive differences in the underlying concept (such as the quality of leadership).
- When empirically assessing the measurement invariance (and thus the cross-country comparability) of a concept that is arguably culturally specific—transformational leadership—we find that cross-country comparisons can be undertaken, although with caution. There is evidence that the concept of transformational leadership is understood in a comparable way across the seven countries included in the analyses. As we find suggestive evidence that cross-country comparisons are possible with even a relatively culturally contingent concept (leadership), cross-country comparisons of more factual questions (for example, “Did you have a performance evaluation last year?”) are plausibly often possible in a valid manner.
- Grouping countries by region and income level removes many of the differences across countries. This suggests that comparisons between countries at similar income levels and in the same world regions can be made with greater confidence.
- Within-country comparisons of transformational leadership suffer from fewer concerns about lack of comparability. Empirically, we find that they can be reliably made across public servants of different genders and education levels, and in different institutions.

## INTRODUCTION

Surveys of civil servants provide insights into core parts of the public administration production function—such as the quality of management and the attitudes (for example, motivation) of employees. As argued by Rogger and Schuster in chapters 1–3 of *The Government Analytics Handbook*, these determinants of public sector productivity are difficult to measure accurately with other data sources. Survey results are typically presented as percentages of public servants who evaluate favorably dimensions of their work environment, management, or themselves—for instance, the percentage of public servants who recommend their organization as a great place to work, or the percentage of public servants who evaluate the leadership of their superior favorably.

How can governments know whether certain percentages—such as 75 percent of public servants who are satisfied with their jobs—are strengths or weaknesses of their public service? Interpreting survey results—and understanding areas for development in the public service—is often greatly aided by comparison. By benchmarking themselves with other countries on the same survey response, governments can understand where their strengths and weaknesses lie. This is one of the founding motivations of the Global Survey of Public Servants (GSPS) initiative (Fukuyama et al. 2022). Similarly, benchmarking internally between groups of public servants—for example, by gender, education, or institution—can help governments understand where, inside government, strengths and weaknesses lie.

However, such benchmarking presupposes comparability in measurement and the survey response process. In other words, it presupposes that respondents understand concepts—such as leadership, motivation, and satisfaction—in the same manner across different countries, government institutions, or groups (for example, men and women) in public service, and that they face similar biases (for example, social-desirability bias) when responding to survey questions. If the same concepts mean different things to different public servants or trigger different response biases in different public servants, valid comparisons are no longer possible, as differences in means might stem from differences in understanding or bias rather than differences in the underlying concept (for example, differences in actual work motivation).

Many public service survey questions are filtered through cultural factors (for example, “My direct superior leads by setting a good example”), individual-level characteristics, like gender (“I am paid at least as well as colleagues who have job responsibilities similar to me”), or both (“I feel sympathetic to the plight of the underprivileged”). If that is the case, then the survey measure lacks *measurement invariance*, which is “a property of a measurement instrument (in the case of survey research, a questionnaire), implying that the instrument measures the same concept in the same way across various subgroups of respondents” (Davidov et al. 2014, 58).

Past research suggests that measurement invariance might affect some measures in surveys of public servants and, in particular, public service motivation (PSM) (Kim et al. 2013; Mikkelsen, Schuster, and Meyer-Sahling 2020). However, what it *means* to be motivated to serve the public in all its dimensions—such as commitment to public values or compassion—is, arguably, highly dependent on cultural factors. As such, concerns about PSM’s lack of cross-country comparability might not travel to other survey questions with less-cultural and more-factual content.

To assess this empirically, this chapter assesses what is arguably a key determinant of public administration effectiveness: the quality of leadership and, in particular, the concept of *transformational leadership*, a style of leadership that inspires and motivates subordinates to go beyond their self-interest and expectation of pecuniary rewards to achieve their goals and an organization’s targets (Jensen et al. 2019; Pearce et al. 2002). Transformational leadership has been found to positively affect performance in public sector organizations across multiple contexts (Hameduddin and Engbers 2021; Pandey et al. 2016; Schuster et al. 2020).

Methodologically, we follow Mikkelsen, Schuster, and Meyer-Sahling (2020, 740) and undertake a measurement-invariance analysis given that “systematic cross-cultural and cross-national measurement-invariance analyses are central to gauge the comparability and generalizability.” We apply the measurement-invariance analysis to an original seven country survey of public servants, in which transformational leadership is measured with exactly the same measurement scale across countries. We assess measurement invariance across countries and within countries across government institutions, as well as across public servants with different genders and education levels.

Our chapter is organized as follows. The chapter begins with a review of the measurement-invariance literature, with a particular focus on its application in the field of public service surveying and on the concept of transformational leadership within the civil service. It then proceeds to describe the approach taken to analyze the measurement invariance of the concept of transformational leadership, including the data set used and the method of analysis: multigroup confirmatory factor analysis (MGCFAs). After that, we present our results—first, for cross-country comparisons and then for within-country comparisons, for civil servants grouped by gender, education level, and organization. We then discuss the theoretical and practical implications of our results and conclude.



## LITERATURE REVIEW

### The Concept of Measurement Invariance

It is common for surveys to aggregate individual questions into larger, overarching constructs. For example, the United States Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS) calculates three subindexes pertaining to some key aspects of public service functioning (“leaders lead,” “supervisors,” and “intrinsic work experience”), each composed by averaging positive responses to five survey questions. These are, in turn, aggregated into an “employee engagement index” (OPM 2019). To take another example, the United Kingdom’s Civil Service People Survey also calculates an “employee engagement index,” tabulated over five questions selected based on factor analysis from pilot surveys (Cabinet Office 2019). Despite similarities in their names and their high-income, English-speaking country settings, however, these two measures cannot be directly compared with each other, due to differences in wording and survey methodology. However, another long-standing concern of survey researchers is the possibility that even exactly the same questions can be interpreted differently by various groups of respondents. Engagement measured with the same battery of questions could still be conceived differently by civil servants in the United States and the United Kingdom due to cultural differences, institutional context, or socioeconomic factors.

Therefore, in order to meaningfully compare a statistical construct, like engagement, motivation, or leadership, and related statistical quantities, like means and regression coefficients, across different groups (or time periods), the construct should first be tested for measurement invariance. Demonstrating the measurement invariance (sometimes also termed equivalence) of a given construct entails showing that it is interpreted in a comparable manner by different sets of respondents. In contrast, “measurement *non*-invariance suggests that a construct has a different structure or meaning to different groups or on different measurement occasions in the same group, and so the construct cannot be meaningfully tested or construed across groups or across time” (Putnick and Bornstein 2016, 71; emphasis added).

Three basic levels of measurement invariance are usually distinguished: *configural*, *metric*, and *scalar* (Vandenberg and Lance 2000). They represent progressively stricter tests for comparability between groups. Figure 24.1, below, provides a schematic representation of the generalized idea behind these concepts by illustrating how each of them hypothesizes the relationship between manifest variables and underlying latent constructs. A more detailed visualization is provided by figures L.1, L.2, and L.3 in appendix L. They demonstrate different levels of invariance using examples of models of transformational leadership in public service that are analyzed throughout this chapter.

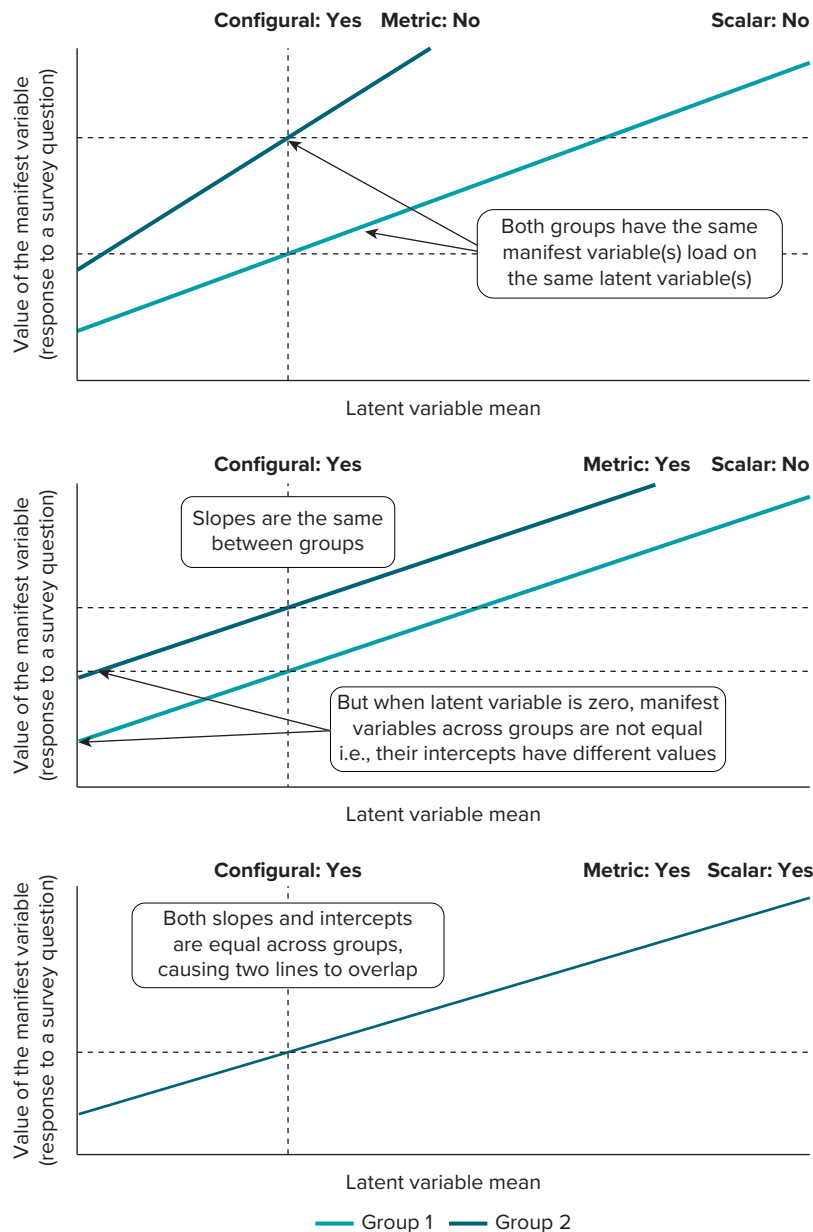
#### Configural Invariance

In the first step toward establishing the comparability of a statistical concept, it needs to be ascertained that it has the same *factor structure* across all groups being compared. This means that in all groups, the same sets of questions are linked to the same sets of underlying constructs, typically termed *latent factors* or *variables* (Kim et al. 2013). This is schematically represented by the top panel of figure 24.1 and, on the example of transformational leadership, by figure L.1 in appendix L. If, in both groups of interest, it can be shown that a model with all observed survey questions loading onto a single latent variable fits the data well, then configural invariance is deemed to hold (upper panel of figure L.1). However, if this hypothesized model is found not to fit the data, then configural invariance cannot be said to hold. One example of such a situation could be where data from one group display a two-latent-factor structure, as in the bottom panel of figure L.1.

#### Metric Invariance

Once the same structure of the items and factors is confirmed across groups, researchers might turn their attention to the equality of *factor loadings* between the groups. Factor loadings can be understood

**FIGURE 24.1** Schematic Visual Representation of the Three Levels of Measurement Invariance: Configural, Metric, and Scalar



Source: Adapted from Cieciuch et al. 2019, 179.

Note: "The X axis represents the latent variable mean; the Y axis represents the response to a survey question item measuring the latent variable. The diagonal represents the function relation between the latent variable and the response to the survey question item in two countries (in unstandardized terms)" (Cieciuch et al. 2019, 179). Here, *countries* is replaced by the more generic term *groups*.

as measures of the strength of the relationship between the observed survey items and the latent factors. Confirming metric invariance is a pre-requisite that "ensures that *structural regression estimates* are comparable across groups" (Mikkelsen, Schuster, and Meyer-Sahling 2020, 4; emphasis added). This is because, in metric-invariant models, differences between survey items are linked to differences in the underlying latent-factor models in the same fashion across all the groups included in the analyses (Steenkamp and Baumgartner 1998). The equal slopes of the lines in the middle panel of figure 24.1 demonstrate this point. In other words, one unit change in the *x*-axis value of the latent variable is associated

with a change in manifest variable values that is the same for both groups. Consequently, only with metric invariance can regression with observed survey items be compared in a meaningful manner (Hong, Malik, and Lee 2003). This focus of metric invariance on the equality of factor loadings across groups is also shown in figure L.2 in appendix L.

### Scalar Invariance

To ensure scalar invariance, not only factor loadings but also the means of the *item intercepts* must be shown to be equal between groups (Vandenberg and Lance 2000). Even when factor loadings suggest that the latent constructs have the same impact upon the value of observed items in all the groups considered, as in the definition of metric invariance, it is still possible for groups to have different values of the intercept—that is, the value of observed items when the latent variable is zero—due to some unobservable characteristics.

Only when the intercepts are the same in all groups, as in the bottom panel of figure 24.1, can the model be said to be scalar invariant. This is also the situation presented in figure L.3 in appendix L. Thus, establishing scalar invariance should precede any attempt at comparing the latent means and intercepts of observed items between the groups. Only once it is established can it be assumed that “cross-national differences in the means of the observed items are due to differences in the means of the underlying construct(s)” (Steenkamp and Baumgartner 1998, 80).

### Measurement Invariance

Although first developed in the mid-20th century in the field of psychology (see, for example, Meredith 1964; Struening and Cohen 1963), measurement invariance has since become a concern in multiple other disciplines. In the field of education, researchers have applied it to better understand the comparability of concepts such as the time management of US undergraduate students (Martinez 2021) or the different subscales of school climate measured in the Georgia School Climate Survey (La Salle, McCoach, and Meyers 2021). The Organisation for Economic Co-operation and Development (OECD) has used it to gauge the comparability of latent factors measured by the Programme for International Student Assessment (PISA) and several other cross-country surveys (Van De Vijver et al. 2019). It has also been analyzed in other contexts as diverse as consumer research (De Jong, Steenkamp, and Fox 2007) and sociology—for example, to better understand concepts such as attitudes toward granting citizenship rights in the International Social Survey Program (Davidov et al. 2018) and German adolescents’ life attitudes (Seddig and Leitgöb 2018). Given its importance and the widespread academic interest in it, public administration researchers, in the face of increasing surveying efforts, have also come to analyze measurement invariance in public service surveys.

### Measurement Invariance in Public Service Surveys

In recent years, multiple researchers have emphasized the importance of studying public administration from a comparative perspective, both to improve researchers’ theoretical understanding and to draw practical lessons (Fitzpatrick et al. 2011; Jreisat 2005). At the same time, surveys have become one of the key methods used to better understand public administration. As emphasized throughout this part of the *Handbook*, surveys allow researchers and policy practitioners to gain insights into dimensions of public administration’s functioning that would otherwise be unmeasurable. Concepts such as job satisfaction or attitudes toward management could scarcely be gauged otherwise. Surveys can also be used to anonymously ask about aspects of civil servants’ work that might otherwise not be talked about—such as perceptions of corruption or workplace harassment. However, the comparability of survey results—both across countries and across demographic groups within the civil service—cannot be taken for granted. Challenges to comparability stem from several sources, like differences in the mode of survey delivery (see chapter 19) or perceived question sensitivity (see chapter 22). Another obstacle is the different phrasing of questions—an issue that initiatives such as the GSPS have recently begun to address (Fukuyama et al. 2022). However, even with all these problems solved, it is not certain that the same survey concepts would be understood in the same way by different groups of civil servants.

This explains the recent turn of several public administration scholars toward analyzing measurement invariance across public service surveys. Kim et al. (2013) test for measurement invariance in PSM—one of the frequently recurring parts of many public service survey questionnaires, which aims to measure respondents' motivation and willingness to serve society. The authors use a PSM index containing questions asked using a 1–5 Likert scale. The tests for configural invariance suggest that PSM has the same structure in 8 out of the 12 countries studied. However, neither metric nor scalar invariance can be detected, meaning that the construct has a different meaning and different levels across countries (apart from the sample restricted to the culturally similar Australia, the United Kingdom, and the United States, for which metric invariance can be detected). Mikkelsen, Schuster, and Meyer-Sahling (2020) expand these results by using a more diverse sample of countries and larger sample sizes within countries. Using survey results from over 23,000 civil servants across 10 countries, they demonstrate that a 16-item PSM scale displays first- and second-order partial metric invariance, apart from the case of two Asian countries studied (Bangladesh and Nepal). Still, their study finds that PSM levels cannot be compared across countries due to a lack of scalar invariance.

### The Concept of Transformational Leadership

In the face of the expansion of measurement-invariance studies within the field of public administration, a relative lack of attention to concepts other than PSM can be discerned. Although PSM is of clear importance and is commonly measured, many other dimensions of work affect civil servants' performance and are regularly included in public service surveys. One key concept, measured in some form in virtually every public service survey, is leadership. Its measurement is most frequently based on past research, scales, and wording from the management science and psychology literature (Tummers and Knies 2016). In particular, the idea of *transformational leadership* has gained traction with public administration researchers (see, for example, Kroll and Vogel 2014; Pandey et al. 2016). It was first developed in the 1970s by Downton (1973) and more fully by Burns (1978), who applied it, together with the contrasting idea of *transactional leadership*, to study political leaders. Whereas transactional leadership is conceived as a leadership style focused on tangible benefits obtained via exchange between a leader and followers (for example, jobs for votes), transformational leadership is chiefly focused on motivating and engaging potential followers to move in a desired direction by conveying a sense of mission, employing compelling argumentation, and using one's own example. Under transformational leadership, followers are inspired to maximize their performance and achieve set goals for the sake of "higher level needs such as self-actualization" (Pearce et al. 2002, 281), attention, and personal development (Nguyen et al. 2017). Bass (1985) extended both conceptions of leadership to the management of organizations. Since then, transformational leadership has been found to be one of the key factors explaining improvements in many dimensions of performance in multiple settings in the private sector, including increased agreement on strategic goals in a large Israeli telecommunications firm (Berson and Avolio 2004), satisfaction with supervisors in Turkish boutique hotels (Erkutlu 2008), and knowledge management in Spanish firms (García-Morales, Lloréns-Monte, and Verdú-Jover 2008). A meta-analysis of 113 primary studies on the topic by Wang et al. (2011) finds transformational leadership to be associated with better performance across the individual, team, and organizational levels.

Moreover, a recent meta-analysis of the PSM and leadership literature, conducted by Hameduddin and Engbers (2021), has found that 50 percent ( $n = 20$ ) of publications concerned with leadership rely on the concept of transformational leadership, making it the most common conceptualization of leadership by public administration scholars. Following this approach, Park and Rainey (2008) establish a positive relationship between transformational leadership and outcomes such as job satisfaction, quality of work, and perceived performance across US federal agencies. Pandey et al. (2016) find its direct and indirect impacts on normative public values. Donkor, Sekyere, and Oduro (2022) further find that higher transformational leadership is linked to higher organizational commitment across 16 Ghanaian public sector organizations. In a survey of over 21,000 civil servants in Chile, Schuster et al. (2020) similarly find transformational leadership to be correlated with higher job satisfaction, motivation, and engagement. Hameduddin and Engbers' (2021) review

of 40 studies finds a link between transformational leadership and PSM—a relationship that holds across a diverse set of countries analyzed.

Transformational leadership was therefore chosen as the survey instrument of focus in the present chapter because of its solid theoretical development, extensive academic research pedigree, and practical importance for public sector performance. Two further reasons can be adduced to explain this choice. First, it is a concept that can usually be mapped onto a single underlying construct. In other words, survey questions about transformational leadership are all aimed at measuring different but related aspects of the same latent factor. This is often not the case with many other sections of public service surveys, like salaries or performance management, which measure many divergent subdimensions—including administrative (for example, salary amount and participation in performance evaluations), motivational (for example, satisfaction with salary and usefulness of performance evaluations), and ethical (for example, salary and performance evaluations' fairness) subdimensions.

Second, there exists a relative imbalance between the large number of studies relying on measures of transformational leadership in the public sector and the lack of research investigating the measurement invariance of this concept. To the best of the authors' knowledge, the only analysis of measurement invariance focused on transformational (and transactional) leadership is a paper by Jensen et al. (2019). However, it presents only a limited test of measurement invariance for transformational leadership, as it focused on full configural and metric invariance, without tests of partial metric invariance or scalar invariance. Jensen et al. (2019) also do not engage in cross-country or cross-cultural analysis of invariance because their sample is composed of respondents from Denmark. The authors focus on invariance across time, sector (including public vs. private), and randomized training groups but not demographic variables, like gender or education, or organizations within the public sector—a focus of the present chapter. Thus, although transformational leadership has gained a well-established position within the public administration literature, only limited attention has been paid to testing the measurement invariance of this concept, which provides the rationale for the analyses contained in the pages below.

## METHODOLOGY

### Data Set

The data used for the analysis in this chapter come from the GSPS initiative. The GSPS is a combined effort of researchers at the World Bank's Bureaucracy Lab, University College London (UCL), the University of Nottingham, and Stanford University that aims to better understand the attitudes and behaviors of civil servants around the globe. Part of the GSPS is focused on making public administration survey questionnaires more comparable. It strives to achieve this by developing and promoting the inclusion of a "core" survey module, which would ask the same set of questions about the principal dimensions of civil service work, such as job satisfaction, work motivation, and leadership, to all the civil servants surveyed.

Seven public service surveys are included in the analyses below. They come from the following countries: Albania, Bangladesh, Brazil, Chile, Estonia, Kosovo, and Nepal.<sup>1</sup> Together, the surveys gathered responses from over 21,000 civil servants. Surveys were delivered both online and in person between 2017 and 2018 and included an extensive set of questions pertaining to multiple aspects of civil service functioning.<sup>2</sup> Importantly for present purposes, the phrasing of questions was exactly the same across countries. In order to ensure that respondents' understanding of the questions would remain unaffected by translation into local languages, the questions were pretested using cognitive interviews with civil servants and iteratively revised (Mikkelsen, Schuster, and Meyer-Sahling 2020). Moreover, each survey strove to include a comparable sample of respondents—that is, central government civil servants who perform general administrative duties.<sup>3</sup> Due to incomplete personnel records on civil servants, the samples are not fully representative. Furthermore, in the in-person surveys, informal quota sampling and in-person surveys based on information from

individual public administration organizations were used (see Mikkelsen, Schuster, and Meyer-Sahling 2020). When possible, the demographics of the survey samples were compared to servicewide values (see table 24A.1), and those comparisons reveal broadly aligned values.

The final advantage of the present choice of surveys is that they represent a diverse set of regional and economic groupings: from South America through Europe to Asia, and from lower-middle-income countries, like Bangladesh and Nepal, through upper-middle-income Albania, Brazil, and Kosovo to high-income Chile and Estonia (see table 24.1). This allows analyses in this chapter to not only focus on differences between groups within each civil service but also to compare invariance across the cross-cultural contexts of different regions and countries.

In the present sample of civil servant surveys, the concept of transformational leadership was measured using the level of agreement with the following three questions, all starting with the prompt “To what extent do you agree with the following statements?”:

1. My direct superior articulates and generates enthusiasm for my organization’s vision and mission (abbreviated as *enthusiasm*).
2. My direct superior leads by setting a good example (abbreviated as *good example*).
3. My direct superior says things that make employees proud to be part of this organization (abbreviated as *pride*).

The responses were measured using a 1–5 Likert scale, where 1 signified “strongly disagree” and 5 “strongly agree.” The basic statistics on each of the variables are presented in table 24.2. A majority of the respondents agree with the question prompts, confirming that their direct superiors generate enthusiasm about the organization’s vision and mission, lead by setting a good example, and make them proud to be a part of the organization. Correlations between the three variables are also very high ( $>0.75$ ), which could be interpreted as an early indication that they indeed measure one underlying concept of transformational leadership.

**TABLE 24.1 Summary of the Seven Public Servant Surveys Used in the Chapter**

	Albania	Bangladesh	Brazil	Chile	Estonia	Kosovo	Nepal
Respondents	3,690	1,049	3,992	5,742	3,555	2,465	1,249
Response rate	47%	Convenience sample	11%	37%	25%	14%	Convenience sample
Mode of delivery	Online	In-person	Online	Online	Online	Online	In-person
Year	2017	2017–18	2018	2016–17	2017	2017	2017–18
Language	Albanian	English, Bangla	Portuguese	Spanish	Estonian	Albanian, Serbian	English, Nepali
Report	Meyer-Sahling et al. (2018d)	Meyer-Sahling et al. (2019)	Pereira et al. (2021)	Schuster et al. (2017)	Meyer-Sahling et al. (2018a)	Meyer-Sahling et al. (2018b)	Meyer-Sahling et al. (2018c)
Region <sup>a</sup>	ECA	South Asia	LAC	LAC	ECA	ECA	South Asia
Income group <sup>a</sup>	Upper-middle income	Lower-middle income	Upper-middle income	High income	High income	Upper-middle income	Lower-middle income
GDP per capita (current US\$) <sup>a</sup>	\$5,246	\$1,967	\$6,797	\$13,232	\$23,027	\$4,347	\$1,155

Source: Original table for this publication.

Note: ECA = Europe and Central Asia; LAC = Latin America and the Caribbean.

a. Based on World Bank data and groupings.



**TABLE 24.2 Basic Statistics on the Three Questions Aiming to Measure Transformational Leadership**

Statistic	Variable		
	Enthusiasm	Good example	Pride
Mean	3.59	3.74	3.40
Median	4.00	4.00	4.00
SD	1.29	1.27	1.28
Skew	−0.63	−0.81	−0.42
Kurtosis	2.29	2.61	2.11
Corr. with <i>enthusiasm</i>	1.00	0.80	0.84
Corr. with <i>good example</i>	0.80	1.00	0.79
Corr. with <i>pride</i>	0.84	0.79	1.00

Source: Original table for this publication.

Note: All variables are measured on a 1–5 Likert scale, where higher values indicate greater agreement. The values shown in the table are aggregated across countries. SD = standard deviation.

## Measuring Invariance

As discussed more broadly in the literature review section, invariance can be measured on three key levels: configural, metric, and scalar. These levels of invariance are tested here using MGCFA. This has been the main method of testing measurement invariance in the past three decades (see, for example, Hofman, Mathieu, and Jacobs 1990; Mikkelsen, Schuster, and Meyer-Sahling 2020; Putnick and Bornstein 2016). It is carried out by setting progressively stricter constraints upon the parameters of the model being evaluated. First, the same model structure is imposed on all groups tested. If the model fit proves satisfactory (see the subsection below for criteria on this), metric invariance is tested by restricting factor loadings to be equal across groups. If the results from the comparison of model fit show that the constrained model is not performing significantly worse than the unconstrained one, then metric invariance can be inferred. Upon finding evidence of metric invariance, the means of the latent construct can be set to be equal across the groups, and, if this extra restriction also does not result in significantly worse model fit, then scalar invariance can be ascertained.

The first set of MGCFA tests pertains to measurement invariance across the seven countries included in the study. First, models for a set of countries grouped by region and income level are fit, before moving to full cross-country models. The results therefore demonstrate the extent to which national context determines how civil servants understand the concept of transformational leadership. The second set of analyses turns toward demographic groups within countries and evaluates whether respondents of different genders (female vs. male), education levels (below university vs. university), and organizations within the public administration interpret the questions about transformational leadership in the same manner. These two levels of analysis—inter- and intracountry—have been the key focus of measurement-invariance research (Vandenberg and Lance 2000).

## Model Fit Indexes

To compare the progressively more restricted measurement-invariance models, one has to calculate how well they fit the data. Three measures are relied upon for this purpose. The first one is chi-square ( $\chi^2$ ). This is a likelihood ratio test that calculates how well the specified model and the associated expected distributions fit the observed data distributions. The  $\chi^2$  value, combined with the model's degrees of freedom, can be used

to calculate the  $p$ -value—the likelihood that the observed deviation from the perfect model is due to chance. However, researchers are in agreement that, because the mathematical formula for its derivation is dependent on the sample size ( $N$ ), this statistic is highly sensitive in large samples and might show statistically significant differences in model fit even when only small deviations from perfect fit are present (Byrne, Shavelson, and Muthén 1989; Cheung and Rensvold 2002; French and Finch 2006; Putnick and Bornstein 2016).

For this reason, two further fit indexes are consulted when comparing model fit. One is the comparative fit index (CFI). Its value is scaled between 0 and 1 and is specifically designed to deal with the limitations of  $\chi^2$ , including its oversensitiveness in large samples (Bentler 1990). The model might be assumed to fit well already when the CFI is above 0.90 (Cheung and Rensvold 2002), but a more restrictive threshold of 0.95 is typically used (Hooper, Coughlan, and Mullen 2008; Hu and Bentler 1999). However, in the measurement-invariance literature, if restricting model parameters leads to a decrease in the CFI of more than 0.01, the invariance is typically rejected (Cheung and Rensvold 2002).

The third and final fit index consulted throughout the analyses is the standardized root mean squared error (SRMR) (see Bentler 1995). The SRMR is calculated on a range from 0 to 1 and can be viewed “as the average standardized residual covariance” of the model variables (Shi, Maydeu-Olivares, and Rosseel 2020, 2). It can range from 0 to infinity, and, typically, absolute SRMR values below 0.05 are indicative of good model fit, although values up to 0.08 are deemed satisfactory (Hu and Bentler 1999). When the fit of models is compared for invariance, increases in the SRMR of more than 0.03 and 0.01 are taken as signaling significant model deterioration in metric- and scalar-invariance models, respectively (Chen 2007). Given the large sample sizes used here, the concern with the overrejection of invariant models by the SRMR raised by Chen (2007) is largely ameliorated.<sup>4</sup>

Therefore, when discussing model fit below, whether in absolute terms or when comparing its fit to another model, changes (indicated with  $\Delta$ ) in all fit indexes are reported (the  $p$ -value of  $\Delta\chi^2$ ,  $\Delta$ CFI, and  $\Delta$ SRMR). The models are estimated in RStudio using the `lavaan::cfa()` function. Given a nonsymmetrical distribution and the ordinal nature of the data (see table 24.2), a diagonally weighted least squares (DWLS) estimator is used for model estimation (Li 2016; Rosseel 2012). Comparisons of model fit ( $\chi^2$ , CFI, and SRMR values), are made using the `semTools::compareFit()` function.

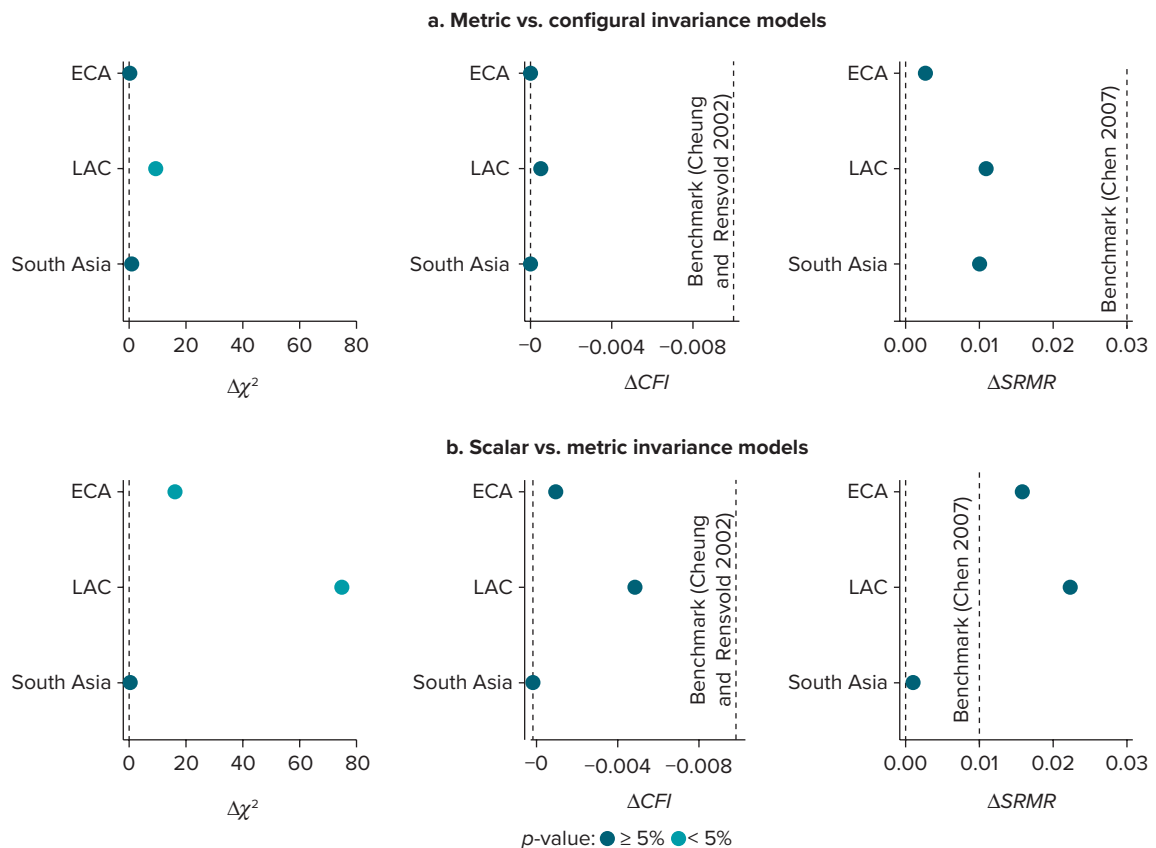
## RESULTS

Measurement invariance is tested first in the cross-country context before moving to within-country invariance across demographic groups (gender and education level) and public administration organizations. We start by fitting the cross-country comparison in groupings of countries based on their income and region before moving to compare individual countries to each other. In each case, configural-, metric-, and scalar-invariance models are tested sequentially, provided that the acceptable fit of a higher-level model is first confirmed.

### Cross-Country Comparison

The analysis begins with models comparing groups of like countries against each other. It is expected that civil servants in similar countries—that is, those at comparable levels of development or in a single geographical region—are more likely to conceive of transformational leadership in the same manner. Such grouping of countries ensures that the inevitable cultural and socioeconomic differences between countries are minimized. By contrast, comparing a high-income European country, like Estonia, and a large, upper-middle-income country in the heart of Latin America, like Brazil, is a much more demanding test of the invariance concept. Therefore, we move to the latter only after establishing that invariance holds within broader groupings of like countries.

**FIGURE 24.2 Measurement Invariance across Countries Classified by Region: Change in Model Fit Indexes across Configural-, Metric-, and Scalar-Invariance Models**



Source: Original figure for this publication.

Note: Regional classifications are based on World Bank data. The Europe and Central Asia (ECA) Region includes Albania, Kosovo, and Estonia; Latin America and the Caribbean (LAC) includes Brazil and Chile; and South Asia includes Bangladesh and Nepal. CFI = comparative fit index; SRMR = standardized root mean squared error.

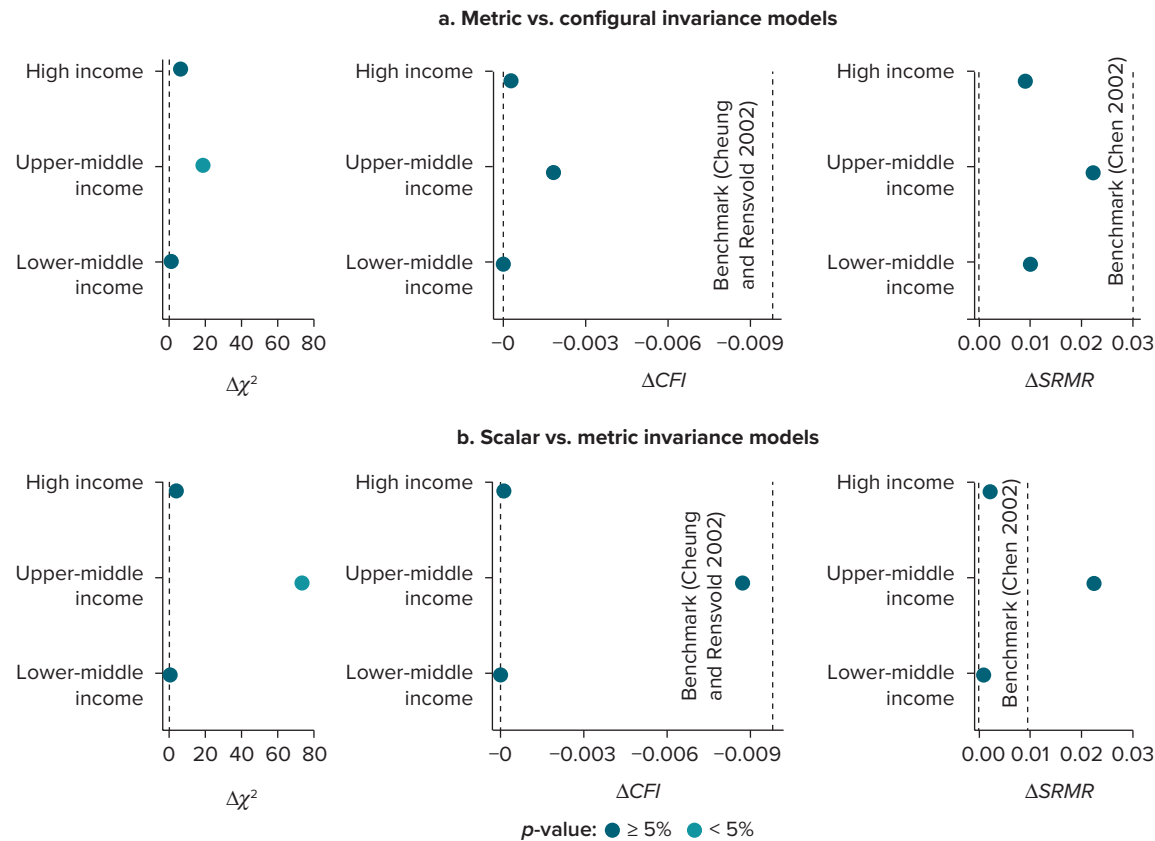
The results of comparing measurement invariance across countries within the same geographical region are shown in figure 24.2. Since all configural-invariance models with only three manifest variables and one latent factor have, by definition, a perfect level of fit, what the figure shows is the change ( $\Delta$ ) in key fit statistics— $\chi^2$ , CFI, and SRMR—between configural- and metric-invariance models and between metric- and scalar-invariance ones. For consistency, the changes in the CFI are reversed, meaning that model fit is deteriorating when going right along the  $x$  axis.

From the figure, it can be seen that metric invariance models fitted across countries from the Europe and Central Asia region (Albania, Estonia, and Kosovo) and for the South Asia region (Bangladesh and Nepal) do not exhibit significantly worse fit on all three indexes. For Latin America and the Caribbean (LAC) (Brazil and Chile), only the  $\Delta\chi^2$  is statistically significant, but, given very low changes in the other two indexes, metric invariance can still be inferred.

Taking the metric-invariant models to the next level and imposing scalar invariance, model fit remains fully acceptable for South Asia. For ECA and LAC, both the  $\Delta\chi^2$  and the  $\Delta SRMR$  point to a significantly worse fit, and, therefore, as with the full cross-country model, scalar invariance can be only tentatively inferred based on the fact that the  $\Delta CFI < 0.01$ .

Figure 24.3 shows the results of the same analyses replicated across income groupings rather than regions. On the basis of the  $\Delta CFI$  and the  $\Delta SRMR$ , all three income groupings exhibit metric invariance. Only a high  $p$ -value for the upper-middle-income group (Albania, Brazil, and Kosovo) points toward a

**FIGURE 24.3 Measurement Invariance across Countries Classified by Income Group: Change in Model Fit Indexes across Configural-, Metric-, and Scalar-Invariance Models**



Source: Original figure for this publication.

Note: Income groupings are based on World Bank data. High-income countries include Chile and Estonia; upper-middle-income countries include Albania, Brazil, and Kosovo; and lower-middle-income countries include Bangladesh and Nepal. CFI = comparative fit index; SRMR = standardized root mean squared error.

different conclusion, but, as explained in the methodology section, this is not taken as sufficient evidence to overrule a good fit based on the CFI and the SRMR.

If such an interpretation is adopted, scalar-invariance models can be estimated for all income groupings. According to all fit indexes, scalar invariance can be inferred across high-income (Chile and Estonia) and lower-middle-income (Bangladesh and Nepal) countries. For the upper-middle-income countries, the  $\Delta\chi^2$  is statistically significant, and the  $\Delta SRMR$  is well above the threshold of 0.01. The absolute value of the SRMR of the scalar-invariance model is also only borderline acceptable, at 0.046. The  $\Delta CFI$ , standing just below 0.009, similarly approaches the threshold of significant deterioration. Therefore, a conclusion of scalar invariance can be drawn only on the basis of the CFI, and, even then, it is not strong. (It should also be noted that the results and conclusions for countries in the lower-middle-income category are exactly the same as for the South Asia category above because those two groups happen to contain the same pair of countries: Bangladesh and Nepal.)

Given the relatively robust evidence of metric and scalar invariance within groupings of comparable countries, we now move to compare all seven countries against each other. When the metric-invariance model is fitted by restricting the factor loadings to be equal across all seven countries, the absolute model fit is still good according to all three fit indexes. The value of  $\chi^2$  is 47.1 ( $df = 12$ ), and the associated  $p$ -value is close to 0. The CFI drops to 0.998, and the SRMR increases to 0.019. Therefore, the change in the latter two fit indexes is well within the limits recommended by the literature. Although the  $\Delta\chi^2$  with a  $p$ -value below 5 percent points toward significantly worse fit, the large sample size and perfect fit of the unrestricted model

make this a less reliable measure. Therefore, metric invariance can be inferred for cross-country comparisons of transformational leadership.

Given this conclusion, a scalar-invariance model can be fitted. It represents a borderline case of significant deterioration. The  $p$ -value of the  $\Delta\chi^2$  is close to 0, and the  $\Delta SRMR$  is 0.018, which is above the threshold recommended for scalar-invariance models by Chen (2007). However, the difference is small, and the absolute model fit (the  $SRMR = 0.037$ ) is still good. Furthermore, the  $\Delta CFI$  of 0.008 can be viewed as acceptable. Therefore, a tentative conclusion of scalar invariance can be reached.

To summarize the above analyses—in a full cross-country analysis, no significant deterioration in the model of metric invariance suggests that researchers and policy practitioners should be able to compare factor loadings and structural regression coefficients across countries. Item intercepts and means of the indicators can also be compared, although cautiously, given that not all fit indexes suggest that the model with equal intercepts fits the data well.

However, comparisons of this type might be more warranted within groups of like countries. There is evidence that cross-cultural differences in understanding of the idea of transformational leadership are (largely) removed by grouping countries according to their geographical regions. Within such groupings, there is clear evidence of metric invariance. With the same caveat as in the full cross-country models, scalar invariance can also be demonstrated for those models. Invariance is even stronger when countries are grouped by their income level. Both high- and lower-middle-income groups exhibit full metric and scalar invariance. The only group where the conclusion of scalar invariance has very little backing is upper-middle-income countries. This is perhaps unsurprising, given that this group can be viewed as the most heterogeneous, and, therefore, differences in the understanding of concepts such as leadership remain substantial.<sup>5</sup>

### Within-Country Comparison: Gender

Turning to intracountry comparisons, gender is the key demographic measure in all public service surveys and also, typically, one of the first lines along which survey results are broken down. The distribution of respondents by gender in the survey sample used here is reported in table 24.3. As can be observed, the gender distribution of civil servants who respond to the surveys varies highly by country. In three out of seven countries, women form the majority of the respondents. The female-to-male ratio varies from approximately 3:1 in Estonia to less than 1:3 in Bangladesh. These cross-country differences are largely consistent with the variation in gender balance across survey populations in countries where personnel records are available (see table 24A.1).

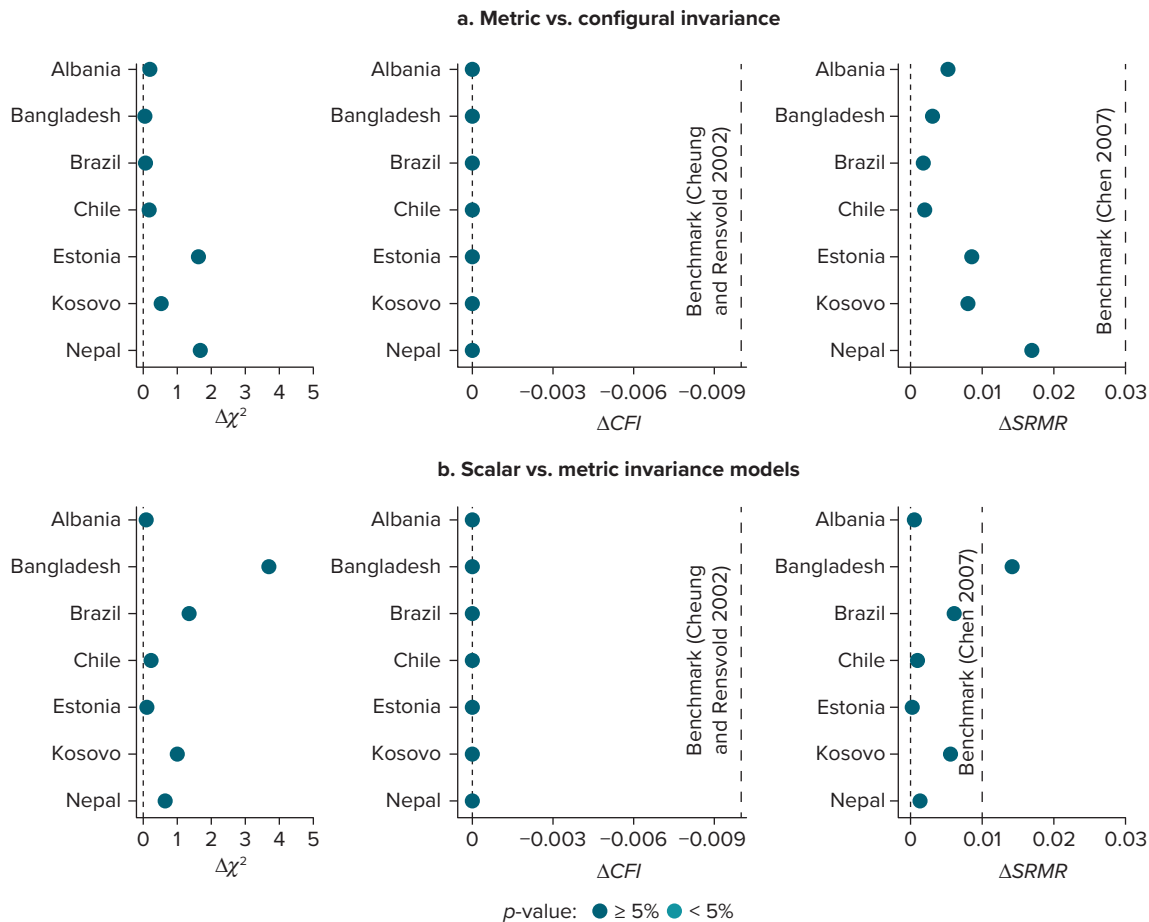
The results of measurement-invariance analyses across gender groups within countries are presented in figure 24.4. Metric invariance—the equality of factor loadings between genders—is obtained with little

**TABLE 24.3** Distribution of Respondents, by Gender

Country	Male	Female	Missing
Albania	1,374 (37.2%)	2,261 (61.3%)	55 (1.5%)
Bangladesh	801 (76.4%)	224 (21.4%)	24 (2.3%)
Brazil	2,268 (56.8%)	1,701 (42.6%)	23 (0.6%)
Chile	2,502 (43.6%)	3,155 (54.9%)	85 (1.5%)
Estonia	845 (23.8%)	2,462 (69.3%)	248 (7.0%)
Kosovo	1,363 (55.7%)	1,028 (42.0%)	57 (2.3%)
Nepal	817 (65.4%)	421 (33.7%)	11 (0.9%)

Source: Original table for this publication.

**FIGURE 24.4 Measurement Invariance across Gender within Countries: Change in Model Fit Indexes across Configural-, Metric-, and Scalar-Invariance Models**



Source: Original figure for this publication.

Note: CFI = comparative fit index; SRMR = standardized root mean squared error.

space for doubt for all seven countries. In none of them are changes in  $\chi^2$  statistically significant, nor are the changes in the CFI and SRMR above their respective thresholds.<sup>6</sup>

As a next step, scalar-invariance models are fitted and compared. Here, the arguments and conclusion remain unchanged. All three fit indexes point to good fit and no significant deterioration of the model after adding equality constraints on item intercepts, which allows us to conclude scalar invariance across genders in all countries considered.

### Within-Country Comparison: Education Level

Like with the analyses focused on gender, this subsection concerning education begins with a demographic overview (table 24.4), which presents the distribution of civil servants by their level of education across countries. Here, the heterogeneity across surveys is even more pronounced than in the case of gender. Whereas in Albania, 92.1 percent of civil servants who responded to the survey had university-level education, and only 5.1 percent had below-university-level education, these proportions are equal in Nepal, and in Chile become almost exactly reversed.

Figure 24.5 demonstrates the results from fitting different levels of invariance models across different education levels in seven countries. As with gender, both metric and scalar invariance can be concluded for all seven countries. None of the changes in the fit indexes come near their respective thresholds.

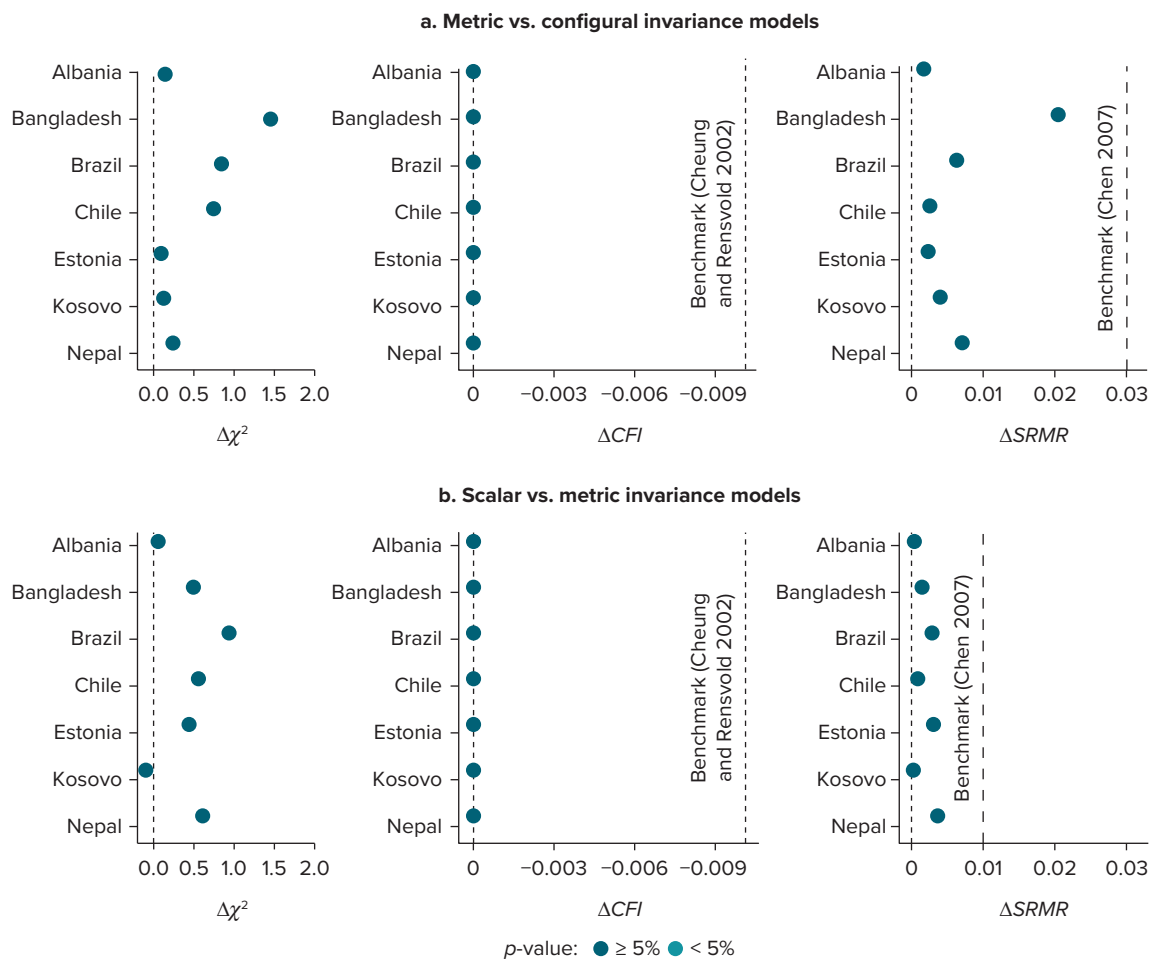


**TABLE 24.4** Distribution of Respondents, by Education Level

Country	University	Below university	Missing
Albania	3,399 (92.1%)	188 (5.1%)	103 (2.8%)
Bangladesh	560 (53.4%)	468 (44.6%)	21 (2.0%)
Brazil	1,964 (49.7%)	1,895 (47.5%)	113 (2.8%)
Chile	586 (10.2%)	5,081 (88.5%)	75 (1.3%)
Estonia	1,898 (53.4%)	1,439 (40.5%)	218 (6.1%)
Kosovo	1,150 (47.0%)	1,261 (51.5%)	37 (1.5%)
Nepal	603 (48.3%)	603 (48.3%)	43 (3.4%)

Source: Original table for this publication.

**FIGURE 24.5** Measurement Invariance across Education Levels within Countries: Change in Model Fit Indexes across Configural-, Metric-, and Scalar-Invariance Models



Source: Original figure for this publication.

Note: CFI = comparative fit index; SRMR = standardized root mean squared error.

## Within-Country Comparison: Public Administration Organization

The final set of invariance models is fitted across public administration organizations. The surveys analyzed here were conducted among several central government organizations in each country (see table 24.5). The number of organizations with more than 50 respondents ranges from 7 in Bangladesh to 27 in Estonia. Across countries, the mean number of respondents per organization varies between 83.9 in Kosovo and 522 in Chile. In the latter country, standing at 1,520, the largest number of respondents per organization is also observed.

Figure 24.6 replicates the measurement-invariance comparisons discussed above for gender and education level. However, here the results are less clear-cut. For metric-invariance models, there is clear evidence to suggest the equality of factor loadings for five out of seven countries. For Kosovo and Nepal, the  $\Delta SRMR$  is, however, just above 0.03, which suggests significant deterioration compared to the configural-invariance model. Yet the  $\Delta\chi^2$  remains small in absolute terms and is also not statistically significant, even though this metric tends to be the most sensitive of the fit indexes. Therefore, metric invariance is concluded for these two countries, albeit with a caveat.

We find similar results when scalar-invariance models are fitted, although here it applies to two additional countries: Bangladesh and Estonia. For these countries, a change in the SRMR points toward significant deterioration in model fit, whereas all other measures suggest acceptable deterioration. Overall, the results suggest that both factor loadings and the means of the transformational-leadership latent factor can be compared across organizations within public administration, but this conclusion is tentative for Kosovo and Nepal, as well as for Bangladesh and Estonia in the case of scalar invariance.

## DISCUSSION

The results of the measurement-invariance analyses of the concept of transformational leadership presented above warrant a tentative two-level conclusion. First, there is strong evidence of metric invariance across countries and tentative evidence of scalar invariance. The latter conclusion can be strengthened if countries are grouped according to region or income level. In that case, full scalar invariance is observed across high-income and South Asian or lower-middle-income countries. Second, transformational leadership appears invariant, both at the level of factor loadings and latent factor means, across gender, broad education level, and organization within public administration in most of the countries studied. The evidence for the

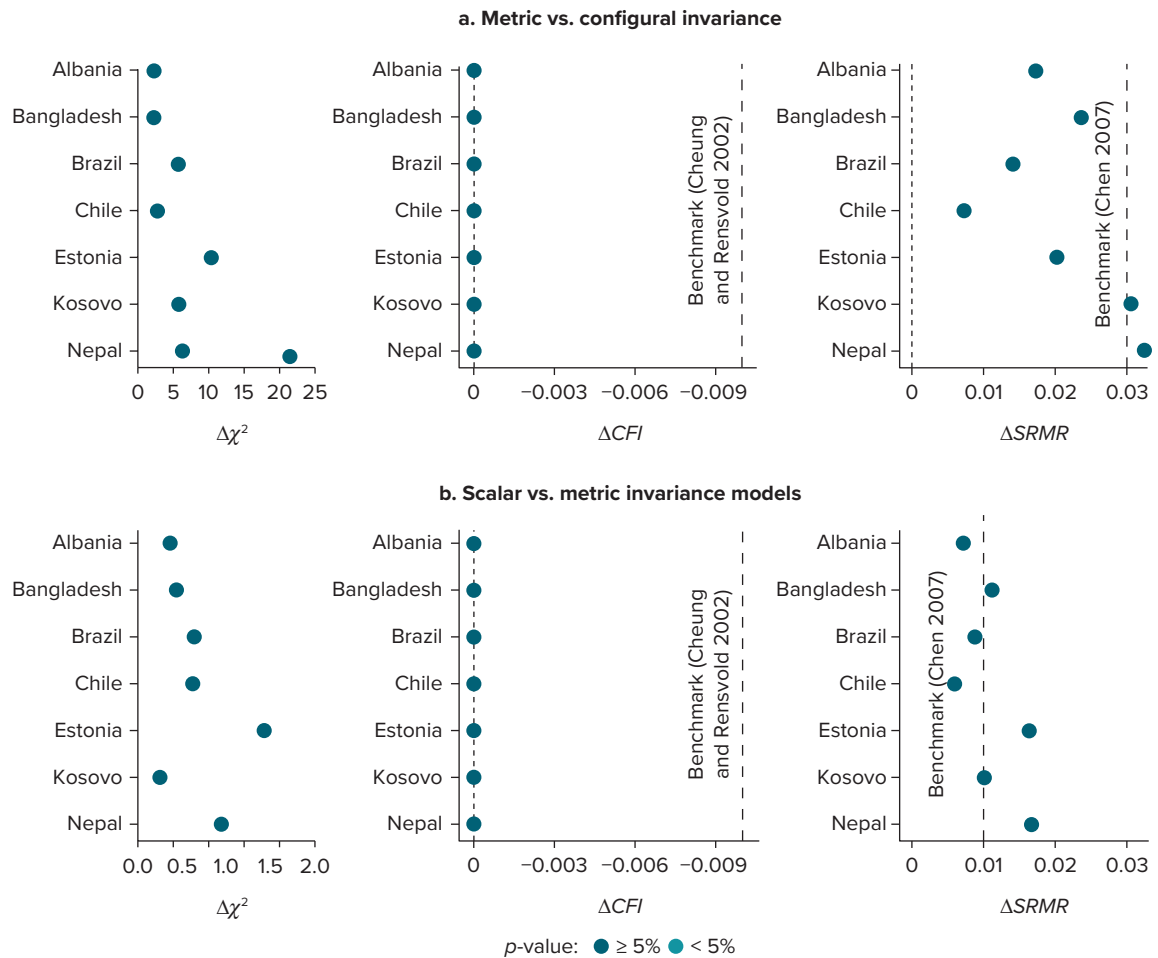
**TABLE 24.5** Distribution of Respondents within Public Administration Organizations

Country	No. of respondents					No. of organizations
	Mean	Median	SD	Min.	Max.	
Albania	215.1	183.0	141.9	83	585	15
Bangladesh	120.0	82.0	70.7	52	218	7
Brazil	292.5	165.0	305.9	57	1,062	12
Chile	522.0	382.0	449.4	87	1,520	11
Estonia	108.1	80.0	70.6	64	331	27
Kosovo	83.9	73.5	30.5	54	150	14
Nepal	97.8	83.5	61.1	55	241	8

Source: Original table for this publication.

Note: Only groups with 50+ observations are included in the analyses. SD = standard deviation.

**FIGURE 24.6 Measurement Invariance across Public Administration Organizations within Countries: Change in Model Fit Indexes across Configural-, Metric-, and Scalar-Invariance Models**



Source: Original figure for this publication.

Note: Only groups with 50+ observations are included in the analyses. CFI = comparative fit index; SRMR = standardized root mean squared error.

first two groups is clear-cut, and, for invariance across organizations, the only caveat that should be raised is the borderline significance of the  $\Delta SRMR$  values in some models.

It is possible that the relatively stronger evidence of invariance within rather than across countries comes from translation differences of the survey items, rather than their differential interpretation. The language of a survey is known to affect the thought and response-forming process of survey respondents, even when care is taken—for instance, through extensive cognitive interviews—to ensure comparable understanding across languages (Peytcheva 2020). Chen (2008) suggests that such difference could also come from a propensity, observable in some cultures, to skew survey responses toward more-neutral options. These possibilities highlight the need not only to standardize the question wording and response scale, as was done here, but also for researchers to retest measurement invariance in public service surveys across further concepts and country settings, including countries with a shared language. Although safeguarding actions were taken to minimize these differences, the results suggest that some residual variation might stem from them.

This unavoidable limitation can, however, serve as a response to another potential criticism of the analyses presented above—namely, the fact that they were focused on only seven countries and a small number of questions concerned with transformational leadership. Including a larger sample of countries would be admittedly desirable, yet it is problematic precisely because the question wording and the wider context of different public service surveys are too dissimilar to warrant inclusion. Most surveys of civil servants include

a leadership section, but they are not directly focused on transformational leadership, and in the cases when they are, their phrasing or response scales make comparisons difficult (chapter 18). Despite including only seven countries, this chapter is, to the best of the authors' knowledge, the first to look at the measurement invariance of the concept of transformational leadership in the public sector in a cross-country context.

## CONCLUSION

Surveying civil servants is often the only feasible way to learn more about their attitudes, behaviors, and work environment. Yet survey results are challenging to interpret without context—comparisons to other countries or previous surveys, or between different demographic groups within the public service. However, researchers and policy practitioners only occasionally pause to statistically assess whether the attitudes and behaviors they want to measure—be they engagement, motivation, or leadership—are understood in the same way by different groups of civil servants.

Drawing on the concept of measurement invariance, this chapter was able to show that when it comes to the differential understanding of the concept of transformational leadership, differences in gender, education level, and organization have a very small impact. In most countries, the three leadership questions relate in the same way to the underlying concept of transformational leadership, and the mean levels of transformational leadership are also comparable across groups. The same can be said when looking across countries, even if the conclusion of the comparability of the mean levels is weakened, to an extent, when comparing countries at different income levels and, in particular, in different regions. This suggests, tentatively, that global benchmarking exercises like the GSPS have a legitimate empirical foundation. Contrasting our results with those of measurement-invariance analyses of PSM—a concept which is arguably even more culturally specific than leadership—suggests, in particular, that questions that are less culturally loaded and more factual—for instance, about management practices rather than culturally specific attitudes—might have a stronger empirical basis for comparison.

Of course, we assess only one measurement scale in our analysis and draw on data from seven countries. Thus, much fertile empirical ground for future cross-country work on measurement invariance remains to further solidify claims about what can be compared across countries and what cannot. At least four further contributions would be especially welcome in the future. First, future investigations should extend analyses of measurement invariance to other recurring topics in public service surveys. Perceptions about work environment, engagement, teamwork, compensation, turnover, performance, meritocratic practices, and harassment are components of many public service surveys. Yet the extent to which they measure the same underlying concepts across different groups of civil servants and across countries is uncertain. A second possible extension of the present analyses would include more countries in the analyses, preferably with heterogeneous geographical and economic features. A third avenue for future work would address the fact that at present, only a limited number of groupings of civil servants have been compared for measurement invariance. This chapter focused on gender, education level, and organization. Including further groupings—by age and tenure level, managerial position, and contract type—would be warranted in future studies. A fourth type of analysis would ascertain intertemporal measurement invariance. Just as the same question can measure divergent concepts across different countries, cultures, or demographic groups, it can be measurement variant across different time periods. Due to changes in social, economic, political, and, in the longer term, cultural conditions, the same survey question might come to be interpreted differently in different time periods, even when asked to the same population.

Along with further investigations of measurement invariance, researchers and practitioners wishing to compare the results of surveys of public servants would be well served by relying, at least in part, on a standardized questionnaire. One such effort is the GSPS initiative, which catalogs 20+ sets of public service survey results, along with their respective questionnaires, section names, and metadata. Including some of the standardized questions would allow for survey results to be more readily compared with other countries'

results and, ultimately, for the establishment of international benchmarks against which civil servants' attitudes and behaviors could be reliably compared. Even when such comparisons are tentative, given concerns with measurement invariance, this certainly trumps comparisons of core concepts (for example, employee engagement) across countries using different measures.

## NOTES

We are grateful to Daniel Rogger and Galileu Kim for helpful comments.

1. Unless justified for other reasons, in all instances countries are listed in alphabetical order. See the GSPS website (<https://www.globalsurveyofpublicservants.org/>) and Mikkelsen, Schuster, and Meyer-Sahling (2020) for further details on the surveys included.
2. Technical limitations, like limited access to electricity, computer, or the Internet, as well as incomplete databases of email records for civil service officials, made online surveying unfeasible in the two Asian countries (Bangladesh and Nepal) included in this chapter.
3. State or local government officials and nonadministrative public sector employees, like teachers, nurses, doctors, policemen, and the military, were thus excluded.
4. It was decided that another commonly employed measure of model fit, the root mean square error of approximation (RMSEA), would not be used in the analyses presented here. The RMSEA was introduced by Steiger and Lind (1980) and extended by, among others, Browne and Cudeck (1993) and Steiger (1998). However, it can be unreliable when comparing just-identified with overidentified models, as is done here. Using Monte Carlo simulations, Kenny, Kaniskan, and McCoach (2015) find that in models with few degrees of freedom, the RMSEA tends to be overinflated and, therefore, falsely points to bad model fit. Moreover, in close-fit models, more restricted models might counterintuitively show a decrease in the RMSEA—that is, better fit—because of the increased number of degrees of freedom (Shi, Lee, and Maydeu-Olivares 2019). Notwithstanding the above, RMSEA values point toward the same broad conclusions as the other three fit indexes consulted in the text.
5. In contrast, the lower-middle-income group is relatively homogeneous, since it is comprised of two South Asian countries.
6. In fact, it can be observed the  $\Delta CFI$  is 0 across all intracountry comparisons. This is because the model fit is close to perfect, and, as a result,  $\chi^2$  is low enough as to be smaller than the number of degrees of freedom. Given the formula used to calculate the CFI, the resulting value of this fit index will always be 1 in those cases (see Bentler 1990).

## REFERENCES

- Bass, B. M. 1985. *Leadership and Performance beyond Expectations*. New York: Free Press.
- Bentler, P. 1990. "Comparative Fit Indices in Structural Models." *Psychological Bulletin* 107 (2): 238–46. <https://doi.org/10.1037/0033-2909.107.2.238>.
- Bentler, P. 1995. *EQS 5* [Computer program]. Encino, CA: Multivariate Software.
- Berson, Y., and B. J. Avolio. 2004. "Transformational Leadership and the Dissemination of Organizational Goals: A Case Study of a Telecommunication Firm." *The Leadership Quarterly* 15 (5): 625–46. <https://doi.org/10.1016/j.leaqua.2004.07.003>.
- Browne, M. W., and R. Cudeck. 1993. "Alternative Ways of Assessing Model Fit." In *Testing Structural Equation Models*, edited by K. A. Bollen and J. S. Long, 136–62. Newbury Park, CA: Sage.
- Burns, J. M. 1978. *Leadership*. New York: Harper & Row.
- Byrne, B. M., R. H. Shavelson, and B. Muthén. 1989. "Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance." *Psychological Bulletin* 105 (3): 456–66. <https://doi.org/10.1037/0033-2909.105.3.456>.
- Cabinet Office. 2019. *Civil Service People Survey 2019: Technical Guide*. London: Cabinet Office, United Kingdom Government. [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/867302/Civil-Service-People-Survey-2019-Technical-Guide.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/867302/Civil-Service-People-Survey-2019-Technical-Guide.pdf).
- Chen, F. F. 2007. "Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance." *Structural Equation Modeling: A Multidisciplinary Journal* 14 (3): 464–504. <https://doi.org/10.1080/10705510701301834>.
- Chen, F. F. 2008. "What Happens If We Compare Chopsticks with Forks? The Impact of Making Inappropriate Comparisons in Cross-Cultural Research." *Journal of Personality and Social Psychology* 95 (5): 1005–18. <https://doi.org/10.1037/a0013193>.

- Cheung, G. W., and R. B. Rensvold. 2002. "Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance." *Structural Equation Modeling: A Multidisciplinary Journal* 9 (2): 233–55. [https://doi.org/10.1207/S15328007SEM0902\\_5](https://doi.org/10.1207/S15328007SEM0902_5).
- Cieciuch, J., E. Davidov, P. Schmidt, and R. Algesheimer. 2019. "How to Obtain Comparable Measures for Cross-National Comparisons." *Kolner Zeitschrift für Soziologie und Sozialpsychologie* 71 (S1): 157–86. <https://doi.org/10.1007/s11577-019-00598-7>.
- Davidov, E., H. Dülmer, J. Cieciuch, A. Kuntzm, D. Seddig, and P. Schmidt. 2018. "Explaining Measurement Nonequivalence Using Multilevel Structural Equation Modeling: The Case of Attitudes toward Citizenship Rights." *Sociological Methods & Research* 47 (4): 729–60. <https://doi.org/10.1177/0049124116672678>.
- Davidov, E., B. Meuleman, J. Cieciuch, P. Schmidt, and J. Billiet. 2014. "Measurement Equivalence in Cross-National Research." *Annual Review of Sociology* 40 (1): 55–75. <https://doi.org/10.1146/annurev-soc-071913-043137>.
- De Jong, M. G., J.-B. Steenkamp, and J.-P. Fox. 2007. "Relaxing Measurement Invariance in Cross-National Consumer Research Using a Hierarchical IRT Model." *Journal of Consumer Research* 34 (2): 260–78. <https://doi.org/10.1086/518532>.
- Donkor, F., I. Sekyere, and F. A. Oduro. 2022. "Transformational and Transactional Leadership Styles and Employee Performance in Public Sector Organizations in Africa: A Comprehensive Analysis in Ghana." *Journal of African Business* 23 (4): 945–63. <https://doi.org/10.1080/15228916.2021.1969191>.
- Downton, J. V. 1973. *Rebel Leadership: Commitment and Charisma in the Revolutionary Process*. New York: Free Press.
- Erkutlu, H. 2008. "The Impact of Transformational Leadership on Organizational and Leadership Effectiveness: The Turkish Case." *Journal of Management Development* 27 (7): 708–26. <https://doi.org/10.1108/02621710810883616>.
- Fitzpatrick, J., M. Goggin, T. Heikkilä, D. Klingner, J. Machado, and C. Martell. 2011. "A New Look at Comparative Public Administration: Trends in Research and an Agenda for the Future." *Public Administration Review* 71 (6): 821–30. <https://doi.org/10.1111/j.1540-6210.2011.02432.x>.
- French, B. F., and H. W. Finch. 2006. "Confirmatory Factor Analytic Procedures for the Determination of Measurement Invariance." *Structural Equation Modeling* 13 (3): 378–402. [https://doi.org/10.1207/s15328007sem1303\\_3](https://doi.org/10.1207/s15328007sem1303_3).
- Fukuyama, F., D. Rogger, Z. Hasnain, K. Bersch, D. Mistree, C. Schuster, K. Mikkelsen, K. Kay, and J. Meyer-Sahling. 2022. *Global Survey of Public Servants Indicators*. <https://www.globalsurveyofpublicservants.org>.
- García-Morales, V. J., F. J. Lloréns-Montes, and A. J. Verdú Jover. 2008. "The Effects of Transformational Leadership on Organizational Performance through Knowledge and Innovation." *British Journal of Management* 19 (4): 299–319. <https://doi.org/10.1111/j.1467-8551.2007.00547.x>.
- Hameduddin, T., and T. Engbers. 2021. "Leadership and Public Service Motivation: A Systematic Synthesis." *International Public Management Journal* 25 (1): 86–119. <https://doi.org/10.1080/10967494.2021.1884150>.
- Hofman, D. A., J. E. Mathieu, and R. Jacobs. 1990. "A Multiple Group Confirmatory Factor Analysis Evaluation of Teachers' Work Related Perceptions and Reactions." *Educational and Psychological Measurement* 50 (4): 943–55. <https://doi.org/10.1177/0013164490504024>.
- Hong, S., M. L. Malik, and M.-K. Lee. 2003. "Testing Configural, Metric, Scalar, and Latent Mean Invariance across Genders in Sociotropy and Autonomy Using a Non-Western Sample." *Educational and Psychological Measurement* 63 (4): 636–54. <https://doi.org/10.1177/0013164403251332>.
- Hooper, D., J. Coughlan, and M. R. Mullen. 2008. "Structural Equation Modelling: Guidelines for Determining Model Fit." *The Electronic Journal of Business Research Methods* 6 (1): 53–60.
- Hu, L.-T., and P. Bentler. 1999. "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives." *Structural Equation Modeling: A Multidisciplinary Journal* 6 (1): 1–55. <https://doi.org/10.1080/10705519909540118>.
- Jensen, U. T., L. B. Andersen, L. L. Bro, A. Bøllingtoft, T. L. Eriksen, A.-L. Holten, C. B. Jacobsen, et al. 2019. "Article Conceptualizing and Measuring Transformational and Transactional Leadership." *Administration & Society* 51 (1): 3–33. <https://doi.org/10.1177/0095399716667157>.
- Jreisat, J. G. 2005. "Comparative Public Administration Is Back in, Prudently." *Public Administration Review* 65 (2): 231–42. <https://doi.org/10.1111/j.1540-6210.2005.00447.x>.
- Kenny, D. A., B. Kaniskan, and B. D. McCoach. 2015. "The Performance of RMSEA in Models with Small Degrees of Freedom." *Sociological Methods & Research* 44 (3): 486–507. <https://doi.org/10.1177/0049124114543236>.
- Kim, S., W. Vandenabeele, B. E. Wright, L. B. Andersen, F. P. Cerase, R. K. Christensen, C. Desmarais, et al. 2013. "Investigating the Structure and Meaning of Public Service Motivation across Populations: Developing an International Instrument and Addressing Issues of Measurement Invariance." *Journal of Public Administration Research and Theory* 23 (1): 79–102. <https://doi.org/10.1093/jopart/mus027>.
- Kroll, A., and D. Vogel. 2014. "The PSM-Leadership Fit: A Model of Performance Information Use." *Public Administration* 92 (4): 974–91. <https://doi.org/10.1111/padm.12014>.



- La Salle, T. P., D. B. McCoach, and J. Meyers. 2021. "Examining Measurement Invariance and Perceptions of School Climate across Gender and Race and Ethnicity." *Journal of Psychoeducational Assessment* 39 (7): 800–15. <https://doi.org/10.1177/07342829211023717>.
- Li, C.-H. 2016. "Confirmatory Factor Analysis with Ordinal Data: Comparing Robust Maximum Likelihood and Diagonally Weighted Least Squares." *Behavioral Research Methods* 8 (3): 936–49. <https://doi.org/10.3758/s13428-015-0619-7>.
- Martinez, A. J. 2021. "Factor Structure and Measurement Invariance of the Academic Time Management and Procrastination Measure." *Journal of Psychoeducational Assessment* 39 (7): 891–901. <https://doi.org/10.1177/07342829211034252>.
- Meredith, W. 1964. "Notes on Factorial Invariance." *Psychometrika* 29: 177–85. <https://doi.org/10.1007/BF02289699>.
- Meyer-Sahling, J., C. Schuster, K. Sass Mikkelsen, C. Pesti, and T. Randma-Liiv. 2018a. *Civil Service Management in Estonia: Evidence from a Survey of Civil Servants and Employees*. London: British Academy, UK Department for International Development Anti-Corruption Evidence Programme. Last updated October 2018. <https://christianschuster.net/Meyer-Sahling%20Schuster%20Mikkelsen%20Pesti%20Randma-Liiv%20Estonia%20Report%20FINAL.pdf>.
- Meyer-Sahling, J., C. Schuster, K. Sass Mikkelsen, H. Qeriqi, and F. Toth. 2018b. *Towards a More Professional Civil Service in Kosovo: Evidence from a Survey of Civil Servants in Central and Local Government*. London: British Academy, UK Department for International Development Anti-Corruption Evidence Programme. <https://christianschuster.net/Meyer-Sahling%20Schuster%20Mikkelsen%20Qeriqi%20Toth%20Kosovo%20Report%20FINAL.pdf>.
- Meyer-Sahling, J., C. Schuster, K. Sass Mikkelsen, T. Rahman, K. M. Islam, A. S. Huque, and F. Toth. 2019. *Civil Service Management in Bangladesh: Evidence from a Survey of More Than 1,000 Civil Servants*. London: British Academy, UK Department for International Development Anti-Corruption Evidence Programme. <https://christianschuster.net/2019.03.10.%20Bangladesh%20FOR%20PUBLICATION.pdf>.
- Meyer-Sahling, J., C. Schuster, K. Sass Mikkelsen, S. K. Shrestha, B. Luitel, and F. Toth. 2018c. *Civil Service Management in Nepal: Evidence from a Survey of More than 1,200 Civil Servants*. London: British Academy, UK Department for International Development Anti-Corruption Evidence Programme. <https://christianschuster.net/2018.12.20.%20Nepal%20FOR%20PUBLICATION.pdf>.
- Meyer-Sahling, J., C. Schuster, K. Sass Mikkelsen, and A. Shundi. 2018d. *The Quality of Civil Service Management in Albania: Evidence from a Survey of Central Government Civil Servants and Public Employees*. London: British Academy, UK Department for International Development Anti-Corruption Evidence Programme. <https://christianschuster.net/Meyer-Sahling%20Schuster%20Mikkelsen%20Shundi%20Albania%20Report%20FINAL.pdf>.
- Mikkelsen, K. S., C. Schuster, and J.-H. Meyer-Sahling. 2020. "A Cross-Cultural Basis for Public Service? Public Service Motivation Measurement Invariance in an Original Survey of 23,000 Public Servants in Ten Countries and Four World Regions." *International Public Management Journal* 24 (6): 739–61. <https://doi.org/10.1080/10967494.2020.1809580>.
- Nguyen, T. T., L. Mia, L. Winata, and V. K. Chong. 2017. "Effect of Transformational-Leadership Style and Management Control System on Managerial Performance." *Journal of Business Research* 70: 202–31. <https://doi.org/10.1016/j.jbusres.2016.08.018>.
- OPM (Office of Personnel Management). 2019. *2019 Office of Personnel Management Federal Employee Viewpoint Survey: Technical Report*. Washington, DC: US Office of Personnel Management, US Government.
- Pandey, S. K., R. S. Davis, S. Pandey, and S. Peng. 2016. "Transformational Leadership and the Use of Normative Public Values: Can Employees Be Inspired to Serve Larger Public Purposes?" *Public Administration* 94 (1): 204–22. <https://doi.org/10.1111/padm.12214>.
- Park, S. M., and H. G. Rainey. 2008. "Leadership and Public Service Motivation in U.S. Federal Agencies." *International Public Management Journal* 11 (1): 109–42. <https://doi.org/10.1080/10967490801887954>.
- Pearce, C. L., H. P. Sims Jr., J. F. Cox, G. Ball, E. Schnell, K. A. Smith, and L. Trevino. 2002. "Transactors, Transformers and Beyond. A Multi-Method Development of a Theoretical Typology of Leadership." *Journal of Management Development* 22 (4): 273–307. <https://doi.org/10.1108/02621710310467587>.
- Pereira, A. K., R. A. Machado, P. L. Costa Cavalcante, A. De Avila Gomide, A. Gomes Magalhaes, I. De Araujo Goellner, R. R. Coelho Pires, K. Bersch, F. Fukuyama, and A. R. Da Silva. 2021. "Government Quality and State Capacity: Survey Results from Brazil." CDDRL Working Paper, Center on Democracy, Development, and the Rule of Law, Stanford University, Stanford, CA. <https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/governancereportbrazil.pdf>.
- Peytcheva, E. 2020. "The Effect of Language of Survey Administration on the Response Formation Process." In *The Essential Role of Language in Survey Research*, edited by M. Sha and T. Gabel, 3–22. Research Triangle Park, NC: RTI Press.
- Putnick, D. L., and M. H. Bornstein. 2016. "Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research." *Developmental Review* 41: 71–90. <https://doi.org/10.1016/j.dr.2016.06.004>.
- Rossee, Y. 2012. "Lavaan: An R Package for Structural Equation Modeling." *Journal of Statistical Software* 48 (2): 1–36. <https://doi.org/10.18637/jss.v048.i02>.
- Schuster, C., J. Fuenzalida, J.-H. Meyer-Sahling, K. Mikkelsen, and N. Titelman. 2020. *Encuesta Nacional de Funcionarios en Chile: Evidencia para un servicio público más motivado, satisfecho, comprometido y ético* [National Survey of Civil Servants in Chile: Evidence for a More Motivated, Satisfied, Engaged, and Ethical Public Service]. Santiago: Dirección Nacional del

- Servicio Civil. <https://www.serviciocivil.cl/wp-content/uploads/2020/01/Encuesta-Nacional-de-Funcionarios-Informe-General-FINAL-15ene2020-1.pdf>.
- Schuster, C., J. Meyer-Sahling, K. S. Mikkelsen, and C. González Parrao. 2017. *Prácticas de gestión de personas para un servicio público más motivado, comprometido y ético en Chile: Evidencia de una encuesta con 20.000 servidores públicos en Chile y otros países*. Santiago: Dirección Nacional del Servicio Civil. <https://documentos.serviciocivil.cl/actas/dnsc/documentService/downloadWs?uuid=60fcd3de-fa9e-4906-9396-c7637b4cd167%20>.
- Seddig, D., and H. Leitgöb. 2018. "Approximate Measurement Invariance and Longitudinal Confirmatory Factor Analysis: Concept and Application with Panel Data." *Survey Research Methods* 12 (1): 29–41.
- Shi, D., T. Lee, and A. Maydeu-Olivares. 2019. "Understanding the Model Size Effect on SEM Fit Indices." *Educational and Psychological Measurement* 79 (2): 310–34. <https://doi.org/10.1177/0013164418783530>.
- Shi, D., A. Maydeu-Olivares, and Y. Rosseel. 2020. "Assessing Fit in Ordinal Factor Analysis Models: SRMR vs. RMSEA." *Structural Equation Modeling: A Multidisciplinary Journal* 27 (1): 1–15.
- Steenkamp, J.-B., and H. Baumgartner. 1998. "Assessing Measurement Invariance in Cross-National Consumer Research." *Journal of Consumer Research* 25 (1): 78–90. <https://doi.org/10.1086/209528>.
- Steiger, J. H. 1998. "A Note on Multiple Sample Extensions of the RMSEA Fit Index." *Structural Equation Modelling* 5 (4): 411–19. <https://doi.org/10.1080/10705519809540115>.
- Steiger, J. H., and J. C. Lind. 1980. "Statistically Based Tests for the Number of Common Factors." Paper presented at the annual Spring Meeting of the Psychometric Society, Iowa City, IA.
- Struening, E. L., and J. Cohen. 1963. "Factorial Invariance and Other Psychometric Characteristics of Five Opinions about Mental Illness Factors." *Educational and Psychological Measurement* 23: 289–98. <https://doi.org/10.1177/001316446302300206>.
- Tummers, L., and E. Knies. 2016. "Measuring Public Leadership: Developing Scales for Four Key Public Leadership Roles." *Public Administration* 94 (2): 433–51. <https://doi.org/10.1111/padm.12224>.
- Vandenberg, R. J., and C. E. Lance. 2000. "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research." *Organizational Research Methods* 3 (1): 4–70. <https://doi.org/10.1177/109442810031002>.
- Van De Vijveri, F. J. R., F. Avvisati, E. Davidov, M. Eid, J.-P. Fox, N. Le Donne, K. Lekvi, B. Meuleman, M. Paccagnella, and R. Van De Schoot. 2019. "Invariance Analyses in Large-Scale Studies." OECD Education Working Paper 201, OECD, Paris. <https://doi.org/10.1787/254738dd-en>.
- Wang, G., I.-S. Oh, S. H. Courtright, and A. E. Colbert. 2011. "Transformational Leadership and Performance across Criteria and Levels: A Meta-Analytic Review of 25 Years of Research." *Group & Organization Management* 36 (2): 223–70. <https://doi.org/10.1177/1059601111401017>.



## CHAPTER 25

# Making the Most of Public Servant Survey Results

## Lessons from Six Governments

Christian Schuster, Annabelle Wittels, Nathan Borgelt, Horacio Coral, Matt Kerlogue, Conall Mac Michael, Alejandro Ramos, Nicole Steele, and David Widlake

### SUMMARY

Governments around the world increasingly implement governmentwide surveys of public servants. How can they make the most of them to improve civil service management? This chapter first develops a self-assessment tool for governments that lays out the range of potential uses and benefits of public servant survey findings, arguing that public servant survey results can improve civil service management by providing tailored survey results to four key types of users (the government as a whole, individual public sector organizations, units within organizations, and the public, including public sector unions); holding government organizations accountable for taking action in response to survey results; and complementing descriptive survey results with actionable recommendations and technical assistance for how to address the survey findings to each user type. To substantiate the tool, the chapter then assesses the extent to which six governments—Australia, Canada, Colombia, Ireland, the United Kingdom, and the United States—make use of public servant survey findings. It finds that five out of six governments provide tailored survey results at both the national and agency levels, yet no government fully exploits all the potential uses and benefits of public servant surveys. For instance, not all governments provide units inside government organizations with their survey results or complement survey results with accountability or recommendations for improvement. Many governments could thus, at a low cost, significantly enhance the benefits they derive from public servant surveys for improved civil service management.

---

Christian Schuster is at University College London. Annabelle Wittels is an independent researcher. Nathan Borgelt is in the Australian Public Service Commission. Horacio Coral is in the National Administrative Department of Statistics Colombia. Matt Kerlogue is in the UK Cabinet Office. Conall Mac Michael is in the Department of Public Expenditure, Government of Ireland. Alejandro Ramos is in the National Administrative Department of Statistics Colombia. Nicole Steele is in the Australian Public Service Commission. David Widlake is in the UK Cabinet Office.

## ANALYTICS IN PRACTICE

- Public servant data can provide important evidence for management improvements in government, but how impactful it is depends on what governments do with it. This chapter contains self-assessment tools for governments conducting surveys of public servants, with a number of relatively low-cost actions governments can take to support evidence-based reforms based on insights from public servant surveys.
- Reporting results has two core aims. The first aim is to make salient key takeaways about the strengths and weaknesses of particular organizations or units. Reporting should thus include coded management reports or appropriately coded front pages of dashboards, which provide an overview of strengths and areas for development. Second, reporting aims to enable users to explore the survey results in a bespoke manner (while ensuring the anonymity of responses). This can be done, for example, through dashboards that allow users to split questions by demographic groups—for instance, by gender or age.
- Reporting results is more impactful when it reaches the different groups that can take action based on them in a tailored manner. These groups include central government agencies (for example, the civil service agency), individual public sector organizations, individual units (or their managers) within public sector organizations, and the public, including public sector unions. Tailored results reports can enable better management responses. For instance, by providing individual public sector organizations and units with tailored survey results, public managers can more easily identify appropriate actions to tackle the specific problems of their organizations or units.
- Reporting results is also more impactful when it includes recommendations to users—such as the managers of units or organizations—on how best to address survey findings, as well as action plans for users to develop their own actions. At low cost, recommendations can be automated at the unit and organizational levels—for instance, by linking training offerings to specific survey results or providing management “checklists” to managers with certain survey results. Moreover, action plan templates can be provided to units and organizations, with suggested methodologies to develop actions based on survey results. Where more resources are available, automated recommendations and action plan templates can be complemented by tailored technical assistance—or human resource management (HRM) consultancy—provided either by a central human resource (HR) unit or an external provider to help managers turn survey findings into improvements.
- To foster the use of results, governments can introduce accountability mechanisms—for instance, through central oversight of actions taken in response to survey findings by government organizations and units, by making (anonymized) survey data available to the public and other users (such as unions) to construct “best place to work” indexes and enhance transparency around staff management in public sector institutions generally, or by introducing survey measures that capture employee perceptions of the extent to which government organizations take action in response to survey findings.

## INTRODUCTION

How can governments make the most of public servant survey results for management improvements? Understanding this challenge is important. Governments around the world increasingly implement governmentwide employee surveys (see chapter 18). Implementing surveys is often costly to governments, not least in terms of the opportunity cost of staff time to respond to the survey (chapter 20). This puts a premium on making the most of public servant survey results—in other words, maximizing the benefits governments derive from public servant survey results for civil service management improvements. Yet the results from

surveys of public servants do not themselves engender change. They require effective dissemination, as well as the capacity and motivation to improve civil service management based on them. This translation process is challenging. In the United States Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS), for instance, only a minority of public servants believe that survey results will be used to make their agency a better place to work (OPM 2021).

How, then, can governments tackle this translation challenge more effectively? This chapter complements the in-depth exploration of the FEVS in chapter 26 of *The Government Analytics Handbook* with a self-assessment framework for governments to use and a case comparison of six governments to identify the range of potential approaches governments can take to maximize management improvement benefits from public servant survey results.

The conceptual starting point for the self-assessment framework consists of a series of theories of change linking public servant survey results to civil service management. The framework posits that public servant survey results can improve civil service management by enhancing the *informational basis* for civil service management improvements, the *capacity* of managers to improve civil service management, and the *motivation* of managers to improve civil service management. Tailored survey results—in the form of dashboards and reports—can improve the informational basis for management improvements for the government as a whole, for individual organizations, and for units within organizations. Publishing survey findings can provide both internal central oversight stakeholders and external stakeholders—such as the public and unions—with information to hold public managers accountable for management improvements, thus motivating managers to act on findings. Finally, complementing descriptive survey results with actionable recommendations and technical assistance in addressing the survey findings can enhance the capacity and ability of managers to pursue management improvements.

The chapter then assesses empirically the extent to which six governments—Australia, Canada, Colombia, Ireland, the United Kingdom, and the United States—make use of this range of potential uses of public servant survey findings.<sup>1</sup> It finds that most governments provide tailored survey results at both the national and agency levels, yet no government fully exploits all the potential uses and benefits of public servant surveys. For instance, not all governments provide units inside organizations with their survey results or complement survey results with accountability or recommendations for improvement. Many governments could thus, at a very low cost, significantly enhance the benefits they derive from public servant surveys for civil service management improvements.

## INFORMATION, MOTIVATION, AND CAPACITY: HOW PUBLIC SERVANT SURVEY RESULTS CAN IMPROVE CIVIL SERVICE MANAGEMENT

The core purpose of implementing public servant surveys is to improve employee management to, ultimately, attain a stronger workforce. For instance, the United Kingdom Civil Service People Survey seeks to inspire action “to increase and maintain . . . levels of employee engagement, and staff wellbeing” (UK Government 2018). How can public servant survey results attain this aim? From a theory-of-change perspective, three mechanisms stand out.

Survey results can enhance the *informational basis* for management improvements, the *motivation* of managers to pursue management improvements, and the *capacity* of managers to pursue improvements.

These mechanisms provide a broad framework for centralized entities to assess their own efforts at inducing public sector action from surveys of public servants. In relation, chapter 26 in the *Handbook* highlights how a complementary architecture within each agency supports these actions. Thus, the two chapters can be seen together as a framework against which public sector analysts interested in generating action can benchmark the institutional environment in which their survey results are disseminated.



## Business Intelligence: Improving the Informational Basis for Management Improvements through Survey Results

Better business intelligence—a stronger informational basis for management decisions—is the first and most obvious use of public servant survey results. As the Australian Public Service Commission puts it, the “results also help target strategies to build Australian Public Service (APS) workplace capability now and in the future” (Australian Public Service Commission 2021b). Or, as the government of Canada lays out:

The objective of the Public Service Employee Survey is to provide information to support the continuous improvement of people management practices in the federal public service. The survey results will allow federal departments and agencies to identify their areas of strength and concern related to people management practices, benchmark and track progress over time, and inform the development and refinement of action plans. Better people management practices lead to better results for the public service, and in turn, better results for Canadians. (Government of Canada 2021)

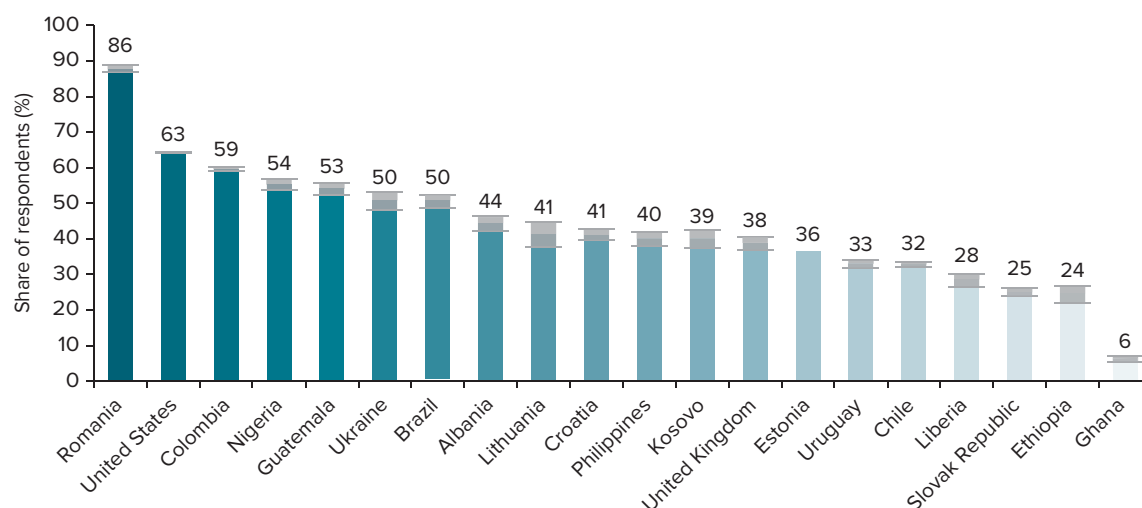
Public servant surveys can provide business intelligence on several aspects of the public administration production function (see chapter 2). They can help in understanding key public servant attitudes and how civil servants experience their work—for example, their job satisfaction or intent to stay in or leave their organization. And they can help in understanding management practices and the organizational environments shaping these public servant attitudes and experiences, such as the quality of leadership or performance management. Having data on both can also help in understanding the drivers of employee attitudes, such as engagement (namely, which management practices are statistically most important to improve engagement). In some countries where personnel databases of the civil service are highly decentralized (and centralized demographic data about the civil service are not available), surveys have also been used to create an overview of the demographic structure of the civil service (for example, India’s Civil Services Survey of 2010), by asking about gender, age, or education, for instance.

A number of users can benefit from this business intelligence. First, this business intelligence can enable governmentwide reforms. Governmentwide public servant survey results can spur improvements to specific management functions if particular government shortcomings are identified. For instance, upon finding in its National Survey of Public Servants that a third of public servants indicated that they entered public service through personal or political connections, the government of Chile drafted new legislation to strengthen the merit basis of public service (Briones and Weber 2020). Governmentwide survey results can also highlight the need to improve management of and for particular groups—for instance, to track diversity and inclusion progress, as in New Zealand’s government (Te Kawa Mataaho Public Service Commission 2021).

Understanding strengths and weaknesses governmentwide is often aided by international benchmarking, when survey measures across governments are comparable. For instance, if a government wants to understand whether it needs to act upon the low pay and benefits satisfaction of its staff, one potential point of reference is the pay and benefits satisfaction of public servants in other countries. The Global Survey of Public Servants (GSPS) enables such benchmarking, as illustrated below (figure 25.1). In Ghana, for instance, 6 percent of public servants are satisfied with their pay, compared to between 24 percent and 86 percent of public servants in other countries, suggesting that pay satisfaction might constitute a particular challenge in Ghana (rather than merely reflecting the general discontent of public servants with their salaries around the world).

For business intelligence from public servant surveys to be intelligible and actionable, it needs to be presented in a manner that increases awareness and understanding of key areas measured by the survey and, in particular, the key priority areas for action in light of the survey results. It also needs to allow governmentwide users to explore topics of interest, such as how survey responses differ by key groups of public servants—for instance, between men and women (cf. Pandey and Garnett 2007). Understanding key areas for action requires reporting results either in a management report or in appropriately coded dashboards, which front-page key areas of strength and development. Complementing management reports with

**FIGURE 25.1** Share of Public Servants Satisfied with Their Pay and/or Total Benefits, Various Countries



Source: Fukuyama et al. 2022.

Note: Years of measurement vary by country. Colors denote the extent of job satisfaction, with darker shades signifying greater job satisfaction. The gray vertical bars denote 95% confidence intervals.

dashboards allows users to easily explore aggregate data splits—for instance, by demographic group. User exploration is also aided by allowing ad hoc requests from central government agencies (such as ministries of finance) for particular tailored survey data analyses that go beyond what is displayed in a dashboard—for instance, particular regression analytics to understand the drivers of gender gaps in different organizations. Finally, central business intelligence is further strengthened when public servant survey results are integrated with other human resources (HR) data sources—for instance, in an HR dashboard that places survey results side-by-side with indicators such as retention, sick leave, number of applicants for public sector jobs, and gender pay gaps.

Second, public servant survey results can enable reforms at the organizational level by disaggregating results to organizational averages, benchmarking organizations in the public sector against each other, and allowing organizations to understand differences in the experiences and responses of different groups inside an organization. Providing organization-level business intelligence matters because differences in employees' experiences between public sector organizations inside a government are often larger than differences between governments (Meyer-Sahling, Schuster, and Mikkelsen 2018). Governmentwide reforms alone thus often miss priorities for improvement in particular public sector organizations. Drawing on its organization-level results, to cite just one example, the Primary Care Division of the Scottish government identified key areas for improvement (including empowerment of staff and team spirit) in its 2012 Civil Service People Survey—in which it scored 54 percent in engagement—and it acted upon the survey findings to increase engagement to 78 percent in 2014 (Cabinet Office 2015). Management reports for each organization, appropriately coded dashboards, which front-page key areas of strength and development for each organization, and dashboards to allow organizations to explore aggregated responses of different demographic groups inside the organization can provide the business intelligence for such organizational improvements.

Third, public servant survey results can enable improvements at the level of units or divisions inside organizations by disaggregating results to the unit level and making them accessible to unit managers through management reports and dashboards.<sup>2</sup> Unit-level reporting is important because differences in key indicators between units inside organizations—such as in the quality of leadership and employee engagement—are often as large as differences between organizations (see chapter 20). The UK Cabinet Office's Social Investment and Finance Team (SIFT), for instance, excelled relative to other teams inside the Cabinet

Office in employee engagement through “tight-loose” leadership—tightness around the mission but delegation in allowing members of the team autonomy to achieve the mission (Cabinet Office 2016).

### **Capacity: Enhancing the Ability of Managers to Undertake Management Improvements**

Descriptive survey results can identify key strengths and weaknesses in staff management in the government, a particular government organization, a unit inside an organization, or a particular demographic group of public servants. By themselves, however, survey results are not prescriptive: they do not identify how best to address survey findings. In other words, they identify strengths and weaknesses but not managerial actions for improvement. It is thus important to complement survey results with either a process to identify improvements or the identification of specific substantive improvements.

Approaches that focus on an improved process can take the form of methodologies to develop action plans, with templates and, potentially, technical assistance (for example, from a civil service agency or a management consultancy) to help government organizations or units undertake improvements. This approach is typical of employee engagement consultancies, which have developed standardized toolkits based on staff survey results (see, for example, Gallup 2022).

The substantive approach couples the presentation of survey results with specific recommendations for improvement based on the survey results to facilitate turning results into action. In country-level reports, these can be qualitative and detailed, based on inferring key management improvements from the data (see, for example, Schuster et al. 2020). At lower levels of disaggregation—for organizations and, in particular, units where hundreds of results reports are needed—recommendations can be automatically coded to be added to the results presentation. For instance, Google’s approach to people analytics flags specific training offerings to managers based on survey results for their units (Penny 2019).

### **Accountability: Motivating Managers to Undertake Management Improvements**

Public servant survey results can make transparent the quality of management in specific units or organizations or in the government as a whole. Where transparency is coupled with accountability for management improvements, it can provide additional motivation to managers to pursue improvements (beyond their intrinsic motivation).

Accountability can come, first of all, from the bottom up: public servant surveys provide employees with a voice to raise concerns about their experiences with and perceptions of management, their team, and their organizational environment. For employees—or public sector unions as their representatives—to hold government organizations accountable for management improvements, results need to be published, at least at an aggregate level. Providing employees with a voice is an explicit objective of most public servant surveys. For instance, the Australian government stresses that their survey “is an opportunity for employees to tell the Australian Public Service Commissioner and Agency heads what they think about working in the APS” (Australian Public Service Commission 2021b). Accountability to employees can be fostered by measuring employee perceptions of the extent to which their organization is taking action to respond to survey findings. For instance, the UK Civil Service People Survey asks respondents about their agreement with the statement “Where I work, I think effective action has been taken on the results of the last survey” (Cabinet Office 2019).

Accountability can also come from the outside—the media, public sector watchdogs, and researchers—when data, including organization-level data, are made public.<sup>3</sup> For instance, the Partnership for Public Service—a US nonprofit—generates the Best Places to Work in the Federal Government index based on published US public servant survey results, benchmarking public sector organizations in the United States and rendering salient organizations that perform poorly (Partnership for Public Service 2021). This type of transparency and publicity about poor performance may, in a poorly performing organization, motivate action to improve its ranking.

Similarly, the media can act as an external accountability mechanism to motivate improvements when data are made public. For instance, in Australia, low staff morale and dissatisfaction with leadership in the Department of Home Affairs made headlines in main news outlets (Doran 2019). Similarly, in Ireland, the media reported that only a small fraction of civil servants thought that poor staff performance was adequately addressed in their departments (Wall 2021).

Researchers can add a further layer of accountability, particularly when anonymized microdata from survey respondents are made available. This precludes the selective reporting of results by allowing researchers to analyze the anonymized raw data. It can thus further improve the aforementioned informational basis for management improvements by fostering a body of research work about a government's public service. To illustrate, a recent review identified 48 research articles using published microdata from the FEVS (Resh et al. 2019). Among these studies, a number have assessed diversity management in the US government based on these microdata. They have found, for instance, that employees in organizations with greater racial diversity tend, all else being equal, to report lower job satisfaction. Yet they have also found that when diversity is managed well, employees in organizations with more racial diversity report greater job satisfaction (Choi 2009; Choi and Rainey 2010). This makes transparent both a potential challenge in the US government (lower job satisfaction in more diverse institutions) and the effectiveness of diversity management as a solution.

Accountability and oversight can, of course, also be internal. For instance, heads of organizations can hold managers of units inside their organizations accountable for improvements based on their results, and central oversight agencies (such as ministries of finance or civil service agencies) can hold public sector organizations accountable for improvements. As detailed below, in the Irish government, a dashboard tracks the actions of each government organization in response to the public servant survey, while Canada uses a management accountability framework (MAF) to assess the progress made by organizations in management practices, including those identified in the employee survey.

In short, public servant survey results can foster management improvements through better business intelligence, greater managerial motivation, and an increased capacity to improve. Governments can maximize each of these uses by generating customized reports for the government as a whole, each organization, and each unit, ensuring that users can both explore aggregate data easily and access key findings for their organization/unit/government.

Governments can also complement descriptive survey results with recommendations, action plans, and methodologies to turn survey results into improvements and accountability mechanisms inside the government and externally—including publishing results and data—to motivate action. The next section will compare the extent to which six governments with long-standing public servant surveys have made use of these approaches to maximize the benefits of public servant survey results.

## TO WHAT EXTENT ARE GOVERNMENTS MAKING FULL USE OF PUBLIC SERVANT SURVEY RESULTS? BENCHMARKING SIX GOVERNMENTS

To what extent are governments making full use of public servant survey results? This section compares the approaches taken by six governments with long-standing (at least three iterations) governmentwide public servant surveys: Australia, Canada, Colombia, Ireland, the United Kingdom, and the United States. It does so by benchmarking the actions taken by each government against each of the potential uses of public servant survey results identified in the previous section of this chapter. Table 25.1 summarizes this comparison and the self-assessment framework, which can be used by other governments to identify actions that could further enhance their use of public servant survey results. Of course, there may be variations within each category across the six governments we have reviewed. For simplicity, we code each country in the framework for each category according to a binary: *exists* vs. *does not exist*.

Looking first at business intelligence, the comparison shows that governments generally produce country-level results reports. With one exception, they also produce agency-level reports (that is,

**TABLE 25.1 Comparing Country Approaches to Making the Most of Survey Results**

	Australia	Canada	Colombia <sup>a</sup>	Ireland	United Kingdom	United States
<b>Information provided to central government</b>						
National results report						
Dashboard for customized queries						
Ad hoc analyses on topics of interest to central government						
Survey results integrated in HR business intelligence platform or regular report with other HR data (for example, turnover or mobility)					Only ad hoc in select agencies	
<b>Information provided to government organizations</b>						
Results report for each agency						
Dashboard with results of agency and internal comparisons						
Rapid-response analyses on topics of interest in response to requests from particular agencies						
<b>Information provided to units inside government organizations</b>						
Results report for each unit within the agency						
Dashboard with results of units and customized queries						
<b>Capacity to take action based on survey results</b>						
National results report with recommendations for management improvement	In accompanying reports	In accompanying reports				
Organizational reports with recommendations for improvement						
Action plan templates and methodologies to help organizations take						
action based on survey findings						
Results presentations and technical assistance to help agencies take action based on survey results						
<b>Accountability: Information made available to the public</b>						
National results report or table						
Dashboard for customized queries						Previously the Unlock Talent dashboard

*(continues on next page)*

**TABLE 25.1 Comparing Country Approaches to Making the Most of Survey Results (continued)**

	Australia	Canada	Colombia <sup>a</sup>	Ireland	United Kingdom	United States
Institutional results reports or dashboards					In a spreadsheet	
Anonymized individual-level microdata		On request		On request		
<b>Bottom-up and top-down accountability for using survey results</b>						
Central government mechanism to hold organizations accountable for acting on results						
Survey measuring whether public servants perceive their organization is taking action to address results						

Source: Original table for this publication.

Note: In the table, green cells indicate Yes and red cells indicate No. To make the analysis tractable, the authors have delineated a binary conception of whether countries undertake the focal practices or not. Though there may be variation within each category and country, this provides a generalized assessment of the information available from public data and clarifications received from countries.

a. Colombia counts on a comprehensive management dashboard that covers human resource management, enables comparisons over time, and contains recommendations for each organization and action plans (DAFP 2022). However, this dashboard currently does not integrate results from Colombia's public servant survey. HR = human resources.

reports for individual government organizations), enabling each organization to understand its strengths and weaknesses based on survey results. There is a greater divergence when it comes to unit-level reports. Australia, Canada, the United Kingdom, and the United States disaggregate data to the unit level, enabling heads of units or divisions inside a government organization to understand their strengths and weaknesses. As this disaggregation to unit-level reports or dashboards multiplies the number of potential users of the data, it is an important low-cost avenue for greater management impact of the survey in countries that currently lack this disaggregation. Governments also differ in the extent to which they create dashboards that allow users to easily explore the results along the margins most interesting to them—for instance, by splitting indicators by demographic groups (such as gender) for the government as a whole or particular organizations. As the creation of such dashboards need not be costly—for instance, if free online platforms such as Tableau Public are used—this represents a second low-cost way for many governments to enhance the business intelligence users derive from survey results. All governments, with one exception, also undertake bespoke analyses of the data for users—for instance, in response to requests from the ministry of finance or other particular organizations with specific interests. Finally, Australia, Canada, and the United States integrate public servant survey results systematically with other HR data—such as data on turnover—in their reporting to generate a more comprehensive overview of HR strengths and weaknesses.

In terms of enhancing the capacity to turn survey results into actions at the national level, only Australia and Canada accompany their descriptive survey results with specific management improvement recommendations in accompanying briefings and reports (though not in the survey results directly). At the agency level, two countries rely on action plan templates to help organizations with a process to turn survey results into action. Finally, in four of the countries, the center of government provides results presentations or technical assistance to individual public sector organizations to help them turn survey results into action.

In terms of external accountability, all countries publish country-level results. All governments except for one also make institution-level reports public. However, only Australia and Canada provide the public



with access to a dashboard to explore the data, while three countries publish the anonymized microdata (and a further two countries make the data available upon request to researchers under certain conditions). Similarly, three governments have institutionalized center-of-government mechanisms for holding government organizations accountable for improvements based on survey results, and only a minority of governments measure the extent to which civil servants believe that their organizations are taking effective action based on survey results. In many countries, stronger external transparency and internal accountability mechanisms to motivate managers to take action based on survey results could thus be considered.

Table 25.1 highlights both commonalities and variations between countries in the extent to which survey results are used—and opportunities to further this use. To make these opportunities more actionable, the next subsections showcase specific examples of how governments approach each of these uses.

First, a brief note on the capacity to undertake these actions is due. While this chapter does not focus on *why* different governments do not adopt some of the potential uses of survey results, a plausible conjecture is the differential organizational setup of public servant surveys across countries. This differential organizational setup generates differences in, for instance, organizational capacity to deliver management reports, dashboards, and bespoke analyses. In the United Kingdom and Australia, data collection is contracted out, as is, for instance, the production of results dashboards. In Colombia, the national statistical agency handles the process, while Canada and Ireland use a hybrid approach whereby surveys are conducted through a partnership between civil service departments and the national statistics agency. In the United States, the survey is conducted by the OPM, which is the US federal civil service department. Where surveys are conducted in-house, the ability to deliver dashboards and coded reports is conditioned by the data analytics staff's capacity in the government agency in charge of the survey.

### Information Provided to the Central Government

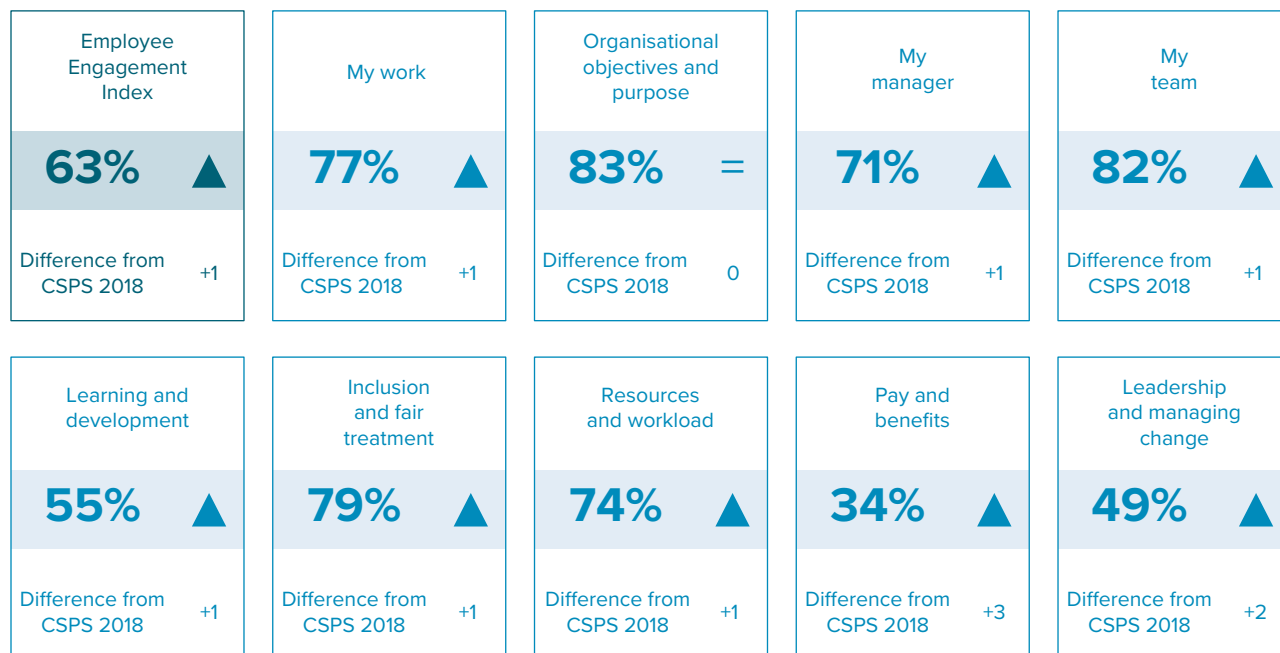
As noted above, all governments generate national results. They do so in different ways, however. In the United Kingdom, a slide deck is produced for the most senior officials (the cabinet secretary, the civil service's chief operating officer, and departmental permanent secretaries) and HR directors in departments. They are also given access to the interactive dashboards so they can explore the results in more detail. In some previous years, a slide deck visually highlighting key findings and showing the progression from the past year was also made public (figure 25.2), and the head of the civil service provided a write-up of highlights (for example, Heywood 2017). This is not currently the case, however. The Colombian government, similarly, presents national results in a slide deck together with a press release with key findings (DANE 2022).

By contrast, Ireland and the United States present national results *reports*. Both highlight up front the most positive and the most challenging results. The Irish report does this by theme (figure 25.3); the US report lists items with the highest and lowest agreement (as key areas of strength and development) (OPM 2021).

As a further means of highlighting key strengths and weaknesses, the Irish report also contains international comparators (figure 25.4)—a practice otherwise underutilized by governments, in light of the comparator data available through the GSPS (Fukuyama et al. 2022).

Finally, Australia presents results not only in a slide deck (Australian Public Service Commission 2021c) and a summary write-up of results (Australian Public Service Commission 2021a) but also in an annual State of the Service Report that integrates employee survey results with other workforce data—for instance, on gender pay gaps, diversity, and mobility—to provide a comprehensive HR diagnostic, often focused on key themes (Australian Public Service Commission 2021d). Figure 25.5 showcases an example figure from the State of the Service Report, which integrates findings from the country's public servant survey with external labor market data to better understand skills shortages in the public sector. Similarly, Canada and the United States integrate HR and survey data in their reporting. For instance, in the United States, employee survey results are, as part of the President's Management Agenda (PMA), provided to the White House together with HR metrics, such as staffing and quit rates. Survey and HR data were also integrated into a

**FIGURE 25.2 Results Report from the UK Civil Service People Survey**

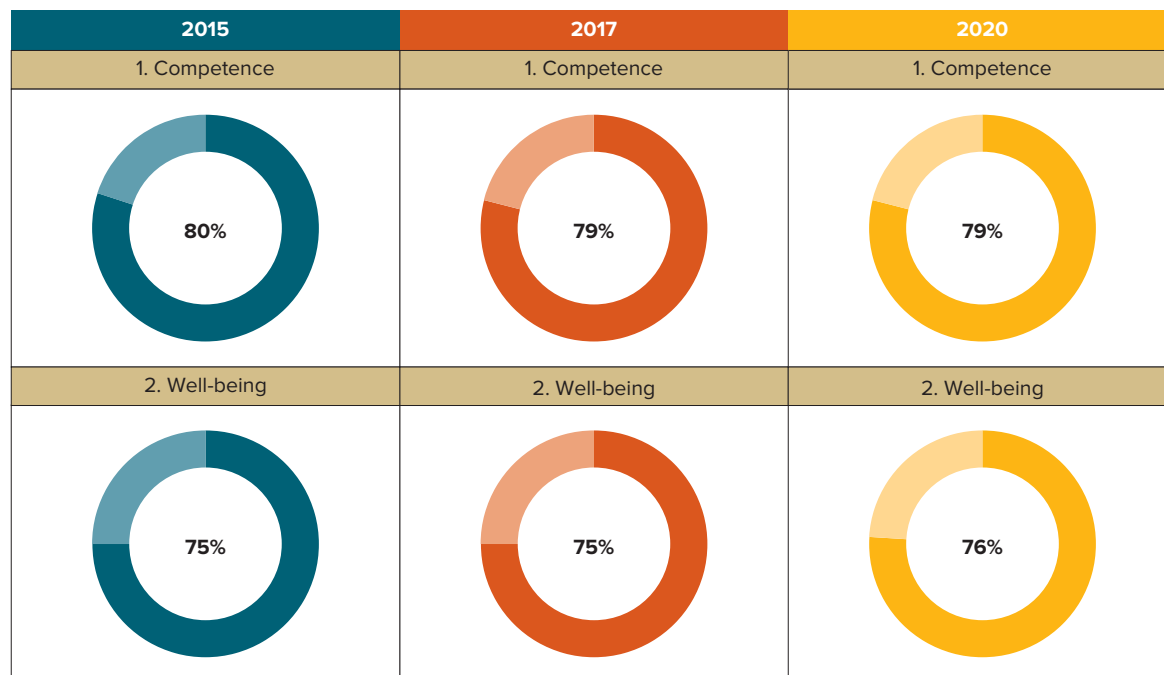


Source: Cabinet Office 2019.

Note: CSPA = Civil Service People Survey.

**FIGURE 25.3 Results Report from Ireland, Top 5 Positive Results**

**Positive Results – Top 5**



Source: Department of Public Expenditure and Reform 2020.

## FIGURE 25.4 International Benchmarking in Results Report from Ireland

### International Benchmark:

In the survey, 33% of staff agreed with the statement ‘I feel that my pay adequately reflects my performance’, which compares to 30% among respondents in the 2017 UK Civil Service People Survey.

Source: Department of Public Expenditure and Reform 2017.

dashboard—Unlock Talent—that allowed users to compare agencies and units in survey results (for example, engagement) and HR data. Funding for the dashboard has run out and, at the time of the writing of this chapter, the US government is developing a replacement.

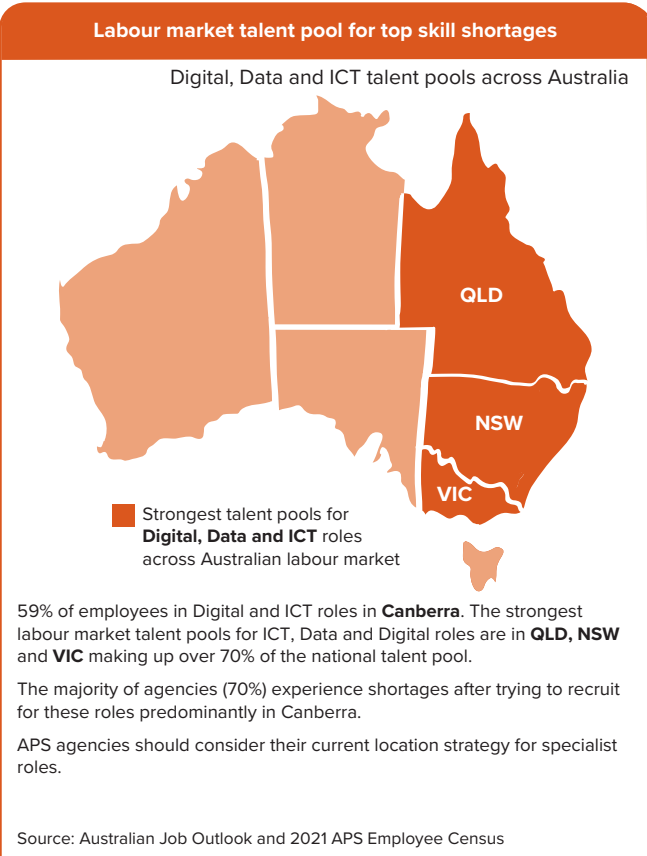
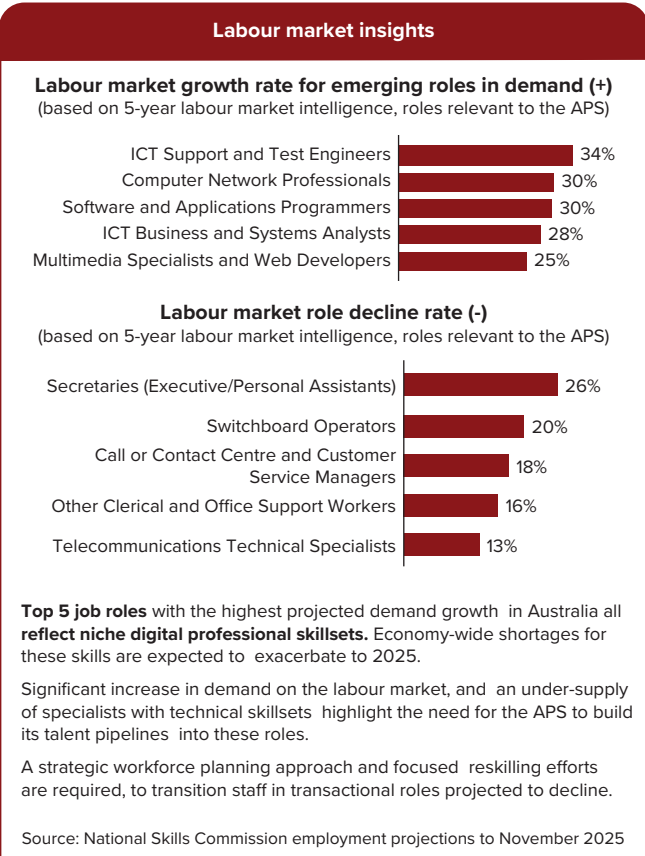
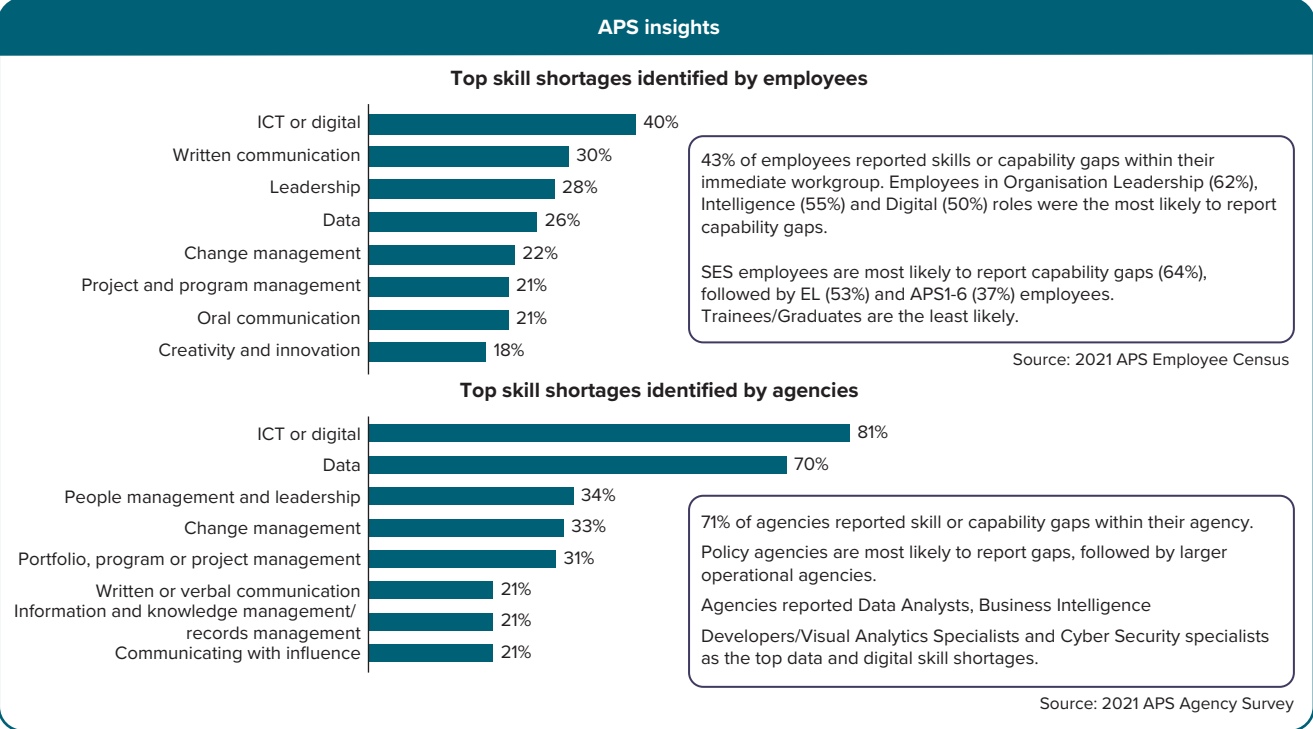
In short, all countries report national results. Four of the six countries do well to visualize highlights up front, giving stakeholders a sense of key strengths and areas for improvement. Ireland also uses international comparisons to further contextualize strengths and challenges, while Australia and Canada are the only countries to systematically integrate employee survey and other workforce data for a more comprehensive, regular HR diagnostic.

Governments also differ in the extent to which they enable government users to further explore data beyond the results report—by making a results dashboard available or conducting on-demand, bespoke analysis of the data. Australia, Canada, the United Kingdom, and the United States use dashboards to enable users to explore the (aggregate) data in a more customized way—for instance, by comparing responses of different demographic groups in different state institutions. These dashboards can be relatively low cost, as in the case of the Employee Viewpoint Survey Analysis and Results Tool (EVS ART) in the United States (see chapter 9, case study 9.3 in chapter 9, and chapter 26) or Canada’s Power BI dashboard (figure 25.6). Canada’s Power BI dashboard allows users to compare indicators, organizations, and trends over time. All data are aggregated as percentages for each response option (for example, the percentage of respondents who answered “strongly agreed” or “agreed”) (Government of Canada 2020b).

Canada also produces dashboards focused on specific groups of public servants—such as Indigenous people, women, persons with disabilities, or LGBTQ+ employees. Figure 25.7, for instance, shows the dashboard for persons with a disability. Canada thus provides users with accessible overviews of results for groups of public servants with particular needs or particularly concerning results.

A subset of governments also conducts more bespoke, on-demand analysis of data. For instance, the Australian Public Service Commission analyzes and reports on employee survey data in bespoke reports for specific purposes. These are typically reports for internal civil service use and consideration but may also comprise reports for public release. Areas from across the civil service that require employee survey results to inform their work and activities can request these from the commission. The commission then prepares responses to these requests for information. In Canada and the United States, analytical reports can be requested by participating agencies. The OPM also publishes a series of special reports—for instance, on women in public service, employee engagement drivers, and millennials in public service (OPM 2022). Ireland also occasionally commissions academics and consultants to provide more in-depth analytical reports to provide further insight into areas that were identified as needing intervention (Department of Public Expenditure, National Development Plan Delivery and Reform 2022).

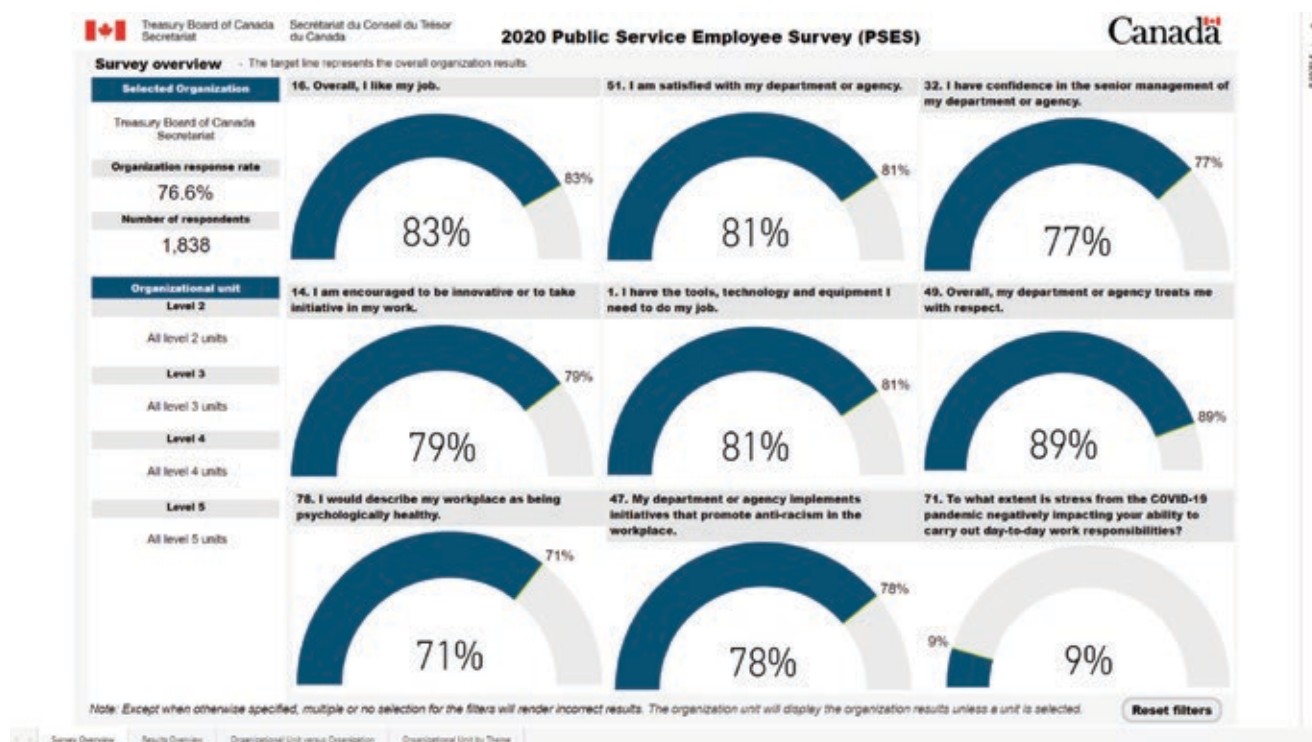
**FIGURE 25.5 State of the Service Report from Australia**



Source: Australian Public Service Commission 2021d.

Note: APS = Australian Public Service; EL = executive level; ICT = information and communication technology; SES = senior executive service.

**FIGURE 25.6** Canada Public Service Employee Survey Dashboard



Source: Government of Canada 2020b (example screenshot).

### Information Provided to Government Organizations

All governments provide results data at the organizational level to participating government organizations. The format and accessibility of these agency-level results, however, differ. In Colombia and the United States, data are presented in tables or data files (see chapter 9, case study 9.3 in chapter 9, and chapter 26 for greater detail on the EVS ART approach). Ireland produces bespoke reports for each agency, accompanied by an “at a glance” dashboard (see more on this below). These reports are descriptive. Agencies are encouraged to draw their own conclusions for programs of change based on the results.

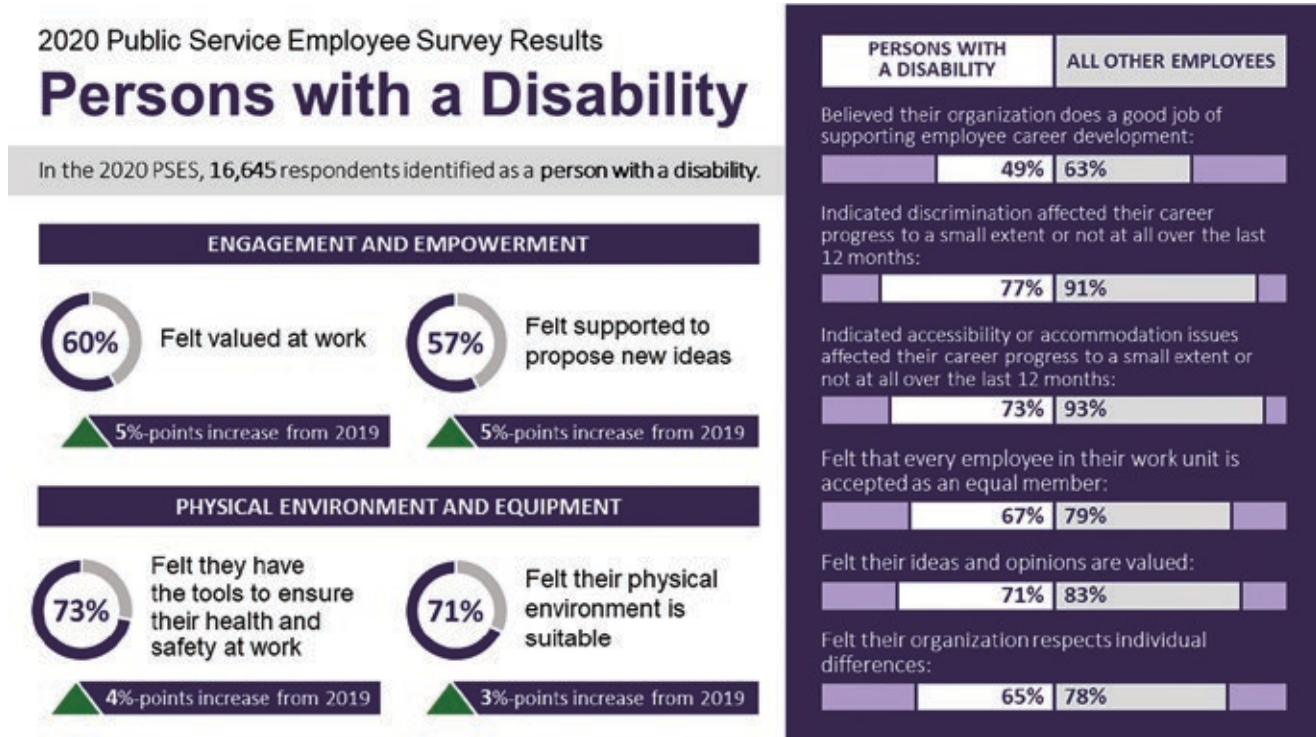
Australia, Canada, and the United Kingdom offer online dashboards to agencies through which they can filter results and explore the parts of the data relevant to them. Dashboards have privacy protection safeguards programmed into them, such as not allowing for cross-tabulations below a certain number of employees or only providing a subset of open-ended responses for teams that are very small. Figure 25.8 visualizes the UK Civil Service People Survey’s (internal) dashboard, which the United Kingdom contracts from Qualtrics. Australia, similarly, uses a contractor to generate an easily accessible online dashboard that allows splits at the agency and subdivision level by, for instance, gender and technical expertise for each agency and subdivision. Canada built its own dashboard with Power BI (figure 25.6).

### Information Provided to Units inside Government Organizations

Canada, Australia, the United Kingdom, and the United States also generate unit-level results—for instance, by generating team-level reports accessible to each team, as in the United Kingdom’s (figure 25.8) and

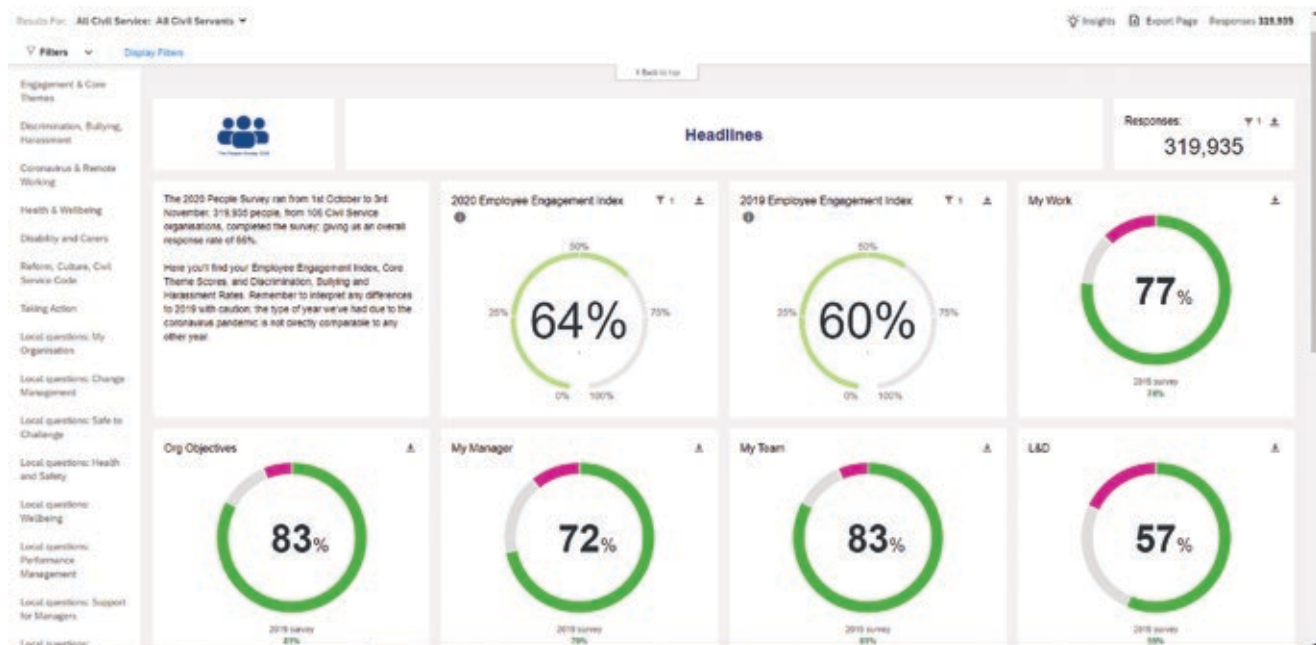


**FIGURE 25.7** Canada Public Service Employee Survey Dashboard for Persons with a Disability



Source: Government of Canada 2020b (example screenshot).  
Note: PSES = Public Service Employee Survey.

**FIGURE 25.8** United Kingdom Civil Service People Survey Results Dashboard for Organizations and Teams



Source: Screenshot of the headlines page of the internal dashboard used by the Civil Service People Survey Team.



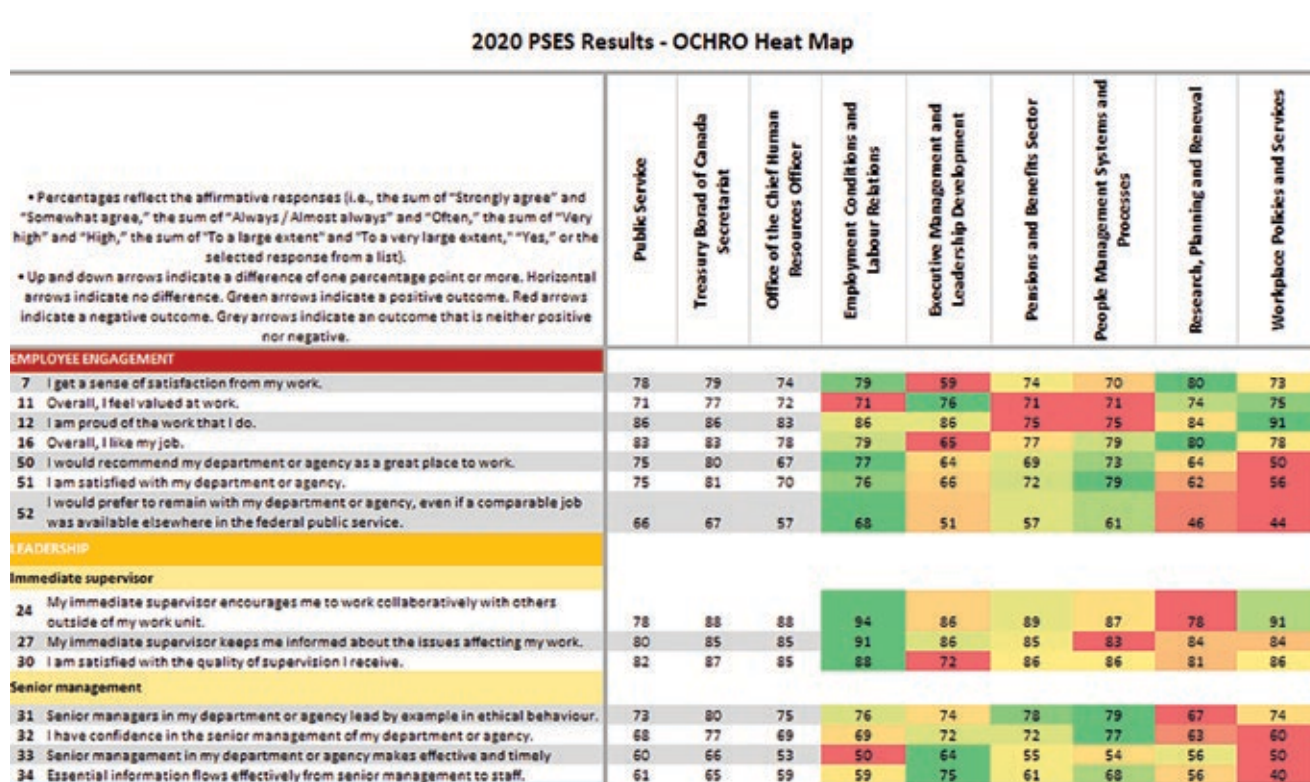
Canada (figure 25.6) dashboards. Canada also provides heat maps for each unit to crystallize strengths and areas for development (figure 25.9).

Generating unit-level results requires generating unit- or division-level identifiers in each organization, which are either linked to the unique survey ID of a survey respondent or selected by survey respondents. These can be collected from central human resources management information systems (HRMIS), where they exist, collected from institution-level HRMIS and appended to the email addresses used to disseminate the survey, or gathered from each government organization manually, with respondents then selecting the unit in which they work when completing the survey.

For instance, in Australia, several agencies choose to map their Australian Public Service Employee Census respondents to their organizational hierarchies. Where an organizational hierarchy has been included, analysis and reporting of results are possible for individual work units within an agency. This includes analysis and reporting for demographic and other groups within an agency or organizational unit. In 2020, just over 60 percent of agencies included an organizational hierarchy in the Australian Public Service Employee Census. How far down an agency chooses to disaggregate its hierarchy typically depends on its size and structure. Most, however, will disaggregate their hierarchies to the lowest practicable level while safeguarding anonymity (for instance, by not reporting results for work units with fewer than 10 respondents).

When agencies provide such disaggregation, reports for agencies and their organizational units are developed and released to those agencies. Representatives within individual agencies have access to the online dashboard, in which they can source their prepared summary reports but also analyze, filter, and compare results for their agency and its constituent organizational units. This portal allows for more interactive descriptive analysis and exploration of results and enables agencies to source more survey results than are made available in the static reports.

**FIGURE 25.9** Canada's Heat Maps for Survey Results of Units



Source: Screenshot of unit report heat map by the Treasury Board of Canada Secretariat.

Similarly, in the United States, disaggregation occurs up to the ninth level of hierarchy in some organizations, multiplying the number of units and teams benefiting from survey results. Lower levels of government are provided with “subagency breakout reports,” which display results for an individual office (the lowest level of the agency) for all core and demographic survey items, and “subagency comparison reports,” which compare all work units within a breakout for all core and demographic survey items.

## Capacity to Take Action on the Basis of Survey Results

As noted, turning survey results into action is facilitated by accompanying descriptive survey results with prescriptive recommendations at the national, organizational, or unit level, where appropriate (for instance, by linking training offers to managers to certain survey results in leadership quality); by presenting results in person to organizations to help them understand them and consider actions in response; and by offering action plan methodologies to agencies or units to take action based on results.

In national survey results reports, governments typically do not include prescriptive recommendations, though recommendations or actions are sometimes included in accompanying publications—for instance, in a blog by the chief executive of the UK’s civil service (Manzoni 2020), a press release by Colombia’s Public Service Department (DAFP 2016), or, perhaps most directly, in Australia’s State of the Service Report, which, as mentioned before, integrates public servant survey data with other HR data sources to analyze key HR themes and suggest ways forward (Australian Public Service Commission 2021a). In Canada, presentations and briefings by the Treasury Board of Canada Secretariat include recommendations.

Australia also explicitly offers organizations action plan templates and methodologies to help them take action based on survey findings. Each agency report includes an action template that encourages managers to map actions against survey outcomes (figure 25.10).<sup>4</sup> This is encouraged by tying the release of survey results to the Australian State of the Service Report, which sets out a strategic mission for the civil service. Senior executives from the national commission are asked to present key points of the report to employees in their state and territory. These presentations typically give a high-level overview of the perspectives and direction of the commission and also include Australian Public Service Employee Census results. Each year, focus groups are held with representatives of agencies, during which the use of the results is discussed. Canada, in turn, has an interdepartmental committee in which best practices are shared and organizations are provided guidance on how to create their plans; however, specific templates are not provided. In the United Kingdom, the Cabinet Office shares with departments a guide to running a workshop to discuss the results as a team and take action, while, in the United States, senior accountable officers have been appointed in past years within agencies, and experts in the OPM have worked closely with them to support the interpretation of employee survey results and develop and assess action plans.

Bespoke consultancy by a central agency to help individual organizations improve management based on survey results remains less systematized across governments. Results presentations at the organizational level occur but are not universal or part of a systematic intervention program by a central government agency to boost management practices and employee engagement based on survey results across line agencies. As mentioned before, follow-up consultancy is a cornerstone of the work of engagement consultancy firms—and thus a missed opportunity—but, of course, also resource intensive. At the same time, governments are not currently making use of lower-cost, automated recommendations based on survey results for organizations or units—for instance, by showing specific training offerings to managers with scores in need of improvement in certain areas. More could thus be done to help organizations and managers turn survey results into management improvements.

## External Accountability: Information Made Available to the Public

All countries make country-level reports or statistics publicly available. Australia and Canada provide dashboards to enable the public to explore data. Colombia and the United Kingdom provide statistical

FIGURE 25.10 Australian Public Service Commission Action Plan Template

CELEBRATE

What things do we do well?

THINK ABOUT HOW WE CAN BUILD ON OUR STRENGTHS AND LEARN FROM WHAT WE ARE GOOD AT.

INVESTIGATE FURTHER WITH OUR TEAMS

Are there any other opportunities coming out of the results that we want to explore further?

HOW COULD WE INVESTIGATE? THROUGH LOOKING AT THE DATA IN MORE DETAIL OR THROUGH DISCUSSIONS WITH STAFF?

OPPORTUNITIES

Areas we need to focus on and turn into action plans:

WHAT ARE THE KEY THINGS WE NEED TO IMPROVE TO MAKE WORKING HERE BETTER?

USE THIS PAGE TO START YOUR LOCAL ACTION PLANS

IDENTIFY AREAS TO CELEBRATE, OPPORTUNITIES FOR IMPROVEMENT AND AREAS WHICH YOU NEED TO INVESTIGATE FURTHER.

PRIORITISE 3 AREAS TO TAKE FORWARD

	PRIORITISE 3 AREASFOR ACTION	TIMESCALES	OWNER	RESOURCES REQUIRED	TARGET/SUCCESS MEASURE
1					
2					
3					

Source: Australian Public Service Commission 2021c.

summaries, which might not be easily accessible for audiences unfamiliar with statistics. The British dashboard is not available to the public. Data for Colombia and the United Kingdom can be accessed in an aggregated format by agency on a government website and downloaded as Excel files. Australia, Ireland, and the United States also publish written reports with overall findings. In Australia, Canada, the United Kingdom, and the United States, the availability of publicly available written reports of individual agencies depends on the participating agencies’ willingness to publish them. Ireland does not publish organization-level reports (Australian Public Service Commission 2021b; Cabinet Office 2021; Government of Canada 2020a; OPM 2020).

In terms of transparency to the public, Australia, Colombia, and the United States publish individual-level microdata to enable researchers and other interested users to explore the data. Canada and Ireland provide these data to researchers upon request (and with certain requirements).

Australia and the United Kingdom provide statistics aggregated at the response and agency levels that can be downloaded, and Ireland provides summary statistics in report form that can be publicly accessed.

Only in the United States is public information from the employee survey drawn on by external actors. In the United States this is the Partnership for Public Service, which compiles the Best Places to Work in the Federal Government rankings of public sector organizations as a means to generate further external accountability and motivation for improvement in survey scores for public sector organizations (figure 25.11).

566

THE GOVERNMENT ANALYTICS HANDBOOK

**FIGURE 25.11 Best Places to Work in the US Federal Government, 2022 Rankings**

Large Agencies		Midsize Agencies	Small Agencies	Agency Subcomponents	
Rank *	Agency			2022	2021
1	National Aeronautics and Space Administration			84.3	85.1
2	Department of Health and Human Services			74.3	74.4
3	Intelligence Community			71.9	73.4
4	Department of Commerce			70.6	73.7
5	Department of Veterans Affairs			68.4	70.2
6	Department of Transportation			68.3	68.0
QUARTILE KEY					
Lower Quartile (0-25%)		Below Median (25-50%)		Above Median (50-75%)	
				Upper Quartile (75-100%)	

Source: Partnership for Public Service 2023 (screenshot, <https://bestplacetowork.org/rankings/?view=overall&size=large&category=leadership&>).

In short, there remain significant opportunities for greater transparency and external accountability for public servant survey results, particularly at the organizational level, in many governments—for instance, by replicating “best place to work” rankings and presenting survey results at the national and organizational levels to stakeholders in a more accessible way.

## Internal Accountability for Using Survey Results

Internal accountability can be top-down (through central oversight) or bottom-up (by employees). Among the countries studied, Ireland has the most-established formal top-down accountability mechanism: it obliges all government departments to map actions taken in response to survey outcomes. After each survey, departments are asked to produce an action plan detailing how they will respond to challenging results within their organizations. The report is organized by thematic area, requiring organizations to state the issue, state the statistic underlying the problem identified, list agreed-upon actions, and list the processes put in place to address them.

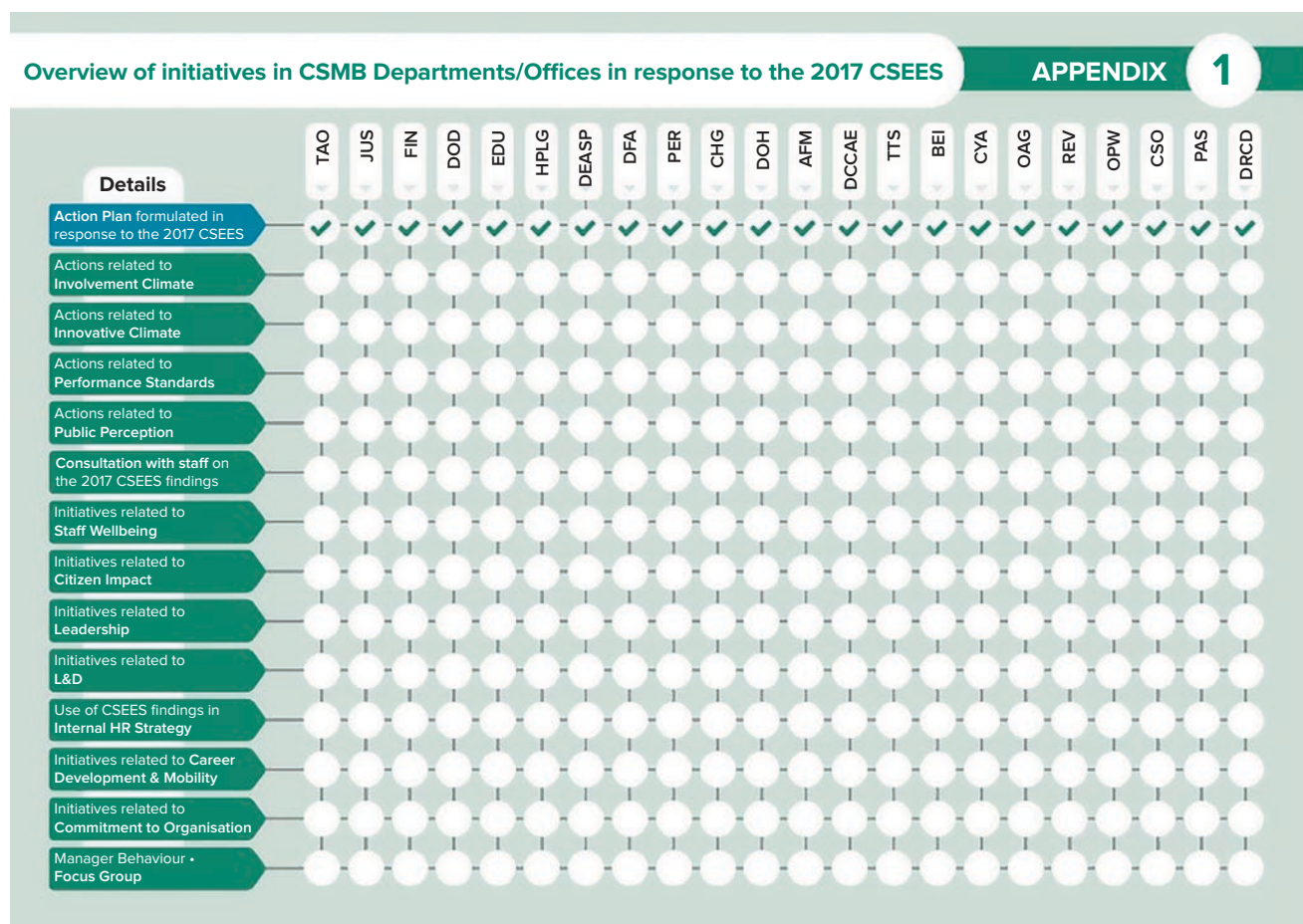
A quarterly update is prepared by the Civil Service Renewal Programme Management Office and then relayed to the Civil Service Management Board. An “at a glance” dashboard allows each head of office or secretary general to chart the progress of his or her organization. An interdepartmental working group provides officials with a forum to share experiences and best practices regarding survey management, driving strong response rates, and responding to their organizational results. In nonsurvey years, the group meets on a quarterly basis to share feedback on responding to departmental results. In survey years, the group meets on a more frequent basis to ensure milestones and targets are met in the run-up to the launch of the survey. Figure 25.12 visualizes the “at a glance” dashboard, which tracks actions taken by departments in response to survey results.

Canada, in turn, leverages the MAF to assess the progress made by an organization in its management practices (in seven areas identified by the survey). The MAF involves three key stakeholders (deputy heads of organizations, the HR community, and the Treasury Board of Canada Secretariat) and enables the Treasury Board to “monitor trends and identify gaps in policy compliance across departments,” among other things, such as including accountability for improvement in poorly performing indicators.

In the United States, survey results are included in the PMA, and agencies are held accountable for action toward organizational change, including employee engagement and related issues, such as diversity, equity, inclusion, and accessibility (see, for example, Donovan et al. 2014). Other governments lack a similarly institutionalized reporting and accountability mechanism for actions taken.<sup>5</sup>



**FIGURE 25.12** “At a Glance” Dashboard, Government of Ireland



Source: Screenshot of Civil Service Management Board “at a glance” dashboard, Government of Ireland.  
 Note: CSMB = Civil Service Management Board; L & D = learning and development.

In terms of bottom-up accountability, in Ireland, Canada, and the United States, questionnaires are typically shared with key employee representative groups and unions before the launch of a survey. For instance, in Canada, extensive consultative engagements with key stakeholders—such as participating departments and agencies and unions—inform questionnaire development, and stakeholders are kept apprised of progress before the survey launch. The United Kingdom and the United States also measure in their surveys whether public servants perceive that their organizations are taking action to address survey results, thus making transparent whether organizations are—in the perception of their staff—acting on the results (and facilitating accountability where staff members do not perceive that their organization is taking effective action). For instance, the US survey inquires whether respondents “believe the results of this survey will be used to make [their] agency a better place to work” (OPM 2021). In short, there remains leeway to strengthen both bottom-up and top-down accountability mechanisms across countries.

## DISCUSSION AND CONCLUSION

This chapter has developed a self-assessment framework to enable governments to identify which additional uses of public servant survey results they could contemplate to maximize the impact on civil service management. It then benchmarked six governments against the self-assessment framework to showcase

the use of the framework, provide further qualitative detail on each of the potential uses of public servant surveys, and provide a state of play for how governments are currently using (or not using) results from public servant surveys.

Our case selection focused on countries with regular governmentwide employee surveys—which, as of now, tend to be Organisation for Economic Co-operation and Development (OECD) member governments. Our findings about the prevalence of different practices should be interpreted accordingly. Non-OECD governments implement governmentwide employee surveys less frequently, though many of the practices we identify in the chapter would certainly be attainable and low in cost for them as well (for example, publishing anonymized microdata from survey results in an Excel file).

The case comparison has shown that all countries we surveyed provided country-level results—including for public consumption—and, for the most part, results to participating agencies.

Reports that provide information on a subagency (that is, a unit or division) level are less common, as are dashboards that allow government organizations and units to explore and filter the data in the way most relevant to them.

Most reports also remain descriptive. Strategic advice and consulting services are typically not included as part of the mission of survey administration teams, nor are automated recommendations tying survey results to specific management actions. However, as some countries (Australia and Ireland) have acknowledged, the demand for bespoke results and advice has increased, and some countries have at least provided action plan templates for organizations to take action.

Countries also differ in the extent of their external and internal accountability mechanisms. Publication of organization-level results is voluntary and selective in most countries, and some do not publish them at all. Three countries (Australia, Colombia, and the United States) publish anonymized individual-level microdata (with Ireland and Canada making the data available upon request). Internal oversight and accountability for taking actions based on results are only formally institutionalized in a dashboard system in Canada's MAE, Ireland, and the United States' PMA, while the United Kingdom and the United States track the extent to which employees believe effective survey action has been taken.

In conjunction, our results suggest that many governments could, at very low cost, significantly enhance the benefits they derive from public servant surveys for civil service management improvements, including by

- Ensuring that results are disaggregated and disseminated to suborganizational hierarchical levels (for example, divisions and units);
- Creating simple dashboards to allow users at different levels of government—and the public, for national and organization-level results—to explore and filter the data according to their needs;
- Coding management reports (or dashboard front pages) such that the key strengths and areas for development of a particular organization or unit are easily identifiable;
- Including action plan methodologies and automated recommendations to users—such as the managers of units or organizations—about how to best address survey findings (automated recommendations can, for instance, contain training offerings tied to specific survey results or management “checklists” for managers with certain survey results);
- Strengthening accountability for results (for instance, through central oversight of actions taken in response to survey findings by government organizations and units, by enabling third parties—or the government itself—to construct “best place to work” league tables of government organizations, and by capturing employee perceptions of the extent to which government organizations take action in response to survey findings);
- Publishing anonymized microdata to encourage research and insight creation by third parties; and
- Standardizing questions to increase comparability with other countries or industry surveys to create better benchmarks of national scores (for example, through the GSPS).



Where further resources are available, governments may also

- Complement agency-level reports with bespoke presentations and consultancy services to agencies to help them improve in response to survey findings,
- Provide insight reports centered around key strategic topics to move the dial on key HR topics with survey results, and
- Integrate staff surveys with other workforce data to generate more holistic HR dashboards and reports on the public service as a whole, as well as particular strategic themes.

## NOTES

1. By *surveys of public servants*, we refer to surveys of employees of government organizations. The coverage of these surveys extends, variously across countries, to the civil service, the public service as a whole—including organizations outside the civil service—or a combination of the two.
2. As with the publication of (anonymized) survey microdata, care needs to be taken to protect the anonymity of survey respondents when disaggregating data to units—for instance, by not reporting unit- or group-level averages with fewer than 10 respondents (cf. OPM 2021).
3. Providing transparency to citizens about the operations of government—including by publishing public servant survey results—is, of course, also an important part of democratic accountability more broadly.
4. The template can be accessed at <https://www.apsc.gov.au/initiatives-and-programs/workforce-information/aps-employee-census-2020>.
5. In Australia, each organization also has a “champion” who fosters survey participation and the use of results from the survey.

## REFERENCES

- Australian Public Service Commission. 2021a. “The 2021 APS Employee Census Overall Results.” Australian Public Service Commission, Australian Government, November 30, 2021. <https://www.apsc.gov.au/initiatives-and-programs/workforce-information/aps-employee-census-2021/2021-aps-employee-census-overall-results>.
- Australian Public Service Commission. 2021b. “APS Employee Census 2020.” Australian Public Service Commission, Australian Government. <https://www.apsc.gov.au/initiatives-and-programs/workforce-information/aps-employee-census-2020>.
- Australian Public Service Commission. 2021c. *Highlights Report: APS Overall*. Canberra: Australian Public Service Commission, Australian Government. <https://www.apsc.gov.au/sites/default/files/2021-12/APS00878%20-%20APS%20Overall.pdf>.
- Australian Public Service Commission. 2021d. *State of the Service Report 2020–21: Reform in the Shadow of COVID*. Canberra: Australian Public Service Commission, Australian Government. <https://www.apsc.gov.au/sites/default/files/2021-11/APSC-State-of-the-Service-Report-202021.pdf>.
- Briones, Ignacio, and Alejandro Weber. 2020. “Un mejor empleo público para un mejor Estado.” *El Mercurio*, February 19, 2020. <https://t.co/u4x6WTzi79>.
- Cabinet Office. 2015. “Case Study: Employee Engagement and Wellbeing: Scottish Government, Primary Care Division.” Cabinet Office and Civil Service, United Kingdom Government, July 2, 2015. <https://www.gov.uk/government/case-studies/employee-engagement-and-wellbeing-scottish-government-primary-care-division>.
- Cabinet Office. 2016. “Case Study: Employee Engagement and Wellbeing: Cabinet Office’s Social Investment and Finance Team.” Cabinet Office and Civil Service, United Kingdom Government, February 18, 2016. <https://www.gov.uk/government/case-studies/employee-engagement-and-wellbeing-cabinet-offices-social-investment-and-finance-team>.
- Cabinet Office. 2019. *Civil Service People Survey: Civil Service Benchmark Scores 2009 to 2019*. London: Cabinet Office, United Kingdom Government. [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/876879/Civil\\_Service\\_People\\_Survey\\_2009\\_to\\_2019\\_Median\\_Benchmark\\_Scores\\_-\\_final.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/876879/Civil_Service_People_Survey_2009_to_2019_Median_Benchmark_Scores_-_final.pdf).

- Cabinet Office. 2021. “Civil Service People Survey: 2020 Results.” Cabinet Office, United Kingdom Government, May 7, 2021. <https://www.gov.uk/government/publications/civil-service-people-survey-2020-results>.
- Choi, Sungjoo. 2009. “Diversity in the U.S. Federal Government: Diversity Management and Employee Turnover in Federal Agencies.” *Journal of Public Administration Research and Theory* 19 (3): 603–30.
- Choi, Sungjoo, and Hal G. Rainey. 2010. “Managing Diversity in U.S. Federal Agencies: Effects of Diversity and Diversity Management on Employee Perceptions of Organizational Performance.” *Public Administration Review* 70 (1): 109–21.
- DAFP (Departamento Administrativo de la Función Pública). 2016. “Sala de prensa: Noticias.” [Press release, December 12, 2016]. Departamento Administrativo de la Función Pública, Government of Colombia. [https://www.funcionpublica.gov.co/noticias/-/asset\\_publisher/mQXU1au9B4LL/content/el-97-5-de-los-servidores-consideran-que-su-trabajo-contribuye-al-logro-de-los-objetivos-de-su-entidad-encuesta-ed](https://www.funcionpublica.gov.co/noticias/-/asset_publisher/mQXU1au9B4LL/content/el-97-5-de-los-servidores-consideran-que-su-trabajo-contribuye-al-logro-de-los-objetivos-de-su-entidad-encuesta-ed).
- DAFP (Departamento Administrativo de la Función Pública). 2022. “Modelo Integrado de Planeación y Gestión.” Departamento Administrativo de la Función Pública, Government of Colombia. <https://www.funcionpublica.gov.co/web/mipg>.
- DANE (Departamento Administrativo Nacional de Estadística). 2022. “Comunicado de prensa: Encuesta sobre ambiente y desempeño institucional Nacional y departamental (EDI-EDID) 2021” [Press Release, June 7, 2022]. Departamento Administrativo Nacional de Estadística, Government of Colombia. [https://www.dane.gov.co/files/EDI\\_nal/2021/Comunicado\\_prensa\\_EDIEDID\\_2021.pdf](https://www.dane.gov.co/files/EDI_nal/2021/Comunicado_prensa_EDIEDID_2021.pdf).
- Department of Public Expenditure and Reform (Department of Public Expenditure, National Development Plan Delivery and Reform). 2017. *Civil Service Employee Engagement Survey*. Dublin: Department of Public Expenditure and Reform, Government of Ireland. <https://www.gov.ie/en/collection/5e7009-civil-service-employee-engagement-survey/#2017>.
- Department of Public Expenditure and Reform (Department of Public Expenditure, National Development Plan Delivery and Reform). 2020. *Civil Service Employee Engagement Survey*. Dublin: Department of Public Expenditure and Reform, Government of Ireland. <https://www.gov.ie/en/collection/5e7009-civil-service-employee-engagement-survey/#2020>.
- Department of Public Expenditure, National Development Plan Delivery and Reform. 2022. “What Is the Civil Service Employee Engagement Survey?” *Civil Service Employee Engagement Surveys*. Department of Public Expenditure, National Development Plan Delivery and Reform, Government of Ireland. <https://www.gov.ie/en/collection/5e7009-civil-service-employee-engagement-survey/#what-is-the-civil-service-employee-engagement-survey>.
- Donovan, Shaun, Beth Cobert, Katherine Archuleta, and Meg McLaughlin. 2014. “Strengthening Employee Engagement and Organizational Performance.” Memorandum for Heads of Executive Departments and Agencies, December 23, 2014. [https://www.whitehouse.gov/wp-content/uploads/legacy\\_drupal\\_files/omb/memoranda/2015/m-15-04.pdf](https://www.whitehouse.gov/wp-content/uploads/legacy_drupal_files/omb/memoranda/2015/m-15-04.pdf).
- Doran, Matthew. 2019. “Bleak Outlook for Home Affairs Morale, as Staff Report Dissatisfaction with Work and Leadership.” *ABC News*, August 29, 2019. <https://www.abc.net.au/news/2019-08-29/bleak-outlook-for-home-affairs-staff-morale/11461442>.
- Fukuyama, Francis, Daniel Rogger, Zahid Husnain, Katherine Bersch, Dinsha Mistree, Christian Schuster, Kim Sass Mikkelsen, Kerenssa Kay, and Jan-Hinrik Meyer-Sahling. 2022. *Global Survey of Public Servants Indicators*. <https://www.globalsurveyofpublicservants.org/>.
- Gallup. 2022. “What Is Employee Engagement and How Do You Improve It?” Workplace, Gallup. <https://www.gallup.com/workplace/285674/improve-employee-engagement-workplace.aspx>.
- Government of Canada. 2020a. “2020 Public Service Employee Survey Results.” Government of Canada. <https://www.tbs-sct.gc.ca/pses-saff/2020/results-resultats/en/bq-pq/index>.
- Government of Canada. 2020b. “Public Service Employee Data Analytics.” Government of Canada. <https://hrdatahub-centredonneesrh.tbs-sct.gc.ca/PSes/Home/Index?GoCTemplateCulture=en-CA>.
- Government of Canada. 2021. “About the 2020 Public Service Employee Survey.” Government of Canada, January 22, 2021. <https://www.canada.ca/en/treasury-board-secretariat/services/innovation/public-service-employee-survey/2020/about-2020-public-service-employee-survey.html>.
- Heywood, Jeremy. 2017. “Civil Service People Survey 2017—The Results.” *Civil Service* (blog). November 16, 2017. United Kingdom Government. <https://civilservice.blog.gov.uk/2017/11/16/civil-service-people-survey-2017-the-results/>.
- Manzoni, John. 2020. “Civil Service People Survey 2019—The Results.” *Civil Service* (blog). March 26, 2020. United Kingdom Government. <https://civilservice.blog.gov.uk/2020/03/26/civil-service-people-survey-2019-the-results/>.
- Meyer-Sahling, Jan-Hinrik, Christian Schuster, and Kim Sass Mikkelsen. 2018. *Civil Service Management in Developing Countries: What Works?* London: UK Department for International Development. <https://christianschuster.net/Meyer%20Sahling%20Schuster%20Mikkelsen%20-%20What%20Works%20in%20Civil%20Service%20Management.pdf>.
- OPM (Office of Personnel Management). 2020. “Data Reports.” OPM Federal Employee Viewpoint Survey, US Office of Personnel Management, United States Government. <https://www.opm.gov/fevs/reports/data-reports>.
- OPM (Office of Personnel Management). 2021. *Federal Employee Viewpoint Survey Results: Governmentwide Management Report*. Washington, DC: US Office of Personnel Management, United States Government. <https://www.opm.gov/fevs/reports/governmentwide-reports/governmentwide-management-report/governmentwide-report/2021/2021-governmentwide-management-report.pdf>.

- OPM (Office of Personnel Management). 2022. "Special Reports." OPM Federal Employee Viewpoint Survey, US Office of Personnel Management, United States Government. <https://www.opm.gov/fevs/reports/special-reports/>.
- Pandey, Sanjay, and James Garnett. 2007. "Exploring Public Sector Communication Performance: Testing a Model and Drawing Implications." *Public Administration Review* 66: 37–51. <https://doi.org/10.1111/j.1540-6210.2006.00554.x>.
- Partnership for Public Service. 2023. *2022 Best Places to Work in the Federal Government Rankings*. Washington, DC: Partnership for Public Service. <https://bestplacetowork.org/rankings/?view=overall&size=large&category=leadership&>.
- Penny, Charlotte. 2019. "Case Study: How Google Uses People Analytics." *Sage*, December 1, 2019. <https://www.sage.com/en-au/blog/case-study-how-google-uses-people-analytics/>.
- Resh, William, Tima Moldogaziev, Sergio Fernandez, and Colin Angus Leslie. 2019. "Reversing the Lens: Assessing the Use of Federal Employee Viewpoint Survey in Public Administration Research." *Review of Public Personnel Administration* 41 (1): 132–62. <https://doi.org/10.1177/0734371X19865012>.
- Schuster, Christian, Javier Fuenzalida, Jan Meyer-Sahling, Kim Sass Mikkelsen, and Noam Titelman. 2020. "Encuesta Nacional de Funcionarios en Chile." Chile Civil Service. <https://www.serviciocivil.cl/wp-content/uploads/2020/01/Encuesta-Nacional-de-Funcionarios-Informe-General-FINAL-15ene2020-1.pdf>.
- Te Kawa Mataaho Public Service Commission. 2021. "Workforce Data—Diversity and Inclusion." Te Kawa Mataaho Public Service Commission, New Zealand Government (accessed December 7, 2021). <https://www.publicservice.govt.nz/research-and-data/workforce-data-diversity-and-inclusion/>.
- UK Government. 2018. "Civil Service People Survey Hub." Civil Service, United Kingdom Government, October 15, 2018. <https://www.gov.uk/government/collections/civil-service-people-survey-hub>.
- Wall, Martin. 2021. "Most Civil Servants Happy with Conditions but Not Promotional Access." *Irish Times*, May 14, 2021. <https://www.irishtimes.com/news/ireland/irish-news/most-civil-servants-happy-with-conditions-but-not-promotional-access-1.4565595>.

## CHAPTER 26

# Using Survey Findings for Public Action

## The Experience of the US Federal Government

*Camille Hoover, Robin Klevins, Rosemary Miller, Maria Raviele, Daniel Rogger, Robert Seidner, and Kimberly Wells*

### SUMMARY

Generating coherent public employee survey data is only the first step in using staff surveys to stimulate public service reform. The experiences of agencies of the United States federal government in using the Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS) provide lessons in the translation of survey results to improvements in specific public agencies and public administration in general. Architecture at the agency level that supports this translation process is critical and typically includes a technical expert capable of interpreting survey data, a strong relationship between this expert and a senior manager, and the development of a culture or reputation for survey-informed agency change and development initiatives. This chapter outlines the way that the FEVS, its enabling institutional environment, and corresponding cultural practices have been developed to act as the basis for public sector action.

### ANALYTICS IN PRACTICE

- Generating coherent survey data that describe the state of the public administration is a vital foundation for inspiring effective reform of the public service. But it is only the first step. Complementary efforts to stimulate the use of that survey data are vital for achieving corresponding change.

---

Camille Hoover is an executive officer and Robin Klevins is a senior management analyst at the National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health. Rosemary Miller is a psychologist and Maria Raviele is a program analyst at the US Office of Personnel Management. Daniel Rogger is a senior economist in the World Bank's Development Impact Evaluation (DIME) Department. Robert Seidner was formerly a performance manager at the US Office of Management and Budget. Kimberly Wells is a managing research psychologist at the US Office of Strategic Workforce Planning.

- Survey questions should aim at action from the beginning by asking about topics that staff and senior leaders find most challenging to the achievement of their mission. Designing questions with the chain of policy influence and action in mind prevents the survey process from being weakened at inception by a poor focus on what is important to public sector stakeholders.
- Public action from surveys of public employees requires at least one technical expert capable of analyzing and interpreting survey results and translating them into a clear action plan for the improvement of a specific unit or organization. This may simply entail using the survey as a launchpad to learn more about the issue from people in the organization. At the scale of many public sector organizations, this survey analyst should be embedded within the individual organization. This will provide them with sufficient time and focus to promote, digest, and translate survey findings as a core component of their work program. Given that many of the improvements in public sector organizations have capable personnel driving them, providing an official with the time required to anchor relevant discussions with colleagues is a necessary component of reform.
- Rich survey data and technical expertise to digest their implications are insufficient for public action. Any survey analyst or team must have a strong relationship with a senior manager who sees the value of the survey data for agency reform. Such a manager acts as a bridge between the technical translation of the survey into a form usable by an organization and the strategic processes required to build momentum for change. Rich survey data can generate political will by identifying or making salient significant inequities, opportunities for improved performance, or problematic parts of the agency. However, the case study outlined in this chapter, concerning the United States federal government, implies that a senior manager must champion change for substantial public action to occur. While the skills of the technical expert are important, the accountability and responsibility for developing a sustainable action plan rest on supervisors and leaders.
- For reform to be sustained, the technical staff and the leaders who are the “change champions” must inculcate and manage a culture of using survey data for public service reform. The easiest way to do this is to rapidly respond to issues identified by surveys, with leaders transparently sharing survey results with the workforce and emphasizing the results they deem most important. Leaders should then show staff how they are further exploring the results and creating initiatives that speak directly to the findings. Visible leadership responses to survey results will generate broader buy-in from agency staff, which will strengthen the credibility of the survey process and catalyze the impact of managerial responses. Changes in public administration typically require a coproduction approach, with both managers and staff moving toward improvements. For example, if staff feel that their capacity to perform is not being sufficiently developed, management must make opportunities for capacity development available and feasible to take up, while staff must take those opportunities and put in the effort required for learning.
- A centralized, governmentwide office in charge of survey design and implementation is useful for several reasons. First, there are important methodological decisions that affect all survey users equally but are costly to negotiate. A centralized team can ensure that surveys are effectively implemented and respond to changing service requirements, relieving frequently overburdened agency analytics teams. Second, ensuring a common platform for comparison catalyzes the usefulness of an agency survey by allowing for cross-agency benchmarking. Interagency comparisons rely upon a set of common measures, with data collected using consistent methodologies and under the same conditions and timeframe. Third, such an office can make choices that serve the public service as a whole, independent of any individual agency manager. For example, publishing data on all units, rather than selectively sharing results, ensures a more accurate representation of reality.
- When this central office does not have the capacity to address the demands of all managers in the public service, the case of the US federal government indicates that complementary efforts from individual officers strengthen the possibility that surveys incite public action. For example, the National Institutes of Health (NIH) Employee Viewpoint Survey Analysis and Results Tool (EVS ART) has facilitated granular

analytics of the Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS), allowing managers across the public service to better understand the implications of the survey for their work.

## INTRODUCTION

Generating coherent survey data that describe the state of public administration is a vital foundation for inspiring effective reform of the public service. One of the best-known influential surveys of public officials is the Federal Employee Viewpoint Survey (FEVS), which is administered by the United States federal government's Office of Personnel Management (OPM).<sup>1</sup> The survey was first fielded in 2002 and has been repeated annually since 2011, generating a panel of agency- and unit-level variables including measures of employee engagement, satisfaction, welfare, cooperation, development, leadership, and performance management. Technically an organizational climate survey, it functions to “assess how employees jointly experience the policies, practices, and procedures characteristic of their agency and its leadership” (OPM 2022). FEVS data are made available to managers of units, and an aggregated and anonymized version of the data is made public.<sup>2</sup>

The continuous, comparable, and public nature of the FEVS data has been a boon to the United States government analysts and researchers alike. As Janelle Callahan (2015, 399) states, “Ten years ago, few in government were talking about what federal employees thought or how the survey information could be used to improve employee satisfaction and commitment, and the performance of federal agencies.” Various qualities of the FEVS have made it influential in debates within the federal government, Congress, and society more broadly. The Government Accountability Office (GAO) frequently uses FEVS data in its assessments of federal agencies.<sup>3</sup> Similarly, the Partnership for Public Service, a nonprofit organization focused on strengthening the US civil service, uses the FEVS to publish its Best Places to Work in the Federal Government index.<sup>4</sup> This index frequently stimulates substantial debate on the public service labor market and its relationship to analogous private sector jobs (see, for example, Brust 2021; Mullins 2021).<sup>5</sup>

The core purpose of the FEVS is to provide public sector managers with direct but anonymized feedback from employees on the “state” of their work units. This may be at the agency level or in teams as small as 10 people (it is FEVS policy not to release data on any subagency work units with fewer than 10 respondents in order to protect the identity of individuals). It provides managers with a snapshot of current strengths, opportunities for improvement, and challenges for the organizations they manage, as well as how these have changed since the last survey. As Thevee Gray of the US Department of Agriculture (USDA) stated in an interview for this chapter, “OPM FEVS has been a great tool to ensure everyone has the same collective information on what is happening in our agency—a great starting point.” The ability for staff to provide management with anonymous feedback about their current experience concerning their work, work environment, management, and leadership ensures a minimum floor of feedback across the federal government. While agencies have their own surveying efforts, the FEVS provides a consistent platform for comparison across time and agencies. In a setting like the public service, where benchmarks of the work environment in other offices provide a crucial complement to more objective but coarse measures, this is a powerful feature of the survey.

The FEVS team provides agency managers with summary results of FEVS data for their units, with some breakdown by demographics, and FEVS individual-level data are released publicly (though fully anonymized).<sup>6</sup> It does not typically provide custom analysis to individual managers. This is a product of the mismatch in the scale of the federal government and the size of the FEVS team—there is simply not enough capacity to provide full-service analytics on demand. This leaves most managers with rich but unstructured survey data to explore. Initiatives such as the National Institutes of Health (NIH) Employee Viewpoint Survey Analysis and Results Tool (EVS ART) have sprung up to support managers in analyzing



the data of particular interest to them. However, generating work-unit-focused or topic-specific knowledge from the FEVS requires an effort to engage with the data themselves, which may seem costly or to be a low priority.

Such data releases provide a platform for public service reform at all levels of government, helping managers to better understand the reality of their management approaches and helping agency heads—who often have relatively short tenures heading large and disparate agencies—to identify priorities for the organization as a whole. Agency responses to the FEVS interact with external stakeholders in three ways. First, agencies can quickly identify their relative performance in personnel management, communicate with and learn from more successful agencies, and feel implicit pressure from their public standing. Second, the OPM, Congress, and the White House can do the same. The GAO explicitly uses the FEVS to make recommendations to Congress about how agencies should be reformed. In both cases, agency managers may feel there are career consequences related to improving their standing in the FEVS.<sup>2</sup> Third, bodies outside the government can monitor the workings of the public service and make recommendations about how it should reform or provide inputs into the change and development process for individual agencies.

Simply producing rich survey data has rarely been sufficient to generate public sector action. Complementary efforts to stimulate the use of these data are vital for achieving corresponding change. This chapter argues that external factors and pressures play a secondary role compared to the internal architecture of an agency's response to the FEVS. The experience of the FEVS in its two-decade-long history is that, though external and internal pressures are highly complementary, three pillars of response are critical for the FEVS to induce public action. First, public action requires a technical expert who is capable of analyzing and interpreting the FEVS data and who has sufficient time and focus to understand the implications of the FEVS for an agency and its work units. The approach of these individuals to promoting, digesting, and translating the FEVS for their agencies has varied, but in all cases, these individuals have been committed to the FEVS as a key tool of management and agency betterment. They can be seen as the spark of public action at the agency level.

Second, the survey analyst must have a strong relationship with a senior leader of the agency who sees the value of and endorses the use of FEVS feedback to inform agency-specific development at all levels of the organization. This is often a frontline senior leader or an executive within an organization below the agency level. Broad change can certainly be initiated at higher levels, but real change must happen on the front lines to create sustained culture change. This “change champion” acts as a bridge between the technical translation of the FEVS into a form usable by an agency and the strategic processes required to build momentum for change. The relationship between the survey analyst and the change champion can be seen as the positive friction that turns the spark into a flame for effective organizational development, change, and, ultimately, public sector reform.

However, without the broader buy-in of agency supervisors and staff working within a culture of responsiveness, such efforts are likely to be in vain. This buy-in begins at the initiation of the survey. If few staff respond to the FEVS, the data will not be seen as representative of broader staff concerns. Similarly, if staff do not believe that management will use their feedback to create change, they will not take the survey seriously. Thus, the credibility of FEVS data as a management tool requires a belief that they will indeed be used as a management tool. Once the data are published, agency change and development initiatives stemming from the FEVS currently require a coproduction approach, with both managers and staff moving toward improvements. For example, if staff feel that their capacity to perform is not being sufficiently developed, management must make opportunities for capacity development available and feasible to take up, and staff must take those opportunities and put in the effort required for learning. A culture of survey-informed action at the agency level is the tinder and kindling of public action.

Where these pillars of action have been in place, the FEVS has become a central pillar of personnel management in the US federal government. Callahan (2015, 399) provides the following example:

The Department of Commerce had [FEVS] subcomponents with the highest employee satisfaction in government and the lowest in 2013, prompting leaders to ask what was going

on and to take action. The U.S. Patent and Trademark Office (USPTO) was the number one agency of 300 subcomponents regarding employee satisfaction and commitment, while the Economic Development Administration (EDA), also in Commerce, ranked last. EDA officials said they began consulting with the USPTO and other organizations to gather best practices and work on improving employee satisfaction. In 2014, EDA was the most improved subcomponent, raising its satisfaction score by 11.8 points.

This chapter aims to describe the enabling environments within the US federal government that have been most prevalent in the translation of FEVS results into changes to the way public administration functions. It begins with a discussion of the key uses of public employee surveys through the lens of the use of the FEVS and then presents an overview of experiences using FEVS data and results for public action that stresses the three features of agency environments outlined above that have led to policy changes and improvements in government administration. The arguments presented here are based on the experiences of the authors—many of whom have played a key role in the development of the FEVS or its translation and use at the agency level over the past decade—and interviews with key stakeholders from across the US federal government.

## THE USES OF SURVEYS OF PUBLIC OFFICIALS

Most surveys of public sector employees intend to improve the quality of the environment in which they work and the processes that they undertake. In turn, work environment or process improvements are intended to improve the actions of the public sector toward the better delivery of public services. While some surveys target aspects of public administration that have direct impacts on service delivery, their intention is frequently the improvement of the administrative environment itself.

As such, survey content typically focuses on aspects of the administration that are widely regarded as meaningful for the quality of the work environment or administrative processes. The features of the work environment a survey assesses will directly determine the potential uses of its results. To have the best chance of informing or inducing reform of the public service, surveys should be designed with a theory of policy influence in mind.<sup>8</sup>

One use of survey results is as a centralized monitoring tool. A centralized personnel management agency may want to track the motivation of employees across the public service to ensure they are being effectively managed by senior leadership. The FEVS was initially implemented after an act of Congress required each agency to survey its employees annually.<sup>9</sup> The act required the collection of perceptions of leadership practices contributing to agency performance, employee satisfaction with policies and practices, work environment, rewards and recognition, professional development and growth opportunities, and organizational mission supports. The required content is included in the FEVS, so agency participation in the survey satisfies their statutory requirements. Incentivizing agency participation in a governmentwide survey also provides leadership with data for shaping policies intended to support federal employees.

The content of such centralized, standardized monitoring surveys will necessarily focus on aspects of administration that are said to be of importance to the quality of the work environment generally. Agencies and units can then be assessed against each other for comparison across the public service. Centralized stakeholders, including oversight entities, such as Congress, can use relative performance on survey measures to identify the worst-performing agencies in a particular area or to identify areas of strength and needs for improvement for individual—or even all—agencies.

For example, after a series of reports and internal surveys identified systemic problems in several national parks in 2016, congressional hearings were held on misconduct and mismanagement in the public service.<sup>10</sup> The agency responded with a range of reforms, including complementing the FEVS with a series of new pulse surveys.<sup>11</sup> The Department of the Interior now uses agency-specific items on the FEVS to monitor

agency reforms related to anti-harassment training and employees' knowledge of their rights and resources related to harassment.

Second, surveys of public officials can be used as a tool for agency personnel management. Without having to rely on centralized intervention or coordination, agency or unit managers can undertake their own assessments of their agencies' work environments. If a survey intends to improve agency management, it will naturally focus on elements of the work environment most relevant to its mission. Some of these elements will overlap with the wider service, but others will deviate. Here lies a key tension of centralized surveys of public employees—between the need for comparability and central control over the focus of the questionnaire, on the one hand, and the contemporary requirements of specific agencies, on the other.<sup>12</sup>

Comparability allows managers to use common benchmarks to better understand where they are performing well or poorly. But if comparability is focused on measures that are not relevant to their current concerns, the value of centralized surveys falls. Within the framework of a standardized survey, the FEVS has looked to counter this by providing agency managers with tailored insights, as resource constraints allow. In 2012, as the utility of providing agencies and units with survey results directly became clear, a series of initiatives were undertaken by the OPM to provide work-unit-level data. The OPM intended to empower agency heads and managers to capitalize on it as a tool for the agency. As the OPM (2012) stated,

Working with the information from the survey, ... an agency can make a thorough assessment of its own progress in its strategic goals and develop a plan of action for further improvement. The OPM FEVS findings allow agencies to assess progress by comparing earlier results with [contemporary] results, to compare agency results with the Governmentwide results, to identify current strengths and challenges, and to focus on short-term and longer-term action targets that will help agencies reach their strategic human capital management goals.

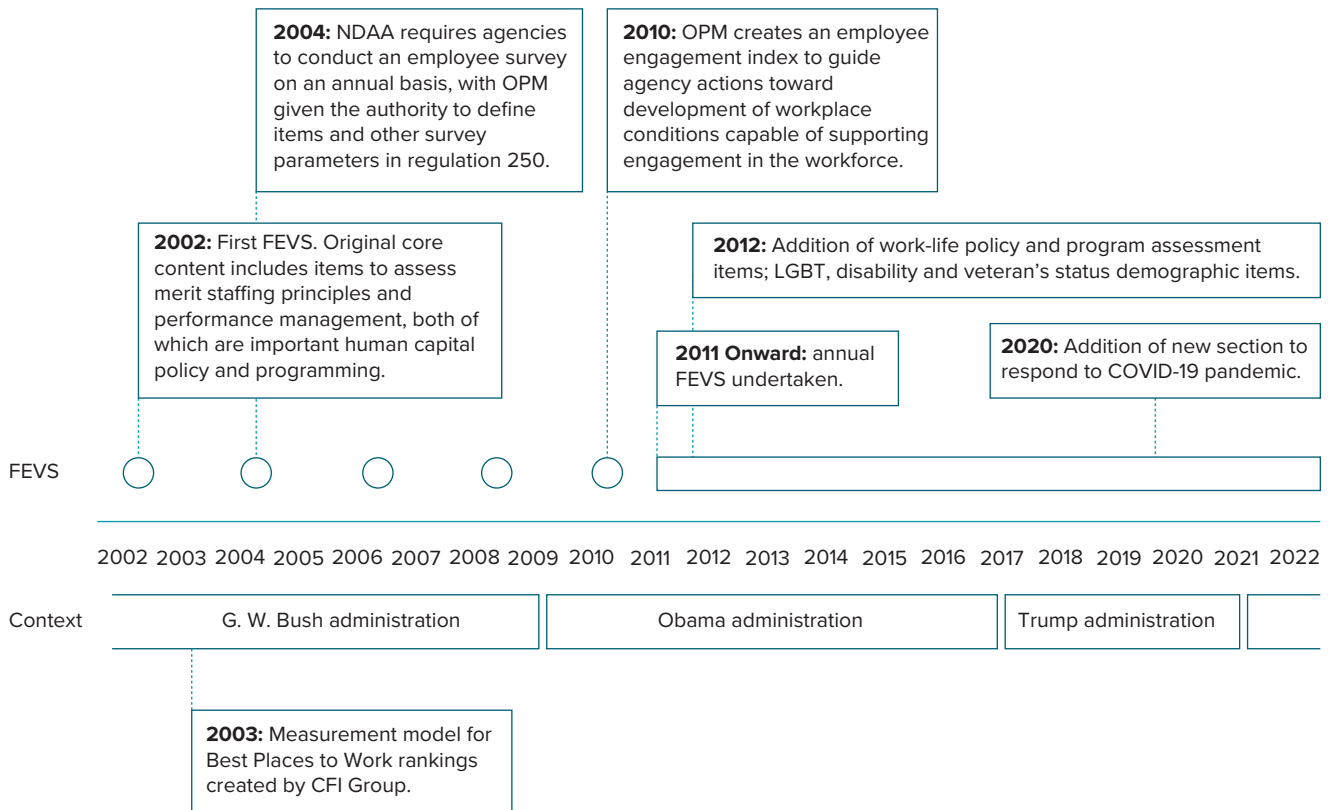
Third, data from public employee surveys can be used as a tool for ensuring the accountability of the government to citizens. In this case, citizens may be less interested in how satisfied or motivated public employees are but more interested in whether they are undertaking their jobs effectively and ensuring the judicious and efficient use of public funds. This implies a third realm of focus for public action to which survey questions may be targeted. Since 2012, the OPM has released anonymized data at the individual level. This has allowed analysts, researchers, the media, and the public to explore the world of the federal government in an unprecedented way.

Figure 26.1 summarizes the use of the FEVS across these three realms over time. The FEVS has at once been an oversight tool for Congress, a key resource for the GAO's large-scale evaluations of government, a means by which the Office of Management and Budget (OMB) can support broad agency functioning, and a rich resource for agencies to use as a core management tool. Each of these drivers of public action has matured and evolved toward an increasingly valuable architecture for the FEVS to impact government functions. These uses of the FEVS have co-evolved, and agency-level responses to the FEVS have been a critical complement to governmentwide policy and program assessments.<sup>13</sup>

The unifying theme of interest across varying federal government stakeholders is organizational effectiveness and performance. In particular, officials must make informed decisions or recommendations, interact with other members of the public service, and effectively deliver their mission to other members of government or the public. Succeeding at these tasks requires sufficient performance. The FEVS is designed as an organizational climate survey—a type of employee survey typically utilized to support organizational change and development initiatives.<sup>14</sup> Climate surveys collect employees' perceptions of management policies and practices, perceptions shown over decades of research to relate to performance. Moreover, employee input on policy enactment provides valuable data for assessing the function of those policies and ensuing practices, serving to guide effective change.

The FEVS contains several variables shown by research to relate to performance. Following an extensive body of research demonstrating the importance of employee engagement to performance, in 2010, the FEVS team introduced an employee engagement index (EEI) (see figure 26.1) that brought together those survey

**FIGURE 26.1** Timeline of the Evolution of the Federal Employee Viewpoint Survey



Source: Original figure for this publication.

Note: CFI = Claes Fornell International; FEVS = Federal Employee Viewpoint Survey; LGBT = lesbian, gay, bisexual, and transgender; NDAA = National Defense Authorization Act; OPM = Office of Personnel Management.

questions related to different aspects of employee engagement. The EEI was subsequently featured in the President's Management Agenda (PMA) and, accordingly, became a focus for agency change and development initiatives and a central part of the FEVS team's reporting and dissemination efforts.

Importantly, the FEVS EEI measures conditions that can lead to the state of engagement. The 15 EEI questions do not directly measure employees' feelings of engagement but rather assess conditions conducive to engagement (for example, effective leadership, meaningful work, and learning and growth opportunities), in keeping with the frame appropriate to an organizational climate survey. Understanding the engagement potential of federal workplaces along the factors of the measure enables one to identify leverage points for developing and sustaining work conditions capable of supporting employee engagement and, consequently, performance. These work conditions can be targeted for reform and provide policy-relevant variables for data collection. With a common measure, offices undertaking centralized monitoring can search for service-wide engagement trends and for work units that are falling behind others in terms of engagement. Agency managers can work to resolve issues with work conditions flagged by surveys, and stakeholders outside of government can monitor the health of their public service through the engagement of public employees.

The legislative foundation of the FEVS questions has limited change in the survey's content over time, although as figure 26.1 points out, changes have been made. Recently, the regulation governing content has been revised, with the number of required questions being reduced from 45 to 16. With this change, an FEVS modernization initiative has resulted in the addition of new content meant to respond to federal government priorities (for example, Executive Order 14035: Diversity, Equity, Inclusion, and Accessibility in the Federal Workplace) and advances in contemporary management theory and research (for example, innovation and organizational resilience).

A major goal for the entire FEVS program is to respond to evolving conditions and priorities. When the public service as a whole faced a significant new challenge, the FEVS responded rapidly. An entirely new and substantial section was added to the survey—for the first time since the development of the FEVS nearly two decades ago—due to the COVID-19 pandemic. Given the nature of the FEVS, OPM leadership felt such an addition would be particularly appropriate to understanding the implications of changes made to governmentwide and agency management practices and policies addressing pandemic challenges. The addition of items to assess responses to the pandemic has given the survey another layer of utility, with results critical to determine responses to future emergencies and inform ongoing discussions about the future of work.

## THE IMPORTANCE OF THE INSTITUTIONAL ENVIRONMENT

FEVS data have provided a window into the public administration of the US federal government. In contrast to many stereotypes of a monolithic bureaucracy, the experience of working in government is hugely diverse. Figure 26.2 presents the EEI across agencies (the solid squares) for 2018. As can be seen, there are substantial differences in how employees across the government perceive the engagement potential of their agencies. This point is amplified by looking at work units within agencies. Stacked vertically around each agency mean are the scores of the departments/work units (level 1) within that agency. In many agencies, we see that EEI scores can range as widely as they do across the public service as a whole.

Variation across and within agencies is a core reason why the institutional environment of an agency or department is so critical in generating public sector reform. The problems facing individual organizations will differ, requiring an organization- or work-unit-specific response that can only be generated if that organization has the right architecture in place to identify problems and build momentum for solutions. The fact that such variation is found in teams with similar budgets, jobs, senior leadership, and history indicates that differences in employee engagement are likely to have unique causes that require specific attention within the organization.

**FIGURE 26.2** Variation across and within Agencies in the Employee Engagement Index



Source: Original figure for this publication based on the FEVS 2019 public data.

Note: Solid squares represent agencies, and dots are scores for department or work units within that agency. FEVS = Federal Employee Viewpoint Survey.

Variation is also at the core of why survey data are so powerful. Rather than making policies based on the general experience of government (perhaps best represented in figure 26.2 by the governmentwide mean), policies can be targeted at those agencies and departments that are most in need. And lessons can be learned from those that are most successful. Thus, the FEVS aims to improve the quality of the management and work environment across agencies and departments by collecting individual employees' perceptions and experiences of their workplaces.

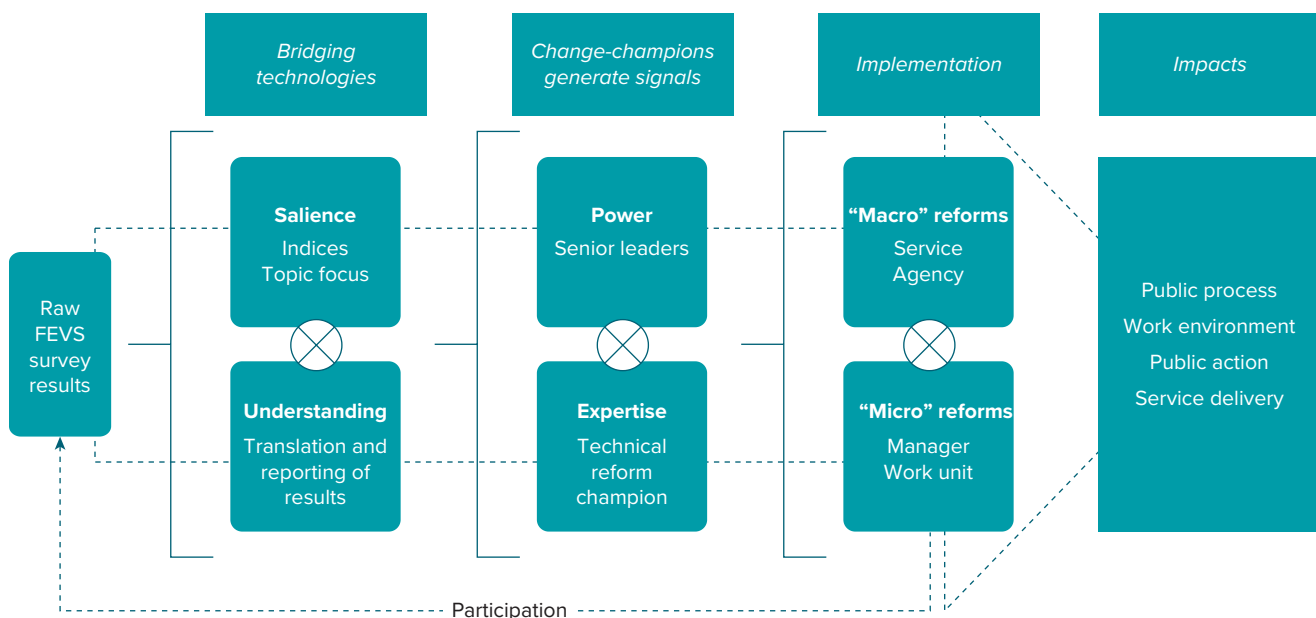
After 20 years of evolution, the FEVS is now built on an increasingly rich institutional scaffold that encompasses statutory requirements for surveying, reporting tools, and centralized initiatives that focus on the weakest performers, as defined by FEVS-based indexes. In many agencies, there are complementary scaffolds that support agency responses to FEVS results, either in reaction to centralized monitoring or as part of an agency-based reform initiative.

The evolutionary process that has occurred in the US federal government has guided agencies toward a structure with a series of key features. Figure 26.3 articulates these features as follows. The first column shows how raw FEVS results require a bridge into the agency where they are translated and their ramifications understood. Given the number of work units in most agencies and the number of questions in the FEVS, the potential complexity of reporting is substantial. Some topics must be made salient, requiring the survey analyst within the agency to appreciate where there is scope for reform and how the survey results might interact with those issues.

That iterative process of mapping results to areas of agency work is not done by the survey analyst alone but happens in collaboration with a senior "change champion." In the second column of figure 26.3, we see how the interaction between *power* and *expertise* within the agency generates the momentum for change, or at least signals it to the wider agency. In the third stage, proposed reforms must be implemented either at a *macro* level, across the public service or agency, or at a *micro* level, by a manager for, perhaps, a single work unit. For many reforms of the public service, a quorum of agency staff must accept the change and invest effort to shift to the new way of working. Together, these columns make up the architecture for impact on public processes, the quality of the work environment, and, eventually, the quality of services delivered.

This static exposition ignores the dynamic nature of these elements. As agency officials observe reactions to the FEVS survey results, the survey itself gains credibility, leading to greater participation in the survey.

**FIGURE 26.3** The Architecture of Policy Action for the Federal Employee Viewpoint Survey



Source: Original figure for this publication.  
Note: FEVS = Federal Employee Viewpoint Survey.



Greater participation, in turn, makes results more representative of the underlying issues, which, in turn, leads to more relevant reform approaches, increasing the credibility of the wider process. In this way, a virtuous circle can be formed. As Tracey Hilliard of the US Department of Health and Human Services (HHS) stated in an interview for this chapter, “Once everyone responds, more managers get [results specific to their work unit], and this ensures problems are less likely to be hidden in averages—a manager can tell what their particular issues are.”

These elements are all necessary to an agency architecture for inducing public action from FEVS raw data. Once these structures are in place, managers receive feedback on their performance and know that senior management is knowledgeable about the areas in which they need to improve. This creates accountability, communication, and a shared understanding for change.

## TRANSLATION OF SURVEY RESULTS THROUGH TECHNICAL EXPERTISE

The FEVS contains a substantial amount of information. For each respondent, there are roughly 85 questions/items (depending on the specific survey, with length varying by year and the track respondents follow). These questions can be assessed by a wide range of groupings, compared to previous years’ trends, or benchmarked against the dynamics of similar variables and groups. Each of these cuts of data can be made for each work unit or aggregated to the departmental, agency, and service levels. For this reason, the potential complexity of analyzing the FEVS is significant.

Similarly, although FEVS results are presented to senior managers in a series of high-level reports, they are also released in a relatively unstructured form to ensure maximum flexibility for managers to analyze those issues most relevant to their teams. As noted above, trying to provide managers with flexibility is one way the FEVS tries to be useful to a diverse public administration. But this confronts managers—many of whom do not have any background in survey data analysis—with the demanding task of making sense of rich but complicated data. That task must fit into their wider work of managing a work unit and undertaking their own portfolio of activities. Frequently, this combination of complexity and constraint prevents managers from fully engaging with the FEVS data. As Stephen Pellegrino of the US Department of Energy stated in an interview for this chapter, “We get a lot of data from OPM, and managers are not going to tease out what is relevant to them.”

Having a colleague whose work program includes time to undertake analysis of the FEVS data and who can identify their relevance for a work unit overcomes the first bottleneck to using the survey to generate reform. Simply having someone who can “translate” the data into practical issues for specific managers ensures the data have meaning for all officials, irrespective of their previous training and inclinations. Mr. Pellegrino noted that he provides his colleagues with simplified answers to the questions they have about the FEVS and only delves into greater detail on methodology for those who request it. As he frames it, “When you get down to a granular level, the statistics don’t matter as much as the story.”

This is particularly true for more senior members of the administration. As Gonzalo Ferro of the US Securities and Exchange Commission (SEC) argued in an interview, “There is a need to help leaders understand the OPM FEVS data for their organization.” For Mr. Ferro, this includes developing data visualizations (such as trend graphs and heat maps) that help managers make sense of the FEVS quickly and efficiently. “At the SEC,” Mr. Ferro reported, “we built a dashboard that makes all of our OPM FEVS data (from 2012 to present) accessible to all of our employees.”

Having a “technical expert” translate and report the results also consolidates the effort to engage with the FEVS at the agency. Potentially, this makes the analysis of the survey more efficient compared with having each manager do complementary work themselves. Such a survey analyst, aware of the priorities and issues of the organization they work for, may also be able to make salient results that speak to priorities. They can link the results to discussions throughout the organization.

In an interview for this chapter, Thevee Gray of the USDA expressed, “For strategic and effective change to happen, it’s important to know how to bridge the gap between the current state and our desired vision. That’s when the survey data plays an important role. It’s vital to know how to interpret the information, understand the culture and speak to both the grassroots and upper management.” In her experience analyzing FEVS data, while a survey expert is critical, so are the data. She continued,

My team and I leverage the FEVS data to help shine the light on issues within an organization and help managers recognize the importance of understanding the collective feedback from their employees. We presented an activity with them where we wrote on a board what they believed they were doing well and then showed them the FEVS data. It was an “aha” moment for all of them. The challenges they identified were completely opposite from employees’ perspective. If we don’t have this data to help guide them, management would focus on completely different issues. They would not be able to effectively close the gaps, wonder why the challenges remain and the needle has not moved in a positive direction.

Additionally, survey results are vital because they provide statistically valid information about what employees think. However, I always share with leadership to probe for what lies behind the survey results. Because as you analyze the data, it doesn’t explain why employees respond to questions as they do, and the reasons will not always be clear. This is why, when assessing the state of organization, the survey data should be used in conjunction with other information.

Ms. Gray worked for the USDA Farm Service Agency, and her work there provides an example of an agencywide initiative arising from the FEVS results. She used the FEVS analysis to identify staff recognition as an area of the work environment that was particularly challenging throughout the organization. Through a series of focus groups and managerial briefings, she and the wider agency came up with a system of celebratory coins themed with harvest-related features. Though the “USDA is not a coin culture—that comes from the military,” it worked effectively in giving managers a low-cost way to recognize excellence within their work units.

Similarly, Tracey Hilliard of HHS argued that it is important to have someone who can work with and interpret the FEVS data at the organizational level: “[The HHS Centers for Disease Control and Prevention] has 10,000 employees, so we created coordinators—two people in each organization that are a point of contact and can interpret the data. They came to meetings twice a month and helped get the data out to the managers to help translation.”

## PARTNERSHIPS FOR POLICY ACTION

In most hierarchical organizations—which, arguably, most agencies of the US federal government are—expertise in FEVS data analytics is not enough to generate change. To generate change, leaders must appreciate the validity and importance of feedback and use this information to make informed strategic decisions, including providing the necessary resources to affect change. In all of the interviews undertaken for this chapter, and in the broader experience of the FEVS team, change has always required buy-in from senior management and the supervisor of the organization. Without buy-in, power will be a bottleneck rather than an enabler. When discussing her experience of trying to generate responses to FEVS results, Thevee Gray argued that “the leadership buy-in was crucial to help shift the needle in a positive direction ... Once you have their buy-in, that cascades through the organization.”

As Ms. Gray pointed out earlier, without the FEVS as a diagnostic tool, management might not tackle the right work environment issues. Thus, in figure 26.3, change arises from the interaction between the survey analyst and the senior manager rather than from the manager alone. Tracey Hilliard suggests, “The survey

didn't change the organization; the leadership did, but they used the survey as their vehicle." As reflected in many of our interviews, the FEVS data do indeed seem to provide managers with new insights into their strengths and weaknesses and the current environment of their work units. Though this was reported to be truer for less-experienced managers and those with some of the worst results, the FEVS was felt to be instructive in general. It is interesting to note that in figure 26.2, those agencies with higher overall engagement scores are also those with the lowest variation in engagement across work units. This is consistent with the idea that agencies in which management has taken engagement seriously ensure that engagement is consistently addressed across the agency.

Once management begins to respond, the FEVS ensures a feedback loop is built into any reform program so that management can continue to measure their success in making relevant changes in the years following reform efforts. Increasingly, over the past 20 years, the FEVS has become the federal government management's tool for getting feedback on the current state of the administration, allowing those who are implementing reforms to course-correct their efforts. And as more senior managers understand the value of the FEVS as a management tool, the peer pressure on the wider cadre of management increases. The front-line supervisor must also understand the value of the FEVS analysis and be the accountable party to take the necessary next steps for reform. Correspondingly, the demand for survey analysts who can support the increased demand for FEVS analytics also increases.

The FEVS enables the team of the survey analyst and the change champion to develop appropriate reforms beyond just ensuring reforms are informed by the survey. The relationship between the survey analyst and the senior manager ensures reforms are based on evidence, but conversely, the FEVS data and evidence make that relationship possible. Stephen Pellegrino points out that without the FEVS data, a survey analyst may not be able to have conversations about areas of weakness with staff and managers across the organization. He suggests that "it's an easy way to start difficult discussions that are otherwise challenging to have—it's the data."

Once a formal position, team, or office is set up to process the FEVS data, it supports the further strengthening of the relationship with senior management. Like in any administration, having an office and personnel dedicated to a topic—in this case, the analysis of the FEVS—increases the salience and acceptability of a message. Karlease Kelly, formerly of the USDA, has presented across the US public service about the idea that consistently reporting FEVS results to managers makes them increasingly likely to accept the results and associated recommendations. The architecture becomes strengthened over time as managers come to view the FEVS as a standard part of their management approach.

## GENERATING A CULTURE OF CREDIBILITY AND RESPONSIVENESS

If the survey analyst provides the spark from the FEVS, and the relationship between the analyst and senior management is the positive friction that turns the spark into a flame for reform, there must be tinder and kindling for public action. Perhaps the key bottleneck to the use of results is the cultural resistance that agencies can have toward capitalizing on diagnostic data, such as the FEVS. The agencies that have been the most successful in using the FEVS for policy action have been those that have created a culture of using the FEVS across the entire staff to complement the basic scaffolding outlined above.

Without the cooperation and effort of the wider body of staff at an agency, any public action is unlikely to succeed. Thus, staff must feel that their efforts—to fill in the survey or support a reform—will be rewarded in some way. This can include simply fulfilling a norm that they expect to fulfill and that they expect others will also conform to. A strong FEVS-based reform culture is one in which all staff members believe that it is the norm to fill in the FEVS, to expect that its results will be responded to by senior management, and to agree that whatever change comes should be adhered to by all staff.

In part, such a culture requires the process of reform to be inclusive. Where cultures of FEVS-based reform have been built, FEVS results and identified focus areas were shared with staff, who were invited to

talk about them openly and often in a variety of venues. These meetings were community focused, diverse, and respectful. The FEVS was not presented as a report card but as a platform for discussion about where to go next. Once challenges were identified and generally agreed upon, staff were involved in changes in a positive way, such as through professional development opportunities. Reform leadership opportunities were created for those with a passion for the subject matter at hand to help lead projects, initiatives, and trainings that had been identified as necessary. A culture of responsiveness to the FEVS has arisen, in part, from credibility built over time. Once the relevant survey analyst and both senior management and frontline supervisors articulated to staff the results and what actions would be taken in response, it often took time for staff to believe this would be a systematic approach. Staff belief in action sometimes took years to develop.

The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) provides an example of how an agency has transformed its work environment by focusing on building a culture of responsiveness to FEVS results. In 2015 and 2016, the NIDDK was scoring toward the middle of the distribution of agencies in its sector. Though it was not one of the overall laggards, its senior leadership wanted it to become a stronger and more effective agency. They focused on key themes coming out of the FEVS and created a campaign based around the motto “You speak, we listen, things happen.”

The NIDDK formulated an approach based on three principles: share FEVS results and analysis broadly and continuously throughout the year, meet with subgroups to better understand their perspectives and feedback, and undertake focus groups and listening sessions to continue the conversation on FEVS results. The FEVS was no longer viewed as a report card looking backward but, instead, as a launchpad for robust conversation moving forward. A clear outreach strategy was combined with regular reporting on how challenges were being targeted. Actions taken by the agency were communicated and tied back to the FEVS results and, more specifically, to “the voice of the people.” Thus, NIDDK staff were given clear signals that senior managers had taken the FEVS seriously and were trying to improve the work environment in response.

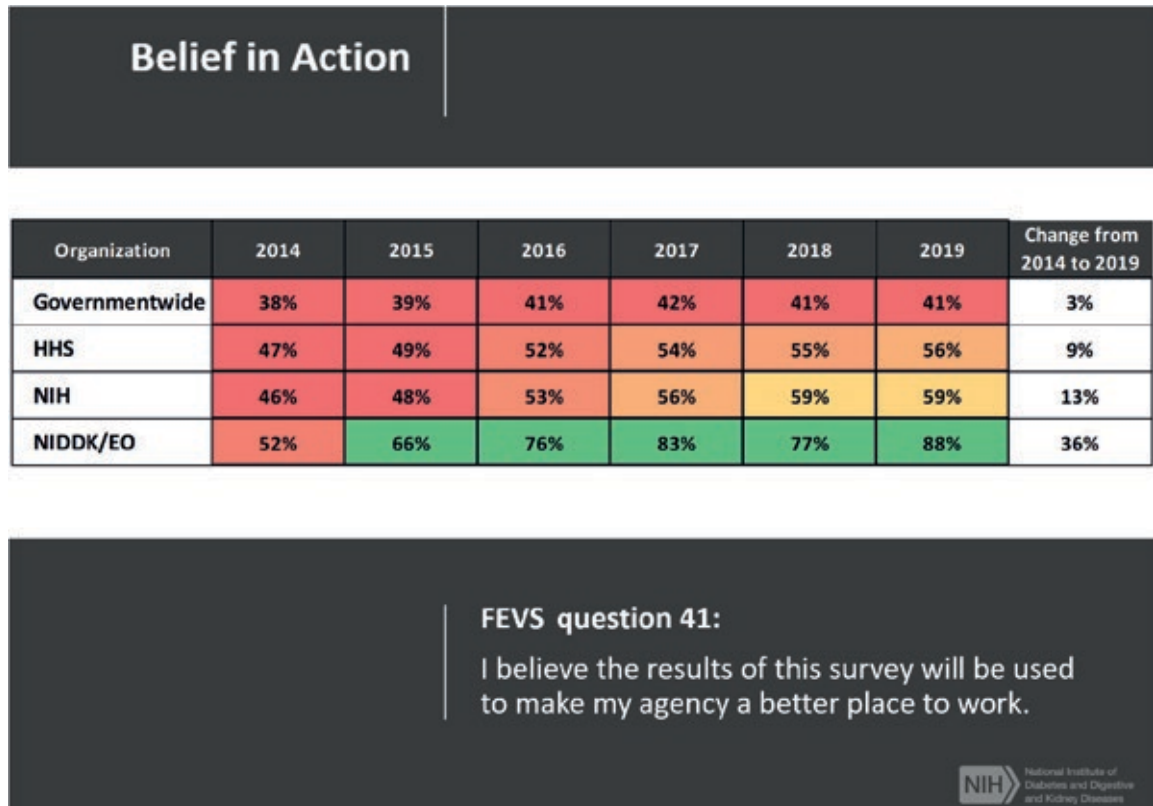
Staff confidence in the efficacy of the FEVS as a management tool grew, and figure 26.4 illustrates the difference this approach made. The figure shows the NIDDK’s trajectory of positive responses to the FEVS item “I believe the results of this survey will be used to make my agency a better place to work.” In contrast to little change in the government as a whole, the NIDDK’s trajectory was substantially positive, changing over 20 percentage points between 2014 and 2019. This also equated to increased survey participation, which climbed from 37 percent in 2014 to 69 percent in 2019, providing even more data for decision-making. Comparator scores are provided for the NIH and for HHS as a whole.

The NIDDK’s increased positive scores related to poor performance are also remarkable and are related to putting both standards and accountabilities into place, in addition to several targeted interventions. Figure 26.5 shows the NIDDK’s results for one of the lowest-scoring questions across the federal government: “In my work unit, steps are taken to deal with poor performers.” In 2015 and 2016, the NIDDK’s results were stagnant, like those of the government and the NIH as a whole. However, with the initiation of strategic initiatives and increased transparency through the launch of the “You speak, we listen, things happen” campaign in 2017, the proportion of positive responses jumped and continued to climb in the following years, indicating an increase in employees’ positive perception of how their agency dealt with poor performers. The NIDDK, using an architecture representative of that outlined in figure 26.3, successfully transformed its staff’s perception of accountability at the organization.

The strategic use of the FEVS has created a ripple effect throughout the institute. As of the 2020 FEVS cycle, the NIDDK’s cumulative scores in the areas of employee engagement, global satisfaction, and “leaders lead” were the highest across all 28 NIH institutes and centers, with positive percentage scores of 91 percent, 88 percent, and 90 percent, respectively.

In some ways, the actions undertaken at the NIDDK increasingly echo across the federal government. Using the FEVS as a critical management tool is becoming the norm, and cultures like the NIDDK’s are being built more widely. This is partly because the culture of the entire public service is being changed by the FEVS. Once disaggregated FEVS results were shared publicly and members of the government and the

**FIGURE 26.4** National Institute of Diabetes and Digestive and Kidney Diseases Staff Responses to Federal Employee Viewpoint Survey “Belief in Action” Question Compared to Organization-Level and Governmentwide Responses



Source: Screenshot of the National Institutes of Health's FEVS dashboard.

Note: EO = executive office; FEVS = Federal Employee Viewpoint Survey; HHS = Health and Human Services; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases; NIH = National Institutes of Health.

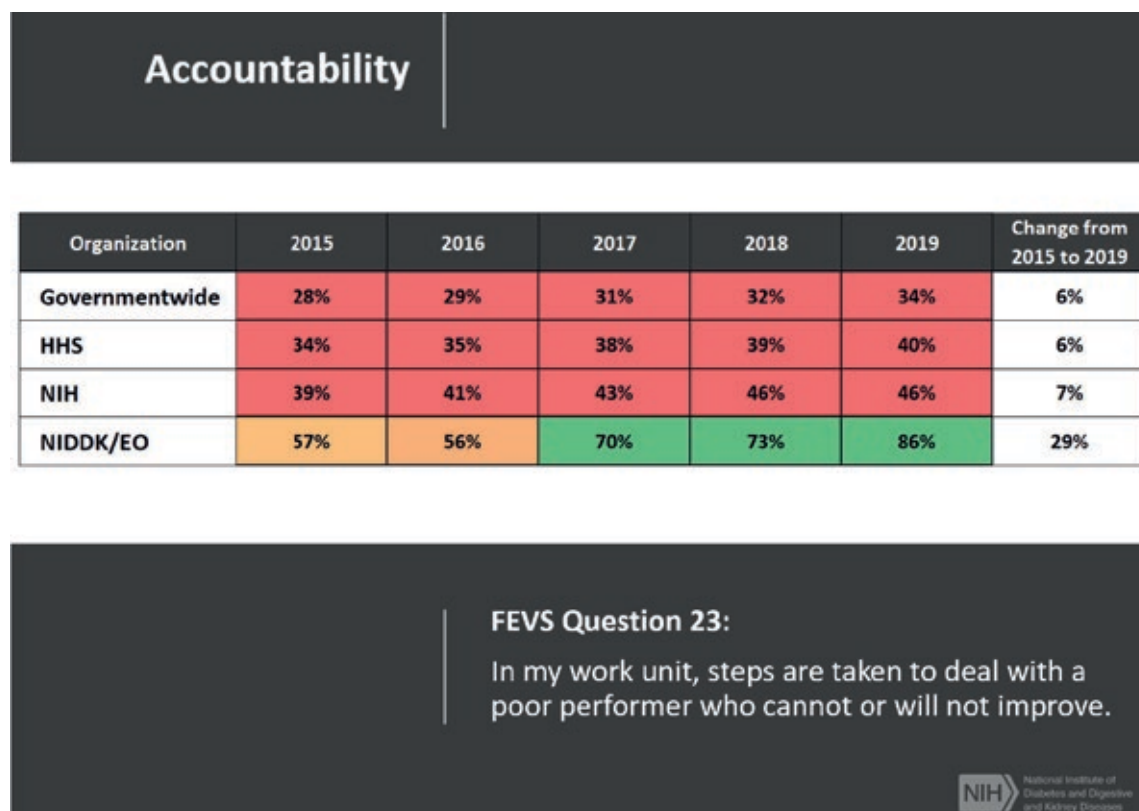
public were able to analyze the raw data by work unit, problem areas in the public service were increasingly difficult to hide within broader averages. This increased peer pressure on managers and changed the nature of recruitment because the quality of the workplace became more transparent. The FEVS data thus generated pressure from within and from outside organizations, making senior leaders more likely to pay attention to its findings.

In contrast to the “top-down” culture change process outlined so far, there are examples of “bottom-up” efforts to build responsiveness to FEVS results. Even when senior management fails to respond to the FEVS, public officials still want to use it as a tool to highlight problems in their organizations. These federal employees may have a passion for improving employee engagement for the benefit of staff and to better support their organizations’ missions, or they may want to improve their own work environments and see the FEVS as a tool to do so. In either case, the FEVS provides them with the ability to draw attention to needs, obtain buy-in for proposals, and measure the impact of the work being done.

A constraint to building a “bottom-up” culture for the use of the FEVS is the ability to communicate throughout an organization. There are often significant hurdles to frontline staff’s agreeing on the key issues presented in the FEVS data and generating a strategy in response. One such issue is the diversity of challenges faced within a single agency, as illustrated in figure 26.2. Thus, at least to date, much of the culture change around the use of the FEVS has arisen from the actions of survey analysts and senior management.



**FIGURE 26.5** National Institute of Diabetes and Digestive and Kidney Diseases Staff Responses to Federal Employee Viewpoint Survey “Accountability” Question Compared to Organization-Level and Governmentwide Responses



Source: Screenshot of the National Institutes of Health's FEVS dashboard.

Note: EO = executive office; FEVS = Federal Employee Viewpoint Survey; HHS = Health and Human Services; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases; NIH = National Institutes of Health.

## THE FEVS IN A WIDER ORGANIZATIONAL AND SOCIETAL ENVIRONMENT

The evolution of the FEVS as a tool for public action at the agency level has interacted significantly with the wider organizational and societal environment of public service. Centralized initiatives have driven agencies to better understand and value the survey as a management tool. Grassroots initiatives have generated novel insights and created a community of users. And external actors have supported the development of reform initiatives by presenting the FEVS data through new lenses. This section paints a picture of some of the most influential aspects of the broader environment in which agencies have worked.

### Influencing from the Center

Senior leaders are most likely to seek out the sorts of relationships with survey analysts that drive effective reform when they are told to do so from the president's office. In 2014, the FEVS became a part of the PMA and was added to leadership performance plans. This grabbed the attention of any senior manager who had not yet taken their FEVS results seriously and gave the FEVS a stronger accountability role. The PMA also encouraged senior managers to learn from agencies that had been the most successful in



developing a quality work environment and personnel management. An initiative collected successful workforce practices from across the federal government and created a platform to share them broadly. As a result, federal leaders, supervisors, and human resources practitioners can now easily review, evaluate, and adopt—or adapt—these proven successful practices with minimal effort. Appendix M.1 provides screenshots from the website.<sup>15</sup>

In a similar vein, the OPM has tried to collate best practices from across the public service so that when agencies determine that they want to improve in a particular area, they have resources to turn to. Appendix M showcases a screenshot of the OPM’s “Successful Workforce Practices” webpage that links to the resources outlined in appendix M.2, as well as other resources. The intention is thus to provide learning resources as well as accountability.

Such initiatives clearly complement agency-specific efforts to respond to the FEVS results. By increasing the salience of and incentives for responding to issues highlighted in the FEVS, the center makes senior officials more likely to set up an architecture like that outlined in figure 26.3. The learning resources provide a menu of options for responding to identified issues.

Much of this thinking was brought together by the “20-20-20 initiative.” The effort, a pillar of President Donald Trump’s first PMA, took aim at the lowest-performing work units in an agency. Trends had shown that while the entire government was improving, these units were falling even further behind. Notably, no one at the leadership level was responsible for focusing on these units. With the culture change now firmly in place, the goal was to improve the lowest 20 percent of an agency’s work units by 20 percent by 2020.

## Influencing across Agencies

Centralized initiatives to share best practices suffer from many of the same issues as centralized surveys. Topics, best-practice methods, and recommendations are all chosen at the center. But cross-agency collaboration and learning in the face of FEVS results have only grown over the past two decades. This learning is related to the facilitation and analysis of the FEVS as well as potential practices that could be put in place to address challenges identified by the survey.

One example of how agencies have tried to support one another in analyzing and responding to FEVS data is the Employee Viewpoint Survey Analysis and Results Tool (EVS ART).<sup>16</sup> In 2015, a small team at the NIH and the NIDDK worked to create a framework that would allow users to translate the enormous amount of survey data they received from the FEVS in a user-friendly and efficient manner. In many ways, it made the task of the first column of figure 26.3 easier by expanding the set of individuals who could undertake that role and potentially widening the number of managers able to analyze FEVS data themselves.

The resulting Excel-based tool, EVS ART, has provided officials across the government with a no-cost, practical, and easy-to-use resource that allows for the easy identification of focus areas with substantial time and cost savings. The team has gifted EVS ART governmentwide, with no usage or licensing fees, and has provided training and support for using it. This has helped to eliminate the duplication of effort because agencies and supervisors no longer need to conduct supplemental analysis. With a simple “copy and paste” motion, EVS ART takes employees’ FEVS feedback and—through a series of hidden pivot tables—translates it into the index measures supplied by the OPM (see screenshots of EVS ART in appendix M.3).

EVS ART contributed to solving two problems. First, it gave survey analysts and the wider community interested in analyzing FEVS data a handy tool for analysis. This reduced the time and cost required to produce disaggregated reports. Second, it showcased the use of the FEVS by other agencies, raising the survey’s profile as a management tool. EVS ART has been generally well received across the public service. As Tracey Hilliard states, “When EVS ART came out, that was wonderful.” It is an example of how grassroots action can complement agency efforts. However, EVS ART is only part of the wider architecture we have outlined, not a substitute for it.

## Influencing from beyond the Public Service

Finally, as hinted at above, external stakeholders have played a role in the development of the FEVS as a tool for public action. Simply by expecting the government to become more analytical, external stakeholders can put pressure on the government to use the FEVS as a tool. However, such an abstract approach is unlikely to gain traction. Instead, most external stakeholders have used the data to draw out interesting perspectives regarding public service that have influenced the debate inside the government about priorities for reform.

The most famous example of this approach is the Partnership for Public Service's Best Places to Work in the Federal Government index. The index ranks government organizations based on FEVS data in terms of the quality of the experience of working for them. As an example of how influential the index was, the December 2015 edition of *MyUSDA: A Progress Report for Employees on USDA's Cultural Transformation* was headlined "USDA Moves Up in Best Places to Work Ranking."

The Best Places to Work index uses a simple idea to motivate agencies to improve their rankings. Agency staff may feel motivated by a desire for their agency to look better in the rankings or to improve the talent pool that seeks employment at the organization. The index brings what is a relatively dry personnel issue in the public sector into the public sphere. Thevee Gray feels that:

the Partnership for Public Service has assisted in gaining the necessary attention for the FEVS. Senior leadership desires for their agencies to be seen as one of the best places to work in the government. As such, it holds them accountable and increases their responsibility to take the FEVS results seriously, knowing they will be published in an influential index and debated publicly.

## CONCLUSION

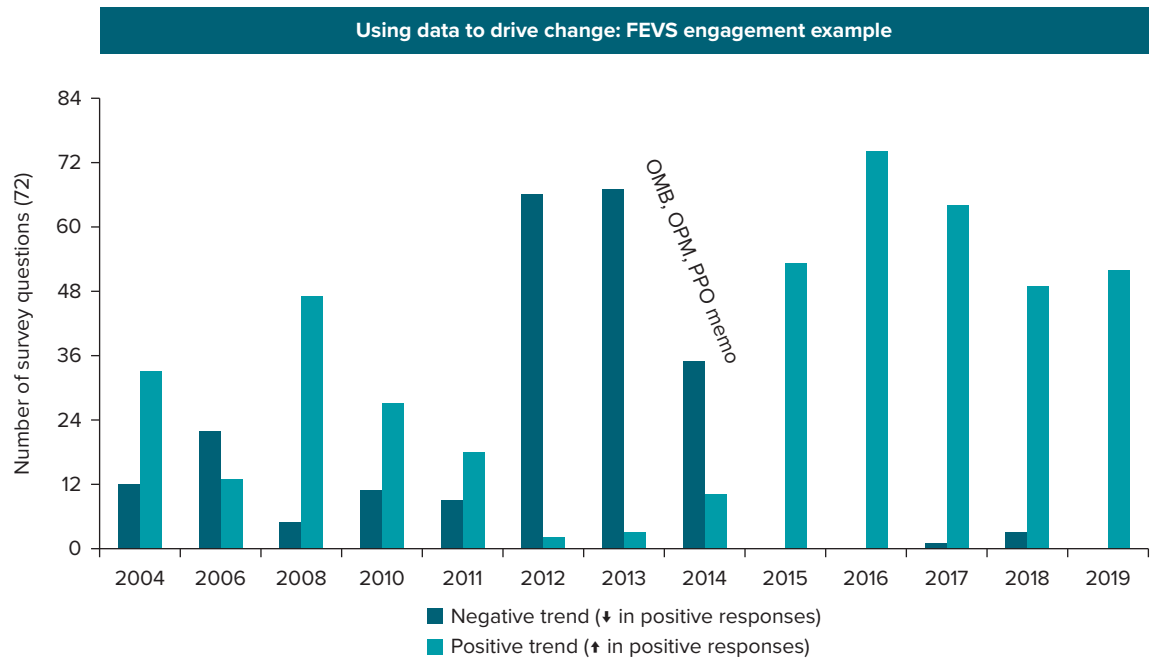
This chapter has argued for the potential power of surveys of public employees for driving public sector reform and improvement in the functioning of government. But it has cautioned that surveys must be embedded within a wider architecture of policy action for impacts to be realized. This architecture requires that a public sector organization have an official capable of translating survey results into actionable advice that a senior manager, working closely with that official, can use to build momentum for reform. And it requires responsiveness to survey results to build a culture that induces wider agency staff to contribute to reform.

Evidence for the arguments laid out in this chapter can be found in a 2014 joint memo from the Executive Office of the President and the OPM (Donovan et al. 2014). In many ways, its guidance closely tracks the arguments made here. After noticing that early adopters, like the Department of Transportation and the Federal Labor Relations Authority, demonstrated rapid improvement in their FEVS scores because of dedicated effort from their senior leaders, the OMB determined that the FEVS and employee engagement deserved elevation as a cross-agency priority goal in the PMA. A joint memorandum signed by OMB, OPM, and White House Presidential Personnel Office leadership with the subject line "Strengthening Employee Engagement and Organization Performance" laid out explicit mandates to agencies that echo the arguments of this paper.

Notably, each agency's career and noncareer leadership needed to take responsibility for changing how they had previously responded or, more realistically, did not respond to employee feedback. For two years after 2014, agencies had senior accountable officials and full-time staff dedicated to analyzing results and creating immediate action plans. The OMB included the results in the "FedStat" meetings held at senior levels as well as in reports to senior White House officials and directly to President Barack Obama.

The results from this effort were quickly apparent in the FEVS data themselves. Whereas the broad trend between 2012 and 2014 had been a declining number of positive responses to the FEVS questions,

**FIGURE 26.6** Trends in Negative and Positive Responses to Federal Employee Viewpoint Survey Questions, 2004–19



Source: US Office of Personnel Management.

Note: FEVS = Federal Employee Viewpoint Survey; OMB = Office of Management and Budget; OPM = Office of Personnel Management; PPO = Public Procurement Office.

the publication of the joint memo led to a positive trend in the majority of the FEVS questions in the following years (as shown in figure 26.6). This happened despite a change of administration, a government shutdown that lasted more than a month, below-market pay adjustments, and the start of the COVID-19 pandemic. Thus, as the FEVS became embedded in a wider architecture for public action, it became a stimulus for reform.

In many cases, where FEVS results were not translated into change in practice, it has been due to the lack of architecture at the agency level to support that translation process. Simply producing survey data, of however high a quality, is rarely enough to drive public action.

Despite the qualities of the FEVS and its related successes, there are legitimate criticisms of the relevance and scope of the FEVS survey questions. The FEVS focuses on drivers of staff engagement and thus has a limited scope in terms of topics. Similarly, given the complexity and breadth of work undertaken in the US federal government, it seems natural that a standardized survey would not be the most effective driver of change across all federal agencies all of the time. But the impacts it has inspired showcase the potential of employee surveys in inducing public action for better government.

## NOTES

The authors are grateful to Corey Adams, Gonzalo Ferro, Thevee Gray, Tracy Hilliard, and Stephen Pellegrino for discussions. Many of the authors have played a key role in the development of the Office of Personnel Management's Federal Employee Viewpoint Survey or in its translation and use at the agency level over the past decade. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of any of the US federal agencies with which several authors are affiliated, including the US Office of Personnel Management.

1. Discussion of aspects of the FEVS can be found in chapter 9,, case study 9.3 in chapter 9, and chapter 25. FEVS data are also used in chapters 19, 20, 21, and 22.
2. For those interested in the specific details of the survey, the OPM releases technical reports each year to accompany the survey report and data. These are available on the OPM FEVS website at <https://www.opm.gov/OPMFEVS/reports/technical-reports/>. A special “research synthesis” in the *Public Administration Review* (Callahan 2015) articulated a series of academic perspectives on the strengths and weaknesses of the FEVS for public service reform and research.
3. See, for example, GAO (2015), which focuses on drivers of engagement and implications for various agencies; GAO (2018), which focuses on the OPM’s delivery of information on performance management; and GAO (2021), which focuses on the US Department of Homeland Security.
4. The current Best Places to Work index is available on the Partnership for Public Service’s website at <https://bestplacetowork.org/about/methodology/>.
5. The publication of the index by an external entity also allows for independent assessments of what drives improvements in the public service work environment, such as Partnership for Public Service and Deloitte (2013).
6. The full set of reports published by OPM can be found in the “Data Reports” section under “Reports” on the OPM FEVS website at <https://www.opm.gov/fevs/reports/data-reports>. Releases are limited to groups of at least 10 officers to safeguard against the identification of respondents.
7. The FEVS also provides the OPM itself with insights into weaknesses in the public service system as a whole that can be targeted without direct agency action. However, this chapter will focus on agency-level reform efforts in response to FEVS findings.
8. There are many potential topics on which a survey could focus, including the physical environment, relationships between colleagues, the quality of management, the engagement of the survey respondent with his or her job, and the most significant challenges the respondent finds to undertaking their work effectively. Chapter 18 provides an overview of the topics the world’s major public servant surveys focus on.
9. National Defense Authorization Act for Fiscal Year 2004, Public Law 108–136, Nov. 24, 2003, 117 STAT. 1641.
10. *Examining Misconduct and Mismanagement at the National Park Service: Hearing before the Committee on Oversight and Government Reform, House of Representatives*, 114th Cong. (2016). <https://www.congress.gov/event/114th-congress/house-event/LC51983/text?s=1&r=100>.
11. In a statement in response to the findings of the hearings, Deputy Director of the National Parks Service Michael Reynolds made reference to actions the agency took to try to improve working conditions for park staff (Reynolds 2016). A webpage outlining the response to the harassment issues can be found on the National Park Service website at <https://www.nps.gov/aboutus/transparency-accountability.htm>.
12. An alternative perspective is that centralized surveying generates greater awareness and appetite for surveys of public servants, thus increasing the likelihood of complementary efforts by managers. Frequently, the FEVS has inspired follow-up surveys by agencies seeking to better understand an area in which they are performing relatively poorly. And the structured survey approach of the FEVS can be complemented by deeper-dive focus groups and listening sessions, which can more deeply explore red flags relevant to a particular agency.
13. The FEVS has gradually changed its content over time to meet the evolving demands of officials using it for policy assessment and organization development initiatives. As outlined in figure 26.1, in 2012, a series of items were added to the FEVS to improve how well it could inform OPM policy evaluations, reports to Congress, and oversight functions, as well as workforce development initiatives within agencies. In 2020, items were added to support policy assessments, including military spouse items and new leave policies for COVID-19 pandemic response. Simultaneously, the performance confidence index was added to support change and development initiatives and action in agencies.
14. Conceptually, organizational climate is a surface manifestation of culture: employees’ perceptions of management practices and policies speak to the values and norms embodied in a culture.
15. The OPM highlights the main features of the successful workforce practices initiative on the “Successful Workforce Practices” page of its website at <https://www.opm.gov/policy-data-oversight/human-capital-management/successful-workforce-practices/>. The website housing the full collection of successful practices is accessible by US government officials at <https://community.max.gov/display/HumanCapital/PMA+Successful+Workforce+Practices+Home>.
16. EVS ART can be accessed by all US government officials at <https://community.max.gov/display/HHS/EVS+ART>. A fuller exposition of EVS ART is provided in chapter 9 and case study 9.3 in chapter 9.

## REFERENCES

Brust, Amelia. 2021. “After 15 Years of Best Places to Work, Data Findings Consistently Point to Engagement Needs.” *Federal News Network*, August 5, 2021. <https://federalnewsnetwork.com/hiring-retention/2021/08/after-15-years-of-best-places-to-work-data-findings-consistently-point-to-engagement-needs/>.

- Callahan, Janelle. 2015. "From Results to Action: Using the Federal Employee Viewpoint Survey to Improve Agencies." *Public Administration Review* 75 (3): 399–400.
- Donovan, Shaun, Beth Cobert, Katherine Archuleta, and Meg McLaughlin. 2014. "Strengthening Employee Engagement and Organizational Performance." Memorandum for Heads of Executive Departments and Agencies, December 23, 2014. [https://www.whitehouse.gov/wp-content/uploads/legacy\\_drupal\\_files/omb/memoranda/2015/m-15-04.pdf](https://www.whitehouse.gov/wp-content/uploads/legacy_drupal_files/omb/memoranda/2015/m-15-04.pdf).
- GAO (US Government Accountability Office). 2015. *Federal Workforce: Additional Analysis and Sharing of Promising Practices Could Improve Employee Engagement and Performance*. Report to Congressional Requesters, GAO-15-585. Washington, DC: US Government Accountability Office. <https://www.gao.gov/products/gao-15-585>.
- GAO (US Government Accountability Office). 2018. *Federal Workforce: Opportunities Exist for OPM to Further Innovation in Performance Management*. Report to the Chairman, Committee on Homeland Security and Governmental Affairs, US Senate, GAO-19-35. Washington, DC: US Government Accountability Office. <https://www.gao.gov/products/gao-19-35>.
- GAO (US Government Accountability Office). 2021. *DHS Employee Morale: Some Improvements Made, but Additional Actions Needed to Strengthen Employee Engagement*. Report to the Chairman, Committee on Homeland Security, House of Representatives, GAO-21-204. Washington, DC: US Government Accountability Office. <https://www.gao.gov/products/gao-21-204>.
- Mullins, Luke. 2021. "The Best Places to Work within the Federal Government, Ranked." *The Washingtonian*, June 29, 2021. <https://www.washingtonian.com/2021/06/29/the-best-places-to-work-within-the-federal-government-ranked/>.
- OPM (Office of Personnel Management). 2012. *2011 Federal Employee Viewpoint Survey: Technical Report*. Washington, DC: US Office of Personnel Management. Accessed April 10, 2022. <https://www.opm.gov/fevs/reports/technical-reports/>.
- OPM (Office of Personnel Management). 2022. "About." OPM Federal Employee Viewpoint Survey, US Office of Personnel Management, United States Government (accessed March 10, 2022). <https://www.opm.gov/fevs/about/>.
- Partnership for Public Service and Deloitte. 2013. *Ten Years of the Best Places to Work in the Federal Government Rankings: How Six Federal Agencies Improved Employee Satisfaction and Commitment*. Boston: Deloitte. [https://www.opm.gov/policy-data-oversight/training-and-development/reference-materials/online-courses/maximizing-employee-engagement/content/common/cw/data/Ten\\_Years\\_of\\_BPTW\\_Rankings.pdf](https://www.opm.gov/policy-data-oversight/training-and-development/reference-materials/online-courses/maximizing-employee-engagement/content/common/cw/data/Ten_Years_of_BPTW_Rankings.pdf).
- Reynolds, Michael. 2016. "NPS Misconduct: Examining Misconduct and Mismanagement at the National Parks Service." Statement of Michael Reynolds, Deputy Director, National Park Service, Department of the Interior, before the House Committee on Oversight and Government Reform, on the National Park Service response to incidents of employee misconduct, September 22, 2016. Office of Congressional and Legislative Affairs, US Department of the Interior. <https://www.doi.gov/ocl/nps-misconduct>.

The background features a series of blue squares arranged in a grid-like pattern that curves from the top left towards the center. Overlaid on this is a stream of binary code (0s and 1s) that also curves across the upper portion of the page. A solid blue horizontal band spans the width of the page, serving as a backdrop for the 'PART 5' text.

## **PART 5**

# Government Analytics Using External Assessments





## CHAPTER 27

# Government Analytics Using Household Surveys

Faisal Ali Baig, Zahid Hasnain, Turkan Mukhtarova, and  
Daniel Rogger

### SUMMARY

This chapter presents a guiding framework for using household survey microdata, readily available to most governments, to develop insights into the structure of the public sector workforce and the qualities of its compensation practices. National statistical authorities frequently collect household surveys with detailed information on labor force participation. These surveys are broadly consistent across time and are developed using globally standardized definitions and classification nomenclature. This offers governments unique insights into the public sector workforce that cannot be derived solely from administrative data sets, including the ability to juxtapose the demographics and skills composition of the public sector workforce to the private sector and assess the relative equity and competitiveness of public sector compensation practices. The chapter provides illustrations of the insights into public sector employment and wages that can be generated by this framework, using examples from the World Bank's Worldwide Bureaucracy Indicators (WWBI). Such insights can inform policy choices related to managing human resources in the public service.

### ANALYTICS IN PRACTICE

- Governments routinely conduct household surveys in order to understand the populations they serve, target public policy, and inform policy debates. Many of these surveys record a dedicated and detailed set of variables on the labor market experiences of people in the country, including whether respondents work in the public or private sector. By collecting comparable data across the two sectors, household surveys provide a foundation for understanding the characteristics of public officials compared with their private sector counterparts—which is not possible through administrative data sources alone. Including precise and coherent indicators of the employment sector in household surveys thus enables their use in understanding the characteristics of public sector workers.

---

The authors' names are listed alphabetically. Faisal Ali Baig is a consultant, Zahid Hasnain is a lead governance specialist, and Turkan Mukhtarova is a consultant in the Governance Global Practice of the World Bank. Daniel Rogger is a senior economist in the World Bank's Development Impact Evaluation (DIME) Department.

- Given the unique nature of the public sector, time-series and international benchmarks and comparisons are critical for understanding the current state of government functioning. The precision and consistency with which household surveys are conducted across time open up the possibility of understanding the longitudinal dynamics of the public sector relative to the private sector. Similarly, ensuring such surveys collect data in line with globally standardized definitions and classification nomenclature enables comparisons across countries.
- Setting up household surveys in this way allows the government to juxtapose any feature of individuals in the public and private sectors that surveys have collected data on. Demographic variables allow for an analysis of how gender-related differences in pay vary across sectors, regions, organizations, and so on. Assessments of the skills composition of the public sector workforce with respect to the private sector identify in what areas the government is competing most intensively for skills with private sector actors and what that competition is doing to wages.
- Taking household surveys as the foundation of analytics around the equity and competitiveness of public sector compensation practices is a relatively low-cost approach to the analytics of the personnel determinants of state capacity. However, the usefulness of these surveys is underpinned by the representativeness of sampling and interviews, which ensure that the resulting data provide a robust estimation of labor force participants. This requires coordination between agencies of public employment and national statistical agencies.
- Detailed data from household surveys on the distribution and remuneration of public employees can help identify more nuanced, targeted, and politically feasible reforms that make explicit the difficult trade-offs in employment and compensation policies. Such an evidence-based approach is necessary because, historically, public sector employment reforms have often occurred in the context of economic crises, with an emphasis on blunt, short-term fixes that can have adverse impacts on long-term growth and welfare, and often create distortions and perverse incentives.

## INTRODUCTION

The effective management of public sector employment and compensation is a vital activity of governments, with broad implications for fiscal sustainability, public sector productivity, and the competitiveness of the overall labor market. The wages of public sector employees consume a significant proportion of government expenditures. Across the world, government expenditures on employee compensation represent, on average, 30 percent of total expenditures (Hasnain et al. 2019). Spending on public sector salaries comes at the opportunity cost of spending on public sector programs.

At the same time, human resources in the public service are essential providers of government services and infrastructure, as well as ensuring the effectiveness of regulations (Arizti et al. 2020; Ingraham, Joyce, and Donahue 2003; Moynihan and Beazley 2016; Rasul and Rogger 2018). The size and nature of public sector wages affect the selection, retention, and motivation of public sector workers, which, in turn, impacts productivity, the amount and quality of government outputs, and public service provision (Finan, Olken, and Pande 2017).

These issues matter not only because they impact the quality of government functioning. The public sector is a large employer, accounting for, on average, 37 percent of global formal employment (Baig et al. 2021). Changes in government wages are likely to produce significant effects on the national labor market and the overall economy, including potentially crowding out recruits in the private sector (Behar and Mok 2013). In many lower-middle-income countries, especially those experiencing fragility, public sector employment is the core ingredient of the political settlement, and wage bill reforms have immediate and often severe implications for political stability, peace, and security (Gifford 2006).

There are thus several important questions about the public sector workforce that governments regularly need to address. What is the appropriate level of employment in the public sector as a whole and for essential workers like public administrators, teachers, and doctors, in particular? Does the public sector pay competitive wages compared to the private sector to attract talent while not crowding out private sector jobs? Does the public sector promote gender equality in employment, both in absolute terms and relative to the private sector? And are public sector pay and employment practices contributing to robust and dynamic labor markets at the national and regional levels?

Answering these questions requires high-quality microdata on public sector employment and compensation and comparable data for the private sector. Utilizing household surveys as a source of information on public employment offers certain advantages over administrative data. These data are routinely collected by national statistical organizations for informing broader policy goals and thus represent a cost-effective tool for government analysts. Household surveys provide a rich, consistent, and regularly updated set of variables for a variety of worker characteristics in the public and private sectors that enable robust, controlled comparisons between the two groups. Such surveys allow data to be drawn from the public and private sectors in a common manner. Thus, these data often represent a richer source of insights than are available from administrative data alone.

This chapter is targeted at government officials, development practitioners, and researchers who aim to gain a better understanding of the structure of the public sector labor market and its implications for the overall labor market. It begins by introducing the advantages of this survey-based framework and key areas for caution. It showcases the main features of the methodology and presents guidance for conducting analysis to delineate trends from these surveys. It goes on to provide guidance on how improvements in the design and conduct of labor force surveys allow for even more granular analysis. Finally, it illustrates the breadth of insights that can emerge from a study of public administration founded on household survey data. The approaches outlined here are a natural complement to those presented in chapter 10 of *The Government Analytics Handbook*, trading off the granularity of data with comparability with the private sector.

## THE POWER OF USING HOUSEHOLD SURVEYS FOR GOVERNMENT ANALYTICS

Nationally representative *household surveys*, collected by national statistical authorities, are some of the most professionally conducted surveys in the world and are frequently supported or improved through consultations with multilateral organizations' data teams, which possess substantial experience in such exercises. By collecting data on a representative sample of the whole or some subset of the population, such surveys provide a window into the lives of those about whom data are collected. When surveys collect data on whether respondents work in the public or private sectors, they provide windows into life in those sectors.

This chapter will focus on employment-related variables in such surveys and thus on the *labor force modules* in household surveys. When these modules are the dominant concern of a survey, the survey is typically classified as a *labor force survey*. In much of the rest of the chapter, we will use these terms interchangeably. However, most of the principles of the discussion carry over to other elements of standard household surveys, such as consumption patterns drawn from consumption modules.

### The Strengths of a Survey-Based Approach

Governments routinely rely on household survey sources to generate headline indicators of the health of labor markets and the overall economy. Insights emerging from these surveys are often used in the design of a wide array of economic and social policies. For example, information on the share of employed (and unemployed) individuals within the labor force frequently has direct consequences on the monetary and fiscal

policies of the government. Unemployment rates are often used as proxies for the vitality of the labor force and are used—in combination with inflation rates—in determining interest rates by central banks.

The quality of these data makes them an attractive foundation for government diagnostics of relative labor market characteristics and dynamics. Specifically, utilizing nationally representative labor force survey data to characterize public and private labor markets offers government and analysts five unique advantages over other data sources.

First, labor force surveys provide a rich, consistent, and regularly updated set of variables for a variety of worker characteristics in the labor market. Given the investments governments have made in methodological rigor, effective implementation, and quality assurance, these surveys are one of the richest available sources of information on population characteristics. Household surveys provide coherent descriptions of the composition of individuals within households, their demographics and qualifications, their consumption behaviors, and the nature and sector of their participation in the labor market, as well as detailed indicators on the industries and occupations they are engaged in and their salaries and other sources of compensation (including in-kind payments, government assistance programs, and social security benefits).

Second, labor force surveys undertake the same measurement approach across the public and private sector labor markets. This is a unique advantage of these surveys for measuring state capacity because these data are collected simultaneously for workers in both sectors from the same sample frame in a coherent manner. Administrative data sources (while being a potentially more accurate and detailed measure of employment and wages in the public sector) only include information on public sector employees and, at times, only the employees of particular ministries or organizations. It is extremely unlikely that any single administrative data set would not only cover workers employed across a diverse set of economic activities (from agriculture to mining, manufacturing, and the services sector) but also include information on both public and private sector labor force participants.<sup>1</sup> Even if such a data set does exist, its data will rarely be consistent with administrative data from other countries, complicating international comparisons.

Third, the granular nature of the underlying data ensures that labor market models are based on representative data sourced from across the economy. These surveys often sample thousands of employed individuals and are based on a meticulously designed sampling frame based on national census data, allowing for a close and accurate approximation of local labor markets. This reduces the assumptions on which analysis is based—the data are allowed to speak for themselves—and allows for decomposition by the characteristics of workers where sample sizes are sufficiently large.

Fourth, household surveys may represent a more complete view of the public sector workforce than even administrative data sets. Public sector administrative data are often too restrictive in defining who is included in their measurement. For example, contract workers have become an essential part of the public sector workforce, working alongside permanent staff in the promulgation of regulations and the delivery of social services. In many developing countries, they also represent a sizeable proportion of the public sector education and health care workforce. However, they are not counted as formal public employees in many administrative systems; that is often the reason for their contractual status. Given that contract workers are often exempt from budgetary limits on spending, their recruitment allows ministries to circumnavigate budgetary conditionalities against overspending on personnel.

This impairs the ability of governments to assess the true size of the public sector workforce. Further, given that these data sets are often unique to individual public sector organizations, the determination of who does or does not constitute a public sector worker may differ across organizations. Both of these factors would bias any estimate of the size of the public workforce, weakening the ability of governments to track wage bill spending. Survey data, on the other hand, are not limited by this distinction. Given that they are collected directly from individuals, who can elaborate on their sector employment, surveys can allow for a better determination of the size and structure of the public sector workforce.

Finally, household survey data are typically collected with research and diagnostic objectives in mind. Administrative data sets are collected for a variety of nonstatistical reasons, such as human resource management, program administration, or other regulatory or enforcement purposes. Therefore, administrative data in their “raw” form may not be suitable for statistical analysis.

## Areas to Be Cautious in a Survey-Based Approach

It is important also to point to the caveats associated with the use of survey data. Labor force surveys can only generate insights into the nature and organization of public sector human resources with rich, complete, and consistent data collection on labor force participants, in general, and the public sector workforce, in particular. Given the self-reported nature of household surveys, respondents may not be able to fully articulate or comprehend nuances within their responses around the nature of their employment. While these surveys are designed and extensively piloted with particular care given to how questions may be interpreted, in order to ensure the quality of responses, there may be lingering imprecision within the variables that define individuals as working in specific sectors.

Utilizing a more broadly defined public sector identifier may make it easier for respondents to accurately answer relevant survey questions and allow for a more comprehensive comparative analysis between the public and private sectors. However, this may make the survey's definition of the public sector unfit for particular purposes. It is often difficult to differentiate between, for example, federal and state employees or those that are employed within specific ministries and those employed within state-owned enterprises.

The second area for caution is the respondents' representativeness of the underlying population of public and private sector workers. Labor force surveys, by definition, sample the working-age population. When designing the sampling frame, surveys often strive to ensure a representative sample in terms of age, gender, and racial and ethnic demographics. Some surveys aim to sample a representative share of employed and unemployed individuals and those not active in the labor force. However, they rarely if ever explicitly attempt to ensure a balanced sample of public and private sector workers. Analysts must thus assess whether the sampling approach might have biased data collection toward one or the other sector's employees or otherwise changed the nature of measurement in either sector.<sup>2</sup>

## SETTING UP THE ANALYSIS

### Capitalizing on Current Household Surveys

What features of a survey are necessary for it to be useful for government analytics? The answer to this question will be determined by the specific analysis intended. This chapter therefore follows the requirements of an analytical framework used by the World Bank to understand public and private sector labor markets in the development of the Worldwide Bureaucracy Indicators (WWBI). The WWBI is a unique cross-national data set on public sector employment and wages that utilizes global repositories of household survey data from 202 economies to present a globally consistent and analytically rigorous set of indicators across five categories: the demographics of the private and public sector workforces, public sector wage premiums, relative wages and pay compression ratios, gender pay gaps, and the public sector wage bill.<sup>3</sup>

WWBI indicators on public employment track key demographic characteristics, including the size of the public sector workforce (in absolute and relative numbers), the age of the workforce, and the distributions of employees across genders, industries, income quintiles, and academic qualifications. Compensation variables capture both the competitiveness of public sector wages (compared to the private sector) and wage differentials across industry or occupation, gender, education, and income quintile within the public and private sectors, as well as pay compression ratios in the public and private sectors.

What features of a survey make it eligible for inclusion in the WWBI, and, more broadly, what features are useful for the analytics of public and private labor markets? Since the WWBI focuses on national aggregates, the survey must be representative at the country level (rather than, for example, including just urban areas). Correspondingly, the survey must have taken a sampling approach that attempts to represent each unit of observation across the country equally. Beyond the WWBI, if analysts are only interested in public



sector labor markets in urban areas, the survey should have appropriately sampled within the requisite conurbations of the country.

A second set of requirements relates to the size and composition of surveys included in the WWBI. Specifically, attention is paid to the sample sizes for major categories of respondents. The ability of the WWBI to properly characterize the public and private sector workforces is dependent on the underlying surveys' possessing large enough samples of these two categories of workers that any estimates would approximate the demographics and compensation of the actual labor forces they model. Within the WWBI framework, surveys with fewer than 200 observations for either labor market, or in which either labor market comprises less than 5 percent of all employed individuals within the survey, are excluded from the analysis. More broadly, any survey should be judged on its ability to enable statistically valid inferences on the underlying population.

Third, the survey should have a sufficient sample size for key variables, so as not to be dropped by the WWBI's quality filters. There are three sets of variables we use for the WWBI, presented in table 27.1. If a survey does not include any of the variables shaded in green in table 27.1 or has greater than 40 percent missing/miscoded observations for any of those variables, the survey is disregarded. If a survey is missing any of the variables shaded in blue or has greater than 40 percent missing/miscoded observations for any of those variables, the specific set of variables related to that module is excluded. The gray variables are additional variables that are not universally used in the construction of the WWBI variables, so we do not require them. However, those variables related to sampling are required if sampling weights were used. Finally, the unshaded variables are frequently used to investigate outliers and so are useful to have if available.<sup>4</sup>

The availability of the variables outlined in table 27.1 provides analysts with a basic setup for labor market analysis. Many such analyses look to compare contemporary results over time or across countries. This requires the availability and harmonization of variables across surveys. The WWBI aims to produce statistics that can be compared across time and space and thus faces issues of the classification of employees, the definition of the public sector, and the formulation of wages.

The classification of employed individuals, paid employees, and public paid employees is based on labor and employment status and sector type. Definitions for total, paid, and formal employment are based on the International Labour Organization (ILO) International Classification of Status in Employment (ICSE), making the WWBI and the ILOSTAT databases cross-compatible (fundamental differences in survey coverage, representation, sample size, and timing notwithstanding). According to the ICSE, total employment is defined as

all those of working age who, during a short reference period, were engaged in any activity to produce goods or provide services for pay or profit. They comprise employed persons "at work," i.e., who worked in a job for at least one hour; [and] employed persons "not at work" due to temporary absence from a job, or to working time arrangements (such as shift work, flexitime and compensatory leave for overtime). (ILO 2013, 6)

Paid employment refers to a subsection of total employment and includes only salaried workers, excluding unpaid or own-account (commission-based) employees, employers, and those that are self-employed. Formal employment is a further subset of paid employment and counts those who are employed in formal occupations (possessing a written contract or having access to benefits like health insurance, pensions, or union membership).

A globally harmonized definition of the public sector is hindered due to issues of comparability emerging from the heterogeneous definition of public employees across countries. To avoid this, the WWBI, as a guiding principle, utilizes the more broadly defined *public sector* as opposed to *general government*, as defined by the International Monetary Fund (IMF) *Manual on Government Finance Statistics* (IMF 2014). Specifically, the public sector consists of all institutional units controlled by the central and subnational governments, as well as public corporations that are engaged in market-based activity. Utilizing this broader definition allows for a cleaner comparison across national surveys.

To make wage data as comparable as possible across surveys, the WWBI denotes only the income associated with the occupation used in the analysis (to which the individual dedicated most of their time in the

**TABLE 27.1 Variables Required for the Inclusion of a Survey in the WWBI**

<b>Metadata variables</b>
Survey ID
Country ID
Year of the survey
Month of the interview
Household ID
Individual ID
Household sampling weight
Strata
Primary sampling units ID
<b>Demographics</b>
Household size
Gender
Age
Urban/rural
Education module application age
Ever attended school
Attending school
Years of education
Level of education (no education, primary, secondary, tertiary)
<b>Labor module</b>
Labor module application age
Labor status <ul style="list-style-type: none"> <li>• Employed</li> <li>• Unemployed</li> <li>• Not in labor force</li> </ul>
Employment status <ul style="list-style-type: none"> <li>• Paid employee</li> <li>• Unpaid employee</li> <li>• Employer</li> <li>• Self-employed</li> <li>• Other, workers not classifiable by status</li> </ul>
Number of additional jobs
Sector of activity (public vs. private) <ul style="list-style-type: none"> <li>• Public sector, central government, army, NGO, state-owned company</li> <li>• Private</li> </ul>
Industry sector classification (minimum one-digit level, but three-digit level is required for occupational decomposition)
Occupational classification (minimum one-digit level, but three-digit level is required for occupational decomposition)
<b>Wage module</b>
Hours of work in the last week
Last wage payment

Source: Original table for this publication, based on World Bank 2021.

Note: The table shows the three sets of variables used. If a survey does not include any of the variables shaded in green or has greater than 40 percent missing/miscoded observations for any of those variables, the survey is disregarded. If a survey is missing any of the variables shaded in blue or has greater than 40 percent missing/miscoded observations for any of those variables, the specific set of variables related to that module is excluded. The gray variables are additional variables that are not universally used in the construction of the WWBI variables, so are not required. ID = identification; NGO = nongovernmental organization; WWBI = Worldwide Bureaucracy Indicators (World Bank).

week preceding the survey) and excludes bonuses, allowances, and other in-cash or in-kind payments from the same job, as well as all additional sources of income (from other jobs) or investments and transfers. Due to the almost complete lack of information on taxes in most household surveys, the wages from the primary job are not net of taxes. For those who are self-employed or own their own businesses, this corresponds to net revenues (net of all costs excluding taxes) or the amount of salary withdrawn from the business.<sup>5</sup>

Wage information in the surveys is reported in each country's local currency units, with a diverse array of periodicity. Great care should be taken to identify the exact frequency of income for each individual within the surveys and convert all wages to a weekly (or another common unit of) wage after accounting for the varying hours worked to ensure credible comparisons across individuals and groups. Additionally, to control for the effect of possibly spurious outliers, the wage variables in the WWBI are winsorized by limiting extreme values in the survey data at the top 0.01 percent level.<sup>6</sup> More broadly, analysts may want to be cautious with wage information that seems like an outlier from the general distribution of a particular survey.

Overall, to be useful for government analytics, existing household surveys should have sufficient coverage of the population and relevant variables, be of sufficient size, and, where comparisons to international surveys are required, have questions appropriately harmonized with international standards. Fitting these criteria, individual country efforts can always be integrated into existing indicators, such as the WWBI, or compared with relevant surveys in other countries of interest.

## Extending Data Collection

What if appropriate household surveys do not exist? Governments, independent organizations, or even individual analysts may be in a position to create and field such surveys. In many instances, project teams from the World Bank have run large, nationally representative household surveys themselves to collect information to aid policy guidance. In India, a private sector organization, the Centre for Monitoring Indian Economy (CMIE), complements the government's labor force survey. By operating the "world's largest household panel survey," with over 2 million individual respondents covering 236,000 households three times a year, the CMIE increases the frequency of up-to-date labor market data for the government and other stakeholders.<sup>7</sup>

Household surveys that are optimized for government analytics could solve the issues with representative sampling identified above by targeting populations of public and private sector workers in a way that ensures an equal probability of inclusion. Such sampling could be done at the subnational level and targeted at those sections of the labor market where the government is particularly prevalent or is aiming to emphasize recruitment. Information not typically collected by household surveys but of substantial interest to those aiming to understand public sector labor markets could be collected, such as information on perceptions of the public recruitment process at different levels of government and how features of the public sector (such as perceived wage and pension benefits) affect respondents' wider labor market choices. Finally, sector variables, such as what specific parts of the government a respondent works in (or its private sector comparator), would allow for analyses that are more precisely targeted at particular job categories.

## INSIGHTS EMERGING FROM HOUSEHOLD SURVEYS

Such a systematic utilization of labor force data can allow for the delineation of unique stylized facts on public sector employment and compensation that can provide valuable insights for governments. This section provides illustrative examples of insights into the (relative) nature of government labor markets emerging from household surveys.

## The Size of the Public Sector in the Overall Labor Market

A foundation stone of government analytics is the size of the public sector as a share of the national, regional, or local labor market. This topic relates to questions about the appropriate size of government and its impacts on private sector labor markets.

The WWBI reveals that the public sector is a major source of employment in most countries; often, it is the single largest employer. More specifically, the public sector accounts for an average of 16 percent of total employment and over 30 percent and 37 percent of paid and formal employment, respectively. The first metric measures the overall labor market footprint of the public sector, while the latter two are more precise measures of the public sector's relative size within the salaried and formal segments of the labor market.

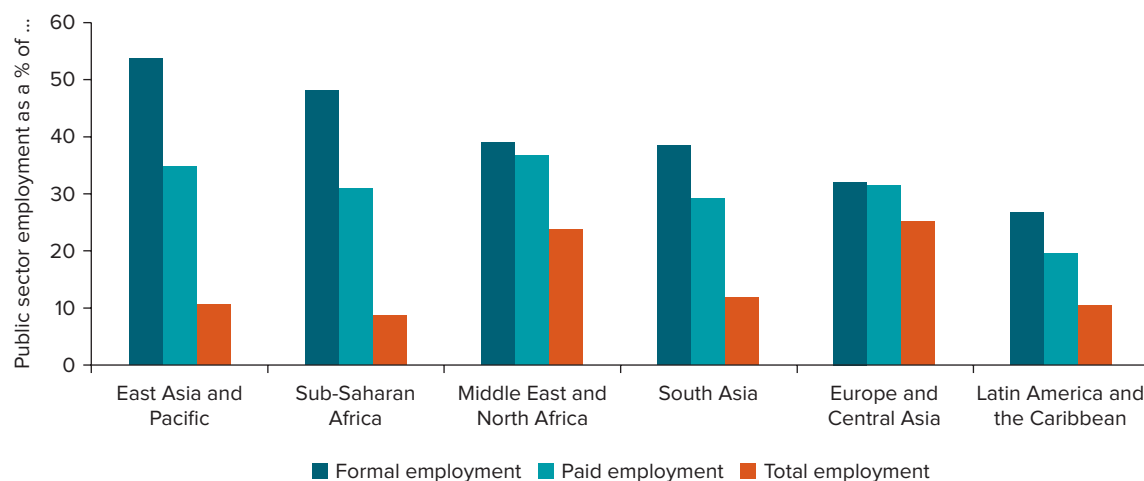
The size and importance of the public sector vary extensively by country income and region (see figure 27.1). While less than 9 percent of the total labor force of the average nation in Sub-Saharan Africa is employed in the public sector, the governments of the Middle East and North Africa employ a quarter of the entire labor force there. This difference is even more stark when looking at formal employment. Such comparisons can be made at the subnational level, allowing the government to develop a sense of how “imposing” its employment is as part of the total stock of formal jobs.

These basic statistics illustrate the wealth of information contained within household surveys that can help governments understand the importance of the public sector, not only as a provider of essential public services but as a key determinant of the health of labor markets, which can help practitioners make better-informed policy decisions.

Further, tracking these indexes over time can help governments understand how the share of the public sector has evolved over time. Figure 27.2 illustrates that, for the world as a whole, a convergence is taking place in terms of the relative size of the public sector. While the public sector's share within total employment has increased, public employment as a share of formal employment has steadily declined over the 18-year period studied.

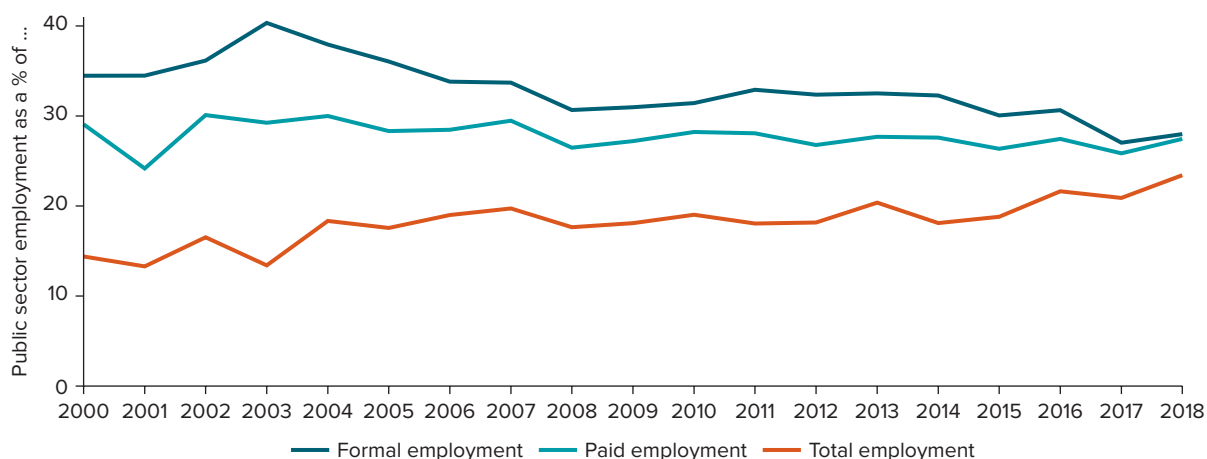
The former is likely due to the fact that as countries develop, their public sectors are called upon to provide more and increasingly complicated services. Conversely, the decline in the public sector's importance within the formal labor force is in part driven by the increased penetration of formal contracting and benefits within the private sector. Regional analysis shows that the relative importance of the public sector within formal employment fell faster and further in middle-income countries than in high- or low-income countries, both of which experienced relatively slower growth rates of labor force productivity and per capita income (Cho et al. 2012).

**FIGURE 27.1** Differences in Public Sector Employment, by Region, 2000–18



Source: Worldwide Bureaucracy Indicators, version 2.0, <https://datacatalog.worldbank.org/search/dataset/0038132>.

**FIGURE 27.2** Relative Size of the Public Sector Workforce, 2000–18



Source: Worldwide Bureaucracy Indicators, version 2.0, <https://datacatalog.worldbank.org/search/dataset/0038132>.

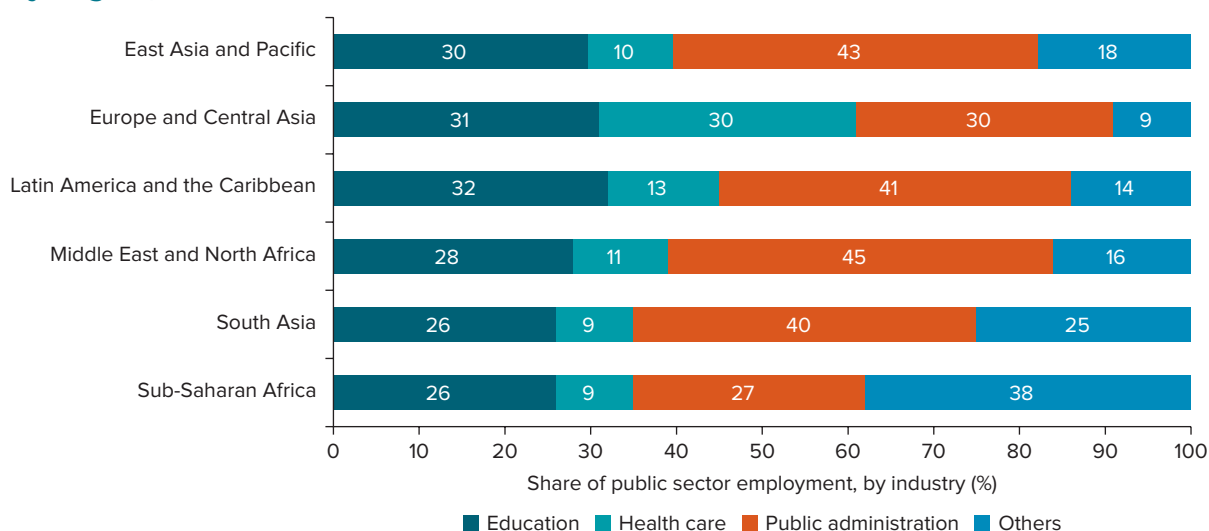
Moreover, household surveys allow policy practitioners to further disaggregate public employment by industry. Countries often have unique legal and occupational classifications for public sector employees, complicating cross-national comparisons. In some countries, all government employees are classified as civil servants, meaning they enjoy distinct legal protections. In others, only management and policy staff are categorized as civil servants, with others, particularly service delivery staff, enjoying fewer privileges and being governed by labor codes similar to private sector employees.

The WWBI reveals that the public administration workforce (which includes individuals responsible for the general administration of the government; the provision of defense, justice, police, and foreign affairs; and the management of compulsory social security) is the single largest segment of the public sector workforce in most countries. On average, 35 percent of the public sector workforce is employed in public administration, followed by the education and health care sectors, which employ, on average, 30 percent and 19 percent of the public sector workforce, respectively. Together, these three industries account for over 80 percent of all public sector employees (figure 27.3). The oversized nature of the health care sector within the Europe and Central Asia region is driven primarily by the extensive public health systems within countries in the European Union.

Additionally, the “other” category in figure 27.3 accounts for public sector employment in all remaining areas of economic activity, ranging from construction and infrastructure to the provision of public utilities, or workers employed within state-owned enterprises other than those involved in public administration, education, or health care provision. Here, countries within the Sub-Saharan Africa region are clear outliers, a phenomenon driven by large public sector penetration in the mining, manufacturing, and services sectors. Given the relatively lower levels of economic development in many countries within the region, this points to the important role that the public sector plays in countries with underdeveloped private sectors. Still, while there may not exist a universal formula for the ideal makeup of the public sector workforce, household surveys can allow a government to benchmark the organization of its public sector workforce across peer countries, or even historically, to track its evolution.

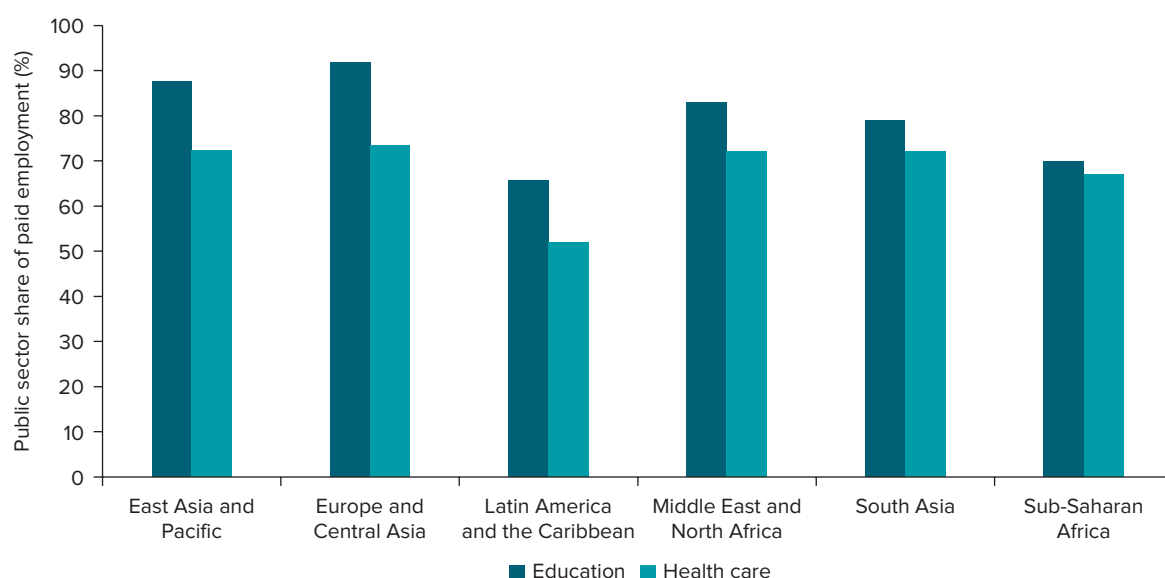
Education and health care workers are essential to a country’s ability to meet its Sustainable Development Goals (SDGs) for the adequacy and universality of health care coverage and education provision. The share of the public sector devoted to the provision of social services differs with country incomes. Looking closely at the education and health care workforce through labor force surveys helps explain the importance of the public sector in the provision of these services. Globally, over three-fourths and two-thirds of the education and health care paid workforce are employed in the public sector, respectively (figure 27.4). This is, in part, driven by the importance that governments across the world place on the provision of education and health care as mandated by the SDGs. It is also partly due to the limited capacity within the private sector to satiate national demand for these services.

**FIGURE 27.3 Areas of Economic Activity as a Segment of Public Employment, by Region, 2000–18**



Source: Worldwide Bureaucracy Indicators, version 2.0, <https://datacatalog.worldbank.org/search/dataset/0038132>.

**FIGURE 27.4 Public Sector Education and Health Care Employment, by Region, 2000–18**



Source: Worldwide Bureaucracy Indicators, version 2.0, <https://datacatalog.worldbank.org/search/dataset/0038132>.

Both these segments of the workforce have seen significant attention in the aftermath of the COVID-19 pandemic, as frontline education and health care providers, academics and researchers, epidemiologists, public health experts, and engineers have been an essential bulwark against the public health crisis. Their importance and contribution cannot be overstated. Household surveys can shed light on the role that the public sector education and health care workforce plays within these two sectors. The WWBI finds substantial variation by region (as illustrated in figure 27.4). While over 91 percent of the education workforce and 73 percent of the health care workforce in the Europe and Central Asia region is employed in the public sector, the Latin America and the Caribbean region employs just under 66 percent and 52 percent of these workers, respectively.



The public sector is an important employer for workers with tertiary degrees. Given the particular focus that household surveys place on collecting information systematically on the academic qualifications of workers in the labor force, using globally harmonized measures of educational attainment, these surveys offer a window into the skills makeup of the public and private sector workforces. Looking at data from the WWBI, which tracks the qualifications of workers internationally, we can see that the public sector has a higher proportion of workers with tertiary degrees. Of public sector workers, 47 percent have a tertiary degree, compared to 21 percent in the private sector. (Figure 27.5 provides a dot plot of countries in the WWBI comparing the national shares of tertiary-educated workers in the public and private sectors.) These differences between public and private sector workers have implications for any comparative analysis between the two labor markets, especially public-private wage differentials.

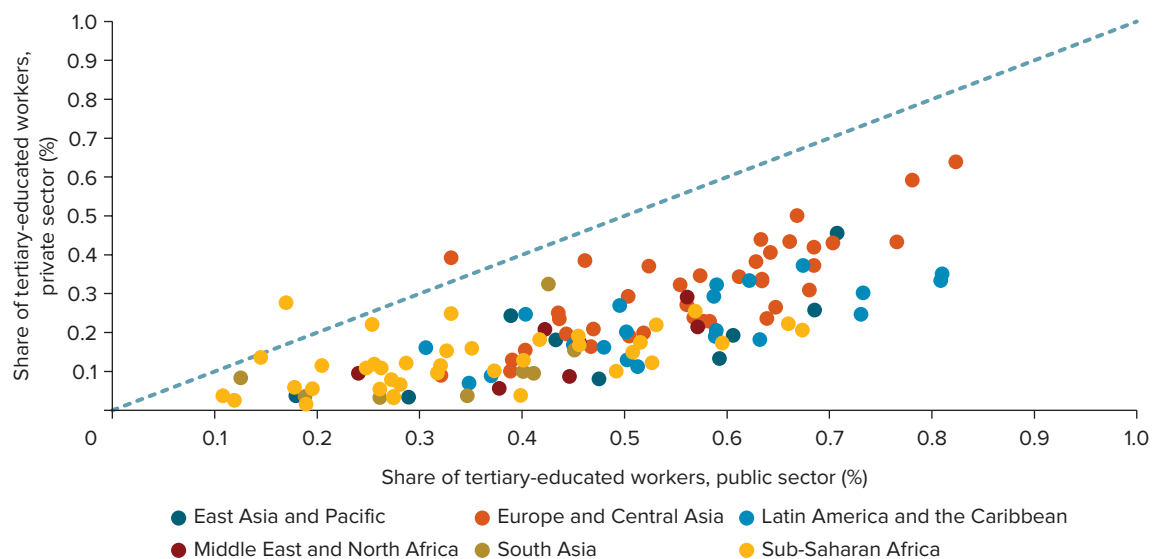
The proportion of public sector workers with tertiary education varies by country income level. In low-income countries, 19 percent of public workers have either no or only primary education, while in high-income countries, this share is negligible. A high proportion of low-skilled workers points to the public sector's serving a social welfare function. A corollary to a high proportion of low-skilled workers is a high proportion of clerical or support jobs. At the other end of the education spectrum, the share of employees with a tertiary degree has increased globally by around 20 percentage points in both sectors over the past decade, but the public sector continues to employ more workers with degrees.

By generating comparative information on the two sectors, possibly over time and across regions and countries, household surveys allow government analysts to understand the broad features of the public sector labor market and the role of the public sector in various national labor markets. A growing body of literature confirms this ability and the importance of the public sector in employing high-skilled workers (Gindling et al. 2020; Grindle and Hilderbrand 1995; Tummers and Knies 2013). Labor force surveys are thus well positioned to enable coherent international comparisons that provide benchmarks to assess a country's current state and dynamics.

## Understanding Gender Discrimination

The public sector is an important source of formal employment for women. The public sector's large labor market footprint means that it can be a strategic leader in changing norms and behaviors and promoting

**FIGURE 27.5 Tertiary Education among Public and Private Sector Workers, by Region, 2000–18**



Source: Worldwide Bureaucracy Indicators, version 2.0, <https://datacatalog.worldbank.org/search/dataset/0038132>.

Note: There is no estimate for the private sector in North America, so the region is not included in the graph.

greater equality in employment in the overall labor market. However, understanding the current state of women's participation in differential labor markets requires detailed information on the quality of gender representation in the public and private sectors.

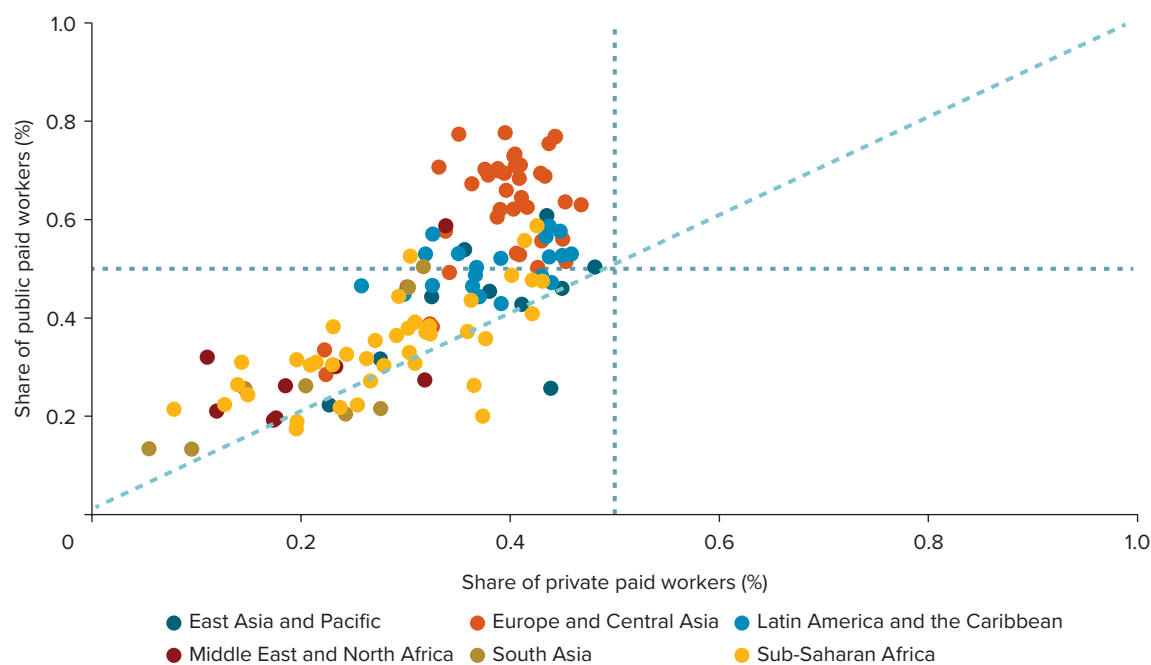
In many developing countries, the public sector, in general, and the education and health care sectors, in particular, have been among the few options for formal employment available to women (Yassin and Langot 2018). Globally, women represent 46 percent of the public sector workforce, compared to 33 percent of the private sector workforce. (Figure 27.6 provides a dot plot of countries in the WWBI comparing the national shares of public and private sector workers who are women.) While men outnumber women in the private sector in all 130 countries for which data are available, women outnumber men in the public sector in 55 countries.

Female representation in the public sector is strongly correlated with country income. A large body of literature finds a U-shaped relationship between female employment in the private sector and economic development (Goldin 1995; Goldin and Polachek 1987).<sup>8</sup> Labor force surveys included in the WWBI help provide evidence for a positive and significant relationship between female participation in the public workforce and country income. Multiple factors influence female participation rates in the labor force.<sup>9</sup> A growing body of literature confirms the positive relationship between more-representative bureaucracies (including through female participation) and improved social and economic outcomes across a wide spectrum, including reductions in gender-based violence (Johnston and Houston 2016), improvements in student performance (Zhang 2019), and improvements in public sector productivity (Andrews et al. 2005; Park 2013).

### The Appropriateness of Public Sector Wages

Public sector wages are an important determinant of personnel quality and motivation and, therefore, a key determinant of state capacity. However, what is the appropriate level and structure of these wages? Answering this question requires an assessment of who makes up the appropriate comparator group for public sector workers. The first option is to directly compare the wages of public and private sector workers

**FIGURE 27.6** Share of Female Workers in the Public versus Private Sector, by Region, 2000–18



Source: Worldwide Bureaucracy Indicators, version 2.0, <https://datacatalog.worldbank.org/search/dataset/0038132>.

Note: The diagonal dashed line represents equity between the two sectors. There is no estimate for the private sector in North America, so the region is not included in the graph.

within a particular country, given that the most likely outside option to employment in the public sector is the corresponding private sector. Estimating public-private wage differentials within a country has been explored in a very large body of academic and policy literature.<sup>10</sup>

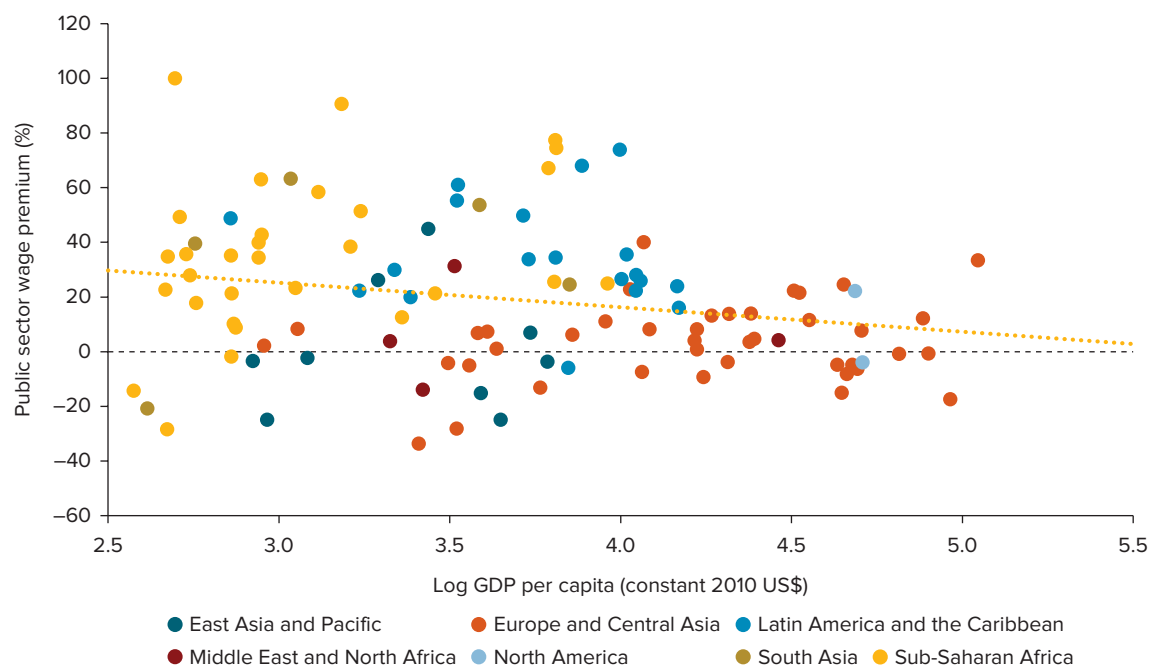
The second approach involves comparing the wages of public sector workers in one country with those of similar workers in other (comparable) countries. Given that these are the closest counterparts to one country's public sector workforce, this is another important method for estimating whether public servants in one country are over- or underpaid. These comparisons are particularly useful in the case of industries or occupations in which workers have transferable skills, such as health care workers who migrate internationally or workers in clerical or managerial positions who rotate within the public sector.

A third option is to compare individuals who perform different tasks or are employed in different occupations within the same country's public sector. This may be useful if public servants are able to move across the service from one organization or region to another.

Household surveys enable each of the above approaches, and such analysis has been undertaken in the WWBI. The data set indicates that public employees in most nations receive a wage premium compared to their counterparts in the private sector. Figure 27.7 shows the premium when the public sector is compared to all private sector salaried employees, irrespective of the type of job and controlling only for worker characteristics (including sex, age, level of education, and location). The figure is ordered by log GDP per capita to provide an indicative sense of premia vary with economic wealth. Public sector workers have approximately 19 percent higher basic wages (excluding allowances and bonus payments) across the 111 countries for which household surveys were sourced, with 80 countries having a positive premium. There is considerable heterogeneity in the size of that premium across countries, varying from a penalty of 33 percent to a premium of 100 percent. The size of the premium is negatively correlated with country income, a finding corroborating academic studies that report higher premiums for developing countries (Finan, Olken, and Pande 2017).

It is important for the government to understand how wage premiums are distributed across worker groups. The public sector wage premium is not uniform and varies by personnel characteristics.

**FIGURE 27.7 Public Sector Wage Premium Compared to Country Income, by Region, 2000–18**



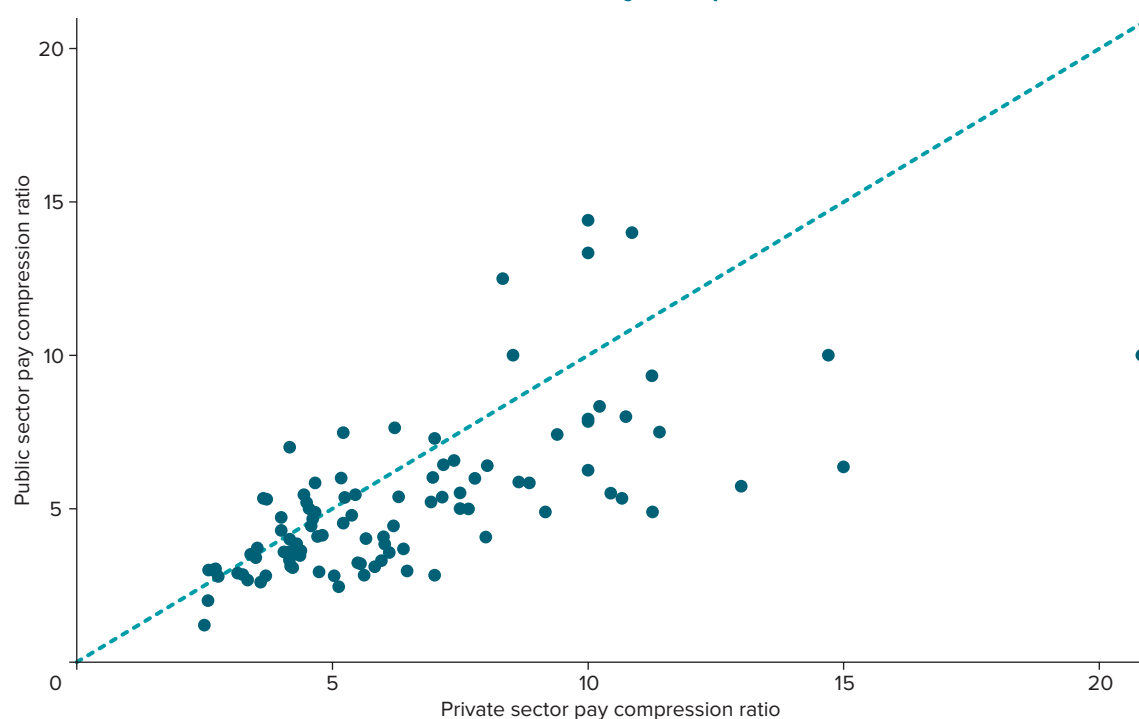
Source: Worldwide Bureaucracy Indicators, version 2.0, <https://datacatalog.worldbank.org/search/dataset/0038132>.

The magnitude of the public sector wage premium depends on an employee's educational qualifications and is lowest for tertiary-educated officials. The main reason that tertiary-educated individuals earn a low, or no, premium compared to private sector workers is the ability to earn greater wages in the private sector. Similarly, the large wage premium for women in the public sector has greater implications for the large gender pay gaps that exist in the private sector.

While the wage differentials between typical public and private sector workers presented above are worthy of attention from public officials in terms of their impact on the competitiveness of wages in the public sector, the public sector workforce represents a specific subset of the national labor force as employment. Public sector workers are concentrated within a handful of industries (public administration, education, and health care) and certain occupational groups (including managerial, professional, and clerical occupations). Therefore, a second, equally important element of the public sector wage structure for government officials is the difference in wages for workers in different segments of the public sector workforce. Studies have shown that workers compare their wages with their peers in an organization, just as they do with the private sector, and wage differentials that are not perceived as justifiable can be demotivating (Borjas 2002). Additionally, wage equity—whether staff in similar jobs, with similar skills and similar performance, are paid equally—impacts worker motivation and productivity and can be a major driver of the wage bill.

Wage dispersion is generally higher in the private sector than in the public sector. One common metric is the wage compression ratio, which is the ratio of the 90th percentile wage to the 10th percentile wage in the salary distribution. This ratio is lower in the public sector for 70 out of 99 countries for which there are data in the WWBI (figure 27.8). The average wage compression ratio for the public sector across 101 countries is 4.9, compared to 6.3 in the private sector. The lower dispersion in the public sector reveals a trade-off between equity and pay competitiveness at the top of the salary distribution that governments manage. Such information can help public sector managers determine new wage schedules aimed at attracting and maintaining a cadre of high-skilled functionaries in the public sector.

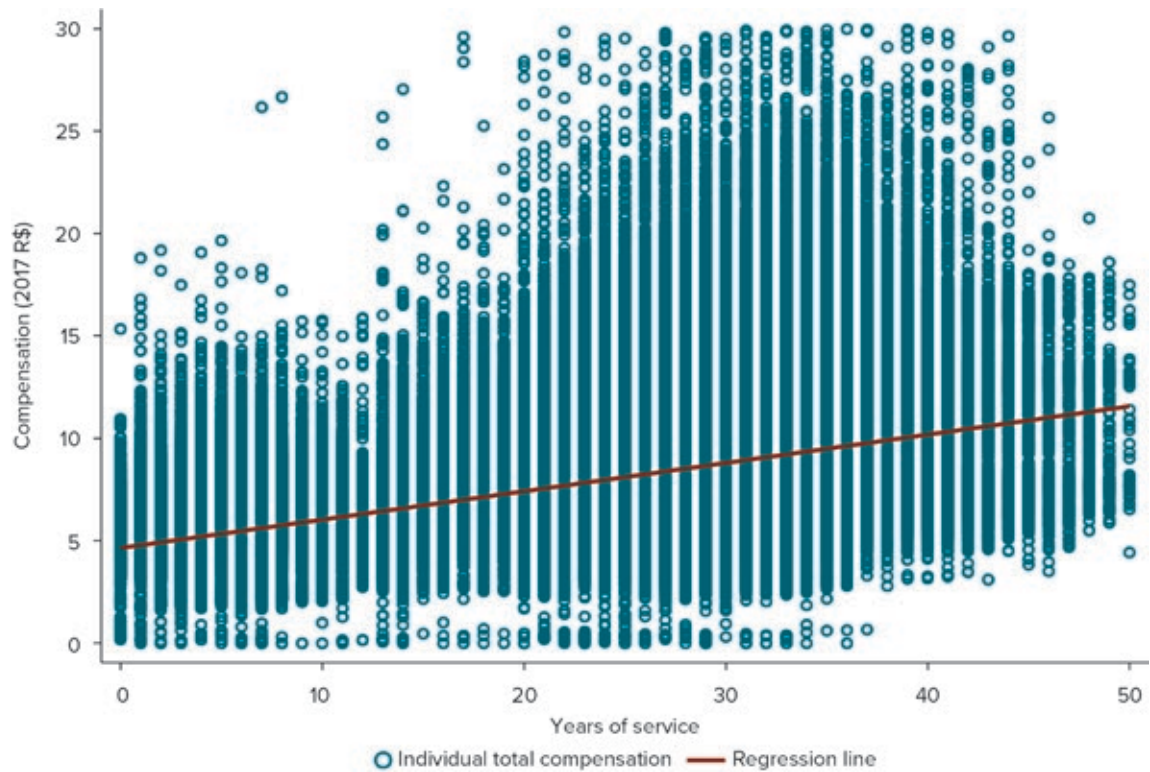
**FIGURE 27.8 Public versus Private Sector Pay Compression Ratios, 2000–18**



Source: Worldwide Bureaucracy Indicators, version 2.0, <https://datacatalog.worldbank.org/search/dataset/0038132>.

Note: Each dot represents a country.

**FIGURE 27.9** Pay Inequity in the Brazilian Public Sector, 2020



Source: World Bank 2020.

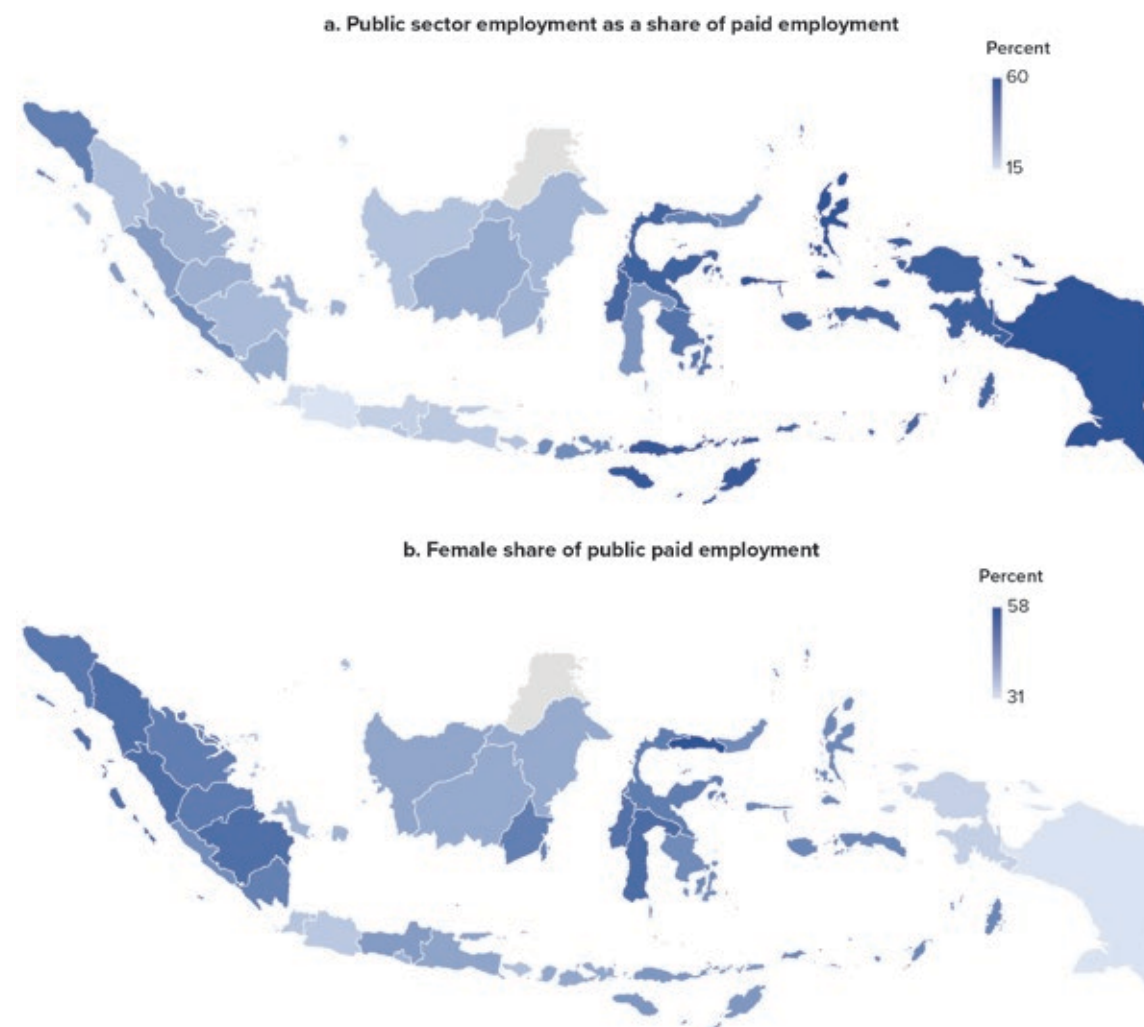
Note: Each dot represents an employee.

Household surveys are also able to provide information on the degree of unexplained variation in wages for individuals employed within similar occupations in the public sector. Figure 27.9 shows that the gross pay received can vary tenfold for workers with similar levels of experience, which is largely a result of non-performance-related payments and not basic pay. While these differences may, in part, be due to personnel demographics (age, gender, or educational qualifications) or the nature of work (job family, industry, or scale), this does point to public sector wages being weakly associated with the experience of workers. Still, these wage differences across employees performing similar tasks and of similar grades but working in different locations or organizations can potentially act as distortions in the workforce.

### Understanding Regional Variation

To further illustrate the power of household surveys, if sample sizes are sufficiently large and survey sampling is appropriately stratified, this analytical approach can be replicated at subnational levels. For example, the team has applied the methodology to analyzing public and private sector labor markets at the provincial level in Indonesia. Two variables from this effort are illustrated for Indonesia's provinces in map 27.1. These efforts allow for a closer understanding of regional disparities in the scale, composition, and compensation of the public and private sectors across administrative divisions within countries. In the case of Indonesia, for example, the public sector comprises almost 60 percent of paid employment in the eastern provinces of East Nusa Tenggara, North Maluku, West Papua, and Papua, compared to less than 15 percent in the western provinces (map 27.1, panel a). The female share of public sector employment (which stands at 44 percent at the national level) is mostly concentrated in the eastern and central provinces (map 27.1, panel b). Stylized facts like these can help shed light on many aspects of the nature of public and private labor markets across subnational units within a country.

**MAP 27.1 Subnational Patterns in Public Sector Employment, Indonesia, 2018**



Source: Original maps for this publication, based on household survey data.

## CONCLUSION

We have presented a microdata-based approach for governments to improve their understanding of the public sector workforce and labor markets. Such understanding helps in the development of empirically grounded public service compensation and employment strategies. We have demonstrated how government analysts can use existing household surveys to generate novel insights into government and how these lead to insights that can allow policy makers to make better fiscal choices. Thus, the range of data that should be included for consideration in human resources management information systems, outlined in chapter 9, includes household surveys. Capitalizing on household surveys for government analytics provides a powerful complement to payroll analysis (as described in chapter 10) and broader budget analytics.

These kinds of analytics matter for the effective management of the state, but they also matter for the impact of the public sector on private sector labor markets. Given the size of the public sector, public sector compensation should be designed in cognizance of its influence on the broader labor market. While public sector wage-setting mechanisms do not mechanically respond to market forces, they should be carefully designed to consider the distributional aspects of wages. Policy makers need to ensure that public sector



wages remain competitive enough to attract and retain high-quality public sector workers while not creating disequilibria in private sector labor markets through queuing and crowding effects. Under an optimal compensation policy, public sector wages will be competitive without being distortionary, and there will not be any shortage of skills in either sector.

We have used a series of examples from the World Bank's WWBI to demonstrate how the use of household survey data can help policy makers gain insight into the current and future state of their government's employment and compensation policies. This approach enables researchers, development practitioners, and policy makers to answer some of the most important questions about the appropriate level and distribution of employment in the public sector; the equity, transparency, and market competitiveness of public sector wages; and their impact on fiscal sustainability, the labor market, and service delivery.

## NOTES

The approach laid out here leverages the methodological and operational guidelines followed by the Bureaucracy Lab in the construction of the World Bank's Worldwide Bureaucracy Indicators (WWBI), a novel cross-national data set on public and private sector employment and compensation practices. The data set was derived from over 1,000 nationally representative household surveys from 202 countries and territories between 2000 and 2020, providing over 300 granular indicators on the composition, demographics, and compensation of public sector workers. However, this chapter goes beyond that effort to showcase how such an approach can be replicated by researchers, development practitioners, and policy makers to gain a better understanding of the personnel dimensions of state capability, the footprint of the public sector within the overall labor market, and the fiscal implications of the public sector wage bill.

1. There are notable exceptions, such as the Brazilian Ministry of Labor and Employment's *Relação Anual de Informações Sociais* (RAIS) data set, which contains information about employees and businesses for 97 percent of the Brazilian formal market.
2. To counteract these two concerns, governments can limit the presence of these biases in two ways. First, they can ensure that the selection of respondents is based on high-quality census data that guarantee that the sample selected is a good representation of the overall population of the country and, more importantly, is a realistic representation of labor force participants in the public and private sectors. Second, they can look to surveys that include tens of thousands (or an even higher number) of respondents to ensure that any potential weakness in sample selection is alleviated through a large sample size.
3. Further details on the construction of the WWBI are available in a technical note (World Bank 2022). The WWBI data set can be accessed online here: <https://datacatalog.worldbank.org/search/dataset/0038132>. WWBI data are displayed in a dashboard viewable at [https://databank.worldbank.org/source/worldwide-bureaucracy-indicators-\(wwbi\)](https://databank.worldbank.org/source/worldwide-bureaucracy-indicators-(wwbi)). The underlying analytical code has also been made available in the World Bank's repository on GitHub at <https://github.com/worldbank/Worldwide-Bureaucracy-Indicators>.
4. The thresholds used by the WWBI are a product of empirical investigation into the robustness of the indicators to different levels of missingness. More details are provided in the various technical reports accompanying distinct versions of the WWBI (see, for instance, World Bank 2022).
5. Certain surveys do include information on work benefits, such as health insurance and social security, but these are not monetized and cannot be added to wages to provide an estimate of total compensation.
6. *Winsorizing* or *winsorization* is the transformation of statistics by limiting extreme values in the statistical data to reduce the effect of possibly spurious outliers.
7. More information about the CMIE is available on its website, accessible at <https://consumerpyramidsdx.cmie.com/>.
8. Female participation is the highest in low-income countries, falling as countries industrialize, and increasing again at high levels of economic development, as the services sector grows.
9. See Jayachandran (2021) for a rich discussion of this literature.
10. See Bales and Rama (2001), Bargain, Etienne, and Melly (2018), Campos et al. (2017), Christophides and Michael (2013), Coppola and Calvo-Gonzalez (2014), Gibson (2009), and Lausev (2014).

## REFERENCES

Andrews, Rhys, George A. Boyne, Kenneth J. Meier, Laurence J. O'Toole, and Richard M. Walker. 2005. "Representative Bureaucracy, Organizational Strategy, and Public Service Performance: An Empirical Analysis of English Local Government." *Journal of Public Administration Research and Theory* 15 (4): 489–504.

- Arizti, Pedro, Daniel J. Boyce, Natalia Manuilova, Carlos Sabatino, Roby Senderowitsch, and Eral Vilá. 2020. *Building Effective, Accountable, and Inclusive Institutions in Europe and Central Asia: Lessons from the Region*. Washington, DC: World Bank. <http://hdl.handle.net/10986/34031>.
- Baig, Faisal Ali, Xu Han, Zahid Hasnain, and Daniel Rogger. 2021. "Introducing the Worldwide Bureaucracy Indicators: A New Global Dataset on Public Sector Employment and Compensation." *Public Administrative Review* 81 (3): 564–71.
- Bales, Sarah, and Martin Rama. 2001. "Are Public Sector Workers Underpaid? Appropriate Comparators in a Developing Country." Policy Research Working Paper 2747, World Bank, Washington, DC. <http://hdl.handle.net/10986/19338>.
- Bargain, Olivier, Audrey Etienne, and Blaise Melly. 2018. "Public Sector Wage Gaps over the Long-Run: Evidence from Panel Administrative Data." IZA Discussion Paper 11924, Institute of Labor Economics, Bonn, Germany. <https://www.iza.org/publications/dp/11924/public-sector-wage-gaps-over-the-long-run-evidence-from-panel-administrative-data>.
- Behar, Alberto, and Junghwan Mok. 2013. "Does Public-Sector Employment Fully Crowd Out Private-Sector Employment?" IMF Working Paper WP/13/146, International Monetary Fund, Washington, DC.
- Borjas, George J. 2002. "The Wage Structure and the Sorting of Workers into the Public Sector." NBER Working Paper 9313, National Bureau of Economic Research, Cambridge, MA.
- Campos, Maria M., Domenico Depalo, Evangelia Papapetrou, Javier J. Pérez, and Roberto Ramos. 2017. "Understanding the Public Sector Pay Gap." *IZA Journal of Labor Policy* 6 (7). <https://doi.org/10.1186/s40173-017-0086-0>.
- Cho, Yoonyoung, David N. Margolis, David Newhouse, and David A. Robalino. 2012. "Labor Markets in Low and Middle-Income Countries: Trends and Implications for Social Protection and Labor Policies." Social Protection and Labor Discussion Paper 1207, World Bank, Washington, DC.
- Christophides, Louis N., and Maria Michael. 2013. "Exploring the Public-Private Sector Wage Gap in European Countries." *IZA Journal of European Labor Studies* 2 (15). <https://doi.org/10.1186/2193-9012-2-15>.
- Coppola, Andrea, and Oscar Calvo-Gonzalez. 2014. "Higher Wages, Lower Pay: Public vs. Private Sector Compensation in Peru." Policy Research Working Paper 5858, World Bank, Washington, DC. <https://doi.org/10.1596/1813-9450-5858>.
- Finan, F., B. Olken, and R. Pande. 2017. "The Personnel Economics of the Developing State." In *Handbook of Economic Field Experiments*, vol. 2, edited by Abhijit Vinayak Banerjee and Esther Duflo, 467–514. <https://doi.org/10.1016/bs.hefe.2016.08.001>.
- Gibson, John. 2009. "The Public Sector Pay Premium, Compensating Differentials and Unions: Propensity Score Matching Evidence from Australia, Canada, Great Britain and the United States." *Economics Bulletin* 29 (3): 2325–32.
- Gifford, Brian. 2006. "The Camouflaged Safety Net: The U.S. Armed Forces as Welfare State Institution." *Social Politics: International Studies in Gender, State and Society* 13 (3): 372–99.
- Gindling, T. H., Zahid Hasnain, David Newhouse, and Rong Shi. 2020. "Are Public Sector Workers in Developing Countries Overpaid? Evidence from a New Global Dataset." *World Development* 126: 104737. <https://doi.org/10.1016/j.worlddev.2019.104737>.
- Goldin, Claudia. 1995. "The U-Shaped Female Labor Force Function in Economic Development and Economic History." In *Investment in Women's Human Capital*, edited by T. P. Schultz, 61–90. Chicago: University of Chicago Press.
- Goldin, Claudia, and S. Polachek. 1987. "Residual Differences by Sex: Perspectives on the Gender Gap in Earnings." *American Economic Review* 77 (2): 143–51.
- Grindle, Merilee S., and Mary E. Hilderbrand. 1995. "Building Sustainable Capacity in the Public Sector: What Can Be Done?" *Public Administration and Development* 15 (5): 441–63.
- Hasnain, Zahid, Daniel Rogger, John Walker, Kerenssa Mayo Kay, and Rong Shi. 2019. *Innovating Bureaucracy for a More Capable Government*. Washington, DC: World Bank.
- ILO (International Labour Organization). 2013. *Resolution Concerning Statistics of Work, Employment and Labour Underutilization*. Adopted by the 19th International Conference of Labour Statisticians, Geneva, October 2–11, 2013. [https://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/normativeinstrument/wcms\\_230304.pdf](https://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/normativeinstrument/wcms_230304.pdf).
- IMF (International Monetary Fund). 2014. *Government Finance Statistics Manual 2014*. Washington, DC: International Monetary Fund. <https://www.imf.org/external/Pubs/FT/GFS/Manual/2014/gfsfinal.pdf>.
- Ingraham, Patricia W., Philip G. Joyce, and Amy Kneeder Donahue. 2003. *Government Performance: Why Management Matters*. Baltimore, MD: Johns Hopkins University Press.
- Jayachandran, Seema. 2021. "Social Norms as a Barrier to Women's Employment in Developing Countries." *IMF Economic Review* 69 (3): 576–95.
- Johnston, Kate, and John Houston. 2016. "Representative Bureaucracy: Does Female Police Leadership Affect Gender-Based Violence Arrests?" *International Review of Administrative Sciences* 84 (1): 3–20.
- Lausev, Jelena. 2014. "What Has 20 Years of Public-Private Pay Gap Literature Told Us? Eastern European Transitioning vs. Developed Economies." *Journal of Economic Surveys* 8 (3): 516–50.
- Moynihan, Donald, and Ivor Beazley. 2016. *Toward Next-Generation Performance Budgeting: Lessons from the Experiences of Seven Reforming Countries*. Washington, DC: World Bank.

- Park, Sanghee. 2013. "Does Gender Matter? The Effect of Gender Representation of Public Bureaucracy on Governmental Performance." *American Review of Public Administration* 43 (2): 221–42.
- Rasul, Imran, and Daniel Rogger. 2018. "Management of Bureaucrats and Public Service Delivery: Evidence from the Nigerian Civil Service." *The Economic Journal* 128 (608): 413–46.
- Tummers, Lars G., and Eva Knies. 2013. "Leadership and Meaningful Work in the Public Sector." *Public Administration Review* 73 (6): 859–68.
- World Bank. 2020. "Brazil: Public Sector Wage Bill and Personnel Management." Unpublished manuscript. Washington, DC: World Bank.
- World Bank. 2021. *Worldwide Bureaucracy Indicators Version 2.0: Codebook and Explanatory Note*. Last updated May 26, 2021. Washington, DC: World Bank. <https://datacatalog.worldbank.org/int/search/dataset/0038132/Worldwide%20Bureaucracy%20Indicators?version=2>.
- Yassin, Shaimaa, and François Langot. 2018. "Informality, Public Employment and Employment Protection in Developing Countries." *Journal of Comparative Economics* 46 (1): 326–48.
- Zhang, Youlang. 2019. "Representative Bureaucracy, Gender Congruence, and Student Performance in China." *International Public Management Journal* 22 (2): 321–42.

## CHAPTER 28

# Government Analytics Using Citizen Surveys

## Lessons from the OECD Trust Survey

*Monica Brezzi and Santiago González*

### SUMMARY

This chapter takes stock of the work of the Organisation for Economic Co-operation and Development (OECD) on defining and measuring what drives people's trust in public institutions. It presents an updated framework on the public governance determinants of trust, demonstrating that competence (for example, responsiveness and reliability) and values (for example, openness, integrity, and fairness) are fundamental public governance levers to improve or harm levels of trust in different institutions, while also recognizing the importance of people's perception of how intergenerational and global challenges are handled by institutions, as well as cultural, economic, and political drivers. In addition, it shows that it is possible, on the basis of this framework, to advance in gathering citizens' evaluations of the functioning of governments and public governance by designing nationally representative population surveys with high standards of statistical quality. The OECD Survey on Drivers of Trust in Public Institutions (OECD Trust Survey) has been specifically designed and tested to capture people's expectations of and experiences with public institutions around the five drivers of trust while controlling for socioeconomic, political, and institutional characteristics. In a handful of countries that, in addition, have conducted in-depth case studies, the evidence resulting from the survey has proven to be a key input for improving policy making, leading to concrete actions for building or strengthening institutional trust. The inaugural cross-country survey was applied to 22 OECD countries at the end of 2021, and the results were made available in June 2022. Consolidating this evidence will be essential to enhancing benchmarking and monitoring of the evolution of policies over time.

## ANALYTICS IN PRACTICE

- It is possible to measure trust and its drivers through household surveys and to produce actionable evidence. Trust indicators have been widely collected by unofficial data producers and are commonly picked up by the media as metrics of government performance. However, indicators on the public governance drivers of trust have been scarcely and inconsistently produced and are not commonly found in population surveys. Over the past years, a body of theoretical and empirical evidence has consolidated, leading to the development of the Organisation for Economic Co-operation and Development Survey on Drivers of Trust in Public Institutions (OECD Trust Survey) as an effort to fill this gap and develop comparative statistics on what drives trust in public institutions. The survey's unique focus on the drivers of trust aims to provide countries with actionable evidence on the strengths and weaknesses of governments and public institutions.
- Bringing together the statistical and policy communities is a key step for ensuring the production of statistics with high standards of quality. Other than their potential sensitivity, there are no good reasons why statistics on trust and its drivers cannot be subject to the same quality standards and requirements that apply to other social, economic, and environmental statistics. The development of the OECD Trust Survey has been led by the OECD Public Governance Committee and has brought together policy makers at the central level of government and representatives from national statistical offices (NSOs), who have reviewed the process for developing and implementing the survey. Bringing together both communities has the advantage of reconciling the development of meaningful statistics with adherence to high requirements and standards of quality. While in most countries, the OECD Trust Survey has been implemented by the OECD via a survey provider, in Finland, Ireland, Mexico, and the United Kingdom, it has been implemented by NSOs.
- The OECD competence-values framework has proven to be a powerful analytical tool to understand the public governance drivers of trust in institutions and provide evidence to public administrations on how to increase their effectiveness to enhance trust. Following the COVID-19 pandemic and the resulting social and economic crisis, a consultative process to revisit the framework on the drivers of trust in public institutions was launched. The process resulted in the enlargement of the framework by adding a section on the perception of public institutions' action on intergenerational and global challenges and further recognition of individual cultural, economic, and political drivers. However, the building blocks of the framework—competence and values—have proven robust, offering a compelling and encompassing framework for understanding and measuring public administrations' work and how they could advance in building trust.
- The drivers of trust vary by institutional actor and level, calling for different types of actions to influence them. The debate on public trust has been dominated by measures of *trust in government* that, to a large extent, are proxies of political trust or trust in the current government. Yet distinctions by the level of government (for example, local government) or between “politicians” and “bureaucrats” have not been rigorously considered. The OECD Trust Survey considers differences across institutions and levels of government and allows for a differentiated analysis of what drives each of them.
- Institutional trust remains a multidimensional construct: drivers and actions to preserve it or restore it are also influenced by context. In addition to the competence and values of public institutions, trust is also influenced by cultural and socioeconomic determinants. Accordingly, baseline levels of trust in a given country are influenced by its history and the current moment. Because cross-country comparisons of trust levels are informative and appealing, contextual elements should be considered when interpreting data and carrying out comparisons. In-depth analysis combining quantitative and qualitative methods at the country level is an important complement to advancing actions for restoring trust.

- Evidence-based regional dialogue and experience-sharing could be powerful tools for strengthening institutional trust. The potential effect of cultural context can be minimized by comparing countries with similar histories, institutional settings, and contexts. Accordingly, evidence generated by the OECD Trust Survey could be a powerful driver of regional dialogue and experience-sharing about actions that could influence public trust in similar contexts.

## INTRODUCTION

The availability of internationally comparable data on how public administrations function and perform has dramatically improved over the past 10 years (OECD 2021b). However, measures of governance outcomes, which can help benchmark the effectiveness of public governance, are still scarce. The reasons behind this are conceptual and methodological in nature. On the conceptual side, measuring governance outcomes implies consideration of democratic standards that go beyond a performance-oriented notion of public management. Methodologically, public surveys of an international scope raise questions regarding the validity and robustness of these measures.

Measurement of governance outcomes includes, among other things, assessing levels of satisfaction with public services or *trust in government*. The inclusion in the United Nations' 2030 Agenda for Sustainable Development of a *governance* Sustainable Development Goal (SDG16) has given new impetus to the development of metrics for the outcomes or results of political institutions and processes and to the expansion of household surveys to measure governance results (United Nations Praia Group 2020). In particular, the United Nations Praia Group on Governance Statistics' *Handbook on Governance Statistics* takes stock of existing practices in governance data collection along eight dimensions of public governance, discusses the steps required to achieve international statistical standards, and, in some cases, proposes common questionnaires to be included in household surveys.

This chapter discusses how population surveys can be used to measure governance outcomes and provide guidance to public administrations on actions to increase their effectiveness. After some examples of existing national citizen surveys, the chapter presents recent developments in measuring people's trust in government and its main determinants. Measures of people's trust in government are commonly used indicators of public administration performance (or results) and are included as one of the eight dimensions of the *Handbook on Governance Statistics*.

This chapter presents results from the application of the Organisation for Economic Co-operation and Development Survey on Drivers of Trust in Public Institutions (OECD Trust Survey) as a key example to better understand opportunities and challenges in using citizen surveys to develop evidence on governance outcomes at a national and international scale. The OECD Trust Survey builds on an analytical framework structured along five key dimensions: responsiveness, reliability, openness, integrity, and fairness. Preliminary results suggest important insights from the survey. We find that different public institutions, from national to local governments, map onto different drivers of public institutional trust. Survey findings can guide governments in prioritizing specific areas to strengthen institutional trust.<sup>1</sup>

## USING CITIZEN SURVEYS TO MEASURE GOVERNANCE OUTCOMES

Governance statistics are fundamental to ensuring that the relationship between the state and its people is inclusive, transparent, and accountable. However, while public governance issues figure prominently in national and international policy discussions, the state of available statistics (in terms of international standards, measurement guidelines, and the availability of comparable information from official sources)



lags well behind the quality of information available in other fields (for example, economic, social, and environmental statistics). As detailed in chapter 27 of *The Government Analytics Handbook*, the World Bank's Worldwide Bureaucracy Indicators are a major exception, aggregating household survey data to measure key public administration productivity inputs (such as public sector wage premiums).

Traditionally, there are two main limitations of public governance statistics. First, most measures refer to the input and processes of governments, while more limited evidence is available on the outputs of governments and the outcomes of public governance, such as the assessment of and satisfaction with public services or trust in government. Second, most governance measures are generated through expert assessments aimed at capturing how governments work. These assessments

typically refer to professionals such as academics, lawyers or civil servants who answer a questionnaire in the area of interest. The resort to experts is thought to be advantageous because they can assess complex topics related to governance through an informed judgment. Expert assessments are frequently combined to create composite governance indexes that summarize multi-concept and complex phenomena into rankings and reference data. There are, however, reasonable concerns on how representative a sample of experts is of the universe of people with knowledge in the matter of interest, as well as about the degree of validity and reliability of data obtained through expert assessments. (United Nations Praia Group 2020, 15; see also González, Fleischer, and Mira d'Ercole 2017; Kaufmann and Kraay 2007)

Broadening the measurement approach to also include citizen surveys can help address both limitations, by shifting the focus to the outcomes of public governance as perceived and experienced by people and by including people's views, through nationally representative population samples, in addition to expert assessments (Fukuyama and Recanatini 2018).

Citizen surveys can also strengthen government accountability. When participating in regular population surveys, citizens are invited to provide feedback on different aspects of public governance, which allows governments to gather input and hear the people's voice beyond electoral processes. Governments and policy makers can also use survey results to better inform policies, identify citizens' priorities and concerns, and assess the support for or impact of different initiatives.

Over the past few years, some developments at international levels, which build on national experiences, have helped to increase the availability of governance statistics. The inclusion of a *governance* goal in the United Nations' 2030 Agenda for Sustainable Development, as mentioned above, and the agreement of the Inter-agency and Expert Group on SDG Indicators (IAEG-SDGs) on several indicators for global reporting in this field (which creates reporting obligations for national statistical offices, or NSOs) are expected to increase the demand for comparable, high-quality evidence on public governance and to broaden measurement approaches. For over a decade, in its flagship report *Government at a Glance*, the OECD has published internationally comparable data on governments' inputs, activities, outputs, and, to the extent possible, outcomes in OECD countries.

The indicators feeding these publications focus on how governments perform from an international perspective, allowing countries to benchmark their performance relative to other countries and over time and providing evidence to policy makers about areas where further progress is needed. Furthermore, initiatives such as the OECD Guidelines on Measuring Trust (OECD 2017a, 2017b) and the Praia Group's *Handbook on Governance Statistics* (United Nations Praia Group 2020) have provided methodological guidance to NSOs on how to develop comparable measures for several dimensions of public governance.

Specifically, the Praia Group, set up in 2014, has mapped in its *Handbook on Governance Statistics* existing measurement initiatives on eight dimensions of public governance: nondiscrimination and equality, participation, openness, access to and quality of justice, responsiveness (satisfaction with services and political efficacy), absence of corruption, trust, and safety and security. The *Handbook on Governance Statistics* highlights the paucity of statistical standards, technical guidelines, and methodological tools that currently exist in many dimensions of governance statistics (for example, on

discrimination, participation in political and public affairs, openness, access to civil justice, satisfaction with services, and forms of corruption other than petty bribery); at the same time, it sheds light on the feasibility of generating comparative evidence on some dimensions of public governance through population surveys, notably on crime victimization, access to criminal justice, political efficacy, and trust in public institutions (United Nations Praia Group 2020).

The *Handbook on Governance Statistics* also highlights some methodological challenges of household surveys—in particular, respondent burden and accessibility standards for specific populations (for example, people with disabilities)—that should be addressed in questionnaire design as well as survey method selection. The authors note that

with a sufficient sample size, survey results may be representative of the population of a country, of a specific area (such as a province or a small locality) or a specific group (e.g., urban/rural populations). However, such disaggregations for governance statistics are not commonly found in national statistical systems and will require special attention from the statistical community (United Nations Praia Group 2020, 14).

Nevertheless, they acknowledge that “household surveys are particularly useful to give voice to marginalized populations in contexts in which the mechanisms to respond to citizens’ demands are not yet consolidated” (United Nations Praia Group 2020, 14; see also AU and UNDP 2017).

One public governance area in which population surveys would provide relevant information to governments is experience and satisfaction with public services. More countries are putting in place regular surveys on public services, although with several challenges. For example, since 2010, the Agency for Public Management and eGovernment in Norway has carried out the Norwegian Citizen Survey. The survey provides a substantial knowledge base for assessing the performance of public services across different sectors and levels of government. The survey is understood as an additional way to engage citizens apart from direct mechanisms, and it addresses perceptions, expectations, and areas of improvement, aiming to develop public outputs and services in a more user-friendly manner, based on citizens’ needs and expectations.

However, the evolution of the survey over time has shown that respondents may find it difficult to answer “customer satisfaction” questions on a large number of services. Since 2018, the survey has included only the largest state-provided services (tax administration, hospitals, police, etc.) and services at the municipal level to which people have regular access (for example, schools). The survey also asks about recent experiences with these services as one criterion that could shape satisfaction levels. The questions address the quality and accessibility of services; satisfaction with information, communication, and consumer orientation; perceived competence, capacity, and trust in the public authority; and general satisfaction (OECD 2022b).

Similarly, the Citizen Experience Survey, conducted since 2019 by the Department of the Prime Minister and Cabinet in Australia, measures public satisfaction, trust, and experiences with Australian public services, with the aim of improving public services and making them more citizen-centered. The survey aims to provide the right evidence for governmental agencies to act on: when data flag “hot spots” in the system that warrant extra attention, senior policy makers and other stakeholders engage in dialogue to identify actions that can help improve services.<sup>2</sup>

Surveys of the general population can also provide useful data about citizens’ perceptions of the quality of governance in the country where they live, an outcome that is usually measured through expert assessments. The General Population Poll (GPP) by the World Justice Project (WJP) asks citizens about their perceptions of and experience with public institutions.<sup>3</sup> The survey inquires about a wide range of topics, from access to public services (including access to information and justice), respect for the law by private and public actors (for example, abuses of power), and civic participation. Data coming from the GPP are included in the WJP Rule of Law Index, which ranks countries according to the quality of their governance, alongside the results of an expert assessment.

## WHY MEASURE PUBLIC TRUST?

People's trust in public institutions has long been considered a key outcome of good governance (OECD 2021b). Accordingly, institutional trust happens when citizens consider the government and its institutions in general, as well as individual political leaders, to be promise-keeping, efficient, fair, and honest (Blind 2007). In this chapter, *trust* is defined as a person's belief that another person or institution will act consistently with expectations of positive behavior (OECD 2017a, 2017b). While there are several trust relationships, this chapter focuses on the interaction between governments and citizens, or *institutional trust*, which refers to people's appraisal of public institutions.

Institutional trust is recognized as an important foundation upon which the legitimacy and sustainability of political systems are built. Trust is the basis of the social contract that allows for the delegation of power and sets the basis for democracies to work. In turn, trust is essential for social cohesion and well-being because it affects a government's ability to govern and enables it to act without having to resort to coercion. The COVID-19 pandemic has underlined the importance of public trust for enhancing and accepting behavioral change in order to achieve a collective objective, as well as its importance for achieving compliance (Bargain and Aminjonov 2020). Recent research has shown that support for future-oriented policies on global challenges, such as climate change, is mediated by people's institutional trust (Fairbrother et al. 2021). In a low-trust context, citizens will prioritize immediate, appropriable, and partial benefits and will induce politicians to seek short-term and opportunistic gains through free-riding and populist attitudes (Gyórfy 2013).

Still, institutional trust remains a complex, multidimensional concept influenced by a wide array of facts, circumstances, experiences, and perceptions (OECD 2013). Trust metrics are often quoted by the press and have the capacity to raise awareness and trigger institutional reactions; however, they remain poorly understood. Furthermore, it is not always clear what lies behind these signals and to what extent it is possible to influence them. The OECD's work has focused on conceptualizing the main determinants of institutional trust and developing comparative evidence on them based on citizen surveys.

The remainder of this chapter explores existing metrics of institutional trust, presents the OECD policy framework and its accompanying measurement strategy for the public governance determinants of institutional trust, and concludes with lessons that could inform a measurement agenda moving forward.

## EXISTING METRICS FOR INSTITUTIONAL TRUST

In some contexts, there is a long tradition of collecting metrics of trust in government. For instance, in the United States, the Pew Research Center has measured "government confidence" since at least 1958. Similarly, the American National Election Study (ANES), a project of the Center for Political Studies at the University of Michigan, has collected survey-based measures of trust, associated with electoral cycles, since at least 1952. However, it is only since the beginning of the 21st century that cross-country comparative statistics of institutional trust have become widely and regularly available. González and Smith (2017) review these metrics and find seven cross-country comparative surveys (commercial and noncommercial) that have regularly collected trust data since 2002. These surveys have different coverage periodicity and work under different criteria of statistical quality. The most widely used of these surveys is the Gallup World Poll because of its extensive country coverage, its time extension, and the annual frequency of its data.

However, existing measures of institutional trust have shortcomings. Some are technical (such as sampling, scale, and level of representativeness for some population groups); others are conceptual (such as the meaning of *government*). In this chapter, we use *institutional trust* to indicate trust in different types of institutions, such as trust in political institutions (for example, the parliament), trust in administrative

institutions (for example, the civil service or public administration) and nongovernmental institutions (for example, the media), and trust in justice and law administration (for example, the police). Accordingly, institutional trust is measured using a general formulation: “Do you have confidence in ...?” or “How much do you trust ...?” followed by a detailed list of institutions. Empirical analysis suggests that, given the wide range of institutions, people’s responses can be grouped into three categories: political and administrative institutions, law-and-order institutions, and nongovernmental institutions (González and Smith 2017).

Based on existing evidence, the OECD has developed the Guidelines on Measuring Trust (OECD 2017a, 2017b). The guidelines mark an advance in providing an analysis of the accuracy of trust measures. Accuracy has two dimensions: *reliability* and *validity*. The *reliability* of a metric is the degree to which repeated measurements of the same thing produce the same results. In this sense, a reliable measure involves minimal *noise*, or random errors in the measurement process. *Validity* is usually analyzed in terms of *face validity* (whether the measure makes sense intuitively), *convergent validity* (whether the measure correlates well with other proxy measures of the same concept), and *construct validity* (whether the measure behaves as theory and common sense dictate).<sup>4</sup> The guidelines have found good evidence on the reliability and construct validity of existing trust metrics, but evidence on their face and convergent validity is scarce and inconclusive, calling for further research in the area. The guidelines also propose modules for measuring institutional trust, including an experimental module on the determinants of public trust that will be discussed later in this chapter.

## MEASURING WHAT DRIVES PUBLIC TRUST IN INSTITUTIONS

### The OECD Framework on Drivers of Trust in Public Institutions

At least three trends emerge from the academic literature for understanding what drives levels of trust in institutions. One theory emphasizes the role of culture and argues that individuals learn to trust or distrust based on early socialization and interpersonal networks, which, in turn, influence their trust in institutions (Tabellini 2008). A second stream of work recognizes the importance of the economic cycle, as well as economic and personal characteristics and preferences (Algan et al. 2018, 2019). Finally, institutional theories focus on the performance and reputation of institutions, both in terms of processes and outcomes, as the key determinants explaining levels of institutional trust (Bouckaert 2012; Rothstein 2013; Van de Walle and Migchelbrink 2020). While institutional trust is probably influenced by a combination of elements driven by culture, economic conditions, and institutions, the OECD’s work on understanding drivers of trust in public institutions has, since 2013, emphasized the importance of high-performing institutions for building public trust.

Understanding the effects of institutions on trust depends on the congruence of people’s preferences (their interpretations of what is right and fair and what is unfair) and their perceptions of the actual functioning of government (Van de Walle and Bouckaert 2003). Other authors have distinguished between *trust in competence*, the ability to deliver on expectations, and *trust in intentions*, performing in good faith according to one’s competence (Nooteboom 2006). These distinctions are extended by Choi and Kim (2012) and Bouckaert (2012), who distinguish between the *logic of consequences*, where trust is derived causally from outcomes, and the *logic of appropriateness*, where trust is based on values, such as integrity and transparency.

Despite the complexity of the subject and the variety of approaches, there is consistency across the literature on institutional trust in at least two key aspects. First, the literature highlights two different but complementary components that matter in understanding and analyzing trust: *competence*—operational efficiency, or the ability, capacity, and good judgment to actually deliver on a given mandate—and *values*—the underlying intentions and principles that guide actions and behaviors. Second, there is consistency in the literature regarding specific attributes that matter for trust, in relation to both competence and values (see table 28.1).

**TABLE 28.1 Deconstructing Citizens' Trust in Public Institutions**

Trust component	Government mandate	Concerns affecting trust	Policy dimension
<b>Competence:</b> <i>The ability of governments to deliver to citizens the services they need, at the quality level they expect</i>	Provide public services	<ul style="list-style-type: none"> <li>Access to public services, regardless of socioeconomic condition</li> <li>Quality and timeliness of public services</li> <li>Respect in public service provision, including response to citizens' feedback</li> </ul>	Responsiveness
	Anticipate change and protect citizens	<ul style="list-style-type: none"> <li>Anticipation and adequate assessment of citizens' evolving needs and challenges</li> <li>Consistent and predictable behavior</li> <li>Effective management of social, economic, and political uncertainty</li> </ul>	Reliability
<b>Values:</b> <i>The drivers and principles that inform and guide government action</i>	Use power and public resources ethically	<ul style="list-style-type: none"> <li>High standards of behavior</li> <li>Commitment against corruption</li> <li>Accountability</li> </ul>	Integrity
	Inform, consult, and listen to citizens	<ul style="list-style-type: none"> <li>Ability to know and understand what government is up to</li> <li>Engagement opportunities that lead to tangible results</li> </ul>	Openness
	Improve socioeconomic conditions for all	<ul style="list-style-type: none"> <li>Pursuit of socioeconomic progress for society at large</li> <li>Consistent treatment of citizens and businesses (vs. fear of capture)</li> </ul>	Fairness

Source: Original table for this publication.

Building on the above, the OECD has put forward an analytical framework that offers an instrumental approach to building citizens' trust in public institutions, facilitating measurement efforts (both based on experience and expectations) and policy attempts to influence trust. The OECD Framework on Drivers of Trust in Public Institutions, developed in 2017 and reviewed in 2021 through broad consultation with academics, policy makers, and civil society, includes four components (Brezzi et al. 2021). The updated framework is presented in table 28.2.

First, the framework places a greater emphasis on capturing trust levels in a larger set of institutions—for example, political parties or intergovernmental organizations—to further recognize the variety of institutions that influence policy making and that can shape people's assessment of public affairs as well as leaders' behavior. In addition, it recognizes the importance of improving the representation of diverse population groups that may be systematically excluded from voicing their views in traditional democratic processes, either due to personal characteristics (for example, geography or socioeconomic background) or because they persistently distrust “the system” and opt out of opportunities to express their voice.<sup>5</sup> The two broad dimensions of public sector competence and values—disentangled in responsiveness, reliability, integrity, openness, and fairness—remain core to the framework, as tested through country studies in the Republic of Korea, Finland, and Norway (OECD 2021a, 2022b; OECD and KDI 2018).

Third, the revised framework presents an “overlay” of the cultural, political, and economic factors that, at both an individual and group level, strongly influence levels of trust in government. Institutional competence and values are, in fact, mediated by individual and group identities, traits, and preferences—including political attitudes. These revisions attempt to emphasize more strongly the role played by political attitudes, including disengagement with the system, in explaining institutional trust.

Finally, the revised framework underlines the role people's confidence plays in the sustainability and effectiveness of policy action to address long-term and global challenges (for example, climate change, fiscal sustainability, digitalization, and inequality) (Brezzi et al. 2021). As the issues tackled by public institutions become increasingly complex, with long-term consequences involving a larger set of governmental and non-governmental actors, greater coordination and the ability of institutions to manage uncertainty and address trade-offs (for example, generational and economic trade-offs) will be key to preserving social cohesion and maintaining institutional trust.



**TABLE 28.2 OECD Framework on Drivers of Trust in Public Institutions**

Levels of trust in different public institutions		
Trust in national government, local government, civil service, parliament, police, political parties, courts, legal systems, and intergovernmental organizations		
Public governance drivers of trust in public institutions		
Competence	Responsiveness	<ul style="list-style-type: none"> <li>• Provide efficient, quality, affordable, timely, and citizen-centered public services that are coordinated across levels of government and satisfy users</li> <li>• Develop an innovative and efficient civil service that responds to user needs</li> </ul>
	Reliability	<ul style="list-style-type: none"> <li>• Anticipate needs and assess evolving challenges</li> <li>• Minimize uncertainty in the economic, social, and political environment</li> <li>• Effectively commit to future-oriented policies and cooperate with stakeholders on global challenges</li> </ul>
Values	Openness	<ul style="list-style-type: none"> <li>• Provide open and accessible information so the public better understands what the government is doing</li> <li>• Consult, listen, and respond to stakeholders, including through citizen participation and engagement opportunities that lead to tangible results</li> <li>• Ensure equal opportunities to participate in the institutions of representative democracy</li> </ul>
	Integrity	<ul style="list-style-type: none"> <li>• Align public institutions with ethical values, principles, and norms to safeguard the public interest.</li> <li>• Make decisions and use public resources ethically, promoting the public interest over private interests, while combatting corruption</li> <li>• Ensure accountability mechanisms between public institutions at all levels of governance</li> <li>• Promote a neutral civil service, whose values and standards of conduct prioritize the public interest</li> </ul>
	Fairness	<ul style="list-style-type: none"> <li>• Improve living conditions for all</li> <li>• Provide consistent treatment of businesses and people regardless of their background and identify (for example, gender, socioeconomic status, racial/ethnic origin)</li> </ul>
Cultural, economic, and political drivers of trust in public institutions		
<ul style="list-style-type: none"> <li>• Individual and group identities, traits, and preferences, including socioeconomic status and interpersonal socialization and networks</li> <li>• Distrust toward and disengagement with the system</li> </ul>		
Perception of government action on intergenerational and global challenges		
<ul style="list-style-type: none"> <li>• Perceptions of government commitment and effectiveness to address long-term challenges</li> </ul>		

Source: Brezzi et al. 2021.

Note: OECD = Organisation for Economic Co-operation and Development.

## The Measurement Strategy

The OECD Framework on Drivers of Trust in Public Institutions is operationalized through a nationally representative population survey: the OECD Trust Survey. An experimental module of questions to measure the five public governance drivers of trust was included in the OECD Guidelines on Measuring Trust (OECD 2017a, 2017b). The statistical feasibility and empirical relevance of the population survey were tested in six countries (France, Germany, Italy, Slovenia, the United Kingdom, and the United States) through the OECD Trustlab in 2018 (Murtin et al. 2018) and in Korea, Finland, and Norway through in-depth country studies (OECD 2021a, 2022b; OECD and KDI 2018). Adjusted and improved versions of these questions have been selected by the European Social Survey (ESS) to be included in their Cross-National Online Survey (CRONOS 2) in 2021.

The measurement approach for the competence and value drivers of institutional trust moves away from perceptions and focuses instead on specific situations. Situational questions present respondents with a stereotypical situation involving the interaction of people with public institutions and inquire about its expected outcome. The deconstruction of situational questions allows for analysis of the kind of behavior under scrutiny. Typical behavioral questions, as used in psychology or sociology, investigate the subjective reaction expected from individuals in a specific situation. Complementary- and confirmatory-mechanism experiments are suggested to see whether individuals stick to their revealed choices.

However, these situational questions are not stereotypical behavioral questions: they don't focus on individual behavior but rather on the positive conduct expected of a third party: in this case, public institutions.



**TABLE 28.3** Examples of Questions on the Determinants of Public Trust

Dimension	Example question
The following questions are about your expectations of behavior of public institutions. Please respond on a scale from 0 to 10, where 0 means very unlikely and 10 means very likely.	
Responsiveness	If many people complained about a public service that is working badly, how likely or unlikely do you think it is that it would be improved?
Reliability	If a serious natural disaster occurred in [country], how likely or unlikely do you think it is that existing public emergency plans would be effective in protecting the population?
Openness	If a decision affecting your community is to be made by the local government, how likely or unlikely do you think it is that you would have an opportunity to voice your views?
Integrity	If a government employee is offered a bribe in return for better or faster access to a public service, how likely or unlikely is it that they would accept it?
Fairness	If you or a member of your family would apply for a government benefit or service (e.g., unemployment benefits or other forms of income support), how likely or unlikely do you think it is that your application would be treated fairly?

Source: 2021 OECD Trust Survey (Nguyen et al. 2022).

For this reason, they measure the *trustworthiness* of a given institution or public agent. Unlike attitudes (passive response) and behaviors (active response), trustworthiness is based on the expectation of positive behavior that lies at the heart of the working definition of trust being considered. In general terms, a situational approach to measuring trustworthiness is based on the following type of question: “If a certain situation happens, how likely or unlikely is it that [public institution] will do [expected positive behavior]?” (see table 28.3).

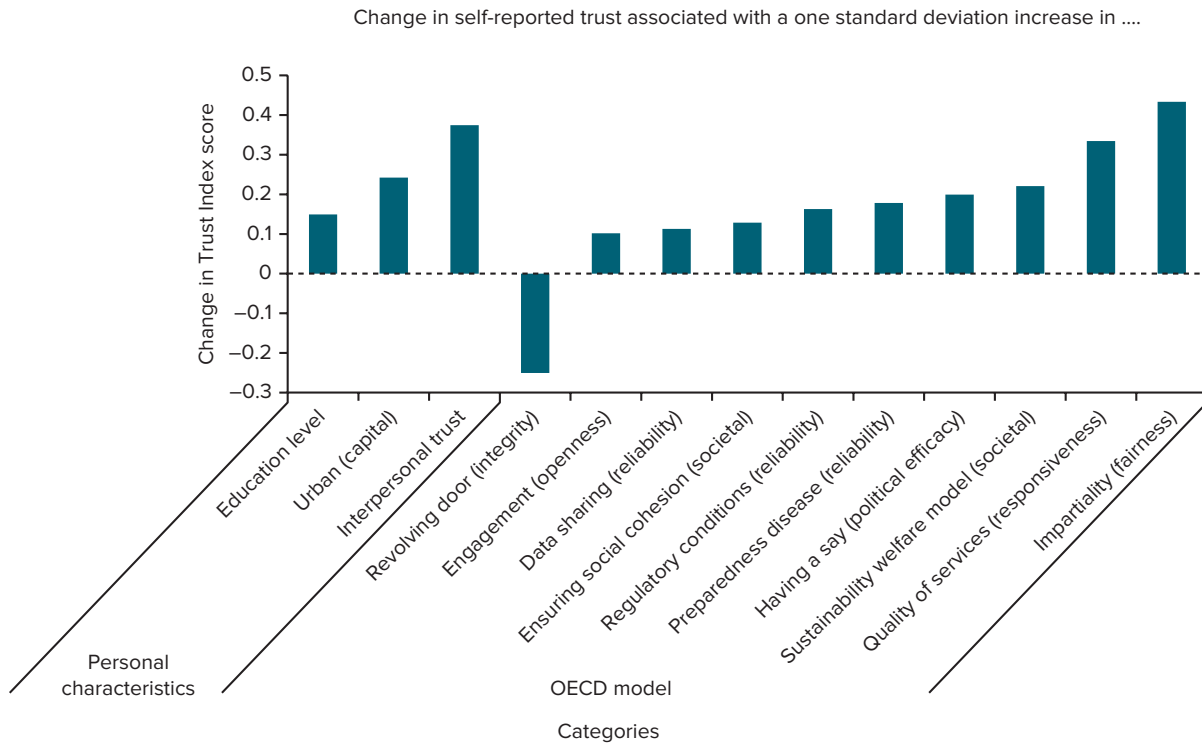
The module of the survey on the five public governance drivers of trust helps practitioners understand not only people’s perception of their government’s responsiveness, reliability, openness, fairness, and integrity but also which of these five components has a larger impact on the level of trust toward specific institutions. This evidence can provide guidance to improve public administration’s effectiveness.

## SOME RESULTS FROM APPLYING THE OECD TRUST SURVEY

Data collected through the OECD Trust Survey go beyond measuring trust levels and aim at identifying what drives trust in public institutions. As an example, in the case of Norway, the OECD Trust Survey shows that the most important determinants of trust in the civil service in Norway are impartial treatment when addressing the administration and responsiveness to people’s demands for service improvement. Other societal (for example, the sustainability of the welfare model) and personal (for example, living in the capital or being more educated) characteristics also have significant, although small, relative correlations (figure 28.1).

Along the same lines, the OECD Trust Survey in Finland finds that the government’s responsiveness and reliability are the main drivers of trust, but drivers of trust vary among institutions. Specifically, in Finland, the trust relationship between people and their institutions is strongly correlated with the perceived high competence of the government and the civil service and more tenuously with values such as integrity, openness, and fairness, most likely because the latter are recognized as entrenched in public sector culture. Figure 28.2 shows the trust payoff if all significant elements pertaining to competence and values increase by one standard deviation. The responsiveness of public services and the reliability of the government in addressing future challenges and providing a stable economic environment have a greater effect on trust in the national government and civil service, while engagement opportunities are more important for explaining trust toward local governments. These data, combined with qualitative analysis and international policy dialogue, have enabled a series of policy recommendations to preserve and strengthen the trust capital in these countries (OECD 2021a, 2022b).

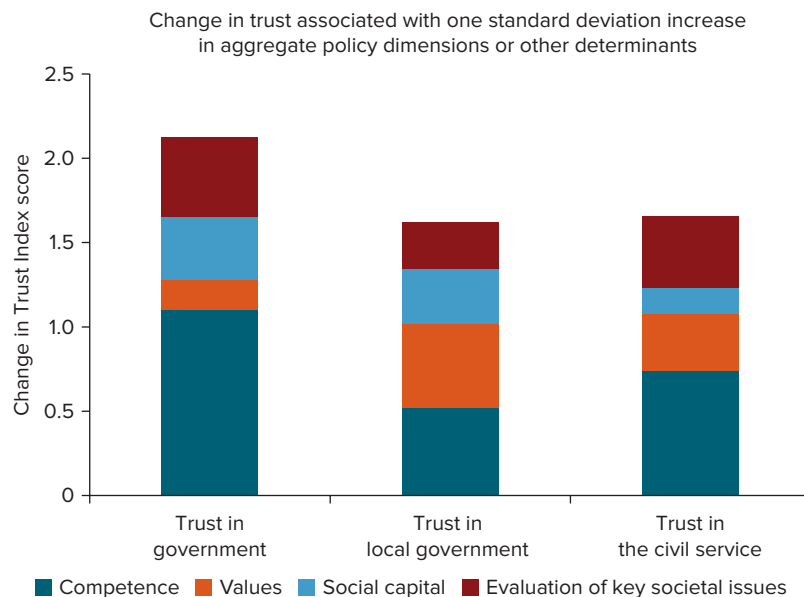
**FIGURE 28.1 Determinants of Trust in the Civil Service in Norway**



Source: OECD 2022b.

Note: This figure shows the most robust determinants of self-reported trust in government in an ordinary least squares (OLS) estimation that controls for individual characteristics. All variables depicted are statistically significant at 99 percent. The policy dimension is shown in parentheses. OECD = Organisation for Economic Co-operation and Development.

**FIGURE 28.2 Trust Payoff Associated with Increase in Public Institutions' Competence and Values in Finland, 2020**



Source: OECD 2021a.

Note: This figure shows the results of a statistical analysis to calculate the most robust determinants of trust in government. The figure shows that if competence increases by one standard deviation, then trust in the national government would increase by 1.10, trust in the local government would increase by 0.52, and trust in the civil service would increase by 0.74. If the values dimension increases by one standard deviation, then trust in the national government would increase by 0.18, trust in the local government would increase by 0.50, and trust in the civil service would increase by 0.34. Other elements, such as social capital (interpersonal trust) and the evaluations of current policies to address future challenges, are also correlated with trust in the different institutions.

Different countries have pursued different strategies to act upon these data. For example, in 2018, the Korean administration set a numerical target for its innovation strategy to improve trust levels. To achieve this target, it also included a series of actions: among others, generalizing open-government principles in different instances of the administration as a review of hiring procedures to ensure the right mix of skills for enhancing innovation within the administration. Likewise, the Finnish government has set up an inter-agency expert group to discuss concrete actions based on the data—for instance, reforming the process for formulating government policies to ensure better coordination for the inclusion of subjects such as climate change, intergenerational justice, and the preservation of social cohesion. In addition, it promotes citizenship education programs and engagement opportunities in policy choices for improving levels of political empowerment.

Based on the body of evidence developed so far, the OECD Trust Survey (based on table 28.2) has been carried out in 22 OECD countries, and the results were published in June 2022 (OECD 2022a). In addition, a number of briefs, country analyses, working papers, and country dialogues will be developed based on the OECD Trust Database.

## CONCLUSION

This chapter has taken stock and presented evidence of the OECD's work in understanding the determinants of institutional trust and improving their measurement. The evidence developed so far sheds light on the feasibility and pertinence of including questions on public trust and its drivers in regular household surveys as well as their relevance for informing policy developments. While politically sensitive, there are no a priori reasons why measures of institutional trust could not be collected regularly and be subject to the same quality standards and requirements that apply to other social, economic, and environmental statistics, such as those produced by NSOs.

The OECD Trust Survey will provide international benchmarks on people's perceptions, evaluations, expectations, and experiences with the public sector and will inform the debate on how to preserve and strengthen democratic values in OECD countries and beyond. It will also allow practitioners to observe levels of trust across different institutions and drivers of trust according to different socioeconomic characteristics. In addition, it will shed light on the relative effects of the determinants of public trust in different contexts, as well as their commonalities and differences. The results will set a course in public governance areas to improve institutional trust. The performance and pertinence of new experimental modules are still to be assessed.

There is, however, room to further develop our understanding and measurement of different trust relationships. Greater attention could be paid to interagency trust or government officials' perceptions of citizens—for example, by analyzing aspects of trust across different public agencies or trust from institutions toward citizens. It may also be possible to incorporate additional aspects that influence institutional trust into the survey as ad hoc modules that respond to specific realities in each context. While these ad hoc modules might make cross-national comparisons more challenging, they would enable countries to have greater flexibility and control over designing strategies to strengthen institutional trust.

## NOTES

The chapter was prepared between October and December 2021. This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries, and to the name of any territory, city, or area.

1. Results from the OECD Trust Survey are available on the OECD's website at <https://www.oecd.org/governance/trust-in-government/>.

2. Data from the Citizen Experience Survey, now known as the Survey of Trust in Australian Public Services, are available on the Australian government's website at <https://data.gov.au/data/dataset/trustsurvey>.
3. The questionnaire for the 2019 General Population Poll can be found on the website of the World Justice Project at <https://worldjusticeproject.org/our-work/research-and-data/wjp-rule-law-index-2021/2020-wjp-rule-law-index-questionnaires>.
4. Other types of validity, referred to as *discriminant validity* and *predictive validity*, are encompassed by construct validity, and the corresponding analysis is presented in the OECD Guidelines on Measuring Trust (OECD 2017a, 2017b).
5. *Distrust* refers to attitudes of insecurity, cynicism, disengagement, contempt, fear, anger, and alienation that may result in withdrawal, defiance, and support for populism. It is different from *mistrust*, which is associated with a positive attitude of critical citizenship that exercises vigilance in judging the components of the political system by being cautious, constructive, and alert. While measures of *trust in government* and the *trustworthiness* of public institutions are becoming more widely tested and used since the release of the OECD Guidelines on Measuring Trust, there are still a certain number of empirical challenges in measuring distrust at the individual level, which may require additional efforts, including, for example, enhanced sampling and focus groups.

## REFERENCES

- Algan, Yann, Elizabeth Beasley, Daniel Cohen, and Marital Foucault. 2018. "The Rise of Populism and the Collapse of the Left-Right Paradigm: Lessons from the 2017 French Presidential Election." Document du travail 1805, Centre pour la Recherche Économique et ses Applications, Paris.
- Algan, Yann, Elizabeth Beasley, Daniel Cohen, and Marital Foucault. 2019. *Les origines du populisme: Enquête sur un schisme politique et social*. Paris: Seuil.
- AU and UNDP (African Union and United Nations Development Programme). 2017. *Governance, Peace and Security (GPS) Data: Stock-Taking Report 2012–15*. Addis Ababa: UNDP. [https://www.undp.org/sites/g/files/zskgke326/files/migration/africa/UNDP-SHASA-GPS\\_web.pdf](https://www.undp.org/sites/g/files/zskgke326/files/migration/africa/UNDP-SHASA-GPS_web.pdf).
- Bargain, Olivier, and Ulugbek Aminjonov. 2020. "Trust and Compliance to Public Health Policies in Times of COVID-19." *Journal of Public Economics* 192: 104316. <https://doi.org/10.1016/j.jpubeco.2020.104316>.
- Blind, Peri K. 2007. "Building Trust in Government in the Twenty-First Century: Review of Literature and Emerging Issues." 7th Global Forum on Reinventing Government: Building Trust in Government, June 26–29, 2007, Vienna, Austria.
- Bouckaert, Geert. 2012. "Trust and Public Administration." *Administration* 60 (1): 91–115.
- Brezzi, Monica, Santiago González, David Nguyen, and Mariana Prats. 2021. "An Updated OECD Framework on Drivers of Trust in Public Institutions to Meet Current and Future Challenges." OECD Working Paper on Public Governance 48, OECD, Paris. <https://doi.org/10.1787/b6c5478c-en>.
- Choi, Sang Ok, and Sunhyuk Kim. 2012. *An Exploratory Model of Antecedents and Consequences of Public Trust in Government*. Unpublished paper. <https://iiatrust.files.wordpress.com/2012/07/an-exploratory-model-of-antecedents-and-consequences-of-public-trust-in-government.pdf>.
- Fairbrother, Malcolm, Gustaf Arrhenius, Krister Bykvist, and Tim Campbell. 2021. "Governing for Future Generations: How Political Trust Shapes Attitudes towards Climate and Debt Policies." *Frontiers in Political Science* 3: 656053. <https://doi.org/10.3389/fpos.2021.656053>.
- Fukuyama, Francis, and Francesca Recanatini. 2018. "Beyond Measurement: What Is Needed for Effective Governance and Anti-corruption Reforms?" In *Governance Indicators: Approaches, Progress, Promise*, edited by Helmut K. Anheier, Matthias Haber, and Mark A. Kayser, 43–70. Oxford: Oxford University Press. <http://dx.doi.org/10.1093/oso/9780198817062.003.0003>.
- González, Santiago, Lara Fleischer, and Marco Mira d'Ercole. 2017. "Governance Statistics in OECD Countries and Beyond: What Exists, and What Would Be Required to Assess Their Quality?" OECD Statistics Working Paper 79, OECD, Paris. <https://dx.doi.org/10.1787/c0d45b5e-en>.
- González, Santiago, and Conal Smith. 2017. "The Accuracy of Measures of Institutional Trust in Household Surveys: Evidence from the OECD Trust Database." OECD Statistics Working Paper 87, OECD, Paris. <https://doi.org/10.1787/d839bd50-en>.
- Györfy, Dóra. 2013. *Institutional Trust and Economic Policy: Lessons from the History of the Euro*. Budapest: Central European University Press.
- Kaufmann, Daniel, and Aart Kraay. 2007. "Governance Indicators: Where Are We, Where Should We Be Going?" Policy Research Working Paper, World Bank, Washington, DC. <https://doi.org/10.1596/1813-9450-4370>.
- Murtin, Fabrice, Lara Fleischer, Vincent Siegerink, Arnstein Aassve, Yann Algan, Romina Boarini, Santiago González, Zsuzsanna Lonti, Gianluca Grimalda, Rafael Hortalá Vallve, Soonhee Kim, David Lee, Louis Putterman, and Conal Smith. 2018. "Trust and Its Determinants: Evidence from the Trustlab Experiment." OECD Statistics Working Paper 89, OECD, Paris. <https://doi.org/10.1787/869ef2ec-en>.

- Nguyen, David, Valérie Frey, Santiago González, and Monica Brezzi. 2022. "Survey Design and Technical Documentation Supporting the 2021 OECD Survey on Drivers of Trust in Government Institutions." OECD Working Paper on Public Governance 53, OECD, Paris. <https://dx.doi.org/10.1787/6f6093c5-en>.
- Nooteboom, Bart. 2006. "Social Capital, Institutions and Trust." CentER Discussion Paper Series 2006–35, CentER Graduate School for Economics and Business, Tilburg University, Tilburg, The Netherlands. <http://dx.doi.org/10.2139/ssrn.903747>.
- OECD (Organisation for Economic Co-operation and Development). 2013. *Government at a Glance 2013*. Paris: OECD Publishing. [https://doi.org/10.1787/gov\\_glance-2013-en](https://doi.org/10.1787/gov_glance-2013-en).
- OECD (Organisation for Economic Co-operation and Development). 2017a. *OECD Guidelines on Measuring Trust*. Paris: OECD Publishing. <https://dx.doi.org/10.1787/9789264278219-en>.
- OECD (Organisation for Economic Co-operation and Development). 2017b. *Trust and Public Policy: How Better Governance Can Help Rebuild Public Trust*. OECD Public Governance Reviews. Paris: OECD Publishing. <http://dx.doi.org/10.1787/9789264268920-en>.
- OECD (Organisation for Economic Co-operation and Development). 2021a. *Drivers of Trust in Public Institutions in Finland*. Paris: OECD Publishing. <https://dx.doi.org/10.1787/52600c9e-en>.
- OECD (Organisation for Economic Co-operation and Development). 2021b. *Government at a Glance 2021*. Paris: OECD Publishing. <https://doi.org/10.1787/1c258f55-en>.
- OECD (Organisation for Economic Co-operation and Development). 2022a. *Building Trust to Reinforce Democracy: Main Findings from the 2021 OECD Survey on Drivers of Trust in Public Institutions*. Paris: OECD Publishing. <https://doi.org/10.1787/b407f99c-en>.
- OECD (Organisation for Economic Co-operation and Development). 2022b. *Drivers of Trust in Public Institutions in Norway*. Paris: OECD Publishing. <https://doi.org/10.1787/81b01318-en>.
- OECD and KDI (Korea Development Institute). 2018. *Understanding the Drivers of Trust in Government Institutions in Korea*. Paris: OECD Publishing. <https://dx.doi.org/10.1787/9789264308992-en>.
- Rothstein, Bo. 2013. "Corruption and Social Trust: Why the Fish Rots from the Head Down." *Social Research: An International Quarterly* 80 (4): 1009–32. <https://dx.doi.org/10.1353/sor.2013.0040>.
- Tabellini, Guido. 2008. "Institutions and Culture." *Journal of the European Economic Association* 6 (2–3): 255–94. <https://doi.org/10.1162/JEEA.2008.6.2-3.255>.
- United Nations Praia Group on Governance Statistics. 2020. *Handbook on Governance Statistics*. <https://paris21.org/news-center/news/new-praia-city-group-handbook-governance-statistics>.
- Van de Walle, Steven, and Geert Bouckaert. 2003. "Public Service Performance and Trust in Government: The Problem of Causality." *International Journal of Public Administration* 26 (8–9): 891–913. <http://dx.doi.org/10.1081/pad-120019352>.
- Van de Walle, Steven, and Koen Migchelbrink. 2020. "Institutional Quality, Corruption, and Impartiality: The Role of Process and Outcome for Citizen Trust in Public Administration in 173 European Regions." *Journal of Economic Policy Reform* 25 (1): 1–19. <http://dx.doi.org/10.1080/17487870.2020.1719103>.

## CHAPTER 29

# Government Analytics Using Measures of Service Delivery

Kathryn Andrews, Galileu Kim, Halsey Rogers, Jigyasa Sharma, and Sergio Venegas Marin

### SUMMARY

Public services, such as primary health care and education, have important consequences for social welfare and economic development. However, the quality of service delivery across the world is uneven. To improve it, practitioners require evidence to understand what is driving outcomes in education and health, such as student learning and the prevalence of chronic diseases. *Measures of service delivery* (MSDs) provide objective measurements of the quality of public service delivery. These indicators offer a granular view of the service delivery system, providing actionable insights into different parts of the delivery chain: from the physical infrastructure to the knowledge of frontline providers. This chapter provides an outline for how to conceptualize, measure, and disseminate MSDs, leveraging the institutional expertise of teams of practitioners at the World Bank. It offers actionable steps and advice that aim to connect practitioners to wider global efforts to improve the quality of public service delivery.

### ANALYTICS IN PRACTICE

- *Measures of service delivery* (MSDs) are objective measures of different parts of the public service delivery system. These indicators provide a granular view of the entire process of service delivery. MSDs measure the quality of the delivery of public services, such as primary health care and education. In addition to measuring welfare outcomes, measures should focus on different parts of the service delivery system, such as physical capital (for example, hospitals and schools) and human capital (for example, the knowledge of practitioners). Management practices play an important role in translating physical infrastructure and human capital into patient and student outcomes.

---

This chapter's authors are World Bank staff. Kathryn Andrews is a health economist in the Health, Nutrition, and Population Global Practice. Galileu Kim is a research analyst in the Development Impact Evaluation (DIME) Department. Halsey Rogers is a lead economist in the Education Global Practice. Jigyasa Sharma is a health economist in the Health, Nutrition, and Population Global Practice. Sergio Venegas Marin is an education economist in the Education Global Practice.



- Designing MSDs allows practitioners to specify the dimensions of service quality and construct measures to identify how well services perform in each one of them. Developing MSDs for primary health care and education requires considering and defining what dimensions will be used to measure the quality of public services. For example, a personnel dimension may measure absence rates for teachers and doctors.<sup>1</sup> Another dimension may be the availability of learning materials in schools and medical supplies in health facilities. There is a variety of conceptual frameworks and indicators to draw from. For health MSDs, practitioners may build on the existing frameworks described by *The Lancet Global Health Commission* (Hanson et al. 2022) or the World Health Organization’s “building blocks” framework (WHO 2010).
- The implementation of MSDs should follow a sequential structure, from defining a conceptual framework around measurements of the quality of public services to disseminating findings to government stakeholders and citizens. Generally, the first step in the implementation process is defining a conceptual framework and securing institutional support. The next step is identifying what will be measured: the indicators of interest; the questions to be asked; and to whom, where, and with what frequency they will be asked. After defining these indicators, practitioners should develop a rollout strategy for the actual data collection, which could include procuring a survey firm or developing a specific management information system for health care or education. After the data are collected, they should be validated, processed, and transformed into indicators. The final step is crucial: the resulting MSDs should be clearly articulated to stakeholders and disseminated widely, both within the government and to citizens.
- While practitioners may develop MSDs independently, the development of objective measurements to improve public services is part of a global agenda. Connecting to this broader movement allows practitioners to learn from other governments’ experiences. Engagement with global partners can also accelerate the design and implementation of MSDs. This global engagement can raise awareness of the relative standing of countries through benchmarking exercises, as well as facilitate knowledge exchange.
- MSDs should be subject to constant revision, as the understanding of quality in service delivery evolves. Indicators should also respond to new and unexpected demands. MSDs should evolve according to the changing policy objectives of stakeholders and citizens. Adaptations in measurement methodologies reflect an ongoing dialogue between policy makers, citizens, and the practitioners responsible for producing these indicators. As the COVID-19 (coronavirus) pandemic has highlighted, moments of crisis may generate demand for additional indicators, such as the availability of vaccines and the impact of school shutdowns on student learning.

## INTRODUCTION

Governments are responsible for the delivery of public services in primary health care and education, the foundations of public health and student learning.<sup>2</sup> It is well established in the development community that these services have immediate and important consequences for citizens who depend on them. Children rely on education services to learn how to read and write (World Bank 2018), and, as the COVID-19 pandemic has highlighted, access to health services can often determine whether patients recover from severe infections (Gatti et al. 2021). However, these same reports highlight the uneven quality of these services (Andrews et al. 2021). While some citizens may receive high-quality services, with positive impacts on welfare outcomes such as learning and health, others do not. What can practitioners do to improve the quality of these services? How can one measure changes in the quality of service? And what policy levers are available to change them?

We first note that public services are the outputs of a complex *service delivery system*. This system includes a range of inputs and processes. First, policies define how the service delivery system is structured and accessed: policies prescribe who is eligible to receive these services and establish processes to select service providers, such as teachers and nurses. These *de jure* policies lay the institutional groundwork for service delivery, but *de facto* inputs are equally important. These inputs include the facilities necessary for the provision of these services, such as schools and clinics, and the materials necessary for daily operations. Human capital is also crucial: these are the practitioners responsible for teaching students, for diagnosis and treatment, and for using their knowledge and skills to provide these services. Finally, a range of processes and management practices—including referral systems, feedback mechanisms, and counseling—translate these physical and human resources into welfare outcomes.<sup>3</sup>

*Measures of service delivery* (MSDs) measure the quality of these different dimensions of service delivery. MSDs account for and measure multiple factors in the service delivery system, providing policy makers with a holistic and granular view of how public services operate. Measurement of these different factors of production allows practitioners to map out conceptually how each part of the production chain is faring and where improvements can be made (Amin, Das, and Goldstein 2007). MSDs not only allow practitioners to measure each part of the chain; they also uncover causal relationships. As noted by Amin, Das, and Goldstein (2007), one of the key contributions of MSDs is allowing practitioners and researchers to measure the impact of a policy intervention in a rigorous way. Given these potential benefits, how can practitioners develop these indicators?

In this chapter, we focus on two examples of MSDs: the Service Delivery Indicators (SDI) Health Survey and the Global Education Policy Dashboard (GEPD). Drawing on both teams' expertise, we present an overview to practitioners on how to develop MSDs for primary health care and education, focusing on the facilities (schools and health facilities) in which these services are provided. Given this scope, we acknowledge there are secondary and even tertiary levels and other public services that stand to benefit from better measurement (for example, social protection and transportation) and that the private sector often provides these crucial services as well.<sup>4</sup> Nonetheless, we hope this chapter serves as an applied example for government practitioners on how to develop, implement, and use MSDs to improve the quality of public services at a foundational level. We outline conceptual frameworks and indicators, as well as how to generate them from survey data. Additionally, we highlight the benefits of connecting to a global agenda to develop and improve these indicators of service delivery.

This chapter is structured as follows. First, we provide a conceptual framework to measure the quality of service delivery. Section three outlines practical steps for implementation, adapted from the experience of our practitioners. Section four outlines the broader global agenda for developing MSDs. Finally, we conclude.

## AN OVERVIEW OF MEASURES OF SERVICE DELIVERY

Multiple global initiatives promote the use of MSDs. These include the World Bank's Learning Poverty indicator (World Bank 2021) and the Primary Health Care Performance Initiative (PHCPI).<sup>5</sup> Table 29.1 provides an abridged list of key global initiatives in generating MSDs, highlighting other related measurement initiatives. For example, the World Health Organization (WHO) Service Availability and Readiness Assessment (SARA) and the United States Agency for International Development (USAID) Service Provision Assessment (SPA) are similar initiatives to the SDI Health Survey. The revamped SDI Health Survey draws on best practices from both the SPA and SARA and goes a step further in the comprehensiveness of its domains of measurement and its patient-centered focus.

**TABLE 29.1 Survey of Global Initiatives in Education and Health Care Delivery Indicators**

Public service	Initiative	Description
Education	Global Education Policy Dashboard (GEPD)	To help countries put an end to learning poverty, the World Bank's Education Global Practice has developed and is supporting countries in the deployment of the GEPD. This new tool offers a strong basis for identifying priorities for investment and policy reforms that are suited to each country's context. It does so by highlighting gaps between what the evidence suggests is effective in promoting learning and what is happening in practice in each system and allowing governments to track progress as they take action to close those gaps.
	Service Delivery Indicators (SDI) Education	The SDI Education initiative collects data on service delivery in school facilities. It helps countries identify areas of progress and areas for improvement with potential lessons for progress within and between countries. Collected in close collaboration with the countries requesting a diagnostic, the data are used to assess the quality and performance of education. Since the initiative's creation, the surveys used have evolved, and existing data sets have been harmonized to allow for country comparisons over time.
	Systems Approach for Better Education Results Service Delivery (SABER SD)	The SABER SD tool was developed in 2016, in the Global Engagement and Knowledge Unit of the Education Global Practice at the World Bank, as an initiative to uncover bottlenecks that inhibit student learning in low- and middle-income countries and to better understand the quality of education service delivery in countries, as well as gaps in policy implementation. This school survey is aligned with the latest education research on what matters for student learning and how best to measure it. Its main purposes are to provide a mechanism to assess different determinants of learning through a diagnostic tool and to uncover the extent to which policies translate into practice.
	Teach Early Childhood Education (ECE)	Teach ECE is a free classroom observation tool that provides a window into one of the less explored and more important aspects of a child's education: what goes on in the classroom. The tool is intended to be used with children ages three to six and was designed to help countries, in particular low- and middle-income countries, monitor and improve teaching quality following the Teach Primary framework.
	Learning Poverty indicator	This indicator brings together schooling and learning indicators: it begins with the share of children who haven't achieved minimum reading proficiency (as measured in schools) and is adjusted by the proportion of children who are out of school (and are assumed not to be able to read proficiently).
	COVID-19: Monitoring the Impacts on Learning Outcomes (MILO)	The MILO project aims to measure learning outcomes in six countries in Africa in order to analyze the long-term impact of COVID-19 on learning and to evaluate the effectiveness of distance-learning mechanisms utilized during school closures. In addition, this project will develop the capacity of countries to monitor learning after the crisis.
	Early Childhood Development (ECD)	The World Bank team has developed a suite of tools to measure childhood development and early learning quality, including the Anchor Items for Measurement of Early Childhood Development (AIM-ECD), a core set of items with robust psychometric properties across contexts for measuring preschoolers' early literacy, early numeracy, executive functioning, and socioemotional development; Teach ECE, an observation tool that captures the quality of teacher-child interactions in preschools (see above); and the ECD COVID-19 Phone Survey to support countries in capturing the impacts of the pandemic on young children and their families.
Health care	Service Delivery Indicators (SDI) Health	SDI Health provides a set of metrics for benchmarking service delivery performance in primary health care. The overall objective of the indicators is to gauge the quality of service delivery in basic health services measured at the health facility. The indicators enable the identification of gaps and the tracking of progress over time and across countries. It is envisaged that broad availability, high public awareness, and a persistent focus on the indicators will mobilize policy makers, citizens, service providers, donors, and other stakeholders for action to improve the quality of services and, ultimately, to improve development outcomes and social welfare.
	Primary Health Care Performance Initiative (PHCPI)	The PHCPI developed a conceptual framework that describes important components of a strong primary health care system. It is intended to guide what should be measured to inform and drive efforts to improve primary health care. The framework is based on evidence about the key characteristics and determinants of strong primary health care systems, building on existing frameworks for health system performance. The selection of our core indicators and the creation of the PHC Vital Signs Profiles were informed by this framework. The data collected through SDI Health Surveys can be used to help create the PHC Vital Signs Profiles.

*(continues on next page)*

**TABLE 29.1 Survey of Global Initiatives in Education and Health Care Delivery Indicators**  
(continued)

Public service	Initiative	Description
Health care (continued)	Service Availability and Readiness Assessment (SARA)	SARA is a health facility assessment tool designed to assess and monitor service availability and the readiness of the health sector and to generate evidence to support the planning and managing of a health system. SARA is designed as a systematic survey to generate a set of tracer indicators of service availability and readiness. The survey's objective is to generate reliable and regular information on service delivery (such as the availability of key human and infrastructure resources); on the availability of basic equipment, basic amenities, essential medicines, and diagnostic capacities; and on the readiness of health facilities to provide basic health care interventions related to family planning, child health services, basic and comprehensive emergency obstetric care, human immunodeficiency virus (HIV), tuberculosis, malaria, and noncommunicable diseases.
	Service Provision Assessment (SPA)	The SPA survey is a health facility assessment that provides a comprehensive overview of a country's health service delivery. The SPA looks at the availability of health services; the extent to which facilities are ready to provide health services (do they have the necessary infrastructure, resources, and support systems?); the extent to which service delivery processes meet accepted standards for quality of care; and the satisfaction of clients with the service delivery environment.

Source: Original table for this publication.

These different approaches to measuring service delivery propose a conceptual framework for service delivery and how it should be measured. The common pillars of these conceptual frameworks are the following policy objectives:

- Set priorities regarding improvements in service delivery.
- Identify strengths and gaps in delivery system performance.
- Identify knowledge gaps, where deeper diagnostics are needed.
- Monitor progress on the quality of services.

These policy objectives should guide practitioners in defining the relevant dimensions of quality they are interested in measuring. For example, a practitioner may prioritize improving the quality of student learning in a school. One potential indicator is Learning Poverty: the share of children who haven't achieved minimum reading proficiency. This indicator helps identify the strengths and gaps in service delivery performance by providing objective benchmarks with which to compare student learning across schools. Deeper diagnostics may be required: are there particular age groups that are more vulnerable to low reading proficiency? Are there gender gaps that may be driving these results? Finally, a monitoring strategy allows governments to identify whether progress has been made. For example, the impact on student learning of a policy change such as improving access to school materials can be monitored by taking a baseline and endline survey measuring the indicator.

As practitioners explore these sets of questions, we recommend that they draw upon the international experience of other teams that have developed both conceptual and methodological frameworks to address them. For example, the *World Development Report 2018* (World Bank 2018) provides an array of tools that focus on measuring student learning deficits across the world, and *The Lancet Global Health Commission on High-Quality Health Systems in the SDG Era* (Kruk et al. 2018) provides guidance on how measurement efforts can improve the quality of health care services.<sup>6</sup> As the *World Development Report 2018* argues, measurement makes visible otherwise “invisible” quality deficits in the delivery of educational services. However, while objective measures are necessary, they are not in themselves sufficient to improve the quality of public services. These indicators “must facilitate action, be adapted to country needs, and consist of a range of tools to meet the needs of the system” (World Bank 2018, 91).

This section provides an overview of how MSDs are conceptualized and measured. We divide our discussion into education and health care, corresponding to two distinct but analogous approaches to measuring the quality of public service delivery. We draw on the experience of teams at the World Bank who have

developed and implemented MSDs, as well as other global efforts that have promoted the use of objective measures to improve the quality of service delivery. We present these conceptual frameworks as concrete examples for practitioners interested in developing their own MSDs. Practitioners should bear in mind that these frameworks are neither exhaustive nor prescriptive. We draw from two programs, the SDI Health Survey and the GEPD, to explain what these conceptual frameworks are, why they came to be, and how practitioners may draw on them to develop their own MSDs.

### **SDI Health Survey Conceptual Framework**

The SDI Health Survey is a nationally representative, health facility–based survey that measures the quality of delivery of primary health care services as experienced by citizens across the world.<sup>2</sup> Since its inception in 2008, the objective of the SDI Health Survey has been to improve the monitoring of service delivery to increase public accountability and good governance, as well as targeted interventions through objective measurement of the quality and performance of health services. SDI Health Surveys have been completed in several countries in Africa, including Kenya, Madagascar, and Mozambique (see figure 29.1), and the survey has recently expanded to South Asia (Bhutan), Europe and Central Asia (Moldova), and the Middle East and North Africa (Iraq).

To accomplish this objective, the SDI Health Survey team originally developed a conceptual framework that allowed practitioners to measure the quality of health service delivery. As outlined in Gatti et al. (2021), the first generation of SDI Health Survey content was based on three dimensions, with corresponding topics and associated indicators, as illustrated in table 29.2.

Recently, there has been a push to reimagine how the SDI Health Survey measures the quality of services. In particular, the conceptual framework has been expanded to focus on processes of care and person-centered outcomes (such as patients' experience, including wait time and expenditures incurred). Additional measures have been included to measure job satisfaction and the broader work environment as experienced by health care providers. Finally, given the increasing salience of public health crises, measurements of facilities have been expanded to gauge levels of preparedness for pandemics and disaster scenarios. We provide an overview of the updated conceptual framework and questionnaire modules in figure 29.2.

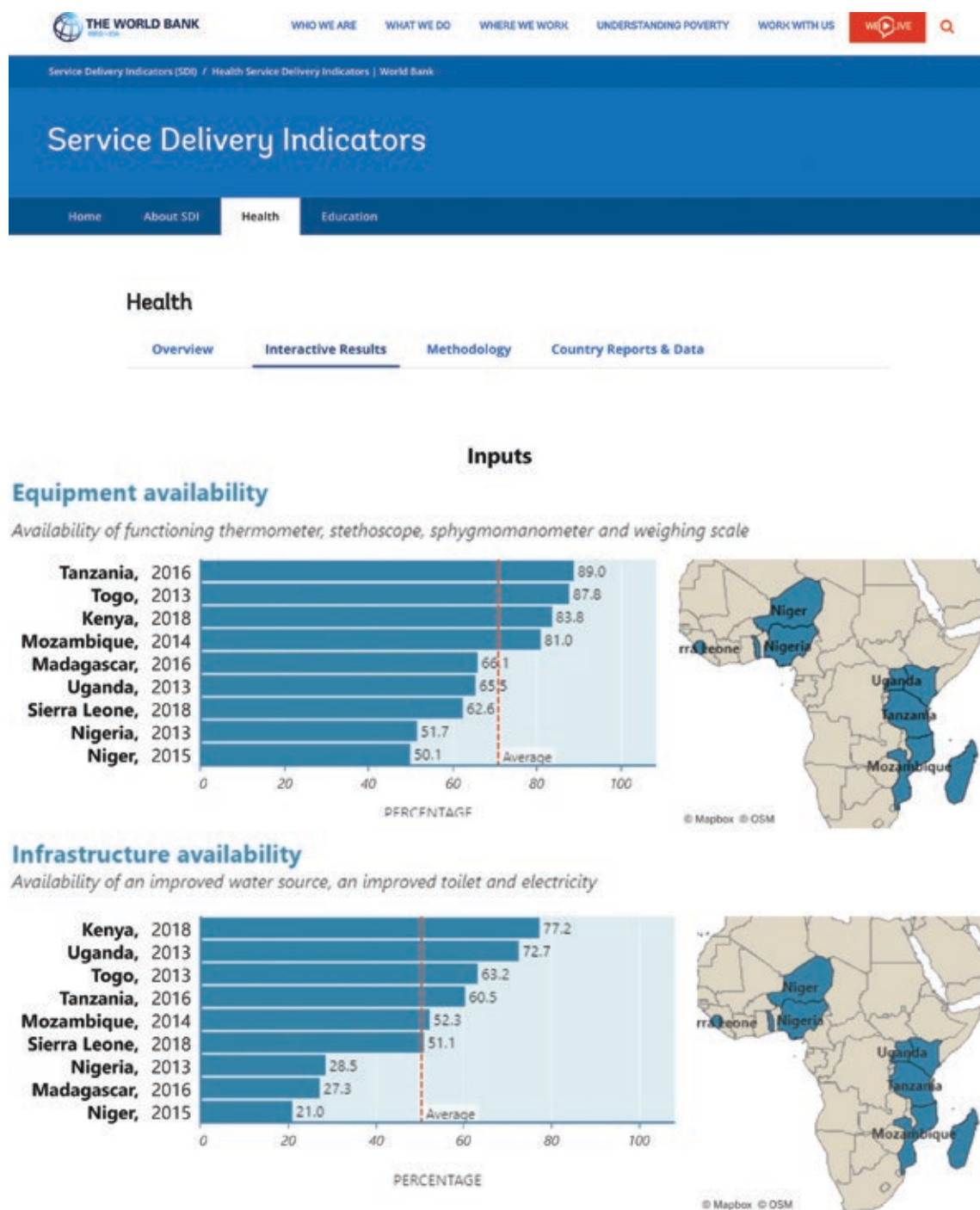
Both table 29.2 and figure 29.2 present the wide array of health service delivery measurement topics available to practitioners. At the same time, they highlight how foundational concepts—facilities, providers, and patients—can provide a basis to measure the quality of service delivery. Practitioners are encouraged to define policy objectives as outlined at the beginning of this section. For example, a practitioner may prioritize increasing equipment and supplies at the facility level. The SDI Health Survey helps identify strengths and gaps in the delivery system. Certain regions may lag behind others in the availability of these inputs, raising questions as to what is driving this limited availability. Deeper diagnostics may suggest a correlation between regions with lower workforce training and lower equipment availability. The impact of a policy intervention to increase the capacity building of staff may be measured by follow-up SDI Health Surveys.

### **The Global Education Policy Dashboard Conceptual Framework**

The GEPD builds on a set of nationally representative surveys that measure the quality of educational services and learning outcomes, including the SDI Education Survey.<sup>8</sup> Since its initial development in 2019, the GEPD has outlined its goal to measure and highlight the key drivers of learning outcomes, connecting all parts of the production chain. It provides a systemic overview of the drivers of learning, focusing on key dimensions of the educational system, such as teachers and the policies overseeing them. GEPD projects have been completed in Rwanda, Jordan, and Peru, with ongoing implementation in other countries in Africa (Ethiopia and Mozambique), as well as plans for expansion into other regions.



**FIGURE 29.1** MSD Health Indicators for a Selection of Countries in Africa



Source: Screenshot of Service Delivery Indicators Health, Interactive Results dashboard, <https://www.sdindicators.org/>.

Note: MSD = measures of service delivery.

The GEPD identifies three key dimensions driving learning outcomes: practices (or service delivery), policies, and politics (figure 29.3). The practices dimension is further divided into four topics: teachers, learners, school inputs, and school management. Because the GEPD also provides indicators for learning, it offers a holistic view of the educational system, connecting outcomes (learning) to their drivers. A total of 39 indicators have been developed to measure these different dimensions. Examples of indicators, as well as their associated topics and dimensions, are provided in table 29.3.<sup>2</sup>

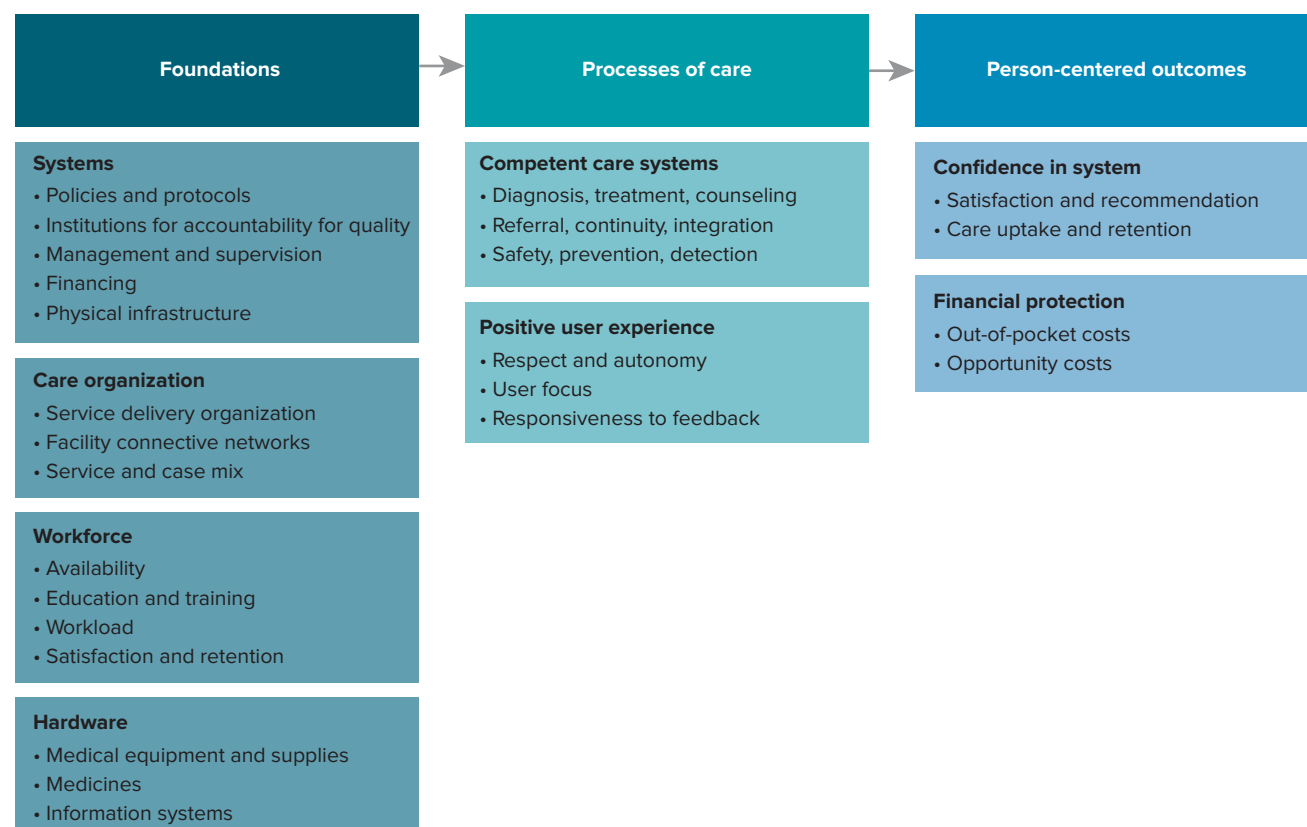


**TABLE 29.2 Indicators in the First Service Delivery Indicators Health Surveys**

Dimension	Topic	Indicators
Provider effort	Provider absence	Share of a maximum of 10 randomly selected providers absent from the facility during an unannounced visit
	Caseload per health provider	Number of outpatient visits per clinician per day
Provider's knowledge and ability	Diagnostic accuracy	Percentage of correct diagnoses provided in a selection of five to six clinical vignettes
	Treatment accuracy	Percentage of correct treatments provided in a selection of five to six clinical vignettes
	Management of maternal and neonatal complications	Number of relevant treatment actions proposed by the clinician
Inputs	Medicine availability	Percentage of 14 basic medicines that were available and in stock at the time of the survey
	Equipment availability	Availability and functioning of a thermometer, stethoscope, sphygmomanometer, and weighing scale
	Infrastructure availability	Availability and functioning of an improved water source, an improved toilet, and electricity

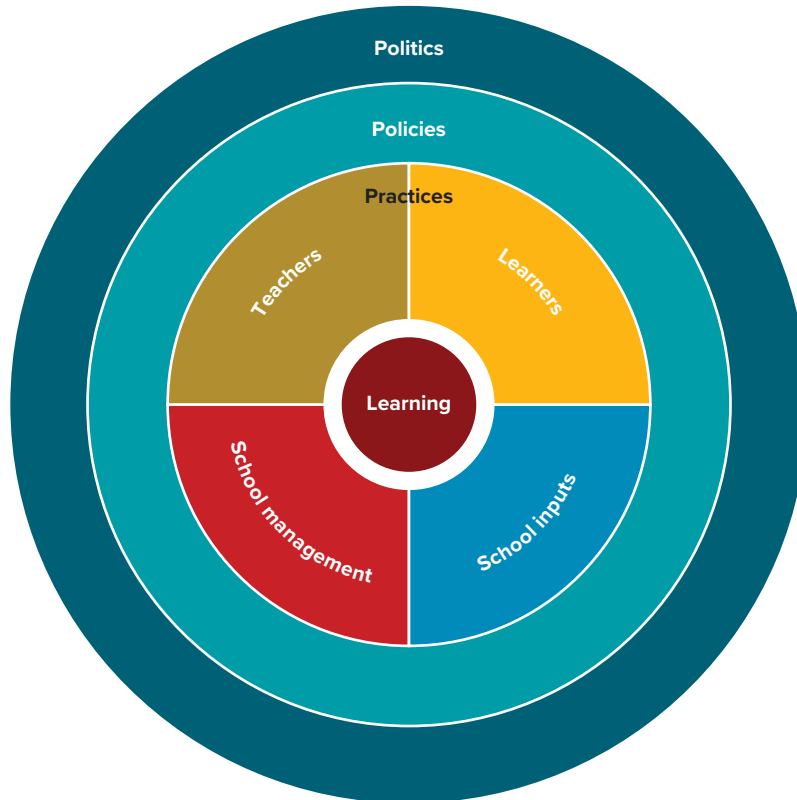
Source: Adapted from Gatti et al. 2021.

**FIGURE 29.2 Updated Conceptual Framework and Questionnaire Modules for the Service Delivery Indicators Health Survey**



Source: Original figure for this publication.

**FIGURE 29.3** Dimensions of the Global Education Policy Dashboard



Source: Global Education Policy Dashboard website, <https://www.educationpolicydashboard.org/>.

We provide an example of how to apply the GEPD indicators to achieve policy objectives. A practitioner may prioritize improving the attraction of teachers. Survey data suggest that while respondents in general perceive recruitment of teachers as meritocratic, the financial incentives of positions are viewed less favorably. A deeper diagnostic shows that these negative perceptions of financial incentives are concentrated among respondents in urban areas, where exit options may be more attractive. As a result, policy makers propose an additional financial bonus for teachers in competitive urban areas. Follow-up surveys monitor the impact of this policy change on the teacher attraction indicator.

As can be seen from the SDI and GEPD examples, contemporary conceptual frameworks provide a holistic assessment of public service delivery. These frameworks include indicators on multiple dimensions, such as workforce, management practices, and welfare outcomes or user experience. As a result, MSDs provide practitioners with granular information on different parts of the delivery chain, enabling targeted interventions. Additionally, granularity allows practitioners to explore causal relationships between dimensions—for example, how worker satisfaction impacts student learning. A holistic assessment of public service delivery therefore serves an important role in better understanding and improving the quality of public service.

## PRODUCING MEASURES OF SERVICE DELIVERY

Equipped with a conceptual framework, the next step for practitioners is the actual measurement of service delivery. Within the scope of this chapter, our primary focus is on facility surveys, although the GEPD includes measurements of additional factors, including politics and policies, which are measured through

**TABLE 29.3 Global Education Policy Dashboard Indicators**

Dimension	Topic	Indicators
Learning	Proficiency on GEPD assessment	Each question on the fourth-grade student assessment is scored as one point. The indicator reports the fraction of students scoring at least 20 out of 24 points on the fourth-grade language assessment and at least 14 out of 17 points on the math assessment.
	Learning Poverty indicator	The Learning Poverty indicator, as reported in the Learning Poverty Database, captures schooling and learning at the end of primary school.
	Teacher content knowledge	This indicator measures the percentage of teachers scoring at least 80 percent correct on the teacher assessment. In this assessment, each question is worth one point.
	Instructional leadership	A score of one to five is assigned based on the presence of four practices as reported by teachers. The four practices, which are given equal weight, are the following: <ul style="list-style-type: none"> <li>• Had a classroom observation in past year</li> <li>• Had a discussion based on that observation that lasted longer than 30 minutes</li> <li>• Received actionable feedback from that observation</li> <li>• Had a lesson plan and discussed it with another person.</li> </ul>
Policies	Teaching—attraction	A score of one to five is assigned based on five factors. Each factor receives an equal weight in terms of possible points (0.8). The factors are the following: <ul style="list-style-type: none"> <li>• Job satisfaction</li> <li>• Community satisfaction</li> <li>• Perceived meritocracy</li> <li>• Financial incentives</li> <li>• Absence of salary delays.</li> </ul>
	School management—evaluation	A score of one to five is assigned based on four factors. Each factor receives an equal weight in terms of points. The factors are the following: <ul style="list-style-type: none"> <li>• Reported evaluation in the past year (1)</li> <li>• Reported multiple evaluation criteria (1)</li> <li>• Reported consequences for negative evaluation (1)</li> <li>• Reported consequences for positive evaluation (1).</li> </ul>
Politics	Characteristics of bureaucracy	A score of one to five is assigned based on four factors. Each factor has been given an equal weight. Each factor is based on a set of three to four questions, each scored one to five. For each factor, the average score across the questions is determined. To construct the total score, the average is taken of the four factor scores. The factors include the following: <ul style="list-style-type: none"> <li>• Knowledge and skills</li> <li>• Work environment</li> <li>• Merit</li> <li>• Motivation.</li> </ul>
	Impartial decision-making	A score of one to five is assigned based on four factors. Each factor has been given an equal weight. Each factor is based on a set of three questions scored one to five. For each factor, the average score across the three questions is determined. To construct the total score, the average is taken of the four factor scores. The factors include the following: <ul style="list-style-type: none"> <li>• Politicized personnel management</li> <li>• Politicized policy making</li> <li>• Politicized policy implementation</li> <li>• Employee unions as facilitators.</li> </ul>

Source: Adapted from GEPD 2021b.

surveys of public officials (discussed in detail in part three of *The Government Analytics Handbook*). Facility surveys are only one of many different, important ways of measuring service delivery. Many things are not measured by visiting facilities: children who do not attend schools are not included, clients who do not visit clinics are not included, and central-governance-level issues, such as national policies or district-level protocols, are not measured. In this section, we provide guidelines on how practitioners can move forward and generate their own MSDs at the facility level.

We divide the production of facility-based MSDs into four stages: design, implementation, analysis, and dissemination. This section is filled with practical advice on how our teams have engaged in the rollout of MSDs (we have drawn especially on GEPD [2021a] and SDI [2019]). Practitioners are invited to adapt these implementation guidelines to their own contexts and needs.

## Design: Stakeholder Engagement and Survey Instrument

The first step is securing engagement and institutional support from MSD stakeholders. Depending on the target public service, these stakeholders may vary. Once the indicators are produced, who will consume these data? Where in public administration would these indicators have a maximal impact? These questions should help practitioners identify relevant actors. For education, stakeholders may include the ministry of education, subnational governments responsible for providing these services, and civil society organizations, such as teacher associations. For health care, potential stakeholders would be the ministry of health, development partners (such as the WHO and UNICEF), and heads of clinics and hospitals, among others. Stakeholder engagement should guide the selection of relevant dimensions and indicators: what specific actions are expected to change because of the MSDs?

Practitioners are encouraged to assess trade-offs: broader coverage in terms of dimensions can come at the expense of depth in particular dimensions—such as teacher skills—that may be of greater interest to stakeholders. Perhaps resources are constrained, and engaging in a full-fledged survey is not feasible. If this is the case, practitioners may have to select a few dimensions that are considered priorities by stakeholders. However, this selectiveness implies a cost: the SDI Health Survey and the GEPD dimensions speak to one another and allow for data triangulation to get a comprehensive picture of the service across key domains. The selection of dimensions may also compromise international benchmarking for mutual learning, which requires comparability between and within countries.

Finally, practitioners should define the level of representativeness of the survey. In some cases, a nationally representative survey will suffice, reducing the burden and cost of implementation. Such statistics are useful for national policy makers to formulate broad changes to the service delivery system as a whole. For example, if there are systemic issues in the distribution of facility inputs, such as schoolbooks, national statistics allow for a broad response. In other contexts, subnational variation in the quality of service may be of interest and is a powerful complement to national statistics. Often, the problems facing subnational regions differ, with some regions requiring facility inputs and others requiring staff training. Gaining this greater degree of granularity requires a different sampling strategy. For example, the GEPD strategy follows stratified random sampling, defining strata as subnational regions and ensuring that all relevant geographical divisions are included. The “GEPD Technical Note” (GEPD 2021b) covers other specifications, such as survey weights and power calculations, as well as data collection and quality checks.

## Implementation: Training and Data Collection

Once the design phase is completed, the production cycle moves to implementation. Both SDI and GEPD teams use in-person surveys to undertake their assessments. There are important benefits to practitioners interested in measuring service delivery quality at the facility level in directly collecting data through field surveys, rather than relying on administrative data alone. Administrative data are often unreliable, in particular for areas of the country in which information systems are not widely available, as is often the case in rural facilities. Additionally, administrative data may be subject to misreporting, whereas enumerators serve as third-party observers. Finally, surveys can complement the development of robust information systems, providing actionable data for a fraction of the time and cost.<sup>10</sup> Data collection may either be done in-house or through public procurement of a survey firm.

Note that only if the accuracy of the data is guaranteed can practitioners generate robust analytical insights through indicators.<sup>11</sup> Otherwise, indicators will reproduce the biases and inaccuracies of the data, providing stakeholders with faulty evidence for policy making. Data accuracy relies on a robust, well-adapted, and piloted survey tool; high-frequency and sense checks; and data validation procedures, such as revisits or callbacks, to samples of the same facilities. These data validation procedures include verifying the time of submission of survey responses, the length of interviews, and systematic missingness in variables, among other checks.<sup>12</sup>

Both the GEPD and the SDI Health Survey engage in the automation of data collection, which is enabled by the use of Survey Solutions, an open-source tool available free of charge.<sup>13</sup> This has improved data quality over the past decade, as electronic data collection technologies have improved and become more pervasive. These technologies make real-time monitoring of data quality—through high-frequency checks and data quality warnings—easier to implement. Chapter 5 of the *Handbook* outlines a variety of protocols to ensure data quality, such as enumerator training and high-frequency checks on every batch of data, typically every day.

One innovation that both the SDI Health team and the GEPD leverage is the combination of announced and unannounced visits.<sup>14</sup> The former allows for a more thorough discussion of topics for which the service delivery providers need to prepare materials and information. Unannounced visits, on the other hand, enable practitioners to identify behaviors or practices that frontline providers or facility managers may have an incentive to either conceal or misreport, as well as those that may be disrupted by conducting a survey. The goal is not to reprimand providers or identify evidence of misconduct but rather to provide a more accurate assessment of the practices that occur during an average day of public service delivery. For example, one of the indicators collected by the SDI Health and Education Surveys is the health care or education provider's presence or absence during an unannounced visit.

### Analysis: Data Validation and Production of Indicators

Once the data have been collected by enumerators, the next step is validating the data, cleaning them, analyzing them, and generating indicators. Data validation should ideally be conducted in an automated fashion, where checks are encoded into relevant software and thus replicable in other settings. For example, the GEPD leverages open-source statistical software called R to validate the data collected in a documented and replicable way.<sup>15</sup> Data validation and processing are often challenging, in particular when in-house data analysis capacity may be more limited. In these contexts, we encourage practitioners to access different resources that document best practices in survey analysis, such as DIME Analytics (Björkefur et al. 2021). Chapter 5 of the *Handbook* provides additional information on this program.

Once data have been thoroughly validated and cleaned, the indicators can be generated. Again, where possible, these indicators should be generated through replicable steps, preferably using open-source software.<sup>16</sup> However, other statistical software, such as Microsoft Excel, may be capable of generating the relevant indicators. The crucial step is that the process for generating the indicators be documented, transparent, and replicable.

### Dissemination: Stakeholder Presentations and Wider Communication Efforts

The final step in the production cycle for MSDs is the dissemination of the results to stakeholders, as well as wider communication efforts. In general, we have found that interactive dashboards are an important component of MSD dissemination. Dashboards allow for intuitive visualization of the different dimensions of public service and empower stakeholders to interact with the data at a greater level of granularity than static reports. For an example, see figure 29.4.<sup>17</sup> Practitioners may click on different dimensions and obtain additional information for particular indicators. Colors allow users to identify where in the public service delivery chain more attention is needed.

An important feature of the GEPD is that it goes beyond the measurement of the indicators, providing visual feedback on each indicator to guide policy making at a granular level.<sup>18</sup> The feedback comprises three colors: red (needs improvement), yellow (caution needed), and green (on target). This visual feedback allows policy makers to immediately identify topics in which additional work is needed, as well as policy areas in which targets have been met (figure 29.4). This action-oriented visualization allows practitioners to design their educational policy with an intuitive and evidence-based approach.

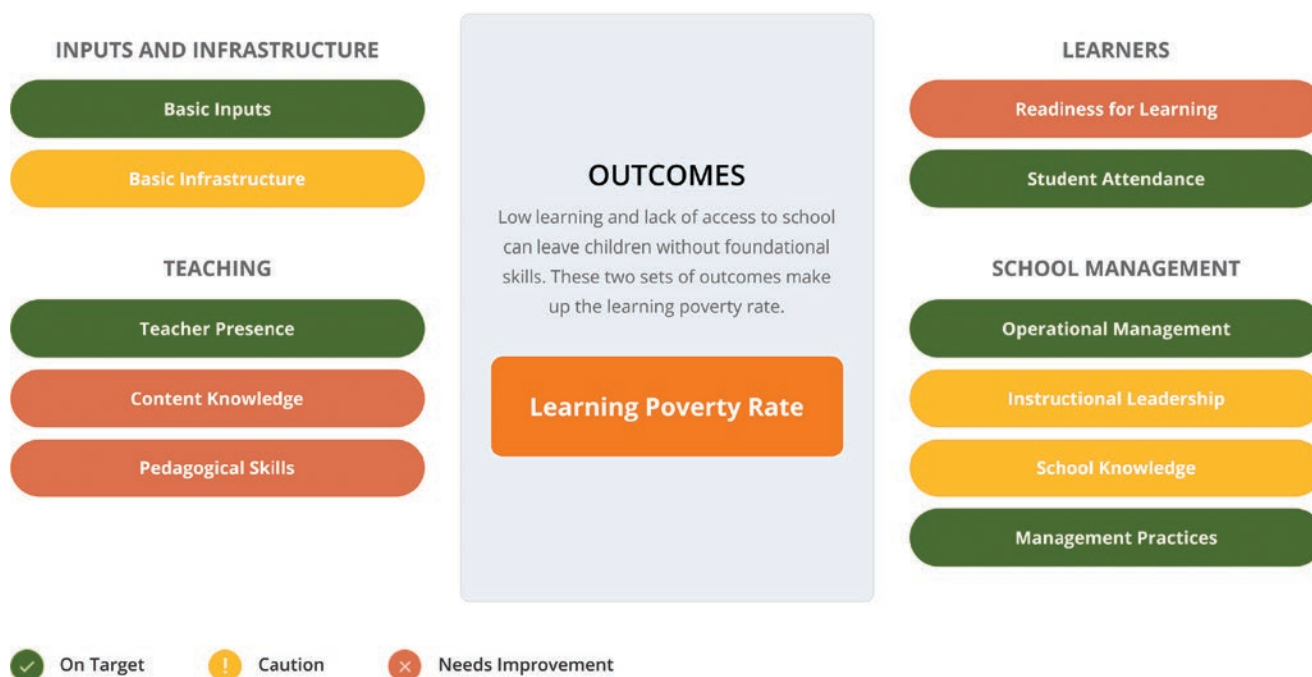
This systematic view, with diagnostics associated with each topic, also allows practitioners to hone in on specific areas that require further development. For instance, in figure 29.4, we find that Peru has done quite

**FIGURE 29.4** Example of Global Education Policy Dashboard

## Peru



To understand learning and participation outcomes, it is important to understand what is happening inside the schools. Here you can find the outcome indicators capturing learning for all (that is, learning combined with access) at the center, ringed by indicators representing the four-main school-level service-delivery factors: teaching, school management, inputs & infrastructure, and prepared learners. Click through the indicators to learn more about them as well as the indicators of deeper systematic characteristics linked with each of them.



Source: Screenshot of GEPD dashboard, <https://www.educationpolicydashboard.org/practice-indicators/per>.

well in management practices at the school level. However, teachers' skills require more attention: in particular, pedagogical skills and content knowledge. Thus, practitioners can prioritize certain areas over others, as they gather granular and actionable evidence on how indicators are faring.

While dashboards may be helpful for consumption within the government, a wider audience can be reached by organizing events where the MSDs are revealed to the public. The main findings of the MSDs should be presented by stakeholders and the implementation team, fostering accountability and transparency. Additionally, facilities that were interviewed could be given both the survey results and actionable steps they can take to improve their indicators, as well as recognition for areas in which they are successful. As noted in chapter 26 of the *Handbook*, governments should move beyond survey indicators by providing immediate feedback to facilities and civil servants.

## A GLOBAL EFFORT TO IMPROVE PUBLIC SERVICES

So far, this chapter has outlined some key considerations in measuring public sector service delivery in primary health care and education, as described by experts in the SDI Health and GEPD teams. These teams are part of a wider community of practice on generating measures to improve public service delivery, and we



encourage practitioners to connect to other global efforts. By engaging with this global community, governments can benefit from knowledge exchange with international organizations, as well as other practitioners pursuing similar initiatives.

Harmonizing surveys and producing indicators that can be benchmarked to other countries provides practitioners with objective standards against which to measure themselves—to help them understand, for instance, whether an enrollment rate is particularly high or low. If multiple countries have fielded similar surveys, a country team can take an indicator and compare it to other countries with similar educational systems and levels of economic development. Additionally, global indicators on public service delivery provide a public good that can be shared and accessed by communities of practitioners across the world.

The SDI and the GEPD are key players in the global movement to generate MSDs, with tools and expertise to help in this endeavor. Connecting to these global initiatives enables practitioners to capitalize on decades of experience, tools, and technical expertise that teams (like those in the World Bank) can offer to help optimize the long-lasting impact of these endeavors. An important exercise that global engagement enables is international benchmarking, which helps practitioners understand how well their countries are performing relative to others. Benchmarking exercises allow countries to quantify how far they may be from the frontier and to learn from the best in class what they can do to improve it.

This global community also makes available teams of education and health service delivery experts that can guide practitioners through the implementation of MSDs and accelerate rollout. Located in the World Bank's Education Global Practice and Health, Nutrition, and Population Global Practice, global experts provide technical assistance to practitioners interested in developing MSDs. While production cycles vary, the entire process from design to dissemination generally takes around one year. The costs vary as well but average US\$100,000–US\$400,000, based on country experiences.

Practitioners are encouraged to reach out to the SDI Health team for further details.<sup>19</sup> The materials and services provided by the SDI Health team include the following:

- Standardized health facility survey materials, field manuals, training materials, and suggestions for adaptation of the survey instrument
- Technical guidance on survey design and sampling strategy
- Assistance with quality control during data collection
- Capacity building for governments to generate and disseminate MSDs.

The GEPD provides similar services. It builds on the MSD framework but also leverages insights from other initiatives within the World Bank, such as the Systems Approach for Better Education Results (SABER) (World Bank 2020) and the Global Survey of Public Servants (GSPS).<sup>20</sup> Both the SDI Health Survey and the GEPD share a focus on capacity building, providing practitioners with the tools and resources necessary to reproduce conceptual and methodological frameworks on the ground. This approach ensures the co-ownership of results and operational relevance.

## CONCLUSION

This chapter has argued that MSDs provide governments with tools to measure and improve the quality of public service delivery. MSDs increase the accountability of governments because stakeholders gain access to objective measures of how public services are operating. Both the GEPD and the SDI Health Survey can provide governments with a systematic overview of service delivery, unpacking welfare outcomes—student learning and health care—as well as the different chains of delivery. We have also presented a step-by-step

guide to producing MSDs, drawing on the experience of teams at the SDI Health Survey and GEPD programs at the World Bank.

However, measurement alone is not enough to improve the quality of public services. MSDs need to be linked directly to stakeholders who can enact change in the delivery system. The broader public has to be made aware of the results as well. Moreover, as emphasized in chapter 4 of the *Handbook*, measurement is not a substitute for the proper management of services. With these caveats in mind, MSDs can allow practitioners and the broader research community to better understand the drivers of health and education outcomes. These efforts are part of a global agenda: we encourage readers to learn more from the publicly available resources listed here and, if interested, to reach out to relevant expert teams.<sup>21</sup>

Improving the quality of service delivery is a complex endeavor. As the COVID-19 pandemic has highlighted, unexpected crises can have profound consequences for the quality of health care and educational services. MSDs should be responsive to these sudden changes, as well as more gradual, evolving needs. Rather than ends in themselves, indicators should be used as tools to improve what ultimately matters: the lives of citizens who rely on public services.

## NOTES

1. This chapter uses the neutral term *absence* rather than *absenteeism*; the focus in the surveys is on the effect of provider absence on the quality of service delivery rather than on assigning blame to providers, who may be absent for reasons that are out of their control.
2. Primary health care, which was enshrined in the World Health Organization's Alma-Ata Declaration, includes essential services in health care, such as prenatal care and basic diagnostics. Primary education refers to pre-secondary education, including primary and middle school.
3. For an overview of these different dimensions in primary health care, see Andrews and Sharma (2021).
4. Additionally, there are often cases in which citizens still lack access to the basic services of health care and education. As a result, any improvements in these services may fail to improve their lives. Ensuring that broad access to these services develops in parallel with improvements in their quality is therefore crucial.
5. Further information about the PHCPI is available at the initiative's website, <https://improvingphc.org/>.
6. We highly recommend chapter 4 of the *World Development Report 2018*, "To Take Learning Seriously, Start by Measuring It."
7. To learn more about this program, see the SDI website, <https://www.sdindicators.org/>.
8. The GEPD School Survey builds on the following surveys: the Service Delivery Indicators (SDI) Survey, on teachers and inputs/infrastructure; Teach, on pedagogical practice; the Global Early Child Development Database (GECDD) and the Measuring Early Learning Quality and Outcomes (MELQO) initiative, on the school readiness of young children; and the Development World Management Survey (DWMS), on management quality. The GEPD also includes data on public officials from the Survey of Public Servants. For additional details, see the GEPD website, <https://www.educationpolicydashboard.org/>.
9. For a complete presentation and discussion of the indicators, see GEPD (2021b).
10. For a broader discussion of information systems, refer to chapter 9.
11. This is a similar argument to the one presented in chapter 9.
12. Examples of these checks for the GEPD can be found in the RMD file "School Data Quality Checks," located in the School folder in the Master Code directory in the GEPD repository in the World Bank's GitHub repository, available at [https://github.com/worldbank/GEPD/blob/master/Master\\_Code/School/school\\_data\\_quality\\_checks.Rmd](https://github.com/worldbank/GEPD/blob/master/Master_Code/School/school_data_quality_checks.Rmd) (latest commit September 24, 2019).
13. For more information, see the Survey Solutions website, <https://mysurvey.solutions/en/>.
14. The GEPD team provides a two-week window in which the visit will take place but does not disclose the precise date.
15. R is freely available at <https://cran.r-project.org/>. The entire GEPD repository and code is available in the World Bank's GitHub repository at <https://github.com/worldbank/GEPD> (latest commit February 23, 2023).
16. The GEPD also makes available the code generating the indicators in the Indicators folder in the GEPD repository in the World Bank's GitHub repository, available at <https://github.com/worldbank/GEPD/tree/master/Indicators> (latest commit January 13, 2023).

17. The dashboard is available on the GEPD website at <https://www.educationpolicydashboard.org/practice-indicators/per>.
18. This is similar to the visual feedback provided by the Employee Viewpoint Survey Analysis and Results Tool (EVS ART), a dashboard described in chapter 9, case study 9.3, and in chapter 26 of the *Handbook*.
19. The SDI Health team can be contacted at [sdi@worldbank.org](mailto:sdi@worldbank.org).
20. For more information on the Global Survey of Public Servants (GSPS), see its website, <https://www.globalsurveyofpublicservants.org/>.
21. The GEPD team can be contacted at [educationdashboard@worldbank.org](mailto:educationdashboard@worldbank.org). The SDI Health team can be contacted at [sdi@worldbank.org](mailto:sdi@worldbank.org).

## REFERENCES

- Amin, Samia, Jishnu Das, and Markus Goldstein, eds. 2007. *Are You Being Served? New Tools for Measuring Service Delivery*. Washington, DC: World Bank.
- Andrews, Kathryn, Ruben Conner, Roberta Gatti, and Jigyasa Sharma. 2021. "The Realities of Primary Care: Variation in Quality of Care across Nine Countries in Sub-Saharan Africa." Policy Research Working Paper 9607, World Bank, Washington, DC.
- Andrews, Kathryn, and Jigyasa Sharma. 2021. "A Revolution in Health Service Delivery Measurement." *Investing in Health* (blog). *World Bank Blogs*, July 1, 2021. <https://blogs.worldbank.org/health/revolution-health-service-delivery-measurement>.
- Björkefur, Kristoffer, Luíza Cardoso de Andrade, Benjamin Daniels, and Maria Ruth Jones. 2021. *Development Research in Practice: The DIME Analytics Data Handbook*. Washington, DC: World Bank. <https://openknowledge.worldbank.org/handle/10986/35594>.
- Gatti, Roberta, Kathryn Andrews, Ciro Avitabile, Ruben Conner, Jigyasa Sharma, and Andres Yi Chang. 2021. *The Quality of Health and Education Systems across Africa: Evidence from a Decade of Service Delivery Indicators Surveys*. Washington, DC: World Bank.
- GEPD (Global Education Policy Dashboard). 2019. "School Data Quality Checks." School, Master Code, GEPD, World Bank. GitHub repository, September 24, 2019. [https://github.com/worldbank/GEPD/blob/master/Master\\_Code/School/school\\_data\\_quality\\_checks.Rmd](https://github.com/worldbank/GEPD/blob/master/Master_Code/School/school_data_quality_checks.Rmd).
- GEPD (Global Education Policy Dashboard). 2021a. "Global Education Policy Dashboard Implementation Brief." Global Education Policy Dashboard, World Bank Group. <https://www.educationpolicydashboard.org/sites/epd/files/resources-documents/GEPD%20Implementation%20Brief.pdf>.
- GEPD (Global Education Policy Dashboard). 2021b. "GEPD Technical Note." Global Education Policy Dashboard, World Bank Group, April 19, 2021. <https://www.educationpolicydashboard.org/sites/epd/files/resources-documents/GEPD%20Technical%20Note.pdf>.
- Hanson, Kara, Nouria Brikci, Darius Erlangga, Abebe Alebachew, Manuela De Allegri, Dina Balabanova, Mark Blecher, Cheryl Cashin, et al. 2022. "The Lancet Global Health Commission on Financing Primary Health Care: Putting People at the Centre." *The Lancet Global Health Commission* 10 (5): E715–E772. [https://doi.org/10.1016/S2214-109X\(22\)00005-5](https://doi.org/10.1016/S2214-109X(22)00005-5).
- Kruk, Margaret E., Anna D. Gage, Catherine Arseneault, Keely Jordan, Hannah H. Leslie, Sanam Roder-DeWan, Olusoji Adeyi, et al. 2018. "High-Quality Health Systems in the Sustainable Development Goals Era: Time for a Revolution." *The Lancet Global Health Commission* 6 (11): E1196–E1252. [https://doi.org/10.1016/S2214-109X\(18\)30386-3](https://doi.org/10.1016/S2214-109X(18)30386-3).
- SDI (Service Delivery Indicators). 2019. "Nuts and Bolts: A Brief Guide for Task Teams." Unpublished document, September 24, 2019, World Bank, Washington, DC.
- WHO (World Health Organization). 2010. *Monitoring the Building Blocks of Health Systems: A Handbook of Indicators and Their Measurement Strategies*. Geneva: WHO Press. <https://apps.who.int/iris/bitstream/handle/10665/258734/9789241564052-eng.pdf>.
- World Bank. 2018. *World Development Report 2018: Learning to Realize Education's Promise*. Washington, DC: World Bank. <https://www.worldbank.org/en/publication/wdr2018>.
- World Bank. 2020. "Systems Approach for Better Education Results (SABER)." Education Brief, World Bank. <https://www.worldbank.org/en/topic/education/brief/systems-approach-for-better-education-results-saber>.
- World Bank. 2021. "What Is Learning Poverty?" Education Brief, World Bank, April 28, 2021. <https://www.worldbank.org/en/topic/education/brief/what-is-learning-poverty>.

## CHAPTER 30

# Government Analytics Using Anthropological Methods

*Colin Hoag, Josiah Heyman, Kristin Asdal, Hilde Reinertsen,  
and Matthew Hull*

### SUMMARY

This chapter aims to present an overview of how anthropologists study bureaucracy and why that approach has value to the World Bank and its interlocutors. Anthropologists are most commonly associated with immersive ethnographic methods, such as participatory observation. In this chapter, we describe those methods and their usefulness, but we also highlight the heterogeneity of the empirical materials that anthropologists draw upon. The chapter makes the case that, while the ethnographic approach of anthropologists might sometimes be perceived as “messy” or “unstructured,” in fact, the efforts of anthropologists are motivated by an abiding concern with empirical rigor—a refusal to ignore any sort of data or to content oneself with a single view of such a multifarious thing as bureaucracy. This is to say that an anthropological approach is a holistic one, which envisions bureaucracy as a rich, multidimensional world.

### ANALYTICS IN PRACTICE

- Study the bureaucratic process or organization you are interested in holistically by observing all aspects of it: engage with the staff who are involved at every level of the organization, from senior officers to low-level staff and contractors, and with different demographic groups; study everyday documents; and watch how officials interact. Observe every part of what they do at work holistically, from their interactions in corridors and meetings to the protocols they observe in their relationships. Processes or organizational outcomes may be shaped by forces outside of those that the analyst presupposes.
- Develop relationships with a variety of people and have open-ended conversations about their work as well as about unrelated issues to understand their values and perspectives.

---

Colin Hoag is an assistant professor at Smith College. Josiah Heyman is a professor at the University of Texas at El Paso. Kristin Asdal is a professor and head of the Centre for Technology, Innovation and Culture at the University of Oslo. Hilde Reinertsen is a researcher at the Centre for Technology, Innovation and Culture at the University of Oslo. Matthew Hull is an associate professor and director of the Center for South Asian Studies at the University of Michigan.

- Engage in participatory observation by observing government activities in person as they unfold. This practice can capture activities that may be so routine that they go unnoticed by public officials and are not self-reported in surveys. Practitioners should talk with those involved in government processes, both public officials and clients, as they participate in them. When attending meetings, practitioners should examine them as a form of social engagement where formal and informal rules preside and the process of forming consensus is revealed. Practitioners should likewise examine not only the content of documents but how they are developed and circulated and how they seek to further broader policy goals.
- Aim to collect the widest practical range and amount of data, both qualitative and quantitative, even if it cannot be easily standardized. Data may be gleaned from studying aspects of everyday life: the way documents are read; meetings are run; and policies, media, and whatever else are perceived as relevant. There is a trade-off in all analyses between standardization and holistic measurement. Ensure a role for holistic measurement in your approach.
- Study the heuristics, interactions, scope for discretion, and microscopic decisions that affect the gap between stated policy goals and the actual work being carried out by public officials. Real-world manifestations of policy depend on the decisions of individual officials about how to interpret broader policy goals. Using the words and actions of officials, practitioners should aim to determine which factors contribute to how officials make decisions and prioritize tasks. In this process, consider that these decisions may result from learned behaviors and rationales that require extensive effort to change.
- Revise research questions and the focus of analysis as the study progresses. As observation and insight reveal new lines of inquiry, practitioners should be open to shifting their questions and methods. The initial research plan (including the research questions and methods) should be considered provisional.

## INTRODUCTION

This chapter aims to present an overview of how anthropologists study bureaucracy and why this approach has value to the World Bank and its interlocutors. Anthropologists are most commonly associated with immersive ethnographic methods, such as participatory observation. In the pages that follow, we describe those methods and their usefulness, but we also highlight the heterogeneity of the empirical materials that anthropologists draw upon.

An anthropological analysis might include an ethnographic rendering of the sights, sounds, spatial organization, and everyday life of a bureaucratic workplace. But it might also include a discourse analysis of policies that emanate from the federal government or “head office,” including the categories of persons and practices that those policies define, as well as a consideration of how such policies articulate with office life and bureaucrats’ professional fears, aspirations, and values. It could include an analysis of documents and other artifacts produced through bureaucratic labor and how these documents move within and beyond political and bureaucratic sites, or of the social interactions that give shape to office life, such as meetings, performance reviews, and dealings with clients. The world outside of the bureaucracy—after-work gatherings, access brokers hanging around the main doors of the office, or an individual bureaucrat’s living situation—might be as relevant to an anthropologist as the world within the office walls.

Anthropologists do not merely study bureaucratic activities at the “street level”; they examine practices from the bottom to the top of organizational hierarchies, as well as those who engage organizations from the outside. Anthropological work is especially attuned to the informal aspects of bureaucratic practices, but it also examines the surprising aspects and effects of formal policies and procedures.<sup>1</sup> This chapter makes the case that, while the ethnographic approach of anthropologists might sometimes be perceived as “messy” or “unstructured,” in fact, the efforts of anthropologists are motivated by an abiding concern with empirical rigor—a refusal to ignore any sort of data or to content oneself with a single view of such a multifarious thing

as bureaucracy. This is to say that an anthropological approach is a holistic one, which envisions bureaucracy as a rich, multidimensional world.

The chapter is organized around four topics that capture this multidimensional approach, and which have been of particular focus to anthropologists. The first section, by Colin Hoag, covers the *everyday life* of bureaucratic institutions, showing why attention to informal practices and the flow of office life can be critical to understanding the workings of formal organizations. Second, Kristin Asdal and Hilde Reinertsen discuss bureaucratic documents, followed by Colin Hoag and Matthew Hull describing the significance of meetings to bureaucratic life. Finally, Josiah Heyman describes the anthropological approach to policy. Throughout, the aim is not to offer an encyclopedic account or to describe the state of the field but rather to show how anthropologists approach bureaucracy and why such an approach has value to policy makers. We hope to show why anthropological sensibilities might offer a meaningful guide for World Bank policy making and for crafting goals and guidelines that are informed by culture, power, and everyday life.

## EVERYDAY LIFE

Anthropology's early interest in the exotic and non-Western has meant that anthropologists have been relatively quiet on the topic of bureaucracy when compared with sociologists and political scientists. This has changed dramatically in recent years, partly out of dissatisfaction with accounts of powerful institutions from the outside. Commenting on research about the state in Africa, for example, the anthropologist Jean-Pierre Olivier de Sardan (2009, 39) explains, "It is easy to get the feeling that, for decades, journalists, politicians and many researchers, both Africans and Africanists, have been engaged in a relentless search for the 'essence' of the African state while neglecting to carry out a concrete analysis of the administrations, public services, bureaucratic system and relations between civil servants and the users of state services." His point is emblematic of a broader commitment among anthropologists of bureaucratic institutions to develop an empirical record of how such organizations function. Their approach seeks to understand bureaucracies as rich lifeworlds rather than organizational charts and protocols. Using anthropological methods of participant observation to understand how bureaucracies work leads to a fundamentally different picture of bureaucracy than is conventionally given by political scientists and public administrators.

A critical concept for anthropologists in general, and certainly for anthropologists of bureaucracy, is *everyday life*. The term refers to all those routine or unremarkable things that are so common they might even go unnoticed by bureaucrats, but which constitute the bulk of bureaucratic activity. Anthropology may have a reputation for caring about the exotic and about major cultural events, such as religious rituals. In fact, many anthropologists focus on the routine and quotidian aspects of culture. In the case of the civil service, anthropologists are likely to be less interested in the pronouncements of top administrators than the flow of ordinary events at the office.

These anthropologists could be said to work in the tradition of scholars of public administration and organizational ethnography from the early 20th century, such as Chester Barnard. Those scholars sought to explain the role of informal practices to the functioning of formal rules. What factors determine why bureaucracies work in the ways they do? Do bureaucrats determine institutional practice, or are they controlled by institutional rules and regulations? How do the effects of bureaucracies correspond to their stated aims? What factors other than formal rules and regulations (for example, institutional history or culture) influence bureaucrats? How do bureaucrats interpret or experience their work—the rules and reforms that guide them; the clients, bosses, or employees with whom they interact; and their own actions?

Accounting for this form of everyday life requires an immersive approach. Anthropologists employ participant observation, a method that entails joining bureaucrats in their work and recording what bureaucrats do, as opposed to simply relying on their verbal or written responses to questions on a survey or interview protocol. This shift of focus helps to remedy well-established problems associated with biased self-reporting.



Instead of recording what bureaucrats *say* they do, anthropologists attempt to record what they *actually* do. Participant observation can appear haphazard and time-consuming, and it indeed entails a substantial amount of simply “hanging around” the office. However, the method opens up an experiential understanding of bureaucracy while also allowing researchers to build rapport with the people they study.

Anthropologists seek to immerse themselves in bureaucratic worlds, aiming to understand what it feels like to do bureaucratic work, based on the assumption that this *feel*—this embodied sense of office life—shapes how the formal rules governing an organization take shape. Are the sights and sounds of an office taxing? For a bureaucrat making repetitive but consequential decisions, this might lead to inconsistency (or, by contrast, a lack of attention and nuance to a given decision). Such an approach also highlights the heterogeneity of office life, rather than presuming that *organizational culture* encompasses all social life. For example, bureaucrats’ perceptions of different spaces within an institution might differ across lines of race, class, or gender.

Anthropological approaches through participant observation strive to provide this kind of texture to account for the institution as a heterogeneous place. An employee’s career trajectory or dissatisfaction with their pay could determine how they appreciate a given policy reform. One employee interviewed during research at the Department of Home Affairs in South Africa, for example, was particularly sour after having been relegated to a position she disliked for a full year. When asked about how she viewed a departmentwide reform initiative called the “turnaround strategy,” the official replied, “But why aren’t they turning around our salaries?” In short, a policy reform that makes perfect sense in strictly public administration terms might fail if it does not account for these everyday factors.

Anthropological approaches might also extend outside the organization to conceive of how office life “spills over” into other social spaces, such as happy hours or holiday parties. By contrast, they might also be interested in describing how outside events, such as party politics, kinship structures, or even football allegiances, shape the lived experience of working at the institution. At the South African Department of Home Affairs (see Hoag 2014), the architecture and materiality of office spaces were important factors that shaped how clients accessed government services. Though rarely (if ever) referenced in government reports about reforms and service delivery, client perceptions of the opacity of the government bureaucracy owed at least partly to the fact that office counters were literally opaque: covered with taped signs and notifications. The loud din of the waiting room made it hard for them to hear the requests of the bureaucrat, and this was a source of frustration for bureaucrats. Instructions for applicants were poorly documented on department websites, meaning that applicants often required multiple trips to the office. Some solicited the support of *agents* or *immigration practitioners*, private citizens who specialized in advising applicants about the process. Bureaucrats had developed relationships with these agents, and those relationships sometimes led bureaucrats to treat applicants preferentially—for example, by excusing mistakes.

## DOCUMENTS

Curiously, among the often-overlooked parts of everyday office life are the materials with which bureaucrats work: bureaucratic *documents*.<sup>2</sup> Indeed, documents are everywhere, but despite their significance, we often speak of them in negative ways: as dusty and dull, piling up on a desk, put on a shelf, or placed in a drawer. Documents tend to symbolize inefficiency, inertia, and pointless bureaucracy. The word *paperwork* itself implies something that stands in contrast to real, hands-on, meaningful work. Yet if we dismiss documents in this way, we risk overlooking their fundamental importance. Not only are documents critical in our individual lives, paperwork is itself a defining feature of modern institutions (Asdal and Reinertsen 2022). Documents, be they physical or digital, are integral to organizational practices, shape organizational culture, and thoroughly shape and reshape our societies.<sup>3</sup>

Sometimes, documents are part of deep controversies. Just think of the reports from the United Nations Intergovernmental Panel on Climate Change (IPCC). These comprehensive reports, their production

process, and their reception and use are subject to intense public and political debate. But documents are also crucial in producing trust, consensus, and agreement. In fact, documents are often also, quite literally, agreements. Other documents, such as governmental budgets, have a perhaps less visible public role but are no less influential. A government budget organizes the political year and determines public spending on roads, schools, hospitals, and all other sectors in a given country. This attests to how documents take part in shaping society. We therefore benefit much from studying documents both *in practice* and *as practice*.

*Practice-based document analysis* (Asdal and Reinertsen 2022) has been developed precisely to capture the significance of documents, both in organizational settings and in society more broadly. This approach delineates six methodological moves for studying documents: document *sites*, document *tools*, document *work*, document *texts*, document *issues*, and document *movements*. These six methodological moves are also simultaneously analytical concepts. In adopting a practice-oriented approach, we see that intense power struggles are in fact taking place in and around documents: Who is allowed to write? Who is the document's sender? Who and what is mentioned in the document, and who and what is not? Who is the recipient, and who is allowed to handle the document? When is it important *not* to write, report, and make a document? Individuals, groups of actors, and issue components may be defined in and out of documents and the issues they concern and shape. Documents are sources of power; they provide opportunities and spaces of action. What is happening in and behind the documents? A practice-oriented document analysis aims at exploring these kinds of questions.

In the following, we will go through these six methodological moves and show how they compose a cohesive analytical framework. But just to make the point clear: even if these elements together make up a whole, this does not mean that they are the only elements or that they are always equally relevant. This depends on the object of analysis and on what one is interested in exploring and analyzing in a specific organizational setting. The different elements are partly overlapping, and they “speak” to one another. When we now go through them in sequence, we will do so by means of the recent case of the COVID-19 (coronavirus) pandemic, to illustrate the many dimensions of documents this method allows us to explore.

As these pages were being written, governments across the globe were struggling to contain the COVID-19 pandemic. Continuously updated risk analyses, swiftly prepared emergency laws, and rapidly changing travel restrictions were but three of the many forms of documents deployed in the effort. The latter two were designed to manage us, as individual citizens, to ensure that we acted in a manner that helped contain rather than spread the virus. Yet newspapers also reported that fake negative COVID-19 test scores could be bought online, enabling individuals to escape quarantine restrictions (and risk being accused of document fraud). These are but a few examples of the many documents involved in the ongoing tracking and handling of the coronavirus. So how might we go about analyzing the coronavirus and the COVID-19 pandemic by way of documents? Indeed, if we start looking at how the pandemic unfolded in practice, we will soon see that documents were involved at every turn. In the following subsections, we analyze the coronavirus and the COVID-19 pandemic by way of the six methodological moves of the practice-oriented method (Asdal and Reinertsen 2022).<sup>4</sup> In short, this will enable us to analyze and demonstrate how documents made the virus governable.

## Document Sites

Documents always exist at specific sites—such as archives, websites, organizations, and bureaucratic offices—and it matters what kinds of sites these are. To understand how bureaucratic institutions have handled the COVID-19 pandemic, we can study what happens inside government offices. Yet we may also extend this site-oriented move to the documents as such. We can consider documents as sites in themselves: sites where medical facts and political decisions are formulated, negotiated, and decided upon. They are sites to which we may go, analytically speaking, to study the pandemic. This means to ask not only “What does this document tell us?” and “What is written here?” but also “What does this document do?” (that is, “What effects or force does it have?”), “What happens here?” and “What are the practices unfolding here?” In so doing, we can, in fact, think of document analysis as a form of fieldwork—a form of document ethnography.

## Document Tools

In a dramatic political situation, such as an unfolding pandemic, what becomes blatantly clear is that documents are tools: they are produced and used for specific reasons, and they are part of larger processes, cases, and institutions. They are written and printed and distributed with the intention that something can and should happen by means of them. Just consider the following three documents: maps displaying countries and regions as green, yellow, and red to signify which travel restrictions are in place; emergency laws equipping the government with extra measures to act in times of crisis; and economic stimulus packages undergoing hard negotiations in the legislature before channeling government funds to industries and public services across the country. Color-coded maps determine who may travel where, emergency laws enable the government to act more independently from the legislature, and stimulus packages help actors and institutions endure a dire economic situation. When we start thinking about documents as tools, we become “tuned in” to investigating what role documents play in a specific situation and how their particular properties affect how they are used and how they shape bureaucratic outcomes—and by extension, societal outcomes.

## Document Work

No document miraculously emerges in its finalized form. Producing and handling documents are, in themselves, forms of labor, craft, and expertise. Documents are always part of specific work practices, including writing, editing, circulating, reading, and use. We can study the various ways in which this work is done by getting close to the people working on and with them. Examples of such document work are the preparation of weekly governmental COVID-19 contamination reports and the updating of public guidelines online. This is often a matter of collective work within larger institutions and bureaucracies. Even though the public faces of a government’s pandemic response are high-level politicians and agency leaders, a host of staffers have been involved across ministries, directorates, and agencies. Their concerted (albeit sometimes conflictual) document work, in which all are involved in drafting separate paragraphs, reports, and reviews, is what, in combination, enabled the full COVID-19 response citizens witnessed through the media.

## Document Texts

Clearly, it is essential to analyze documents as texts. This is the content, the very material, that document work concerns itself with. Paying attention to the document as text includes analyzing its genre—the textual, rhetorical, narrative, and visual properties that together make up the document in front of us. Think, for example, of the guidelines for quarantine and isolation that everyone was obliged to follow, and which have been critiqued for being hard to understand and easy to misinterpret. What rhetorical situation do the guidelines establish? How do they try to explain their topic and convince their readers? In analyzing guidelines as texts, we can look at their author, intended recipient, style, structure, layout, illustrations, graphs, numbers, and references. What are the combined effects of these elements? How do they seek to produce authority and trust? Why and how did they succeed, or not?

## Document Issues

Documents are sources of information about the specific issues in which we are interested, shaping how issues are understood and acted upon. To understand what the COVID-19 pandemic is about, we retrieve documents from the government, from researchers, and from the media. Yet documents also take an active part in forming the issue itself—as, for instance, a situation that is *under control* or *out of control*, as a *global* issue or a *national* question, as an issue that is *closed* or one which is *uncertain* and open for discussion. They act upon the issue and thus have a transformative capacity that we as analysts should not only acknowledge but actively investigate. In a special case such as the COVID-19 pandemic, this potential for intervention and

transformation is readily visible. Yet in less tangible and acute issues as well, such as the implementation of environmental regulations or nature conservation, documents are key to how issues are rendered governable and regulated.

## Document Movements

Documents are seldom lying still. They are often “on the move,” circulating within and across sites. In the case of the COVID-19 situation, this is true to the extent that new rules, regulations, stimulus packages, and the discussion thereof have moved throughout government bureaucracies ceaselessly. Moreover, the virus itself is made manageable by how it enters into documents; thus, this move into documents is what makes it accessible and “workable.” Documents build upon one another, enabling the pandemic issue to move through the government into the public—and often back again, for new iterations of research and regulation. Furthermore, patterns of document movement reflect and even constitute the effective organization of bureaucratic institutions.

Both alone and in combination, these six methodological moves make it possible to analyze documents as valuable sources for understanding the workings of bureaucratic organizations and beyond. In the case of the COVID-19 pandemic, they may help open up the material to rich analyses by getting closer to the role of bureaucratic institutions in the situation. In drawing viruses and documents together, we may thus understand them both better. This is true for any issue in which documents are involved: by better understanding how documents operate and the effects they cause within and beyond bureaucracies, we are also better equipped to notice and appreciate their significance.

## MEETINGS

Meetings are crucial sites of bureaucratic social activity. Indeed, the projects of an organization are often constituted as systems of meetings (Brown and Green 2017). However, the mundane quality of meetings tends to obscure their cultural content. Meetings have been idealized as “the locus and embodiment of ideas of appropriate, transparent decision-making” (Brown, Reed, and Yarrow 2017, p.11), but anthropologists of bureaucracy approach meetings as cultural phenomena, with particular social rules, ritual qualities, spatial and temporal framings, and consequences (Schwartzman 1989).

While analysis of meetings might initially focus attention on the content that is decided at meetings, such as the consequences or the political content of the discussion, anthropologists also pay attention to the meeting itself as a form of social engagement. As a kind of event that requires the face-to-face presence of people, meetings are a form of social life difficult to study without ethnographic work. Anthropologists recognize that organizations are unstable, in spite of their projected coherence. Employees aspire toward promotions and come and go from the organization, and their roles are often contested. An organization’s (or unit’s) goals may be inconsistent or understood differently across the hierarchy, and technologies consistently reroute the ways that people report their work or interact socially with one another. Meetings are a key site where those ambiguities are stabilized and consensus is established or enforced, even if only temporarily.

In an analysis of National Science Foundation grant application reviews, for example, meetings serve as important sites for the collective socialization of reviewers regarding how to assess applications against guidelines (Brenneis 1999). That is, they serve to guide bureaucrats in what they should take note of and value. Meetings can also be a means to generate a shared understanding of which phenomena bureaucrats should overlook, as when meetings of Mexican environmental regulators produce official ignorance of burning and firewood cutting (Mathews 2011). Similarly, in European Union policy meetings, bureaucrats succeed in crafting policy from different national member states’ interests by withholding and strategically concealing the political content of meetings (Thedvall 2013).

Because of the regularity of meetings, which often occur in specific locations at regular intervals and with predictable formats and ways of speaking, they should be understood as having a ritual quality that anthropology is well positioned to interpret. In some contexts, the performances of authority and agreement are more important than efforts to present information and establish common understanding. For example, at meetings in Lesotho between conservation bureaucrats and the rural people who were targets of their conservation efforts, bureaucrats' performance of authority and the public's performance of consent to regulation were as important as the actual transfer of information about a policy (Hoag 2022). Organizational meetings typically combine formal and informal aspects of organizations, opening in some moments to social conflict while containing it within the framework of normal business activity. Meetings often conclude with a mechanism for, if not the reconciliation of, conflict, the redirection of conflict into activities that might resolve it—often further meetings.

## POLICY

*Policy* is a governing statement by an organization, from a nation-state to a local clinic, shaping its actions in its environment, both social and biophysical.<sup>5</sup> Yet often public statements are abstract and ideal, far from actual conduct on the ground. Blame for this gap partly rests on the organization that enacts the policy. The bureaucracy fairly or unfairly takes the blame for any failure to fulfill official aspirations. Sometimes this finger-pointing holds some truth, but bureaucratic failure and limitations are by no means the only causes, and even when bureaucracy does play a role, this only raises questions of why and how it does so, and whether it might change.

It is important not to reduce issues of administration simply to questions of identifying and applying objective scientific management. While skillful management does matter, even the choice to manage for some priorities and not others is inherently a political decision. In Mexico, for example, COVID-19 vaccination has been provided first—and with great publicity—in small, rural, and often indigenous sites. This was in some ways a suboptimal decision—mass vaccination in large cities would have reached more people, more quickly, and at a lower cost. But it was explicitly a political decision to show concern for people and places that historically have been devalued and stigmatized, places that in the past, Mexican bureaucrats would have reached slowly, if at all. There is no unambiguous, scientifically right way for bureaucracies to choose their means and ends; decisions should always be understood as political.

Without delving deeply into the study of policy whys and wherefores, a few basic observations help. Even when overly ambitious or misleading, policy statements are performative and need to be understood as such. Such statements present an ideal to diverse publics (including the bureaucracy) about what should be, with the proviso that they may be taken in a wide range of ways. They may be accepted as legitimate aspirations, even if incompletely achieved. But often there is a glaring gap between formal statements of policy and the real thrusts of organizational action, covered by rhetorical adherence to the formalities of official policy.

In observations of the work of United States immigration and border officers at the US-Mexico border, officers consistently said their overriding policy goal was to interdict terrorists. But in their actual work, they never encountered potential terrorism, whereas their operations were aimed clearly at Latin American labor migrants and asylum seekers. The verbal performance of protecting the erstwhile vulnerable homeland against terror, a hard-to-question goal, justified morally debated and ambiguous duties in reality. Policy formalisms, then, have a complex relationship to actual activities on the ground.

While policy, in one regard, is those means and goals that are publicly announced, policy can also be considered as choices shaping what actually is carried out. What if the real-world implications of bureaucratic assignments and practice constitute well-understood and tacitly accepted policy that deviates from formal statements? The just-cited example of US policy at the US-Mexico border is a clear instance. This dual view of policy as rhetoric and policy as enacted requires observing the actual work routines and the



accumulated choices—to do and, importantly, not to do—of bureaucrats in performance, not taking those features as failures but as social facts in themselves. To explore these phenomena, the notion of *street-level bureaucracy* is helpful. Street-level bureaucrats are officials who interact with the public directly or carry out the actual activities of the organization. Beyond the strict definition of *street-level*, we need also to consider lower levels of management in the organization who work in close proximity to such officers. Policy, whatever is declared, always passes through the hands of street-level bureaucrats and is enacted according to their ideas and actions.

Policies, rules, and the like are inevitably written in general terms. It would be hopeless to write them to cover all possible people and situations. Instead, officials must exercise discretion in when and how to apply actions. One person might be arrested, another ignored; one person might be awarded a valued document, another denied. Discretion is not just the arbitrary impulse of willful bureaucrats. It has specific political, social, and cultural features. For example, officials who award US temporary visiting visas to Mexicans at the US-Mexico border—a desired asset for shopping and family reunions—have guidelines that aim to prevent visa misuse for unauthorized residence or work. But general guidelines need to be applied to the circumstances of diverse applicants. Many different factors enter into judgment—for example, age, with older people deemed more trustworthy. A particularly important factor of discretionary favor is wealth because people with wealth are thought to be less likely to want to cross over from Mexico to work in the United States and are generally looked on as more worthy. They dress cleanly and neatly, present themselves in polite but relaxed ways, and often speak English. That is, they resemble US officers, even outclassing them. The idea that a wealthy person does not seek unauthorized residence or employment is often correct, but by no means always; there is a notable population of wealthy unauthorized residents inside the United States, especially in border communities. Discretion, then, deserves attention as a crucial leverage point in how policy is rendered into reality by a massive aggregation of rules of thumb, interactions, discretion, and microscopic decisions.

Woven into discretionary decisions, street-level bureaucrats often ration their efforts and outcomes, whether positive or negative. They rarely have sufficient resources (personnel, time, equipment, sites, goods and funds to distribute, and so on) to enact the entire policy in all cases, either the policy as formally stated or as tacitly understood. Rather, they prioritize action and (by implication) inaction. When all individual actions are aggregated, the allocation of rewards and punishments constitutes a *de facto* policy. The bases of this allocation are complex but discernible with close attention to the words and actions of officials. Often encountered are criteria of socially interpreted personal or moral worthiness and belonging to insider versus outsider groups. Also relevant, of course, is the cost of various actions to bureaucrats and organizations in terms of scarce time, resources, and so forth.

Bureaucrats do not act in such ways alone. Rather, they interact with a wide variety of counterparties: upper executives, political bosses, publics of varying degrees of influence and respect, other organizations, rivals or collaborators, biophysical organisms, processes and flows, and so forth. Alberto Arce and Norman Long (1993) observe in rural Mexico, revealingly, systematic maneuvers and misinterpretations in the interactions between an idealistic development project engineer and agropastoralists grounded in a local tradition of defiance, mistrust, and subterfuge. A negotiated appearance of collaboration on all sides finally breaks down in project failure. This actually reinforces the interpretations on both sides perfectly. Agency managers view the breakdown as evidence that the engineer's proposed peasant-oriented innovations were wrong all along, and peasant leaders view it as another instance of incompetence and failure by central authorities.

There are wider lessons in this specific case. First, understanding bureaucrats means not only learning about their internal ideas and routines but also seeing them as diverse toolkits for interacting with counterparts, from clients through outside visitors to funders and political bosses. Each kind of interaction, when encountered, provides only partial information about an organization, since the same bureaucrat might deal very differently with another kind of counterparty. This applies, not least, to outside development experts. Second, these webs of relations are suffused with power dynamics: some clients may be deferential supplicants, others well-wired operators with higher status and better connections than bureaucrats themselves. Policy on the ground, then, must be interpreted and enacted in this diverse and unequal web of relationships.



Interpretive judgments of people, situations, and action or inaction require human intelligence. Bureaucrats might seem rigid or unresponsive, and often they are, but they rely on learned thought routines that have worked adequately for current or past sociopolitical fields. Routines only change when those fields meaningfully alter and new work patterns are available to be learned. A well-socialized official also learns to justify their work routines, favoritism, lacunae, and approximations of and deviations from formal policy through the skilled use of rationales, labels, rhetoric, and other language games. The complete web of these learned behaviors, ideas, and words is an *organizational culture*. *Culture* is a tricky term, sometimes insightful but sometimes glib, neutering inquiry. Not all organizations have strong cultures (though some do), not every organization member shares an identical culture or shares it with equal intensity, and organizational cultures are not free-floating isolates but rather are part of wider webs of inequality and power. The best reason to introduce the term *organizational culture* is that, to understand the refraction of official policy into practice, we need to take seriously the everyday working frameworks of bureaucrats. To change policy into practice, then, external mandates or sporadic trainings are not enough—workable new frameworks and reasons to use them need to be introduced. That is a very long and hard process.

## CONCLUSION

Anthropological approaches to bureaucracy are diverse, but at their heart, they include certain commonalities. First, they are holistic, seeking to understand bureaucratic work from a variety of perspectives, including those of bureaucrats differently positioned within a workplace hierarchy and across lines of race, ethnicity, class, and gender, and of the publics who interact with the organization.

Second, they are immersive, leveraging the method of participant observation to understand the look and feel of bureaucratic life, including how outcomes of bureaucratic rules are configured by the process of their execution, as well as how informal practices correspond to formal rules.

Third, they envision bureaucracies as social worlds within which a bureaucratic practice is not merely a reflection of bureaucrats' self-interest or psychology but also their socialization into an organizational culture. Whether examining the factors that inform policy implementation, the role of documents in organizing bureaucratic work, or the social role of meetings in the workplace, anthropologists aim to develop a rich account of the multitude of factors that shape how bureaucratic work is understood and carried out.

## NOTES

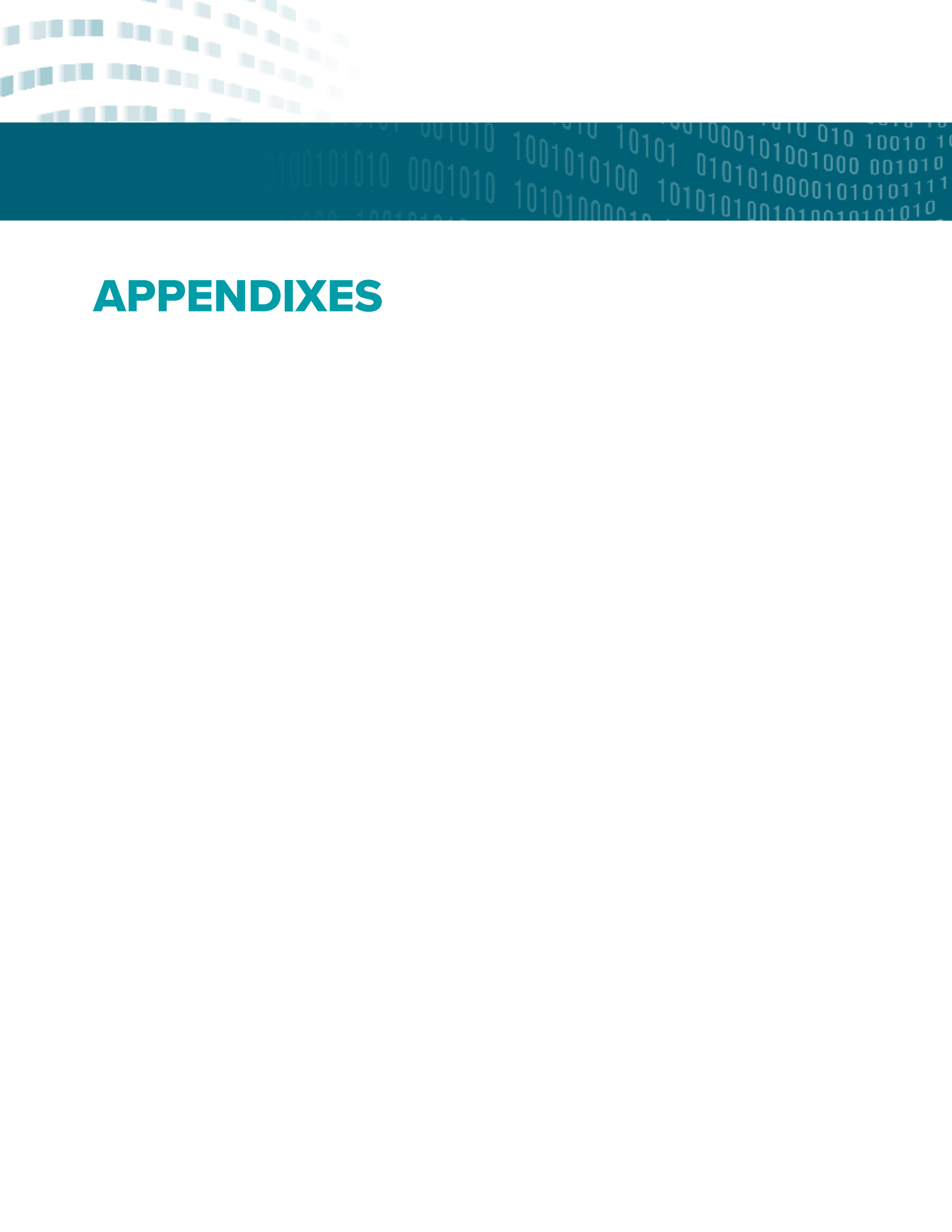
The authors are grateful to Daniel Rogger and other World Bank staff for planning and editorial guidance. The documents section was written by Kristin Asdal and Hilde Reinertsen, and we are grateful to SAGE Publications for allowing us to reprint text from Asdal and Reinertsen's book, *Doing Document Analysis: A Practice-Oriented Method* (2022). The policy section was written by Josiah Heyman. The sections on everyday life and meetings were written by Colin Hoag.

1. For overviews of the anthropological literature on bureaucracies, see Bierschenk and Olivier de Sardan (2014, 2021); Heyman (2004); Hoag and Hull (2017). The following are some key anthropological studies of bureaucracy: Ferguson (1994); Gupta (2012); Hetherington (2011); Heyman (1995); Hull (2012a); Mathur (2016); Lea (2008).
2. This section was written by Kristin Asdal and Hilde Reinertsen and builds on chapters 1 and 7 of Asdal and Reinertsen (2022); see also Asdal (2015).
3. Other instructive anthropological studies of the role of documents in bureaucracies are Hetherington (2011); Hull (2012a); Mathur (2016); Riles (2000). For a review of the anthropological literature on bureaucratic documents, see Hull (2012b).
4. These sections are based on chapter 7 of Asdal and Reinertsen (2022).
5. This section was written by Josiah Heyman.

## REFERENCES

- Arce, Alberto, and Norman Long. 1993. "Bridging Two Worlds: An Ethnography of Bureaucrat-Peasant Relations in Western Mexico." In *An Anthropological Critique of Development*, edited by Mark Hobart, 179–208. London: Routledge.
- Asdal, Kristin. 2015. "What Is the Issue? The Transformative Capacity of Documents." *Distinktion: Journal of Social Theory* 16 (1): 74–90.
- Asdal, Kristin, and Hilde Reinertsen. 2022. *Doing Document Analysis: A Practice-Oriented Method*. Thousand Oaks, CA: SAGE Publications.
- Bierschenk, Thomas, and Jean-Pierre Olivier de Sardan. 2014. "Studying the Dynamics of African Bureaucracies: An Introduction to States at Work." In *States at Work: Dynamics of African Bureaucracies*, edited by Thomas Bierschenk and Jean-Pierre Olivier de Sardan, 3–33. Boston: Brill.
- Bierschenk, Thomas, and Jean-Pierre Olivier de Sardan. 2021. "The Anthropology of Bureaucracy and Public Administration." In *The Oxford Encyclopedia of Public Administration*, edited by B. Guy Peters and Ian Thynne. Oxford, UK: Oxford University Press. *Oxford Research Encyclopedia of Politics*. <https://doi.org/10.1093/acrefore/9780190228637.013.2005>.
- Brenneis, Donald. 1999. "New Lexicon, Old Language: Negotiating the 'Global' at the National Science Foundation." In *Critical Anthropology Now: Unexpected Contexts, Shifting Constituencies, Changing Agendas*, edited by George E. Marcus, 123–46. Santa Fe, NM: School of American Research Press.
- Brown, Hannah, and Maia Green. 2017. "Demonstrating Development: Meetings as Management in Kenya's Health Sector." *Journal of the Royal Anthropological Institute* 23 (1): 45–62.
- Brown, Hannah, Adam Reed, and Thomas Yarrow. 2017. "Introduction: Towards an Ethnography of Meeting." *Journal of the Royal Anthropological Institute* 23 (1): 10–26.
- Ferguson, James. 1994. *The Anti-Politics Machine: Development, Depoliticization, and Bureaucratic Power in Lesotho*. Minneapolis: University of Minnesota Press.
- Gupta, Akhil. 2012. *Red Tape: Bureaucracy, Structural Violence, and Poverty in India*. Durham, NC: Duke University Press.
- Hetherington, Kregg. 2011. *Guerrilla Auditors: The Politics of Transparency in Neoliberal Paraguay*. Durham, NC: Duke University Press.
- Heyman, Josiah McC. 1995. "Putting Power in the Anthropology of Bureaucracy: The Immigration and Naturalization Service at the Mexico-United States Border." *Current Anthropology* 36 (2): 261–87.
- Heyman, Josiah McC. 2004. "The Anthropology of Power-Wielding Bureaucracies." *Human Organization* 63 (4): 487–500. Republished in *The Anthropology of Organisations*, edited by Alberto Corsin Jimenez, 239–54. Aldershot, UK: Ashgate, 2007.
- Hoag, Colin. 2014. "Dereliction at the South African Department of Home Affairs: Time for the Anthropology of Bureaucracy." *Critique of Anthropology* 34 (4): 410–28.
- Hoag, Colin. 2022. *The Fluvial Imagination: On Lesotho's Water-Export Economy*. Oakland, CA: University of California Press.
- Hoag, Colin, and Matthew Hull. 2017. "A Review of the Anthropological Literature on the Civil Service." Policy Research Working Paper 8081, World Bank, Washington, DC. <https://openknowledge.worldbank.org/handle/10986/26953>.
- Hull, Matthew S. 2012a. *Government of Paper: The Materiality of Bureaucracy in Urban Pakistan*. Berkeley, CA: University of California Press.
- Hull, Matthew S. 2012b. "Documents and Bureaucracy." *Annual Review of Anthropology* 41 (1): 251–67.
- Lea, Tess. 2008. *Bureaucrats and Bleeding Hearts: Indigenous Health in Northern Australia*. Sydney: University of New South Wales Press.
- Mathews, Andrew S. 2011. *Instituting Nature: Authority, Expertise, and Power in Mexican Forests*. Cambridge, MA: MIT Press.
- Mathur, Nayanika. 2016. *Paper Tiger: Law, Bureaucracy and the Developmental State in Himalayan India*. Cambridge, UK: Cambridge University Press.
- Olivier de Sardan, Jean-Pierre. 2009. "State Bureaucracy and Governance in Francophone West Africa: An Empirical Diagnosis and Historical Perspective." In *The Governance of Daily Life in Africa: Ethnographic Explorations of Public and Collective Services*, edited by Giorgio Blundo and Pierre-Yves Le Meur, 39–72. London: Brill.
- Riles, Annelise. 2000. *The Network Inside Out*. Ann Arbor, MI: University of Michigan Press.
- Schwartzman, Helen B. 1989. *The Meeting: Gatherings in Organizations and Communities*. New York: Springer.
- Thedvall, Renita. 2013. "Punctuated Entries: Doing Fieldwork in Policy Meetings in the EU." In *Organisational Anthropology: Doing Ethnography in and among Complex Organisations*, edited by Christina Garsten and Anette Nyqvist, 106–19. London: Pluto Press.





# APPENDIXES



## APPENDIX A

# Checklist for Using Expansive and Qualified Measurement for Informed Problem Solving and Learning in Public Administration

## Chapter 4 Appendix

### SUMMARY OF THE TOOLKIT

This toolkit provides a checklist for public policy practitioners who want to use expansive and qualified measurement for informed problem solving and learning. It consists of a series of four key questions that a practitioner should consider before embarking upon solving any measurement problem. Those questions correspond to the four principles for an expansive, qualified data suite, presented in chapter 4. They are ordered in a broadly sequential manner, starting from the overall organizational culture, through the adequacy of quantitative and qualitative types of data for the problem at hand, to the holistic approach to measurement and management, including long-term reforms. Each key question is complemented with a series of subquestions that pertain to more specific concerns and necessary actions, and are, where necessary, clarified with short examples. Although not all of the questions may be applicable and some may require action in certain contexts, it is crucial for a practitioner to have at least considered their relevance. Thinking through them and taking action where necessary is the key to ensuring that the measurement activity undertaken will remain closely anchored to a performance problem that matters to stakeholders. Going about the measurement task in this way should also increase the opportunities for evidence-based learning.



**TABLE A.1 Checklist for Using Expansive and Qualified Measurement for Informed Problem Solving and Learning in Public Administration**

Key question to consider	Subquestions
<b>To what extent is there an organizational culture in place that supports professional, politically insulated measurement tasks?</b>	<ul style="list-style-type: none"> <li>● Are professional principles of data collection, analysis, and curation communicated to all staff, across all levels of the organization?</li> <li>● To what extent is the necessary technical and human capital available at all points in the chain of data measurement?</li> <li>● Are staff at all levels shielded from undue political pressure?</li> <li>● Are there accountability and reporting mechanisms in place to prevent undue political pressures or other behaviors going against the work ethos?</li> <li>● To what extent are the results of data measurement communicated to the public in an open, clear, and transparent way?</li> </ul>
<b>Is measurement focused on the right indicators given the problem at hand?</b>	<ul style="list-style-type: none"> <li>● Are the quantifiable indicators operational and do they accurately measure the specific performance problem at hand?</li> <li>● Are metrics of success measuring the root performance problem (e.g., if low literacy among 10-year-olds is the problem, measuring the number of new schools built will not be sufficient)?</li> <li>● Is success judged based on the problem being fixed or only on output/process metrics (e.g., children learning vs. building schools/passing compulsory education legislation)?</li> <li>● Is the implementation process and related measurement flexible enough (where feasible) to shift if the initial data show no progress on fixing the underlying performance issue?</li> <li>● Is measurement focused in the priority areas (as set by the administration)?</li> <li>● Can most of the key factors specified in the theory of change be gauged in a quantitative manner?</li> <li>● Are there any obstacles to collecting quantitative data in a comprehensive and representative fashion (e.g., active conflict zones, natural disasters, digital poverty/illiteracy, lacking infrastructure, cultural taboos)?</li> <li>● What complementary qualitative methods might be used if the problem (or key aspects of it) are inherently nonquantifiable (e.g., norms and networks, interpersonal trust, quality of life, leadership, creativity)?</li> </ul>
<b>How might qualitative data supplement quantitative measurement so that they are as accurate a capture of intervention success as possible?</b>	<ul style="list-style-type: none"> <li>● Do qualitative data confirm that the intervention is working as intended?</li> <li>● Do qualitative data support, qualify, or refute the conclusions drawn from quantitative data? (If the latter, revisit the theory of change and both sources of data to discern the reason for any discrepancy.)</li> <li>● Are insights gained from qualitative data incorporated into the theory of change, and the inferences drawn from it?</li> <li>● Can the collection of qualitative data be institutionalized and developed in the future (either in its own right or as a prelude to collecting quantifiable data)?</li> </ul>
<b>Does management extend beyond measurement and leave space for judgment and discernment, where necessary?</b>	<ul style="list-style-type: none"> <li>● How will the choice of indicators to measure affect the incentives of the actors involved in the problem-solving process?</li> <li>● Are there any important aspects of the problem being ignored only because they cannot be quantified?</li> <li>● Are there strategies to manage the nonquantifiable aspects of the problem?</li> <li>● Are there open channels for receiving qualitative data, in particular in the form of feedback from front-line workers and intended beneficiaries of the policy at hand (e.g., focus group discussions, case studies, semi-structured interviews, review and reflection points)?</li> <li>● Is the policy implementation and evaluation process open to change based on the received quantitative and qualitative feedback?</li> <li>● Is there a movement towards delegating high levels of discretion and flexibility to front-line staff, especially in complex situations?</li> <li>● Is there a movement towards hiring or including staff with high levels of context-specific knowledge and experience?</li> <li>● Are the channels for receiving quantitative data and qualitative feedback being institutionalized?</li> <li>● Are new policy programs drawn up based on the lessons learned from quantitative and qualitative data collected and analyzed in the past?</li> </ul>

Source: Original table for this publication.

## APPENDIX B

# Framework for Evaluating the Ethics of Measuring and Tracking Public Sector Workers

## Chapter 6 Appendix

What is considered as ethical and morally right can be very context-dependent. However, there are questions that can provide general guidance on how measurement can be conducted in an ethically sound manner if asked regularly at key junctures of data collection, analysis, and reporting. The following 10 questions can be used to probe key aspects of epistemological (*what can and what do we want to know?*) and moral (*what moral considerations apply?*) concerns surrounding data collection in the public sector:

1. What do we want to know?
2. Why do we want to know this?
3. How granular does a measure need to be in order to be useful?
4. What time horizons for data collection are functionally adequate and ethically defensible?
5. To what extent do data need to be linked in order to yield useful insights?
6. Do the gains in knowledge merit data collection?
7. What safeguards exist to prevent harming public sector values, individual freedoms, and dignity?
8. Is the approach to data collection accountable and explicable?
9. Who has a say in what gets measured?
10. Are there times and opportunities built in for review?

The process of answering these questions and ensuring that the answers they provide have an impact can help to create reflective and respectful practices of measurement.

## WHAT DO WE WANT TO KNOW?

Clarifying the goal of data collection helps to narrow the amount and types of data that are collected. In many instances, questions can be answered with less data than what is feasible to be collected. For example, while it is possible to collect data on all user activity on work computers of all employees, it is questionable whether this helps to answer whether they use the technological tools made available to them productively. As another case in point, it would be possible to ask civil servants in surveys as how corrupt they rate their colleagues, but it is questionable whether this would yield data that help to stem corruption. Knowing what the goal of measurement is needs to form the start of an organization- or public-administration-wide data strategy. Where data are already being collected, answering the *what do we want to know* question can help to prune data collection efforts, which can reduce risks associated with data leakage, reidentification, and invasions of privacy.

## WHY DO WE WANT TO KNOW THIS?

It is tempting to collect data simply because they can be collected and leave the task of figuring out what to do with the data for later. Undeniably, collecting data indiscriminately holds the potential for innovation because use cases might only become apparent at a later stage. However, the usability of data is closely tied to how they were collected. Vidgen, Hindle, and Randolph (2020) propose a framework that can help public service providers map data collection efforts to values. Aside from considerations typically covered by ethics guidelines (e.g., fairness and equity), the authors encourage decision-makers to ask themselves how the solution they seek defines them as an organization and how that maps on what the organization aspires to be and become. This logic can be extended to data collection. How an organization handles data should align with aspirations of virtue. Data should be representative of the population of interest and informative to stakeholders.

## HOW GRANULAR DOES A MEASURE NEED TO BE IN ORDER TO BE USEFUL?

The *Handbook* argues for the importance of microdata. While it is a truism that more detailed data provide more information, individual case data also come with more individual-level noise and an increased risk of (re)identification. Practitioners should ask themselves whether they require, for instance, data on every individual in a team, their exact age, location, and years of employment. In many cases, broader bands such as age ranges or team-level information (coarsening) might provide enough or even better insights as individual noise is averaged out.

Where individual-level data are necessary and sensitive details must be included to provide value, statistical techniques can be used to prevent reidentification.

Apart from coarsening, data can be suppressed (e.g., only display cell values over a certain count), swapped, and infused with noise or synthetic data created that are used instead of the original data (for an instructive overview, see Schmutte and Vilhuber 2020). All techniques involve trade-offs with regards to protecting privacy and affording reproducibility. The refrain from earlier sections thus continues: technological solutions alone will not suffice; ethical data strategies will require balancing trade-offs, which should take place in a consultative and transparent manner.

## WHAT TIME HORIZONS FOR DATA COLLECTION ARE FUNCTIONALLY ADEQUATE AND ETHICALLY DEFENSIBLE?

In popular discourse, the “right to be forgotten” has been at the center of debate on the temporal dimension of data. The right to be forgotten is the concept that an individual has the right for personal data held on them to be deleted. It is enshrined in the European Union’s General Data Protection Regulation (GDPR, chapter 17) and several jurisdictions have adopted similar laws (Cofone 2020). News items on the topic often dwell on debates of people having done embarrassing things online or were featured by name in local news stories. They want their data to be removed to avoid jeopardizing future career opportunities (Selin Davis 2021) that could be endangered by public knowledge of their online pasts. While these examples might be amusing, the time dimension of data poses serious challenges for public administrations.

Storing data for longer than its immediate use mechanically increases the risk of data leakage and breaches. Given that political control over the civil service changes periodically, it is also ethically questionable whether current political principals should have access to data on noncurrent public sector employees. To provide protection for public sector employees and instill trust regarding data usage, time horizons for data storage and types of uses should thus be enshrined in measurement protocols.

However, applying one-size-fits-all data storage rules (e.g., many organizations categorically only store data for up to 3 or 5 years) to data handling implies missing out on many of the innovation- and productivity-boosting characteristics of data. For example, the Australian civil service have data on file about civil service employees going back to the 1970s. This enables them to understand how changes in structure and policy correlate with changes in workforce composition. Many governments now also use machine learning that relies on historical data. Chatbots, for instance, continuously improve their function based on data saved from previous interactions. However, if an employee requests for their data to be deleted or if the organization has a policy that data on employees should be deleted once they leave the organization, the model on which the chatbot operates needs to be retrained. This is costly in terms of manpower, time, and disruptions caused to service (e.g., see for a more technical discussion Izzo et al. 2021). For accountability, public administrations also need to plan for how it can be verified that data are deleted thoroughly and securely (Sommer et al. 2020).

Not only what, why, and how questions but also those concerning for how long must form the core of a framework for data collection in public administrations.

## TO WHAT EXTENT DO DATA NEED TO BE LINKED IN ORDER TO YIELD USEFUL INSIGHTS?

The question to what extent data need to be linked follows a similar logic as questions about the granularity of data. For instance, having a centralized human resources database for all government staff that links recruitment, performance, and development data could unlock valuable insights for staffing forecasts and the reform of recruitment and training practices. However, there is a strong argument for keeping separate addresses or health records associated with individual staff as the linking of such data increases the risks associated with data breaches or misuse of data.

Employment data can include names, addresses (both digital and physical), dates of birth, social security numbers, or other sensitive IDs. While it might seem innocuous to use such data for marketing purposes, the linking of names, addresses, and birth dates lends itself to identity fraud. Making such data public can also put staff in physical danger. Persons who protect their addresses because of previous experiences with stalking, domestic violence, or other infringements of physical safety might find that offenders can now easily find them.

Data request protocols could help to prompt administrators who want to collect or analyze data to weigh such risks against the public value that they deliver.

## DO THE GAINS IN KNOWLEDGE MERIT DATA COLLECTION?

Research ethics commonly decide whether the dangers posed by data collection are merited by weighing the potential harms and benefits. In the case of public sector employees, there are five types of gains that can coincide or be mutually exclusive: benefits can accrue to (1) the individual on whom the data are collected; (2) the organization to which the individual belongs; (3) the acting government; (4) the narrowly defined stakeholder group of the organization including user groups and politicians in charge of the relevant policy portfolio; and (5) society at large. In an ideal scenario, all five benefits are present. Often, however, trade-offs will be necessary.

Research ethics typically highlight that harm to individual participants of the research must be avoided. The only exception is where societal benefits might outweigh considerations of harm. For example, if data are not used to benefit society at large, the small infringements on employees' time, privacy, and dignity might not be justified (Resnik 2020).

Frontiers of data collection bring to light new but related conundrums. In some cases, there is a strong case for linking data (see earlier section). That is, for example, the case for being able to link attitudinal data collected via surveys with behavioral data. Behavioral data—such as whether and when a member of staff leaves an organization, gets promoted, moves teams, which cases they handled, and how well they fared—when linked with survey data can help to validate measures. It can help to understand whether patterns in sentiment observed in surveys relate to changes in employment and performance.

Importantly, how such data are collected and how they are used must figure into cost and benefit calculations. Using data on performance can be done in a manner that increases control without passing on benefits to employees. For example, finding ways of motivating employees without increases in pay should not lead to forcing them into economic insecurity. In other words, cost-benefit calculations are inherently value-based. An organization can use insights to pressure employees to become ever more efficient without regard for their physical and mental well-being and economic security. Or the organization can pass gains in efficiency on to workers, for instance, by having shorter working days or awarding extra days of leave to volunteer or to work on their pet projects.

As the next questions probe, whether adequate safeguards are in place and whether decisions can be made collectively can help to provide for more equitable and consensus-based decisions on what counts as a cost and what as a benefit.

## WHAT SAFEGUARDS EXIST TO PREVENT HARMING PUBLIC SECTOR VALUES, INDIVIDUAL FREEDOM, AND DIGNITY?

Safeguards such as anonymizing data, blurring pictures, using encryption, saving parts of a data set in different databases, keeping access to data on-site, and many more mechanisms can be used to make it less likely that data are accessed by those who should not access them or that details are disclosed that should not be disclosed. None of these strategies will be fruitful, however, if the values that they aim to safeguard are not cherished by stakeholders of the data process.

Data collection should not be regarded as a burden or breach of individual dignity and rights to privacy. Engaging with the communities on whom data are collected before this happens is thus crucial. For example, many employees might be willing to provide data on their health if they trust that the information will be

used to offer them better health care coverage and the creation of healthier working environments. The opposite might be the case if they fear that organizations will use the data to minimize costs associated with unhealthy staff, making it less likely that such staff gets hired, promoted, or retained.

Organizations can build internal ethics review boards with rotating membership or collaborate with academics or professionals who have access to external ethical review boards. Most importantly, however, checks for safeguards should be part of every step of the ethics process. This will involve ensuring that those who collect, handle, store, analyze, and report the data have internalized ethical research approaches. As tasks are shared, this might not be a straightforward matter. Fostering a climate of reflective and open exchange will be necessary to avoid turning ethical review processes into red tape. In the private sector, the benefit of fostering an ethical work culture is often justified in terms of gains in productivity or innovation—a link that is often tenuous (Riivari and Lämsä 2014). The link for public sector organizations is much clearer: fostering a culture of ethical research and data handling goes hand in hand with fostering concern for creating a positive impact on society.

## **IS THE APPROACH TO DATA COLLECTION ACCOUNTABLE AND EXPLICABLE?**

Accountability in public service should not only extend to citizens and political principals but also to the very people who run public administrations. Providing for accountability in data collection is instrumental to building trust. What data and why and how they are collected, analyzed, and used should be communicated to staff in a clear and understandable manner. As Morley et al. (2020) stress, explicability—being able to explain motives and what is done—is more important than complete transparency in cases where staff might not have the necessary technical knowledge to decipher sampling approaches, algorithms, database systems, or similar.

## **WHO HAS A SAY IN WHAT GETS MEASURED?**

Ethical challenges facing public administration are particularly protracted as they involve balancing concerns of a large and diverse group of stakeholders. If administrations yield too much to one group, they can find themselves politicized or captured by interest groups. If they adapt too little, they risk becoming technocracies. Mapping who are the communities affected by data collection efforts ought to be integral to the planning process. Stakeholder consultations have become standard practice for policy making in many countries (Gramberger 2001). However, consulting stakeholders does not equate to providing voice to them. This becomes especially evident if there are imbalances in power. For instance, senior staff working in central ministries might dominate decisions on what gets measured about casual workers employed in rural parts of the administration. This is problematic in terms of equity as much as data quality. If those working on the ground and most knowledgeable about what is to be measured are not adequately consulted, measures might miss factors that are predictive of targets such as improved performance, well-being, and staff retention.

As an extension of this question, one should also consider: who collects the data and for whom? In many cases, data collectors are also public sector employees. This means that even when formal controls for privacy and confidentiality are in place, having a team in government that knows more about other staff can skew power balances. As Resnik (2020) points out, data collectors could by accident reveal sensitive information about teams and people on whom they collect data. They might also be pressured by managers or political principals to disclose information informally.



In some cases, government data might be collected on behalf of or by an external party such as a donor. Data collection efforts conducted by the Organisation for Economic Co-operation and Development, United Nations, World Bank, other development banks, big bilateral donors, or academic institutions might introduce different types of power imbalances. Threats can manifest in terms of potential reputational damage, financial losses, or declines in credibility.

Increasing the diversity in who has a say in data collection and ensuring that the results in the end benefit society at large need to be at the core of evaluating whether the identity of the data collector poses an ethical dilemma.

## ARE THERE TIMES AND OPPORTUNITIES BUILT IN FOR REVIEW?

Decision-making power should be distributed and accountable, yet not at the expense of innovation and responsiveness to changing demands and needs of stakeholders.

Setting up data collection efforts following a time- and labor-intensive ethics review process might deter recurring review and innovation. Organizations might stick to reusing measures because they had been approved and gaining renewed consensus is seen as too burdensome and politically risky. This can lock in power imbalances and leave organizations with aging instruments, unfit to measure new developments. Instead, organizations should create data collection frameworks that allow for quick and frequent review. For instance, several public universities have developed ethics boards with rotating members, to share the burden, along with frequent but short meeting times. Applications are prefiltered by the expected gravity of ethical concerns, so those that likely need less in-depth discussion can be reviewed more quickly.

## REFERENCES

- Cofone, I. N., ed. 2020. *The Right to Be Forgotten: A Canadian and Comparative Perspective*. London: Routledge.
- Gramberger, M. 2001. *Citizens as Partners: OECD Handbook on Information, Consultation and Public Participation in Policy-Making*. Paris: OECD Publishing.
- Izzo, Z., M. A. Smart, K. Chaudhuri, and J. Zou. 2021. "Approximate Data Deletion from Machine Learning Models." *Proceedings of Machine Learning Research* 130: 2008–16. <http://proceedings.mlr.press/v130/izzo21a/izzo21a.pdf>.
- Morley, J., L. Floridi, L. Kinsey, and A. Elhalal. 2020a. "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices." *Science and Engineering Ethics* 26 (4): 2141–68.
- Resnik, D. B. 2020. "Ethical Issues in Research Involving Employees and Students." In *The Oxford Handbook of Research Ethics*, edited by Ana S. Iltis and Douglas MacKay (online edition). <https://doi.org/10.1093/oxfordhb/9780190947750.013.40>.
- Riivari, E., and A.-M. Lämsä. 2014. "Does It Pay to be Ethical? Examining the Relationship between Organisations' Ethical Culture and Innovativeness." *Journal of Business Ethics* 124 (1): 1–17.
- Schmutte, I. M., and L. Vilhuber. 2020. *Balancing Privacy and Data Usability: An Overview of Disclosure Avoidance Methods*. New Haven, CT: Yale University Press.
- Selin Davis, L. 2021. "'The Right to Be Forgotten': Should Teens' Social Media Posts Disappear as They Age?" *The Washington Post*, June 14. <https://www.washingtonpost.com/lifestyle/2021/06/14/should-what-children-post-online-come-back-haunt-them-later-life/>.
- Sommer, D. M., L. Song, S. Wagh, and P. Mittal. 2020. "Towards Probabilistic Verification of Machine Unlearning." Working paper. <https://doi.org/10.48550/arXiv.2003.04247>.
- Vidgen, R., G. Hindle, and I. Randolph. 2020. "Exploring the Ethical Implications of Business Analytics with a Business Ethics Canvas." *European Journal of Operational Research* 281 (3): 491–501.

## APPENDIX C

# Stages in the Creation of Expenditure Data

## Chapter 11 Appendix

This appendix offers a detailed description on the expenditure stages and what the stages mean for the interpretation of data. While there are some differences among countries with a commonwealth and francophone heritage, the majority of steps are comparable across countries. One notable difference across systems can be observed in the production and approval of the payment order.

### AUTHORIZATION AND APPORTIONMENT

After a budget is formulated and approved, ministries and agencies receive authorization to spend money. This can be given annually, but it is often given for shorter periods of time, such as on a quarterly basis for goods and services, or even monthly. Some systems require a daily clearance of the controllership function to say that all intended spending is consistent with intended purpose. Once the budget is approved, funds are apportioned to specific spending units. *Use as expenditure proxy:* Data on authorization or apportionment are not expenditure data. Especially in environments with lacking budget credibility, using authorization or apportionment data as proxies for expenditure is likely to be a misleading. Even in countries where aggregate budget and expenditure are close, there may be important differences at the agency level making authorization data a poor proxy.

### COMMITMENT

In the commitment stage, expenditure decisions are made. This often involves a future obligation to pay, such as placing an order or awarding a contract for the delivery of specific goods or services. The commitment only becomes a liability (obligation to pay) if these goods and services are delivered as per the contract's provisions. Payment does not have to occur within the same fiscal year, which is often the case with large investment expenditures or framework contracts. Commitments should only be made if there are associated appropriations and enough cash available to cover the cost. Financial management information systems (FMIS) typically have commitment controls built in that would block a commitment unless these preconditions are met. These controls help to avoid overspending and an accumulation of arrears. For personnel

expenditures, the commitment should correspond to the amount of compensation or contributions due. This also holds for commitments to transfers. *Usefulness as expenditure proxy*: Use of commitments can be a good proxy for spending in an accrual accounting system (see box C.1). There may be considerable differences to cash accounting, especially for potentially large capital spending or framework contracts.

### BOX C.1 Cash Accounting versus Accrual Accounting

Cash-based accounting recognizes expenditure as payments when it is paid and income when it is received. In contrast, accrual-based accounting requires that

- expenditure and liabilities are accounted for when goods and services are delivered or some other event has occurred, even if payments have not been made; and
- revenue and receivables are recorded when goods are sold or some other relevant event has occurred, even if proceeds have not been received.

The accrual basis requires governments to pay more attention to areas they may have ignored—in particular, accounting for their real assets, changes in long-term unfunded, and being conscious that an expense is created as when a service or good is received, thus not providing an artificial distinction that may give rise to payment arrears.

Sources: Potter and Diamond 1999; Premchand 1995.

Note: An in-depth treatment of accrual accounting can be found in Premchand (1995).

## VERIFICATION

Once goods and services are acquired and delivered, the goods or services rendered need to be verified against the original contract, ideally at the time of delivery. For some items, like personnel expenditures or transfers, there is no need for separate verification.

## PAYMENT ORDER

Upon the delivery and verification of goods and services, a payment order is forwarded to a public accountant who makes payments. At this stage there are important differences between francophone and anglophone budget systems. In francophone systems, there is (typically) a clear separation of duties between the authorizing officer (*ordonnateur*) and the public accountant, who decides whether or not to make a payment (a payment can be rejected due to irregularities). The public accountant does not report to the authorizing officer. Increasingly, however, spending authority has been delegated to line ministries in most francophone settings. In anglophone budget systems, on the other hand, financial control is largely assigned to line ministries, along with accountability for irregularities. The accounting officer in charge (generally the permanent secretary of a line ministry) has the authority to make expenditure commitments and issue payment orders. This approach is less cumbersome and gives more flexibility to the line ministry during budget execution.

## PAYMENT

Bills are paid upon receipt of a payment order, either by cash, check, or electronic funds transfer. Processing the transaction is generally done through the FMIS, as is all accounting and reporting. Reporting is done against all segments in the chart of accounts. *Usefulness as expenditure proxy:* The payment stage is the actual point when cash payments are made. Audited financial statements are the most credible expenditure data points. However, this will not reflect outstanding liabilities when goods and services were received but have not yet been paid for and, therefore, will not project a fully accurate financial position.

It is important to be clear about nomenclature as terms are often used interchangeably when different things are meant. It may be appropriate to draw on different stages in the expenditure chain according to context. However, regardless of what is used, it is important to be clear about what was used and what the implications for the analysis are.

## USING MICRODATA TO ASSESS THE QUALITY OF FINANCIAL MANAGEMENT

Analysis of microlevel expenditure data can also be informative about the quality of financial management in an administration. This has been a more classical lens of analysis when using expenditure data, so this appendix provides examples of such analysis in line with the discussion in this chapter.

### Producing a Detailed Transactions Profile in Pakistan

Expenditure microdata were used in Pakistan to populate a transactions profile and inform the quality of expenditure data and expenditure practices more broadly. As a federation, Pakistan has different expenditure management systems and utilization patterns across provinces. At the federal level, FMIS coverage is very low, at about 10 percent. Provincial coverage is higher, ranging from 44 to 72 percent (table C.1).

The system holds valuable information on the quality of the data and expenditure management practices that an analyst can use to inform policy.

Expenditure control is better supported at the provincial level than the federal level. At the federal level, the main business of government is being done outside of the system. This means FMIS budget and commitment controls are only applied to 10 percent of the budget, which raises important questions with regard to

**TABLE C.1** Share of FMIS System Coverage, Pakistan, 2016/17

System	Share of total expenditures (%)	Share of salaries (%)	Share of pensions (%)
Federal	9.8	35.3	11.0
Punjab	43.6	84.8	89.9
Sindh	72.2	96.1	48.9
Khyber Pakhtunkhwa	67.5	98.0	58.1
Baluchistan	68.9	93.7	90.5

Source: Hashim et al. 2019.

Note: FMIS = financial management information system.

the quality and integrity of the remaining 90 percent of the expenditure data, including the majority of salary and pension payments, which are handled manually outside the system. The low percentages for payroll and pensions at the federal level are because the defense department does not use the FMIS. The pension figures for Sindh and Khyber Pakhtunkhwa are low because not all pension payments in Sindh and Khyber Pakhtunkhwa are done through the FMIS and still follow the legacy pension payment order process, which can be brought on system.

Assessing the transaction profile in more detail reveals that most nonsalary transactions processed by the system are low value transactions (<PR\$10,000) that do not make up a large share of the budget. As such, the expenditure data reported by the FMIS and the spending it controls is less significant than the expenditure that is transacted manually. The highly skewed transactions profile (many low transactions make up a small share of the spending, and few large transactions make up the majority of spending), begs the question why large transactions are not captured. Analysis can then be done specifically on what items are missed and could be gradually introduced for better system coverage, expenditure control and data integrity. Capturing debt servicing, the capital budget and transfers were shown to make a big difference at the federal level (table C.2), despite not covering many transactions (in terms of numbers).

The same analysis at the provincial level shows that loans/transfers and debt service transactions are not routed through the FMIS and development spending is transacted through assignment accounts. Checks written by the accounting officers of these agencies are posted in the FMIS ex post. To improve this, the analysis suggests that such payments could be brought into the workflow of the FMIS instead.

Not integrating transactions into the FMIS workflow does not mean they are not reported. They are periodically posted to the FMIS on an ex post basis. This is true for spending from many agencies. While posting transactions may suffice for getting expenditure reports, they would not have been subject to adequate budget control. This means that for some expenditure (e.g., debt or even for the subsidies to the state-owned enterprises and the expenditure carried out by state accounting enterprises) no FMIS-based budget control is being exercised. While it is understood that this is supposed to be done by the various entities that are generating the transactions in an offline mode, this process leaves open the possibility that the expenditure can exceed the budget allocation and undermine the aforementioned principles.

Understanding what incentives drive countries to make use of the FMIS is important. If it is a technical problem, it can be addressed through investments to improve the quality and comprehensiveness of the data. However, it may also be that stakeholders do not want all spending to be subject to FMIS internal controls, avoid a paper trail or data provenance, and be purposefully opaque in the process, leaving them more flexibility than would otherwise be possible. This undermines expenditure data attributes discussed earlier. The FMIS budget coverage estimate (table C.1) can serve as a measure of revealed preference and it would be valuable to calculate this annually and make publicly available such that all parties can be held to account.

**TABLE C.2 Budget Items Not Routed through the FMIS at the Federal Level, Pakistan, 2016/17**

Item	Share of the budget (%)
Debt servicing (domestic and foreign)	30.6
Capital budget	18.4
Transfers and subsidies to state-owned enterprises	4.8
Loans and transfers to province and others	1.7
Postal department	0.1
Foreign affairs	0.4

Source: Hashim et al. 2019.

Note: FMIS = financial management information system.

## Searching for the Cause of Inefficient Payment Delays in Cambodia

In Cambodia, the analysis of government expenditure microdata revealed a different picture. The total amounts processed through the FMIS represent a high percentage of the total domestically financed budget for 2017. This shows that the coverage of the FMIS is relatively high (table C.3). The transactions profile also reveals that the total number of transactions processed at the national level is very small: only 8.7 percent compared to 91.3 percent at the provinces.

An analysis of the granularity of the transactions suggest that the system is predominantly used for drawing out advances from the treasury, which are then processed offline. This then constitutes a wider control problem with the payment approval process in the country. Line ministries and spending units resort to taking advances, since it takes an inordinate time to process payment requests through the treasury (estimated to take about three weeks to get a payment request processed through treasury).

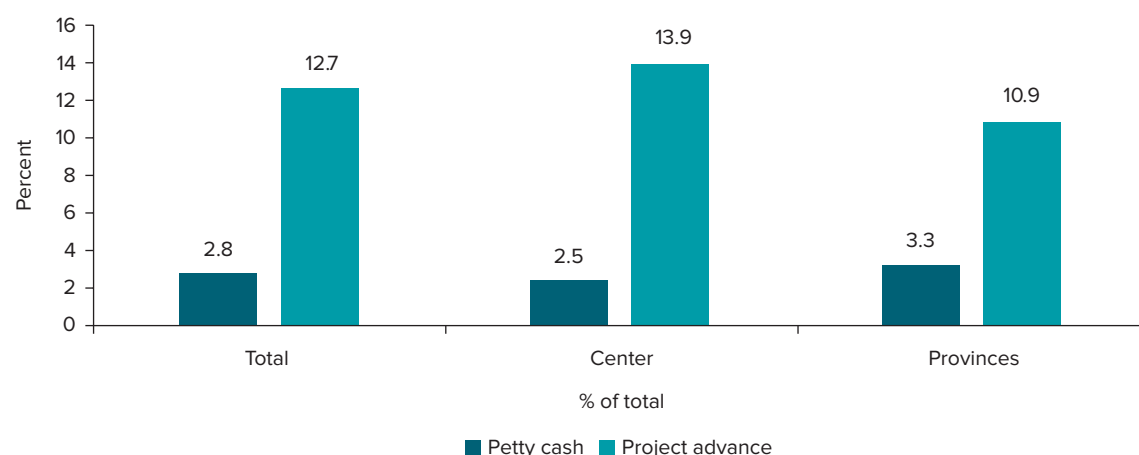
The petty cash and project advance data (figure C.1) show that the total amounts are a relatively modest share of total disbursements. The petty cash process, when limited to small amounts and a low transaction threshold, can be a good way to provide easy access to small amounts of money to spending units like schools and health clinics and thereby support local expenditure management and service delivery by extension. However, at the central level some petty cash transactions were quite large (6% of petty cash transactions were worth more than US\$100,000). At such high levels, this constitutes an accountability concern and undermines the quality of expenditure data as it becomes difficult to determine how this money was actually spent.

**TABLE C.3 Total Amount Processed through the FMIS and Values They Represent, Cambodia, 2016/17**

	Total number of transactions	Share of total number (%)	Total amount (riels)	Share of total amount (%)
National level	19,004	8.70	12,794,095	62.9
Provincial treasuries	198,470	91.30	7,530,580	37.1
Total	217,474	100	20,324,676	100

Source: Hashim et al. 2019.

**FIGURE C.1 Advances as a Share of Total Spending, Cambodia, 2016/17**



Source: Hashim et al. 2019.



In addition to petty cash advances, line agencies also receive program advances for line items that were not eligible for petty cash advances. These were mostly used for investment expenditure. The analysis showed that cash advances and program advances together made up half of the nonsalary part of the recurrent budget of many line agencies. Excessive use of these instruments points to delays and complications in the regular payment process. It also implies inefficient cash management and are a risk factor for misuse of funds. Use of these payment modalities also avoids the ex ante control of payments that is part of Cambodia's expenditure control system. It also means that the expenditure reports from the FMIS will not reflect the full expenditure at all levels before advances have been settled, making them difficult to use for informed policy analysis.

From a detailed analysis of granular transactions data, one can conclude that the process of using advances for such a large share of the recurrent spending is necessitated by an excessively onerous payment process. Agencies adopt these alternative methods to avoid any delays. However, the adoption of this coping mechanism is consequential for overall expenditure management. However, it will be difficult to avoid these issues until the root cause of inefficient payment processes is addressed.

The Cambodia case shows how an understanding of the expenditure data production can help inform problems in expenditure management and functions of government. Even though the FMIS budget coverage estimate (table C.3) was high, the large advances are an accountability and efficiency concern. Incentives for budget managers may be to perpetuate such a system as the status quo extends to them a lot of control over resources with limited oversight.

## REFERENCES

- Potter, Barry H., and Jack Diamond. 1999. *Guidelines for Public Expenditure Management*. Washington, DC: International Monetary Fund.
- Hashim, Ali, Moritz Piatti-Fünfkirchen, Winston Cole, Ammar Naqvi, Akmal Minallah, Maun Prathna, and Sokbunthoeun So. 2019. "The Use of Data Analytics Techniques to Assess the Functioning of a Government's Financial Management Information System: An Application to Pakistan and Cambodia." Policy Research Working Paper 8689, World Bank, Washington, DC.
- Premchand, Arigapudi. 1995. *Effective Government Accounting*. Washington, DC: International Monetary Fund.

# APPENDIX D

## Traditional Public Procurement Indicators

### Chapter 12 Appendix

**TABLE D.1 Public Procurement Indicators**

Indicators	Data inputs	Description
Dimension 1: Time effectiveness of the tendering process		
Preparation period (2-1) Bid submission period (3-2) Tender evaluation period (4-3) Approval period (5-4) Contract signing period (6-5) Contract effective period (7-6) Contract validity period (8-7) Contract duration (9-8)	1. Initiation date 2. Publishing date 3. Submission deadline 4. Evaluation signing date 5. Evaluation approval date 6. Contract signing date 7. Contract effective date 8. Contract validity/end date 9. Contract end/revised date	<b>Data source: Tender or contract data</b> All periods need not all be covered; however, some key periods should be included: 1. Preparation period 2. Tender evaluation 3. Contract signing period 4. Contract duration
Dimension 2: Accessibility and openness of the tendering process		
Accessibility during submission	1. Number of documents sold 2. Number of bid submissions 3. Number of responsive bids 4. Number of bids withdrawn 5. Tender security amount and validity	<b>Data source: Bid and tender data</b>
Accessibility through eligibility	1. Technical qualifications needed 2. Past experience needed 3. Bid security requirement 4. Firm turnover requirements 5. Length of eligibility criteria (number of words) 6. Length of assessment criteria (number of words)	<b>Data source: Tender requirements and qualifications</b>
Accessibility through location	1. Location of procurement 2. Regional eligibility for contract	<b>Data source: Tender data</b>
Accessibility through publication	1. Advertisement of tender in newspapers 2. Newspaper publication date 3. Advertisement of tender on website 4. Portal publication date	<b>Data source: Tender data</b>

(continues on next page)

**TABLE D.1 Public Procurement Indicators (continued)**

Indicators	Data inputs	Description
Dimension 3: Efficiency in procurement size and consolidation		
Procurement size	<ol style="list-style-type: none"> <li>1. Estimated value</li> <li>2. Lot value</li> <li>3. Number and type of lots</li> <li>4. Quantity of items in lots</li> <li>5. Unit price</li> </ol>	
Dimension 4: Procurement planning efficiency		
Expected preparation period Expected bid submission period Expected tender evaluation period Expected approval period Expected contract signing period Expected contract effective period	<ol style="list-style-type: none"> <li>1. Planned initiation date</li> <li>2. Planned publishing date</li> <li>3. Planned submission deadline</li> <li>4. Planned evaluation signing date</li> <li>5. Planned evaluation approval date</li> <li>6. Planned contract signing date</li> </ol>	<b>Data source: Annual procurement plans</b>
Expected contract validity period Expected contract duration	<ol style="list-style-type: none"> <li>1. Planned contract effective date</li> <li>2. Planned contract validity/end date</li> <li>3. Planned contract end/revised date</li> </ol>	
Budget allocation	<ol style="list-style-type: none"> <li>1. Total budget for procuring entity</li> <li>2. Total budget for procurement</li> <li>3. Planned value of procurement by item</li> </ol>	<b>Data source: Budget</b>
Dimension 5: Transparency and integrity of bidding, evaluation, and approval		
Transparency in process	<ol style="list-style-type: none"> <li>1. Time for bid preparation</li> <li>2. Use of the "emergency" exemption</li> <li>3. Opt-out of a centralized procedure</li> <li>4. Relaunch of annulled procedure</li> <li>5. Call for tender modifications</li> <li>6. Complexity of tender description</li> <li>7. Length of product description</li> <li>8. Single bidder tender</li> </ol>	<b>Data source: Tender data</b>
Transparency in evaluations	<ol style="list-style-type: none"> <li>1. Share of excluded bids</li> <li>2. Time for handling complaints</li> <li>3. Weight of nonprice evaluation criteria</li> </ol>	<b>Data source: Bid and evaluation data</b>
Transparency in awards	<ol style="list-style-type: none"> <li>1. Notification of award date</li> <li>2. Where the notification was published</li> </ol>	<b>Data source: Contract data</b>
Dimension 6: Competition and collusion		
Process	<ol style="list-style-type: none"> <li>1. Open procedure</li> </ol>	<b>Data source: Tender data</b>
Participation	<ol style="list-style-type: none"> <li>1. Number of bidders</li> <li>2. Share of incumbent bidders</li> <li>3. Share of SME bidders</li> <li>4. Share of WOE bidders</li> <li>5. Share of local bidders</li> <li>6. Share of international bidders</li> </ol>	<b>Data source: Bid data</b>
Competition	<ol style="list-style-type: none"> <li>1. Winning rebate</li> <li>2. Supplier is an incumbent firm</li> <li>3. Supplier is an SME firm</li> <li>4. Supplier is a WOE firm</li> <li>5. Supplier is a local firm</li> <li>6. Supplier is an international firm</li> </ol>	<b>Data source: Contract data</b>
Market	<ol style="list-style-type: none"> <li>1. Winning advantage of incumbent</li> <li>2. Winning advantage of local firms</li> <li>3. Market share of successful bidders</li> <li>4. Market concentration</li> </ol>	<b>Data source: Bid and contract data</b>

(continues on next page)

**TABLE D.1 Public Procurement Indicators (continued)**

Indicators	Data inputs	Description
Dimension 7: Contract implementation		
Number of variations by type	1. Contract variation date 2. Nature of variations	<b>Data source:</b> Either as wide in tender-level data or as long in variations/modifications-level data
Number of modifications/cancellations	1. Modification date 2. Cancellation date 3. Type of modification/cancellation	
Time overruns	1. Planned contract end date 2. Revised contract end date	
Cost overruns	1. Planned contract amount 2. Revised contract amount	
Data needed to show heterogeneity in indicator/outcomes		
Purchaser	1. Contracting authority name 2. Location of contracting authority 3. Contracting authority officer name and designation on tender	
Procurement	1. Category of procurement process (open, DC, RFP) 2. Type of procurement (goods, works, services) 3. Procurement product: a) Description/name of works b) Product code (e.g., CPV), if any	
Evaluation committee and approving authority	1. Name, designation, office of evaluation committee member and chair 2. Name, designation, office of approving authority	
Firms	1. Name and address of firm 2. Proprietor name, if any 3. History and status of debarment 4. Tax and registration details	

Source: Original table for this publication.

# APPENDIX E

## Tools to Assess Quality of Government Processes

### Chapter 13 Appendix

**TABLE E.1** Tool to Assess Quality of Government Process, Ghana

1. UNIT DETAILS		
(i)	Assessor name	<b>Dropdown list OR “Other” if not on list → specify name</b>        <b>Opened: Closed (if applicable):</b>
(ii)	Assessor ID	
(iii)	Date	
(iv)	Organization	
(v)	Division/unit	
(vi)	File ID/Reference number	
(vii)	File subject (as stated on cover or best guess)	
(viii)	File date coverage (as stated on cover or best guess)	

2. QUALITY OF PROCEDURE		
(i)	Can you describe what <b>percentage</b> of the following elements are complete?	
	(a) How complete is the file ladder? (Each transfer should be documented.)	1 = 0–19% 2 = 20–39% 3 = 40–59% 4 = 60–79% 5 = 80–100%
	(b) Does each step in the file ladder have dates? (Each transfer is associated with a date.)	1 = 0–19% 2 = 20–39% 3 = 40–59% 4 = 60–79% 5 = 80–100%
(ii)	Are folios within the file organized and numbered consecutively?	1 = 0–19% 2 = 20–39% 3 = 40–59% 4 = 60–79% 5 = 80–100%

(continues on next page)

**TABLE E.1 Tool to Assess Quality of Government Process, Ghana (continued)**

(iii)	Where applicable, are minutes, memos? and other necessary records?	1 = 0–19% 2 = 20–39% 3 = 40–59% 4 = 60–79% 5 = 80–100% 9 = Not applicable
(iv)	What proportion of incoming correspondence has an organizational stamp/date/ signature?	1 = 0–19% 2 = 20–39% 3 = 40–59% 4 = 60–79% 5 = 80–100% 9 = Not applicable
(v)	What proportion of outgoing correspondence has a dispatch stamp/date/ signature?	1 = 0–19% 2 = 20–39% 3 = 40–59% 4 = 60–79% 5 = 80–100% 9 = Not applicable
(vi)	Are there any of the following discrepancies in the file?  (a) Duplicates  (b) Drafts  (c) Irrelevant materials (misfiling)  (d) Miscellaneous items (including leaflets)	  1 = Yes 2 = No  1 = Yes 2 = No  1 = Yes 2 = No  1 = Yes 2 = No
(vii)	In general, to what extent does the file organization adhere to government procedure? (Give an overall score from 0 to 100.)	1 = 0–19 (worst) 2 = 20–39 (poor) 3 = 40–59 (neutral) 4 = 60–79 (good) 5 = 80–100 (best)

3. QUALITY OF CONTENT		
(i)	How would you characterize the <b>quality of content</b> you have in the file along the following margins?  (a) Background to issues     (b) Clearly outlining what courses of action are available or taken     (c) The file is organized in a logical flow (where applicable, with an issue arising, being treated consecutively, and then resolved).	  1 = Very poor 2 = Poor 3 = Neither poor nor good 4 = Good 5 = Very good  9 = Not applicable  1 = Very poor 2 = Poor 3 = Neither poor nor good 4 = Good 5 = Very good  9 = Not applicable  1 = Very poor 2 = Poor 3 = Neither poor nor good 4 = Good 5 = Very good  9 = Not applicable

(continues on next page)



**TABLE E.1 Tool to Assess Quality of Government Process, Ghana (continued)**

<b>(d)</b>	Choices are based on evidence in file.	1 = Very poor 2 = Poor 3 = Neither poor nor good 4 = Good 5 = Very good  9 = Not applicable
<b>(e)</b>	Clarity on who should take action at each stage.	1 = Very poor 2 = Poor 3 = Neither poor nor good 4 = Good 5 = Very good 8 = I don't know 9 = Not applicable
<b>(f)</b>	Proportion of relevant materials have a clear deadline.	1 = Very poor 2 = Poor 3 = Neither poor nor good 4 = Good 5 = Very good  9 = Not applicable

#### 4. HOLD-UPS BY OTHER UNITS/DIVISIONS

<b>(i)</b>	Is there evidence that other units or divisions have had to act on the file?	1 = Yes 2 = No
<b>(ii)</b>	If yes: i) specify each unit; ii) write the date transferred and received; and iii) please take a photo of the file ladder.	Unit providing input: ____ Date transferred: ____ Date received: ____ Unit providing input: ____ Date transferred: ____ Date received: ____ Unit providing input: ____ Date transferred: ____ Date received: ____

#### 5. ANSWERING THIS QUESTIONNAIRE

<b>(i)</b>	To what extent do you feel that you have all the information you need to assess the quality of decisions in the file?	<input type="checkbox"/> I have all the information I need to make a judgment on the quality of decision-making in the file. <input type="checkbox"/> I am missing information, but it is not critical to the decision. <input type="checkbox"/> I struggle to make a judgment on the quality of decision-making because of the limited information presented in the file.
<b>(ii)</b>	Did you encounter any of the following challenges in judging the quality of the file? <b>Select all that apply.</b>	1 = The file was poorly organized 2 = Little information provided in file 3 = Lack of coherence in the file 4 = Poor level of legibility 5 = Subject matter is technical/difficult to judge 6 = Other (Please, specify _____): 7 = No challenges encountered 9 = Not applicable

Source: Government of Ghana.

**THANK YOU FOR YOUR TIME AND CONSIDERATE ANSWERS**

**TABLE E.2 Tool to Assess Quality of Government Process, Liberia**

1. ADMINISTRATIVE DETAILS			
Q.#	Question	Answer options	[PROGRAMMER Instructions] / ENUMERATOR Instructions
(i)	Enumerator name		[Text]
(ii)	Enumerator ID		[Numeric]
(iii)	Date		[Date]
(iv)	File identifier/number (that contains PMS form) from front of file	01 = Number _____ 02 = Letter _____ 03 = File/folder name _____ 800 = No filing system >> Describe location of file	[Allow choice of ONE answer option. If select 01, allow numeric answer; If select 02–03, allow text; If select 800, prompt Enumerator to enter description of file location] / Enter file number, letter or file name describing location of PMS form reviewed.
(v)	Organization/ministry/ agency/institution		[Allow choice of ONE answer option. Add in predefined dropdown list]
(vi)	Department	[See Excel file for Department lists per MAC] 900 = Department unknown	[Allow choice of ONE answer option. Add in predefined dropdown list, where departments shown depend on the MAC chosen. Also include option “Department unknown”....]
(vii)	Division/unit	[See Excel file for Unit lists per MAC and Department] 700 = Other (unit not in list) >> Type unit name _____ 900 = Unit unknown	[Allow choice of ONE answer option. Add in predefined dropdown list, where units shown depend on the MAC and department combination chosen. Also include option “Unit unknown.” IF select 700, prompt Enumerator to enter Unit name.]
(viii)	Name of employee/ appraisee	<b>Other (name not in list)</b> → Type name	[Add in predefined dropdown list, where specific names shown depend on unit selected. Also include....]
(ix)	Position/job title	<b>Director/manager/supervisor/ technician/support staff</b> 800 = Job title not written	
(x)	Name of supervisor		
(xi)	Which forms have been completed for employee in 2017? <b>Tick <u>all</u> that apply</b>	01 = Form 1 – Employee performance planning and midyear progress review form 02 = Form 2 – Employee self-assessment form  03 = Form 3 – Performance appraisal form	[Allow MULTIPLE selection of listed options. ONLY show the questions 1 – xii to 1 – xvi that correspond with the answer selected here. Example: if only select option 01, then only ask Enumerator to fill.  in Q. 1 – xii and 1 – xiii.] / Select all PMS forms for the 2017 cycle that you can find for the employee. If no forms are found for an employee, then do not fill in a survey for that person.
(xii)	<b>Form 1 – Objectives and Indicators:</b> Date of employee performance planning	<b>Day / Month / Year</b> 800 = No date visible on form	[Allow date entry and “no date visible” option.] / See if date was entered anywhere on form when the first two columns (Objectives and Indicators) were completed. This should have been earlier in 2017. If no date is written, then select “No date visible on form.”

(continues on next page)

**TABLE E.2 Tool to Assess Quality of Government Process, Liberia (continued)**

(xiii)	<b>Form 1 – Progress report and assessments:</b> Date of midyear review (supervisor sign-off)	<b>Day / Month / Year</b> 800 = No date visible on form	[Allow date entry and “no date visible” option.] / See date entered at the bottom of the sheet. If only the Objectives and Indicators columns are filled in, then enter the date seen at the bottom in question (xii) and select “No date visible on form” for this question (xiii). If no date is written, then select “No date visible on form.”
(xiv)	<b>Form 2 - Employee self-assessment:</b> Review date	<b>Day / Month / Year</b> 800 = No date visible on form	[Allow date entry and “no date visible” option.] / See “Review date” at the top of the form and “Date” at the bottom. If two different dates are given, then record the “Review date.” If no date is written, then select “No date visible on form.”
(xv)	<b>Form 2 - Employee self-assessment:</b> Period under review	<b>Month / Year – Month / Year</b> 800 = No date visible on form	[Allow date entry and “no date visible” option.] / See “Period under review” at the top of the form. If no date is written, then select “No date visible on form.”
(xvi)	<b>Form 3 – Annual performance appraisal:</b> Review date	<b>Day / Month / Year</b> 800 = No date visible on form	[Allow date entry and “no date visible” option.] / See “Reviews” date at the bottom of the Part C of the form. If no date is written, then select “No date visible on form.”

[Programmer note: (xii) – (xvi) should follow if corresponding forms are ticked in (xi). Similarly, questions in section 2 corresponding to specific forms should follow if relevant forms are available.]

2. QUALITY OF PROCEDURE FOR INDIVIDUAL FORMS			
Q.#	Question	Answer options	[PROGRAMMER Instructions] / ENUMERATOR Instructions
<b>Form 1 – Employee Performance Planning and Midyear Progress Review Form</b>			
(i)	<b>Form 1 - OBJECTIVES</b> <i>EXIST: How many <u>different/unique</u> categories of objectives are on form?</i>	0 = No objectives listed → skip to question 2 – v. 1 = One objective 2 = Two objectives 3 = Three objectives 4 = Four objectives 5 or more = Five objectives or more  9 = Document is missing → skip to questions on Form 3 (2 - xiii)	[Allow choice of ONE answer option.  IF 0 is entered, then skip to Q. 2 v. IF 9 is entered, then skip to Q. 2 xiii.] / Count the number of objectives written on Form 1, in column 1.
(ii)	<b>Form 1 - OBJECTIVES</b> <i>SMART: How many of the objectives are <u>specific</u>?</i>	Number _____	[Numeric value allowed] / Count the number of objectives that are SPECIFIC in their language, then enter the number.
(iii)	<b>Form 1 - OBJECTIVES</b> <i>SMART: How many of the objectives are <u>measurable</u>?</i>	Number _____	[Numeric value allowed.] / Count the number of objectives that are MEASURABLE in their language, then enter the number.
(iv)	<b>Form 1 - OBJECTIVES</b> <i>SMART: How many of the objectives are <u>timebound</u>?</i>	Number _____	[Numeric value allowed.] / Count the number of objectives that are TIME-BOUND in their language, then enter the number.

(continues on next page)

**TABLE E.2 Tool to Assess Quality of Government Process, Liberia (continued)**

(v)	<b>Form 1 - TARGETS</b> EXIST: <i>How many lines of performance indicators are on form?</i>	<b>Number</b> _____ 0 → skip to Q. 2 xiii (questions on Form 3)	[Numeric value allowed. IF 0 is entered, then skip to Q. 2 xiii and section on Form 3] / See targets written in column 2 on Form 1, titled "Performance indicators," with one target per row. Count the number of targets written and enter the number.
(vi)	<b>Form 1 - TARGETS</b> RELATE: <i>How many performance indicators are directly linked to objectives?</i>	<b>Number</b> _____	[Numeric value allowed.] / Count the number of targets that are LINKED TO THE OBJECTIVES, then enter the number.
(vii)	<b>Form 1 - TARGETS</b> MEASURABLE: <i>How many performance indicators are measurable/can be verified by supervisor?</i>	<b>Number</b> _____	[Numeric value allowed.] / Count the number of targets that are MEASURABLE, then enter the number.
(viii)	<b>Form 1 - MIDYEAR</b> REPORTS: <i>For how many of the objectives has the employee given a progress report?</i>	<b>Number</b> _____ 800 = "Achievement progress report" column not filled in.	[Numeric value allowed and answer option 800.] / Look at column 3, "Achievement progress report." For how many of the objectives has the staff recorded their progress?
(ix)	<b>Form 1 - MIDYEAR</b> REPORTS: <i>How many of the objectives were met/achieved?</i>	<b>Number</b> _____ 800 = "Achievement progress assessment"  Column not filled in. → skip to Q. 2 xi	[Numeric value allowed and answer option 800. IF answers 800, then skip to Q. 2 – xi on development needs.] / Read the supervisor's comments in column 4, "Achievement progress assessment," and count how many objectives he/she says that the staff <u>ACHIEVED</u> the <u>OBJECTIVE</u> .
(x)	<b>Form 1 - MIDYEAR</b> REPORTS: <i>Did the supervisor give employee recommendations on how to meet their objective?</i>	1 = Yes 2 = No	[Allow choice of ONE answer option.] / Read the supervisor's comments in column 4, "Achievement progress assessment." Does he/she give the staff any <u>ADVICE</u> on how to achieve the objective or improve their work?
(xi)	<b>Form 1 - MIDYEAR</b> REPORTS: <i>Did the supervisor identify the development needs of the employee?</i>	1 = Yes 2 = No	[Allow choice of ONE answer option.] / Look at section 2, column 1 "Development needs." Did the supervisor fill in this column, listing the skills that the staff needs to develop?
(xii)	<b>Form 1 - MIDYEAR</b> REPORTS: <i>Did the supervisor recommend HOW (activities) to build the employee's capacity?</i>	1 = Yes 2 = No	[Allow choice of ONE answer option.] / Look at section 2, column 2, "Capacity-building activities." Did the supervisor fill in this column, advising the staff HOW they should develop needed skills?

(continues on next page)

**TABLE E.2 Tool to Assess Quality of Government Process, Liberia (continued)**

Form 3 – Performance Appraisal Form			
(xiii)	<b>Form 3 - ENDYEAR REPORTS:</b> Annual appraisal reports on objectives in performance plan (and provides ratings of objectives)	0 = No objectives reported on 1 = Small proportion of objectives reported on 2 = Half of objectives reported on 3 = Most objectives reported on 4 = All objectives reported on  9 = Document is missing → skip to questions on ratings (2 – xvii)	[Allow choice of ONE answer option. IF 9 is entered, then skip to Q. 2 xvii.] / Look at the objectives written on Form 1, column 1 and compare these to the objectives reviewed on Form 3, column 1, “Key objectives.” Of the objectives originally listed on Form 1, please document to what extent progress on these objectives are reported on in Form 3.
(xiv)	<b>Form 3 - FEEDBACK QUALITY:</b> Comments are <b>substantive</b> and <b>provide a quality assessment</b> of officers contributions (even if discussion is that officer had to do work not in key objectives)	0 = No comments substantive 1 = Small proportion of comments substantive 2 = Half of comments substantive 3 = Most comments substantive 4 = All comments substantive  800 = Section 1, column 4; section 3 and Part C “remarks” are not filled in.	[Allow choice of ONE answer option. IF select 800, then skip to Q. 2 xvi.] / Look at section 1, column 4 (“Achievement assessment”) and Part C (“General remarks”) – I. What proportion of their comments did the supervisor give the <u>staff</u> feedback on <u>how well/poorly</u> they did their work?
(xv)	<b>Form 3 - CONSTRUCTIVE FEEDBACK:</b> Annual appraisal provides feedback on areas of weakness (development needs/ capacity-building needs) and if these have been addressed	0 = No comments constructive 1 = Small proportion of comments constructive 2 = Half of comments constructive 3 = Most comments constructive 4 = All comments constructive	[Allow choice of ONE answer option. IF 9 is entered, then skip to Q. 2 xvii.] / Look at section 1, column 4 (“Achievement assessment”) and Part C (“General remarks”) – I. What proportion of their comments did the supervisor give the staff feedback on what <u>skills and behavior as a worker</u> they need to improve upon?
Overview of Forms 1–3			
(xvi)	Are all compulsory sections in all <u>available</u> forms filled out?	1 = Yes 2 = No	[Allow choice of ONE answer option] / Document if you have recorded all the information requested in this survey, from Forms 1–3; based on the PMS forms that staff had filled in.
(xvii)	What ratings are given to the achievements?  (a) Midyear review	Key objective 1 = Key objective 2 = Key objective 3 = Key objective 4 = Key objective 5 = Key objective 6 = Key objective 7 = Key objective 8 = Key objective 9 = Key objective 10 =	[Allow numeric entry next to EACH of the objectives identified in Q. 2 i IF 1 was selected in Q. 2 i → show only one objective. IF “5 or more” was selected in Q. 2 i → ask the Enumerator to enter scores for 10 objectives, but only make it mandatory for the first 5.] / Look at Form 1, section 1 table, column 5 “Rating,” and enter the score between 1 and 5 given for each objective. IF there are, e.g., 6 objectives but the tablet asks you for 10, then enter the ratings given for the first 6 and skip the remaining objectives.

(continues on next page)

**TABLE E.2 Tool to Assess Quality of Government Process, Liberia (continued)**

	(b) Annual appraisal	Key objective 1 = Key objective 2 = Key objective 3 = Key objective 4 = Key objective 5 = Key objective 6 = Key objective 7 = Key objective 8 = Key objective 9 = Key objective 10 =	[Allow numeric entry next to EACH.] / Look at Form 3, section 1 table, column 5, "Rating," and enter the score given for each objective. IF there are, e.g., 6 objectives but the tablet asks you for 10, then enter the ratings given for the first 6 and skip the remaining objectives.
--	----------------------	---	---

3. SCAN CONFIRMATION			
Q.#	Question	Answer options	[PROGRAMMER Instructions] / ENUMERATOR Instructions
(i)	Have you taken a picture of all available Forms 1–3 for the employee?	1 = Yes 2 = No	

[PROGRAMMER NOTE: If respondent responds "No" to 4.1, please state "Please take a picture of the appraisal form where possible."]

4. ANSWERING THIS QUESTIONNAIRE			
Q.#	Question	Answer options	[PROGRAMMER Instructions]
(i)	To what extent do you feel that you have all the information you need to assess the quality of this form?	<input type="checkbox"/> I have all the information I need to make a judgement on the quality of this form. <input type="checkbox"/> I am missing information, but it is not critical to the decision. <input type="checkbox"/> I struggle to make a judgement on the quality of this form because of the limited information presented in the form/file.	[Allow choice of ONE answer option]
(ii)	Did you encounter any of the following challenges in judging the quality of the form?	1 = The form was poorly organized 2 = Little information provided in file 3 = Lack of coherence in the form 4 = Poor level of legibility 5 = Subject matter is technical/difficult to judge 6 = Some pages were missing 7 = Other (Please specify: _____) 8 = <b>No challenges encountered</b>	[Allow choice of MULTIPLE answer options. IF select 7, then prompt for written comment on challenges faced. IF select 8, then ]

**THANK YOU FOR YOUR TIME AND CONSIDERATE ANSWERS**



**TABLE E.2 Tool to Assess Quality of Government Process, Liberia (continued)**

**[Programmer note:**

For each form, draw a random uniform number (0,1) and for  $X < 0.25$  prompt Enumerator to undertake module “VAL” as follows: “This appraisee has been randomly chosen for a follow-up interview. Please go to the corresponding unit and ask to speak with them.”]

VAL. APPRAISAL VALIDATION			
Q.#	Question	Answer options	[PROGRAMMER Instructions] / ENUMERATOR Instructions
<b>FIND:</b> Please go to the corresponding unit and ask to speak with appraisee (NOT supervisor)			
0.	Were you able to find employee/appraisee?	1 = Yes 2 = No → <b>END INTERVIEW</b>	[Allow choice of ONE answer option]
<b>INTRODUCTION</b> Please state, “We are an independent research team working with the CSA, LIPA, and GC to understand the state of the appraisal process in this organization. We wanted to ask you some brief questions about your experience of the appraisal process over the past year.”			
1	Would you be willing to answer a few questions?	1 = Yes → skip to question (3.i) 2 = No → skip to question (2)	[Allow choice of ONE answer option IF select 1 → skip to Q. VAL 3 i IF select 2 → skip to Q. VAL 2]
2	Would you mind telling us why you do not want to answer a few questions on your experience of the PMS?	1 = I do not have the time to do it 2 = Not interested in the goal of this project 3 = Do not feel comfortable to answer questions 4 = Other, specify _____	[Allow choice of ONE answer option]
3.i	Were you involved in any appraisal or review process in the past year?	1 = Yes → skip to question (3.iii) 2 = No	[Allow choice of ONE answer option IF select 1 → skip to Q. VAL 3 iii]
3.ii	Can I confirm that your supervisor has not undertaken any planning or review of your work program with you?	1 = No review process at all → <b>END INTERVIEW</b> 2 = Some review process (though it may not be understood as appraisal process)	[Allow choice of ONE answer option IF select 1 → End Interview]
3.iii	Did you have at least one meeting with your supervisor regarding how your work has been going?	1 = Yes 2 = No → skip to question (3.v)	[Allow choice of ONE answer option IF select 2 → skip to Q. VAL 3 v]
3.iv	Roughly when was this meeting?	Month/Year	[Date]
3.v	Did you receive constructive feedback?	1 = Yes 2 = No	[Allow choice of ONE answer option] / If respondent seems unsure of what “constructive feedback” means, then help clarify this with some examples.
3.vi	Did you agree on a capacity-building program where relevant?	1 = Yes 2 = No 9 = Not applicable. None was needed.	[Allow choice of ONE answer option] / ONLY select “Not applicable” if respondent says that there was no area where they needed to improve.
3.vii	Are there any other notes you'd like to make about the appraisal process of this unit? Your answers will be kept confidential.		[Text]

**THANK YOU FOR YOUR TIME AND CONSIDERATE ANSWERS**

**TABLE E.2 Tool to Assess Quality of Government Process, Liberia (continued)**

X. UNIT-LEVEL ASSESSMENT (to be done ONCE for each UNIT after all individual appraisal forms have been reviewed)			
Q.#	Question	Answer options	[PROGRAMMER Instructions] / ENUMERATOR Instructions
<b>CHECKLIST:</b> For this unit, have you: - reviewed all available appraisal forms? - taken pictures of all available appraisal forms? - planned visits to the unit to discuss the appraisal process with a random set of officials?			
1	Organization/ministry/agency/institution	[Dropdown list]	[Allow choice of ONE answer option. Add in predefined dropdown list]
2	Division/unit	[Dropdown list] 700 = Other (unit not in list) >> Type unit name _____ 900 = Unit unknown	[Allow choice of ONE answer option. Add in predefined dropdown list, where units shown depend on the MAC and department combination chosen. Also include option "Unit unknown." IF select 700, prompt Enumerator to enter unit name.]
<b>(i) When reviewing the whole set of appraisal forms for a unit/all those filled in by an appraiser, were there any of the following discrepancies in the set of appraisal forms for the unit?</b> Note: large units sometimes include multiple appraisers. When answering the questions below, consider similarities (etc.) between all appraisal forms filled in by the same appraiser.			
(a)	All appraisal forms give the same scores	1 = Yes 2 = No	[Allow choice of ONE answer option]
(b)	Limited distribution of appraisal scores	1 = Yes 2 = No	[Allow choice of ONE answer option]
(c)	Section 1 <b>objectives</b> are very similar across forms	1 = Yes 2 = No	[Allow choice of ONE answer option]
(d)	Section 1 <b>employee achievement</b> notes are very similar across forms	1 = Yes 2 = No	[Allow choice of ONE answer option]
(e)	Section 1 <b>employee achievement</b> notes are in same handwriting across forms	1 = Yes 2 = No	[Allow choice of ONE answer option]
(ii)	Are there any other notes you'd like to make about the set of appraisal forms of this unit?		[Text]

Source: Government of Liberia.

**THANK YOU FOR YOUR TIME AND CONSIDERATE ANSWERS**

# APPENDIX F

## Further Details of Analysis

### Chapter 15 Appendix

#### APPENDIX F.1 DETAILS ON TEXT ANALYSIS IN BEST, HJORT, AND SZAKONYI (2017)

This appendix provides some details of the procedure used in Best, Hjort, and Szakonyi (2017) to categorize procurement purchases into groups of homogeneous products. They proceed in three steps. First, they transform the raw product descriptions in our data into vectors of word tokens to be used as input data in the subsequent steps. Second, they develop a transfer learning procedure to use product descriptions and their corresponding Harmonized System product codes in data on the universe of Russian imports and exports to train a classification algorithm to assign product codes to product descriptions. They then apply this algorithm to the product descriptions in the procurement data. Third, for product descriptions that are not successfully classified in the second step, either because the goods are nontraded or because the product description is insufficiently specific, they develop a clustering algorithm to group product descriptions into clusters of similar descriptions.

Once the data are grouped into products, they create the main outcome of interest unit prices in three steps. First, they standardize all units to be in SI units (e.g., convert all lengths to meters). Second, for each good, they keep only the most frequent standardized units (i.e., if a good is usually purchased by weight and sometimes by volume, they keep only purchases by weight). Third, they drop the top and bottom 5% of the unit prices for each good since in some cases the number of units purchased is off by an order of magnitude, spuriously creating very large or very small unit prices due to measurement error in the quantity purchased.

#### Preparing Text Data

The first step of the procedure “tokenizes” the sentences that will be used as inputs for the rest of the procedure. They use two data sets of product descriptions: (1) the universe of customs declarations on imports and exports to and from Russia in 2011–13; and (2) the product descriptions in the procurement data. Each product description is parsed in the following way, using the Russian libraries for Python’s Natural Language Toolkit<sup>1</sup>:

1. Stop words are removed that are not core to the meaning of the sentence, such as “the,” “and,” and “a.”
2. The remaining words are lemmatized, converting all cases of the same word into the same “lemma” or stem. For example, “potatoes” becomes “potato.”
3. Lemmas two letters or shorter are removed.

We refer to the result as the *tokenized* sentence. For example, the product description “NV-Print Cartridge for the Canon LBP 2010B Printer” would be broken into the following tokens: [cartridge, NV-Print, printer, Canon, LBP, 3010B]. Similarly, the product description “sodium bicarbonate - solution for infusion 5%, 200 ml” would result in the following tokens: [sodium, bicarbonate, solution, infusion, 5%, 200 ml].

## Classification

In the second step of the procedure, they train a classification algorithm to label each of the sentences in the customs data with one of the  $H_C$  labels in the set of labels in the customs data set,  $H_C$ . To prepare the input data, each of the  $N_C$  tokenized sentences  $\mathbf{t}_i$  in the customs data set is transformed into a vector of token indicators and indicators for each possible bigram (word-pair), denoted by  $\mathbf{X}_i \in \chi_C$ . Each sentence also has a corresponding good classification  $g_i \in g_C$ , so we can represent the customs data as the pair  $\{\mathbf{X}_C, \mathbf{g}_C\}$  and we seek to find a classifier  $\hat{g}_C(\mathbf{x}) : \chi_C \rightarrow H_C$  that assigns every text vector  $\mathbf{x}$  to a product code.

As is common in the literature, rather than solving this multiclass classification problem in a single step, they pursue a “one-versus-all” approach and reduce the problem of choosing among  $G$  possible good classifications to  $G_C$  binary choices between a single good and all other goods, and then combine them (Rifkin and Klautau 2004). We do this separately for each 2-digit product category. Each of the  $G_C$  binary classification algorithms generates a prediction  $p_g(\mathbf{x}_i)$ , for whether sentence  $i$  should be classified as good  $g$ . They then classify each sentence as the good with the highest predicted value:

$$\hat{g}_C(\mathbf{x}_i) = \arg \max_{g \in g_C} p_g(\mathbf{x}_i). \quad (\text{F.1.1})$$

Each binary classifier is a logistic regression solving

$$\min_{\mathbf{w}_g, a_g} \frac{1}{N_C} \sum_{i=1}^{N_C} \frac{1}{\ln 2} \ln \left( 1 + e^{-y_{gi} \cdot (\mathbf{w}_g \cdot \mathbf{x}_i + a_g)} \right), \quad (\text{F.1.2})$$

where

$$y_{gi} = \begin{cases} 1 & \text{if } g_i = g \\ -1 & \text{otherwise} \end{cases}.$$

The minimands  $\hat{\mathbf{w}}_g$  and  $\hat{a}_g$  are then used to compute  $p_g(\mathbf{x}_i) = \hat{\mathbf{w}}_g \cdot \mathbf{x}_i + \hat{a}_g$  with which the final classification is formed using equation (F.1.1). The procedure is implemented using the Vowpal Wabbit library for Python.<sup>2</sup> This simple procedure is remarkably effective; when trained on a randomly selected half of the customs data and then implemented on the remaining data for validation, the classifications are correct 95% of the time.

Having trained the algorithm on the customs data set, they apply it to the procurement data set wherever possible. This is known as transfer learning (see, for example, Torrey and Shavlik 2009). Following the terminology of Pan and Yang (2010), the algorithm  $g_C$  performs the task  $T_c = \{H_C, g_C(\cdot)\}$ , learning the function  $G_C(\cdot)$  that maps from observed sentence data  $X$  to the set of possible customs labels  $G_C$ . The algorithm was trained in the domain  $D_C = \{\chi_C, F(\mathbf{X})\}$ , where  $F(\mathbf{X})$  is the probability distribution of  $\mathbf{X}$ . This algorithm then

needs to be transferred to the domain of the procurement data set,  $DB = \{X_B, F(X)\}$ , so that it can perform the task  $T_B = \{H_B, gB(\cdot)\}$ .

The function to be learned and the set of possible words used are unlikely to differ between the two domains—a sentence that is used to describe a ball bearing in the customs data will also describe a ball bearing in the procurement data, so  $\chi_C = \chi_B$  and  $h_C(\cdot) = h_B(\cdot)$ . The two key issues are, first, that the likelihoods that sentences are used are different in the two samples, so that  $F(\mathbf{X})_C \neq F(\mathbf{X})_B$ . This could be because, for example, the ways that importers and exporters describe a given good differs from the way public procurement officials and their suppliers describe that same good. In particular, the procurement sentences are sometimes not as precise as those used in the trade data. The second issue is that the set of goods that appear in the customs data differs from the goods in the procurement data, so that  $H_C \neq H_B$ . This comes about because nontraded goods will not appear in the customs data but may still appear in the procurement data.

To deal with these issues, they identify the sentences in the procurement data that are unlikely to have been correctly classified by  $\hat{h}_c$  and instead group them into goods using the clustering procedure described in the following section. They construct two measures of the likelihood that a sentence is correctly classified. First, the predicted value of the sentence's classification  $\hat{g}_C(\mathbf{x}_i)$  as defined in (F.1.1). Second, the similarity between the sentence and the average sentence with the sentence's assigned classification in the *customs* data used to train the classifier.

To identify outlier sentences, they take the tokenized sentences that have been labeled as good  $g$ ,  $\mathbf{t}_g = \{\mathbf{t}_i : \hat{g}_C(\mathbf{x}_i) = g\}$  and transform them into vectors of indicators for the tokens  $\mathbf{v}_{gi}$ .<sup>3</sup> For each good, they then calculate the mean sentence vector in the customs data as  $\mathbf{v}_g^C = \sum_{\mathbf{v}_{gi}, \mathbf{x}_i \in \mathbf{X}^C : \mathbf{v}_{gi} = \mathbf{t}_i} \mathbf{v}_{gi}$ . Then, to identify outlier sentences in the procurement data, they calculate each sentence's normalized cosine similarity with the good's mean vector,

$$\theta_{gi} = \frac{\bar{S}_g - s(\mathbf{v}_{gi}, \mathbf{v}_g)}{\bar{S}_g}, \quad (\text{F.1.3})$$

$$\text{where } s(\mathbf{v}_{gi}, \mathbf{v}_g) \equiv \cos(\mathbf{v}_{gi}, \mathbf{v}_g) = \frac{\mathbf{v}_{gi} \cdot \mathbf{v}_g}{\|\mathbf{v}_{gi}\| \|\mathbf{v}_g\|} = \frac{\sum_{k=1}^{K_g} t_{gik} t_{gk}}{\sqrt{\sum_{k=1}^{K_g} t_{gik}^2} \sqrt{\sum_{k=1}^{K_g} t_{gk}^2}}$$

is the cosine similarity of the sentence vector  $\mathbf{v}_{gi}$  with its good mean  $\mathbf{v}_g$ .<sup>4</sup>  $K_g$  is the number of tokens used in descriptions of good  $g$ , and  $\bar{S}_g = \sum_{i=1}^{|\mathbf{t}_g|} s(\mathbf{v}_{gi}, \mathbf{v}_g)$  is the mean of good  $g$ 's sentence cosine similarities. Sentences are deemed to be correctly classified if their predicted value  $\hat{g}_C(\mathbf{x}_i)$  was above the median and their normalized cosine similarity  $\theta_{gi}$  was above the median.

## Clustering

The third step of the procedure takes the misclassified sentences from the classification step and groups them into clusters of similar sentences. These clusters are then used as the good classification for this group of purchases. To perform this clustering, we use the popular k-means method. This method groups the tokenized sentences into  $k$  clusters by finding a centroid  $C_k$  for each cluster to minimize the sum of squared distances between the sentences and their group's centroid. That is, it solves

$$\min_c \sum_{i=1}^N \|f(\mathbf{c}, \mathbf{t}_i) - \mathbf{t}_i\|^2, \quad (\text{F.1.4})$$

where  $f(c, \mathbf{t}_i)$  returns the closest centroid to  $\mathbf{t}_i$ . To speed up the clustering on such a large dataset the algorithm is implemented by mini-batch k-means. Mini-batch k-means iterates over random subsamples (in this case of size 500) to minimize computation time. In each iteration, each sentence is assigned to its closest centroid, and then the centroids are updated by taking a convex combination of the sentence and its centroid, with a weight on the sentence that converges to zero as the algorithm progresses (see Sculley 2010 for details).

The key parameter choice for the clustering exercise is  $k$ , the number of clusters to group the sentences into. As is common in the literature, this is made using the silhouette coefficient. For each sentence, its silhouette coefficient is given by

$$\eta(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}}, \quad (\text{F.1.5})$$

where  $a(i)$  is the average distance between sentence  $i$  and the other sentences in the same cluster, and  $b(i)$  is the average distance between sentence  $i$  and the sentences in the nearest cluster to sentence  $i$ 's cluster. A high value of the silhouette coefficient indicates that the sentence is well clustered: it is close to the sentences in its cluster and far from the sentences in the nearest cluster. They start by using a  $k$  of 300 for each 2-digit product category. For 2-digit product categories with an average silhouette coefficient larger than the overall average silhouette coefficient, they tried  $k \in \{250, 200, 150, 100, 50, 25, 10, 7\}$ , while for product categories with a lower than average silhouette coefficient they tried  $k \in \{350, 400, 450, 500, 550, 600, 650, 700, 750, 800, 850, 900, 950, 1,000\}$  until the average silhouette score was equalized across 2-digit product codes.

## APPENDIX F.2 CONSTRUCTION OF ITEM VARIETY MEASURES IN BANDIERA ET AL. (2021)

This appendix describes the methods used to construct the item variety measures used in Bandiera et al. (2021). The idea behind the methods is to use data that allow us to hold constant all the features of the good that can affect its price in the control group.

The first three measures address these issues through manual grouping of attributes and using hedonic regressions to reduce the dimensionality of the measures. They begin by manually grouping attributes to ensure common support and avoid overfitting. Most of the attributes used are categorical and so they group values. For values that occur less than three times in the experiment's control group or only in the treatment group, they either group them together with similar values (using contextual knowledge and extensive googling to find similar values) or if similar values are not available, set them to missing. Observations with all attributes missing after this cleaning are dropped. Ensuring that each group appears at least three times avoids overfitting, and ensuring that the groups are observed in both the control and treatment groups ensures common support. These groups then form the  $X_{igto}$  controls used in the hedonic regressions (??).

The fourth measure, *machine learning*, develops a variant of a random forest algorithm to allow for non-linearities and interactions between attributes that the hedonic regression ?? rules out and also to perform the grouping of attributes' values in a data-driven way. For this, they do much lighter cleaning of the data only harmonizing spellings. They then train a random forest algorithm for each item, averaging 500 trees to form predicted prices.

After training each tree in the control group, the algorithm places each observation in the treatment groups into its corresponding leaf. It first places all treatment group observations that only have attributes that are sufficient to place it into a unique leaf in the tree. Then, for observations that have an attribute that



prevents it from being placed into a leaf, the algorithm selects all leaves the observation could be placed into given the attributes that *can* be used, and then for each attribute that cannot be used, replaces that attribute with the category in the same treatment group with the closest average, but that does appear in the control group. Once every observation is placed into a leaf, the average price among control group observations in the leaf is then that tree's predicted price. Averaging the 500 trees generates the machine learning measure of item variety.

## NOTES

1. Documentation on the Natural Language Toolkit (NLTK) can be found at <http://www.nltk.org/>.
2. See <http://hunch.net/~vw/>.
3. Note that these vectors differ from the inputs  $X_i$  to the classifier in two ways. First, they are specific to a certain good, and second, they omit bigrams of the tokens.
4. Note that the cosine similarity ranges from 0 to 1, with 0 being orthogonal vectors and 1 indicating vectors pointing in the same direction.

## REFERENCES

- Bandiera, Oriana, Michael Carlos Best, Adnan Qadir Khan, and Andrea Prat. 2021. "The Allocation of Authority in Organizations: A Field Experiment with Bureaucrats." *Quarterly Journal of Economics* 136 (4): 2195–242. <https://doi.org/10.1093/qje/qjab029>.
- Best, Michael Carlos, Jonas Hjort, and David Szakonyi. 2017. "Individuals and Organizations as Sources of State Effectiveness." NBER Working Paper 23350, National Bureau of Economic Research, Cambridge, MA. <https://doi.org/10.3386/w23350>.
- Pan, Sinno Jialin, and Qiang Yang. 2010. "A Survey on Transfer Learning." *IEEE Transactions on Knowledge and Data Engineering* 22 (10): 1345–59.
- Rifkin, Ryan, and Aldebaro Klautau. 2004. "In Defense of One-Vs-All Classification." *Journal of Machine Learning Research* 5: 101–41.
- Sculley, D. "Web-Scale K-Means Clustering." In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*, 1177–78. New York: Association for Computing Machinery. <https://doi.org/10.1145/1772690.1772862>.
- Torrey, L., and J. Shavlik. 2010. "Transfer Learning." In: *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, 242–64. Hershey, PA: IGI Global. <https://doi.org/10.4018/978-1-60566-766-9.ch011>.

# APPENDIX G

## Further Details of Analysis

### Chapter 19 Appendix

#### A FORMAL TEST OF THE MODAL DIFFERENCE BY ORGANIZATION

Further to the analysis summarized in figure 19.3 in chapter 19, we conduct more formal tests by fitting an ordinary least squares (OLS) model with survey mode and organization dummies as predictors and then using  $F$ -tests to compare to the same model where the mode and organization variables are interacted. We find that interaction leads to statistically significant improvement of model's fit across all three indices. The results confirm what the descriptive analysis suggested: some organizations are significantly more prone to mode effects. For example, all five organizations with the largest mode differences for each index (as shown in figure 19.4 in chapter 19) have interaction effects that are significant at the 5 percent level. Fitting a mixed model with random effects for survey mode for each organization also confirms that the mode effects are indeed present and that their strength differs across organizations (see figure G.3).

#### FURTHER INVESTIGATIONS INTO THE ROLE OF INDIVIDUAL CHARACTERISTICS

The findings outlined in chapter 19 suggest that certain demographic characteristics may make individuals more susceptible to mode effects on certain survey topics but not others. However, the scarcity of clear patterns for the control variables and the low overall explanatory power indicated by  $R^2$  suggests that unobservable characteristics may play a significant role in the degree to which that individual experiences survey mode effects. To further assert the role of mode effects and remove the concern of spuriousness due to unobservable variables, we retested the model in several ways. Overall, the results confirmed the mode effects for the management and motivation indexes but also showed that those effects for the ethics index merit some qualifications. First, guided by figure 19.7 in chapter 19 and the differences between groups of respondents it showed, models of table 19.3 in chapter 19 were reestimated adding interaction effects between survey mode and some key demographic controls. The results are shown in table G.6. They confirm the significance and size of *independent* mode effects for management and motivation index. For both of those, the mode effects, as well as the aforementioned coefficients for age and job tenure, remain substantially unchanged. None of the added interactions enter the regressions significantly. In contrast, the survey mode effects for the ethics index loses its significance and size, which are however picked up by the interactions. They confirm what figure 19.7 in chapter 19 visually hinted at and what we discussed in its context, namely that managers provide more positive answers to ethical questions in face-to-face mode compared to

online mode. Almost exactly the same pattern exists for civil servants as compared to contractual workers. It is however missing for gender.

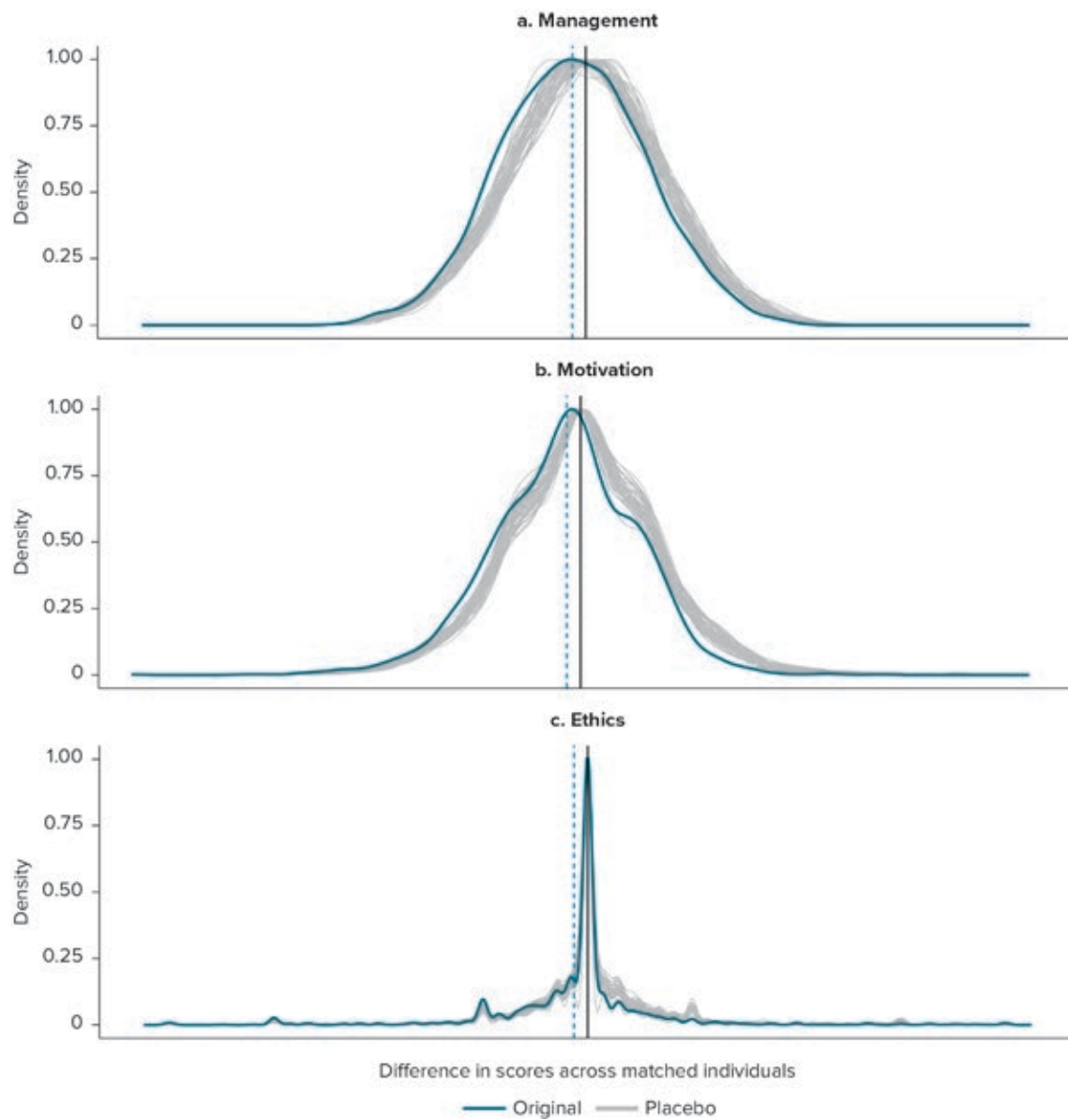
As a second robustness check, we adjusted  $p$ -values from table 19.3 in chapter 19 for multiple testing (Romano and Wolf 2016). With many predictor variables it is always possible that some coefficients will come up as significant due to chance alone. We use a stepdown Romano-Wolf adjustment of  $p$ -values to account for that concern. Those values are presented in table G.7. They suggest that in the models without any controls,  $p$ -values of mode effect are well below the standard 5 percent significance level across all three indices. Including the standard set of controls does not affect those conclusions; all  $p$ -values point to highly statistically significant mode effects for all indices.

Table G.7 also shows the results of our third robustness check, rerandomized  $p$ -values. To calculate those, the treatment status was randomly reassigned and the mode effects based on model specifications in table 19.3 in chapter 19 recalculated 1,000 times. The  $p$ -value produced corresponds to the proportions of simulations in which the mode effect was greater than under the original treatment assignment. In none of the simulations were the mode effects larger, suggesting that the effects in our original sample are a reflection of true differences.

As a fourth robustness check, we simulate treatment reassignments 50 times, each time reestimating the PSM model and plotting the distribution of difference between responses of online-mode individuals and their matched face-to-face (F2F) counterparts. Figure G.1 displays the results. The simulated distributions for the management and motivation index are clearly more symmetric and have a peak closer to 0 compared to the distribution under the original treatment assignment, which is positively skewed. This confirms the finding from rerandomization  $p$ -values, which indicated that obtaining mode effects larger than in our sample is highly improbable. The graph for the ethics index requires a bit closer inspection, but the fatter right tail of the simulated distribution (i.e., a greater share of matched individuals with *positive* difference) is the feature responsible for smaller mode effects in the simulated data.

Finally, we try to account for the unobserved characteristics of some individuals that make them particularly sensitive to measurement. The strong potential effect of those variables is hinted at by the tails of distributions in figure 19.6 in chapter 19, which show that differences between matched individuals reach several standard deviations in some cases. Although the distribution of those extreme differences appears to be approximately symmetrical, they might matter considerably for our results. Given that it is possible to isolate those individuals most sensitive to measurement outcomes, table G.1 repeats the analysis in table 19.2 in chapter 19 but with a restricted sample that excludes individuals that exhibit the most extreme mode effects at the top and bottom 5th percentiles. We see that restricting the sample at that level ameliorates many of the most substantial mode effects in panel 3. The means at all three levels of analysis move somewhat in this restricted sample, but the change is not consistent across the levels and indices, in addition to not being substantial enough to affect our key conclusions.

**FIGURE G.1** Distribution of Survey Mode Differences across Matched Individuals, Alternative



Source: Original figure for this publication.

Note: The figure shows densities of differences between matched individuals in the original sample and compares them to placebo distributions that we obtain by randomly reassigning the treatment status across individuals, based on 50 rerandomizations. Blue dashed lines show the original sample mean differences.

**TABLE G.1 Mean Modal Differences, by Level of Analysis, Restricted Sample**

	Mean	Minimum	Minimum	p25	p50	p75
<i>(1) National level</i>						
Management index	-0.271	—	—	—	—	—
Motivation index	-0.326	—	—	—	—	—
Ethics index	-0.159	—	—	—	—	—
<i>(2) Organizational level</i>						
Management index	-0.270	-1.973	0.948	-0.537	-0.085	0.150
Motivation index	-0.320	-1.429	0.590	-0.554	-0.351	0.005
Ethics index	-0.146	-0.702	0.680	-0.289	-0.173	-0.035
<i>(3) Individual level</i>						
Management index	-0.298	-2.615	1.817	-1.098	-0.274	0.532
Motivation index	-0.340	-2.806	1.715	-1.143	-0.312	0.364
Ethics index	-0.200	-1.970	1.576	-0.394	0.000	0.000

Source: Original table for this publication.

Note: In this table, observations from the bottom or top of the 5th percentile of the difference at the individual level are removed. Panel (1) shows the full-sample differences in the means of the indices between online and face-to-face survey modes ( $\hat{x}_{online} - \hat{x}_{f2f}$ ) after leaving only the individuals whose matched difference falls between the 5th and 95th percentile. Panel (2) repeats those calculations at the level of each organization and summarizes their values. Panel (3) shows the distribution of differences in index values (falling between the 5th and 95th percentile) between individuals matched on the following variables: organization, job tenure, organization tenure, public administration tenure, pay grade, employee status (civil servant vs. contractual staff), age, gender, and education level. Propensity-score matching estimators impute the missing potential outcomes for each subject by using the average of the outcomes of similar subjects that receive the other treatment. Observations are matched using nearest-neighbor matching and the probability of treatment is calculated using a logit model. In the case of a tie, observations are matched with all ties.

**TABLE G.2 Cox-Weibull Hazard Model of Online Survey Breakoff**

	Dependent variable	
	Model 1	
	(1)	(2)
Age	-0.031** (0.014)	-0.025* (0.014)
Education: Undergraduate	0.553 (0.736)	0.507 (0.738)
Education: Master's	1.309* (0.728)	1.158 (0.730)
Education: PhD	0.242 (1.007)	-0.125 (1.012)
Gender: Male	-0.175 (0.207)	-0.173 (0.210)
Status: Civil servant	0.712 (0.623)	0.213 (0.634)
Tenure in position	0.020 (0.019)	0.009 (0.019)
Tenure in organization	0.009 (0.019)	0.044** (0.022)
Tenure in public administration	-0.007 (0.020)	-0.016 (0.022)

(continues on next page)

**TABLE G.2 Cox-Weibull Hazard Model of Online Survey Breakoff (continued)**

		Dependent variable	
		Model 1	
		(1)	(2)
Pay grade		–0.012 (0.042)	–0.009 (0.042)
Average age at organization		—	–0.166* (0.091)
Average education at organization		—	1.208 (0.951)
Average gender at organization		—	1.788 (1.320)
Average job tenure at organization		—	–1.102 (1.950)
Average organization tenure at organization		—	0.304** (0.134)
Average public administration tenure at organization		—	–0.117* (0.060)
Average pay grade at organization		—	–0.116 (0.092)
Constant		—	–0.140 (0.124)
Observations		2,504	2,504
R <sup>2</sup>		0.016	0.031
Score (logrank) test	56	36.072*** (df = 10)	77.845*** (df = 18)

Source: Original table for this publication.

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01.

**TABLE G.3 OLS Results: Number of “Don’t Know,” Refusal, and Missing Responses**

	Dependent variable			
	“Don’t know” rate	“Refuse” rate	Missing rate	Total nonresponse rate
	(1)	(2)	(3)	(4)
Survey mode: Online	0.042*** (0.002)	1.793*** (0.139)	1.619*** (0.111)	7988*** (0.276)
Age	–0.0002 (0.0001)	–0.010 (0.010)	0.029*** (0.008)	–0.001 (0.019)
Gender: Male	–0.006*** (0.002)	–0.494*** (0.154)	–0.299** (0.123)	–1.411*** (0.305)
Education: Undergraduate	–0.002 (0.004)	–0.470 (0.350)	–0.725*** (0.280)	–1.386** (0.695)
Education: Master’s	–0.007 (0.004)	–0.812** (0.354)	–1.007*** (0.283)	–2.588*** (0.703)
Education: PhD	–0.008 (0.006)	–0.587 (0.492)	–1.051*** (0.394)	–2.518** (0.979)

(continues on next page)



**TABLE G.3 OLS Results: Number of “Don’t Know,” Refusal, and Missing Responses (continued)**

	Dependent variable			
	“Don’t know” rate	“Refuse” rate	Missing rate	Total nonresponse rate
	(1)	(2)	(3)	(4)
Status: Civil servant	0.006 <sup>+</sup> (0.004)	−0.133 (0.292)	−0.518 <sup>***</sup> (0.234)	0.044 (0.581)
Pay grade	0.001 <sup>**</sup> (0.0003)	0.045 <sup>+</sup> (0.027)	0.030 (0.022)	0.171 <sup>***</sup> (0.055)
Tenure	0.0004 <sup>**</sup> (0.0002)	0.047 <sup>***</sup> (0.013)	0.020 <sup>*</sup> (0.011)	0.110 <sup>***</sup> (0.027)
Organization tenure	−0.0001 (0.0002)	−0.012 (0.013)	0.018 <sup>+</sup> (0.011)	−0.002 (0.027)
Public administration tenure	−0.00001 (0.0002)	0.010 (0.014)	−0.087 <sup>***</sup> (0.011)	−0.077 <sup>***</sup> (0.028)
Constant	0.035 <sup>***</sup> (0.009)	1.056 (0.680)	5.832 <sup>***</sup> (0.545)	9.104 <sup>***</sup> (1.353)
Observations	4,787	4,787	4,787	4,787
R <sup>2</sup>	0.117	0.046	0.071	0.174
Adjusted R <sup>2</sup>	0.115	0.044	0.069	0.172

Source: Original table for this publication.

Note: <sup>\*</sup>p<0.1; <sup>\*\*</sup>p<0.05; <sup>\*\*\*</sup>p<0.01.

**TABLE G.4 OLS Results: Organizational Characteristics and Mean Survey Differences**

	Dependent variable		
	Management index	Motivation index	Ethics index
	(1)	(2)	(3)
Age	0.110 (0.076)	0.085 (0.057)	−0.011 (0.053)
Gender: Male	−2.134 (1.416)	−0.520 (1.067)	−0.236 (0.988)
Education level	0.055 (0.759)	−0.365 (0.572)	−1.441 <sup>**</sup> (0.530)
Status: Civil servant	1.305 (1.481)	0.104 (1.116)	−0.252 (1.033)
Pay grade	−0.058 (0.121)	−0.055 (0.091)	−0.104 (0.084)
Tenure in position	0.071 (0.107)	−0.012 (0.081)	−0.037 (0.075)
Tenure in organization	−0.095 (0.075)	−0.023 (0.057)	−0.010 (0.052)
Tenure in public administration	0.025 (0.098)	−0.024 (0.074)	−0.007 (0.068)
Meritocracy	0.026 (0.486)	−0.213 (0.367)	0.268 (0.339)

(continues on next page)

**TABLE G.4 OLS Results: Organizational Characteristics and Mean Survey Differences (continued)**

	Dependent variable		
	Management index	Motivation index	Ethics index
	(1)	(2)	(3)
Willingness to answer sensitive questions	–1.378 (0.859)	–0.368 (0.647)	0.705 (0.599)
Willingness to sit through survey	1.221 (1.028)	0.743 (0.775)	–1.210 (0.717)
Number of organization employees	0.0004 (0.0004)	0.0003 (0.0003)	–0.0003 (0.0003)
Constant	–5.865 (4.330)	–1.971 (3.264)	2.998 (3.020)
Observations	43	43	43
$R^2$	0.337	0.191	0.347
Adjusted $R^2$	0.072	–0.133	0.086

Source: Original table for this publication.

Note: OLS = ordinary least squares. \*p<0.1; \*\*p<0.05; \*\*\*p<0.01.

**TABLE G.5 OLS Results: Propensity Score Matched Estimates**

	Mean	Minimum	Maximum	p25	p50	p75
<i>(1) Hierarchical variables and organizational fixed effects</i>						
Management index	–0.267	–4.784	3.946	–1.218	–0.204	0.671
Motivation index	–0.312	–8.365	4.988	–1.247	–0.312	0.623
Ethics index	–0.202	–7.879	7.879	–0.563	0.000	0.000
<i>(2) Organizational fixed effects only</i>						
Management index	–0.247	–4.427	2.560	–1.048	–0.133	0.591
Motivation index	–0.348	–7.915	2.159	–0.985	–0.231	0.481
Ethics index	–0.212	–7.735	1.563	–0.364	0.111	0.303

Source: Original table for this publication.

Note: The values show difference in index values between online and face-to-face (F2F) modes between matched individuals ( $\hat{x}_{online} - \hat{x}_{F2F}$ ). Panel (1) matches individuals on the following variables: organization, job tenure, organization tenure, public administration tenure, pay grade, and employee status (civil servant vs. contractor). Panel (2) matches the individuals on organization only. Propensity-score matching estimators impute the missing potential outcomes for each treated subject (online mode) by using an average of the outcomes of interest of similar subjects that received control group treatment (F2F mode). Observations are matched using nearest-neighbor matching. The probability of treatment is calculated using logit model. In the case of ties, the value for an individual is calculated as the mean difference between his outcome and that of all matched individuals. OLS = ordinary least squares.

**TABLE G.6 OLS Results: Individual Characteristics and Mean Survey Differences**

	Dependent variable		
	Management index	Motivation index	Ethics index
	(1)	(2)	(3)
Survey mode: Online	−0.259*** (0.091)	−0.228** (0.091)	0.045 (0.101)
Age	0.009*** (0.002)	0.011*** (0.002)	0.002 (0.002)
Gender: Male	−0.005 (0.046)	−0.061 (0.046)	−0.092* (0.049)
Education: Undergraduate	0.052 (0.073)	−0.092 (0.074)	−0.063 (0.083)
Education: Master's	0.044 (0.074)	−0.062 (0.075)	−0.149* (0.084)
Education: PhD	−0.113 (0.103)	−0.003 (0.103)	−0.107 (0.116)
Status: Civil servant	−0.116 (0.073)	0.034 (0.073)	0.146* (0.077)
Pay grade	−0.019*** (0.006)	−0.007 (0.006)	−0.018*** (0.007)
Managerial status: Manager	−0.480*** (0.063)	0.082 (0.063)	0.046 (0.067)
Tenure	−0.011*** (0.003)	−0.008*** (0.003)	−0.001 (0.003)
Organization tenure	0.009*** (0.003)	0.003 (0.003)	−0.005* (0.003)
Public administration tenure	−0.002 (0.003)	−0.003 (0.003)	−0.001 (0.003)
Survey mode × Gender	−0.016 (0.063)	−0.095 (0.063)	0.045 (0.069)
Survey mode × Status	0.018 (0.094)	−0.097 (0.095)	−0.276*** (0.105)
Survey mode × Managerial status	0.024 (0.090)	0.009 (0.090)	−0.271*** (0.097)
Constant	−0.023 (0.146)	−0.161 (0.147)	0.209 (0.160)
Observations	4,787	4,734	3,991
R <sup>2</sup>	0.040	0.043	0.023
Adjusted R <sup>2</sup>	0.037	0.040	0.019

Source: Original table for this publication.

Note: OLS = ordinary least squares.

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01.

**TABLE G.7 OLS Results: Propensity Score Matched Estimates**

	Romano-Wolf <i>p</i> -value	Randomization inference <i>p</i> -value
<i>Results from unconditional models</i>		
Management index	0.0010	0.0028
Motivation index	0.0010	0.0003
Ethics index	0.0010	0.0008
<i>Results from conditional models</i>		
Management index	0.0010	0.0018
Motivation index	0.0010	0.0003
Ethics index	0.0010	0.0008

Source: Original table for this publication.

Note: OLS = ordinary least squares. *p*-values are based on 1,000 rerandomizations.

**TABLE G.8 Survey Items, by Index**

Index	Variable text	
Management index	PEM4.3.1	I am satisfied with my salary.
	PEM4.3.3	My work performance has had an influence on my salary in the public administration.
	PEM4.3.4	I am paid at least as well as colleagues who have job responsibilities similar to me in my institution.
	PEM4.3.5	I am paid at least as well as colleagues who have job responsibilities similar to me in other organizations at the same administration level.
	PEM4.3.6	It would be easy for me to find a job outside the public sector that pays better than my current job.
	PEM4.3.7	It would be fair to tie part of my salary (including salary supplements) to my performance.
	PEM4.3.8	Tying salaries (including salary supplements) to performance would improve morale and motivate people to perform better.
	REC3.4.1	Which of the following criteria, in your opinion, help you get a promotion to the next professional grade? Job performance, such as reaching job targets and goals
	REC3.10.1	Which of the following criteria, in your opinion, help you get a promotion to a higher management level? Job performance, such as reaching job targets and goals
	REC3.14.1	Which of the following factors help employees get a promotion in your institution? Job performance, such as reaching job targets and goals
	HRM3.1.1 (F2F only)	Direct superior communicates effectively the institution's vision and mission to employees.
	HRM3.1.2	Direct superior leads by setting a good example.
	HRM3.1.3	Direct superior says things that make employees proud to be part of this institution.
	HRM3.1.4	Direct superior holds subordinates accountable for using ethical practices in their work.
	HRM3.1.5	Direct superior communicates clear ethical standards to subordinates.
	HRM3.1.8	Direct superior personally cares about me.

(continues on next page)

**TABLE G.8 Survey Items, by Index (continued)**

Index	Variable text	
	HRM3.2.2	I lead by setting a good example.
	HRM3.2.3	I say things that make employees proud to be part of this institution.
	HRM3.2.4	I hold my subordinates accountable for using ethical practices in their work.
	HRM3.2.5	I communicate clear ethical standards to my subordinates.
	PEM4.2.3	The last bonus you received was distributed based on performance criteria established at the level of the institution.
	PEM1.2	Have your objectives and performance objectives been set and discussed with you before your last performance evaluation?
	PEM1.4	Has your performance assessment/evaluation report been shared with you/shown to you after it was written?
	PEM1.5	Has your superior discussed the results of your last performance evaluation with you after filling in your performance evaluation report?
	PEM1.6	Was this discussion useful for you to improve your performance?
	PEM1.11.1	My performance indicators measure well the extent to which I contribute to my institution's success
	PEM1.11.2	My superior has enough information about my work performance to evaluate me.
	PEM1.11.3	My superior evaluates my performance fairly.
	PEM1.17.1	Performance evaluation is taken seriously in my institution
	PEM1.17.2	I feel pressure to give all members of my team the highest rating.
	PEM1.17.3	I feel pressure to give all members of my team the highest rating.
	PEM1.17.4	The work climate would be negatively affected if I do not give everyone a high performance rating.
	PEM1.17.5	I fear that employees take legal action if I give them a low performance rating.
	PEM1.17.6	I fear that employees turn to public sector unions for help if I give them a low performance rating.
	PEM1.17.7	I have tools to address underperformance among my employees.
	PEM2.1.1	My institution has a clear set of objectives and targets.
	PEM2.1.2	I have a good understanding of my institution's goals.
	PEM2.1.3	The targets and objectives of my institution are used to determine my work schedule and goals.
	PEM2.1.4	When I arrive at work each day, I know what my individual roles and responsibilities are in achieving the institution's goals.
	PEM2.1.5	My institution uses indicators for tracking performance against organizational targets.
Motivation index	AWE2.4.1	Overall, I am satisfied with my job.
	AWE2.4.2	I do not feel a strong sense of belonging to my institution.
	AWE2.4.3	I am willing to do extra work for my job that isn't really expected of me.
	AWE2.4.4	I put forth my best effort to get my job done regardless of any difficulties.
	AWE2.4.5	I stay at work until the job is done.
	AWE2.4.6	I am proud of the work that I do.
	AWE2.4.7	My job is very interesting.

(continues on next page)

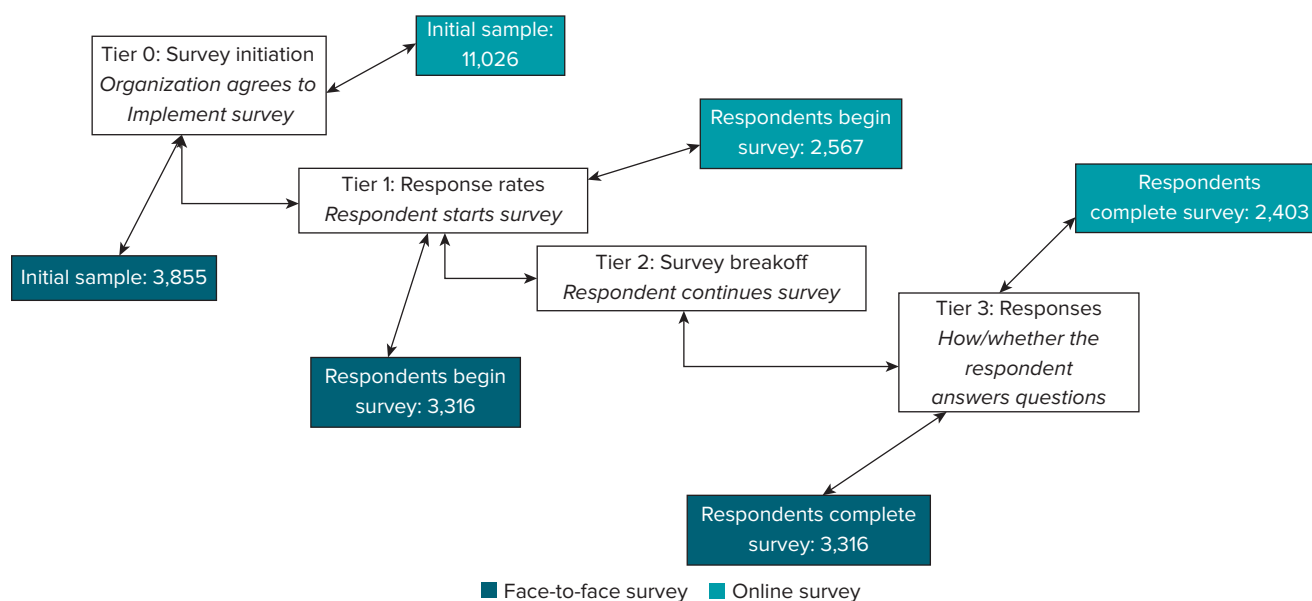
**TABLE G.8** Survey Items, by Index (*continued*)

Index	Variable text	
Ethics index	<b>How frequently do employees in your institution undertake the following actions?</b>	
	AWE4.1.1	Accepting gifts or money from companies.
	AWE4.1.2	Accepting gifts or money from citizens.
	AWE4.1.3	Bending the rules slightly as a favor to a friend.
	AWE4.1.4	Bending the rules slightly to help a poor person in need.
	AWE4.1.5	Observing unethical behavior among colleagues.
	AWE4.1.6	Reporting a colleague for not behaving ethically.
	AWE4.1.7	Pressure other employees to not speak out against unethical behavior.

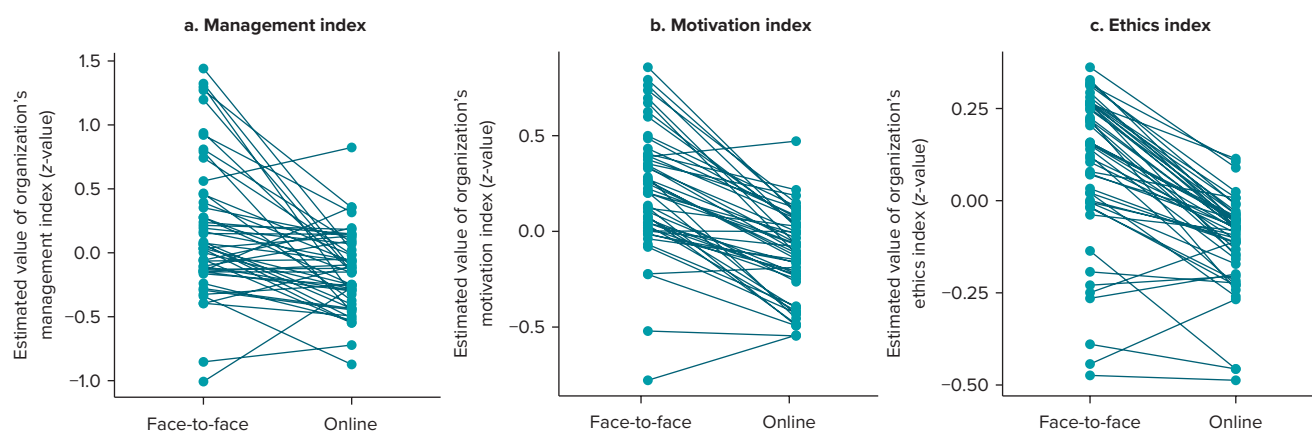
Source: Original table for this publication.

Note: F2F = face-to-face.

**FIGURE G.2** Survey Response and Breakoff Rates, by Mode



**FIGURE G.3** Mixed Model Results: Estimated Intercepts and Random Effects of Survey Mode Dummy across Organizations







## REFERENCE

Romano, Joseph J., and Michael Wolf. 2016. "Efficient Computation of Adjusted  $p$ -Values for Resampling-Based Stepdown Multiple Testing." *Statistics and Probability Letters* 113: 38–40.

# APPENDIX H

## Further Details of Analysis

### Chapter 20 Appendix

#### APPENDIX H.1 SUMMARY TABLES OF STATISTICS FOR CHOSEN INDICATORS AND SAMPLING PROPORTIONS

**TABLE H.1** Romania: Statistics for Chosen Indicators and Sampling Proportions

Proportion	Motivation index			Leadership index			Performance index		
	Mean (95% CI)	Median	SD	Mean (95% CI)	Median	SD	Mean (95% CI)	Median	SD
Full sample	4.488 (±0.012)	4.5	0.462	3.817 (±0.022)	4	0.853	4.713 (±0.016)	5	0.591
95%	4.488 (±0)	4.5	0.462	3.817 (±0.001)	4	0.853	4.713 (±0.001)	5	0.591
90%	4.488 (±0)	4.5	0.462	3.817 (±0.001)	4	0.853	4.712 (±0.001)	5	0.591
85%	4.488 (±0)	4.5	0.462	3.817 (±0.001)	4	0.853	4.713 (±0.001)	5	0.59
80%	4.488 (±0)	4.5	0.462	3.817 (±0.001)	4	0.853	4.712 (±0.001)	5	0.591
75%	4.488 (±0)	4.5	0.461	3.817 (±0.001)	4	0.853	4.712 (±0.001)	5	0.591
70%	4.488 (±0)	4.5	0.462	3.817 (±0.001)	4	0.853	4.713 (±0.001)	5	0.59
65%	4.488 (±0)	4.5	0.462	3.817 (±0.001)	4	0.853	4.712 (±0.001)	5	0.59
60%	4.488 (±0)	4.5	0.461	3.817 (±0.001)	4	0.853	4.712 (±0.001)	5	0.591
55%	4.488 (±0)	4.5	0.461	3.817 (±0.001)	4	0.853	4.713 (±0.001)	5	0.59
50%	4.488 (±0.001)	4.5	0.462	3.817 (±0.001)	4	0.853	4.712 (±0.001)	5	0.591
45%	4.487 (±0.001)	4.5	0.462	3.817 (±0.001)	4	0.853	4.713 (±0.001)	5	0.59
40%	4.487 (±0.001)	4.5	0.462	3.816 (±0.001)	4	0.853	4.712 (±0.001)	5	0.592
35%	4.488 (±0.001)	4.5	0.461	3.817 (±0.001)	4	0.853	4.712 (±0.001)	5	0.591
30%	4.488 (±0.001)	4.5	0.462	3.817 (±0.001)	4	0.853	4.712 (±0.001)	5	0.591
25%	4.488 (±0.001)	4.5	0.461	3.817 (±0.001)	4	0.853	4.712 (±0.001)	5	0.592
20%	4.488 (±0.001)	4.5	0.461	3.819 (±0.002)	4	0.851	4.714 (±0.001)	5	0.588
15%	4.488 (±0.001)	4.5	0.461	3.818 (±0.002)	4	0.853	4.714 (±0.001)	5	0.589
10%	4.488 (±0.001)	4.5	0.463	3.816 (±0.002)	4	0.854	4.712 (±0.002)	5	0.59
5%	4.488 (±0.002)	4.5	0.46	3.818 (±0.003)	4	0.851	4.711 (±0.002)	5	0.593

Source: Original table for this publication.

Note: CI = confidence interval; SD = standard deviation.

**TABLE H.2 Chile: Statistics for Chosen Indicators and Sampling Proportions**

Proportion	Motivation index			Leadership index			Performance index		
	Mean (95% CI)	Median	SD	Mean (95% CI)	Median	SD	Mean (95% CI)	Median	SD
Full sample	4.195 (±0.008)	4.25	0.625	3.582 (±0.013)	3.75	0.994	3.392 (±0.012)	3.4	0.84
95%	4.188 (±0.005)	4.25	0.628	3.574 (±0.007)	3.75	0.994	3.4 (±0.007)	3.5	0.84
90%	4.188 (±0.005)	4.25	0.626	3.573 (±0.008)	3.75	0.994	3.4 (±0.007)	3.5	0.84
80%	4.188 (±0)	4.25	0.627	3.574 (±0)	3.75	0.994	3.401 (±0)	3.5	0.84
75%	4.188 (±0)	4.25	0.627	3.574 (±0)	3.75	0.994	3.401 (±0)	3.5	0.84
70%	4.188 (±0)	4.25	0.627	3.574 (±0)	3.75	0.991	3.4 (±0)	3.5	0.81
65%	4.188 (±0)	4.25	0.627	3.574 (±0)	3.75	0.994	3.401 (±0)	3.5	0.84
60%	4.188 (±0)	4.25	0.627	3.574 (±0.001)	3.75	0.994	3.401 (±0)	3.5	0.84
55%	4.188 (±0)	4.25	0.627	3.574 (±0.001)	3.75	0.994	3.401 (±0)	3.5	0.84
50%	4.188 (±0)	4.25	0.627	3.574 (±0.001)	3.75	0.994	3.401 (±0.001)	3.5	0.84
45%	4.188 (±0)	4.25	0.627	3.574 (±0.001)	3.75	0.991	3.4 (±0.001)	3.5	0.84
40%	4.188 (±0)	4.25	0.627	3.574 (±0.001)	3.75	0.994	3.4 (±0.001)	3.5	0.84
35%	4.188 (±0)	4.25	0.627	3.574 (±0.001)	3.75	0.994	3.401 (±0.001)	3.5	0.84
30%	4.188 (±0)	4.25	0.627	3.574 (±0.001)	3.75	0.994	3.4 (±0.001)	3.5	0.84
25%	4.188 (±0)	4.25	0.627	3.574 (±0.001)	3.75	0.994	3.401 (±0.001)	3.5	0.84
20%	4.188 (±0.001)	4.25	0.627	3.574 (±0.001)	3.75	0.995	3.4 (±0.001)	3.5	0.84
15%	4.188 (±0.001)	4.25	0.627	3.574 (±0.001)	3.75	0.995	3.401 (±0.001)	3.5	0.84
10%	4.188 (±0.001)	4.25	0.627	3.573 (±0.001)	3.75	0.995	3.401 (±0.001)	3.5	0.84
5%	4.188 (±0.001)	4.25	0.627	3.574 (±0.002)	3.75	0.994	3.4 (±0.002)	3.5	0.84

Source: Original table for this publication.

Note: CI = confidence interval; SD = standard deviation.

**TABLE H.3 Liberia: Statistics for Chosen Indicators and Sampling Proportions**

Proportion	Management index		
	Mean (95% CI)	Median	SD
Full sample	3.337 (±0.032)	3.333	0.612
95%	3.335 (±0.001)	3.333	0.61
90%	3.335 (±0.001)	3.333	0.61
85%	3.335 (±0.001)	3.333	0.61
80%	3.334 (±0.001)	3.333	0.61
75%	3.335 (±0.001)	3.333	0.61
70%	3.334 (±0.001)	3.333	0.61
65%	3.334 (±0.001)	3.333	0.61
60%	3.333 (±0.001)	3.333	0.61
55%	3.333 (±0.001)	3.333	0.61

(continues on next page)

**TABLE H.3 Liberia: Statistics for Chosen Indicators and Sampling Proportions (*continued*)**

Proportion	Management index		
	Mean (95% CI)	Median	SD
50%	3.332 (±0.001)	3.333	0.61
45%	3.331 (±0.001)	3.333	0.61
40%	3.331 (±0.002)	3.333	0.609
35%	3.329 (±0.002)	3.333	0.61
30%	3.328 (±0.002)	3.333	0.61
25%	3.326 (±0.002)	3.333	0.609
20%	3.323 (±0.002)	3.333	0.611
15%	3.317 (±0.002)	3.333	0.61
10%	3.309 (±0.003)	3.333	0.61
5%	3.29 (±0.004)	3.333	0.61

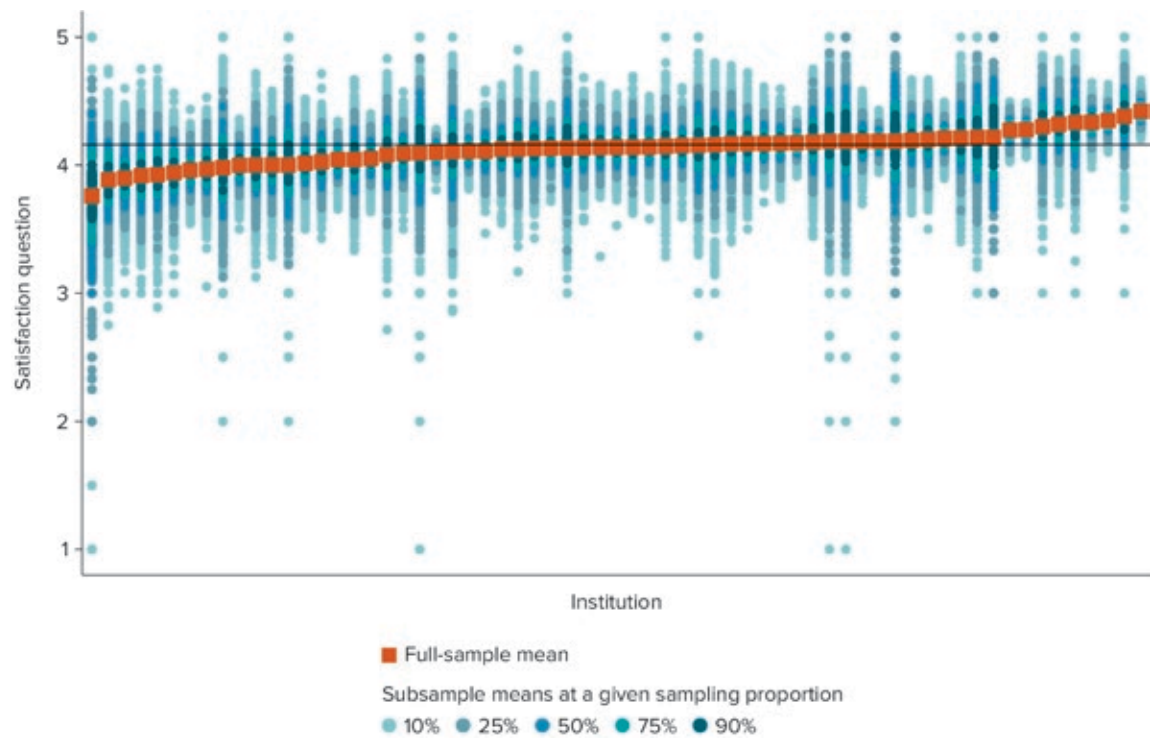
Source: Original table for this publication.

Note: CI = confidence interval; SD = standard deviation.

## APPENDIX H.2 CHANGES IN INSTITUTIONAL MEANS ACROSS SIMULATIONS

Institutions within public administration often form one of the key points of interest for survey designers. Capturing mean values at the institutional level allows to compare discrete and relatively confined work units against each other, identify underperformers, and focus the reform efforts on them—potentially based on the lessons drawn from the institutions found to be performing above average. The analysis can also show that the immediate work environment, such as inside a particular civil service institution, determines an employee’s engagement and attitudes in a crucial manner. Therefore, it is understandable why surveys measure those factors precisely at the institutional level. At the same time, institutions are necessarily much smaller in size than the civil service as a whole; therefore, any estimates pertaining to them are less precise. The figures below show how much the institutional-level means would vary for our key outcomes of interest if the surveys are conducted in exactly the same manner but with smaller sample sizes.

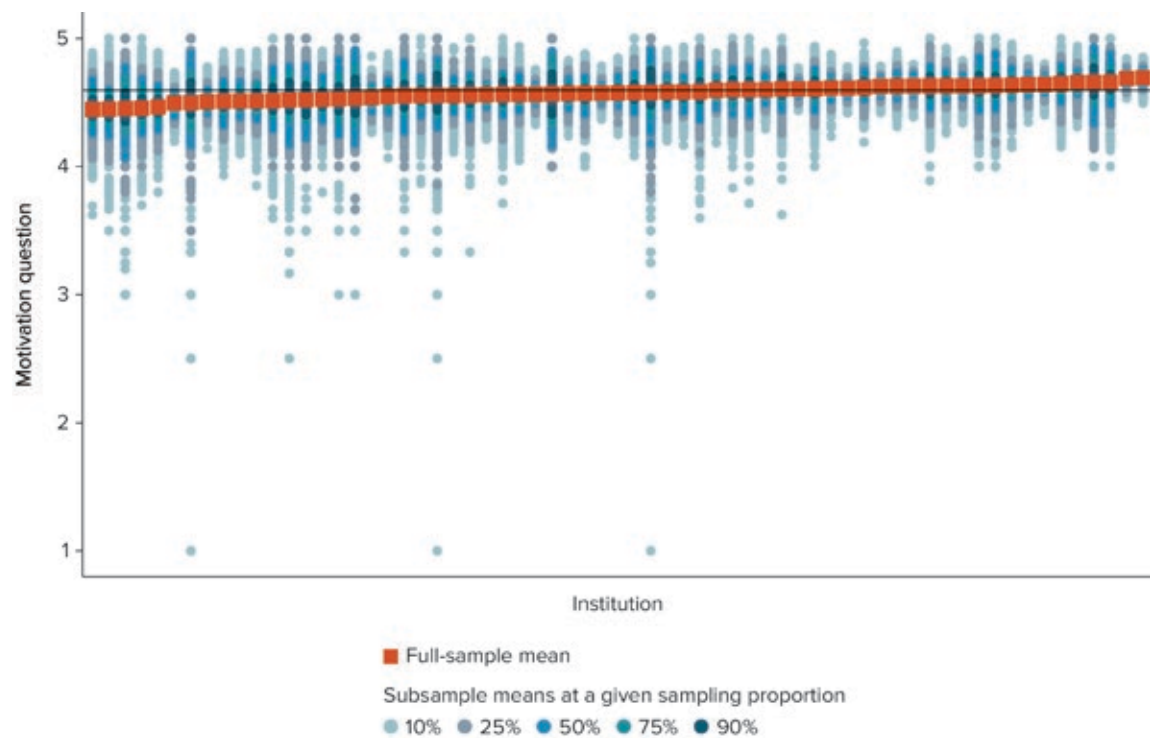
**FIGURE H.1 Chile: Satisfaction Question**



Source: Original figure for this publication.

Note: Institutional means are actual and simulated at different sampling proportions across 1,000 simulations.

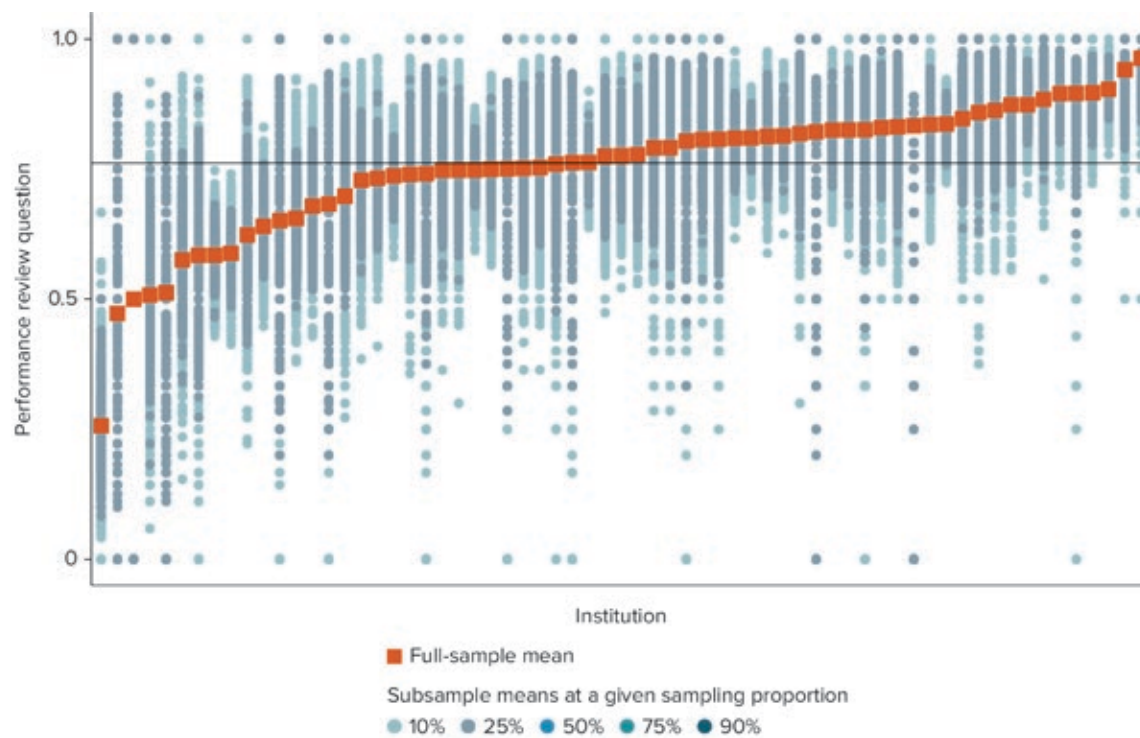
**FIGURE H.2 Chile: Motivation Question**



Source: Original figure for this publication.

Note: Institutional means are actual and simulated at different sampling proportions across 1,000 simulations.

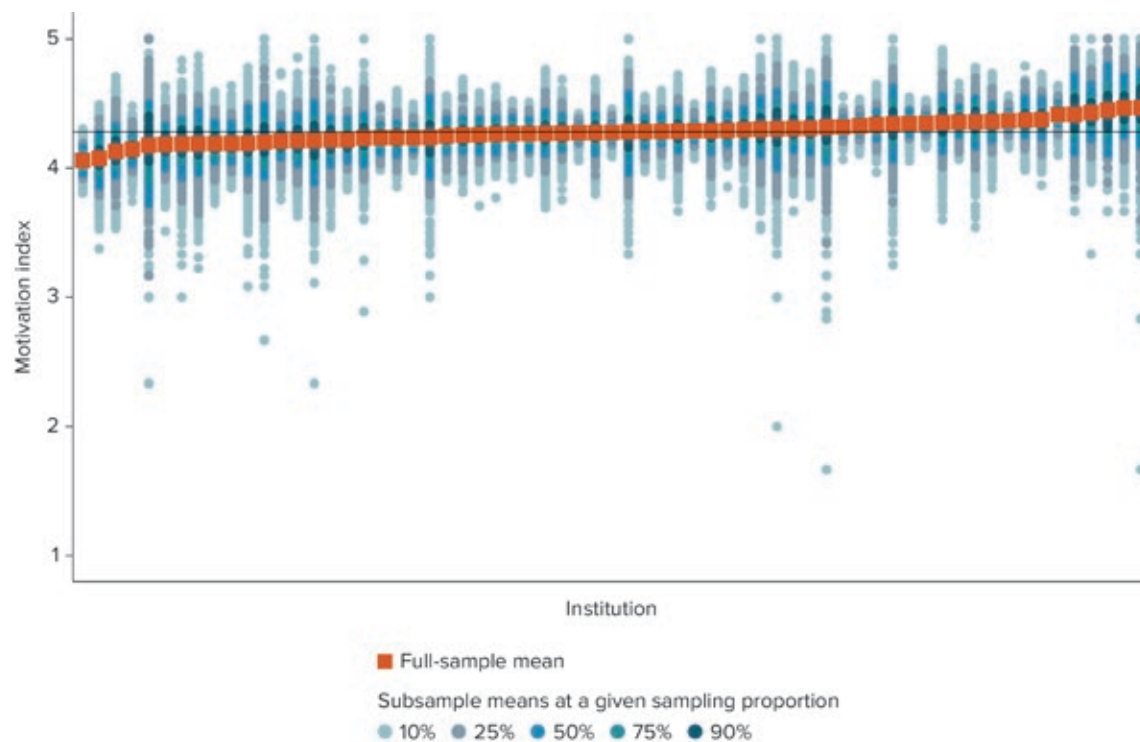
**FIGURE H.3** Chile: Performance Review Question



Source: Original figure for this publication.

Note: Institutional means are actual and simulated at different sampling proportions across 1,000 simulations.

**FIGURE H.4** Chile: Motivation Index

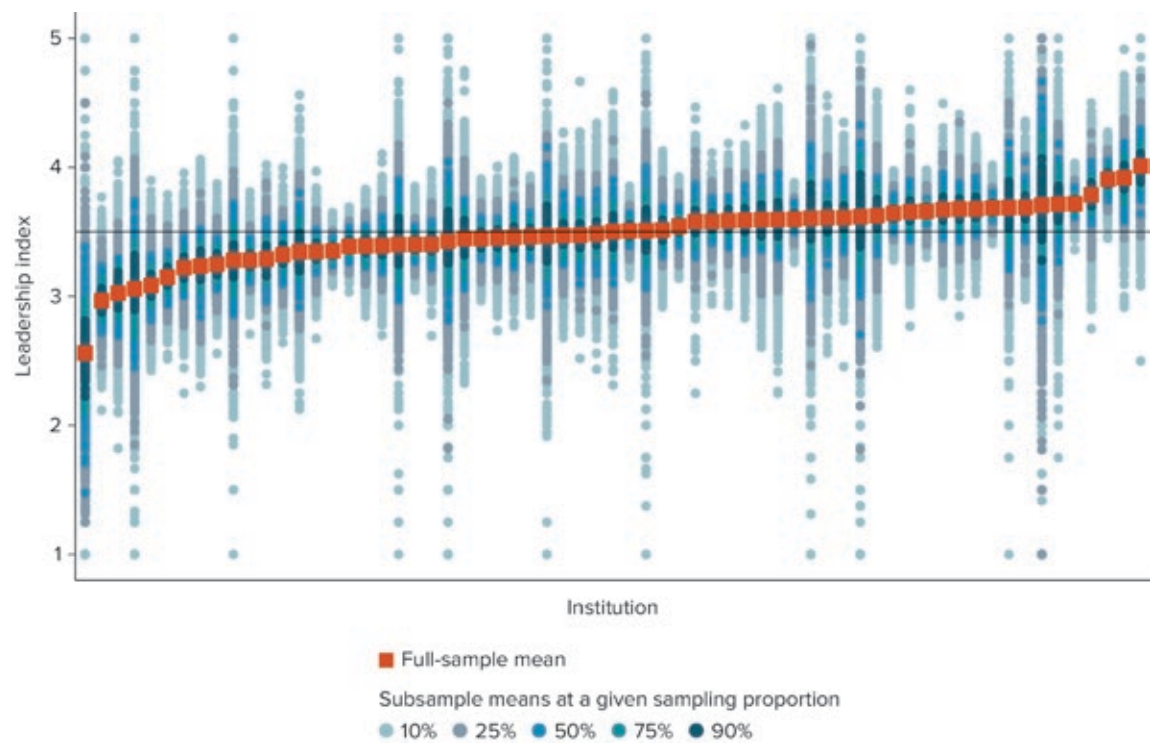


Source: Original figure for this publication.

Note: Institutional means are actual and simulated at different sampling proportions across 1,000 simulations.



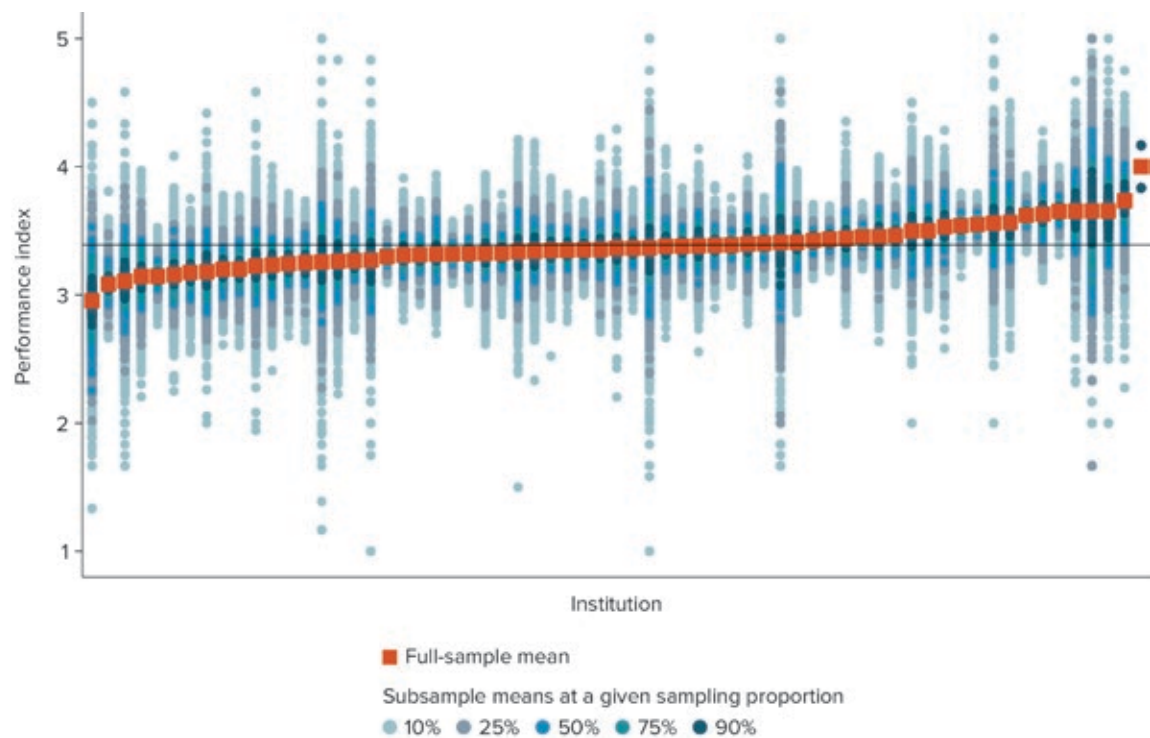
**FIGURE H.5** Chile: Leadership Index



Source: Original figure for this publication.

Note: Institutional means are actual and simulated at different sampling proportions across 1,000 simulations.

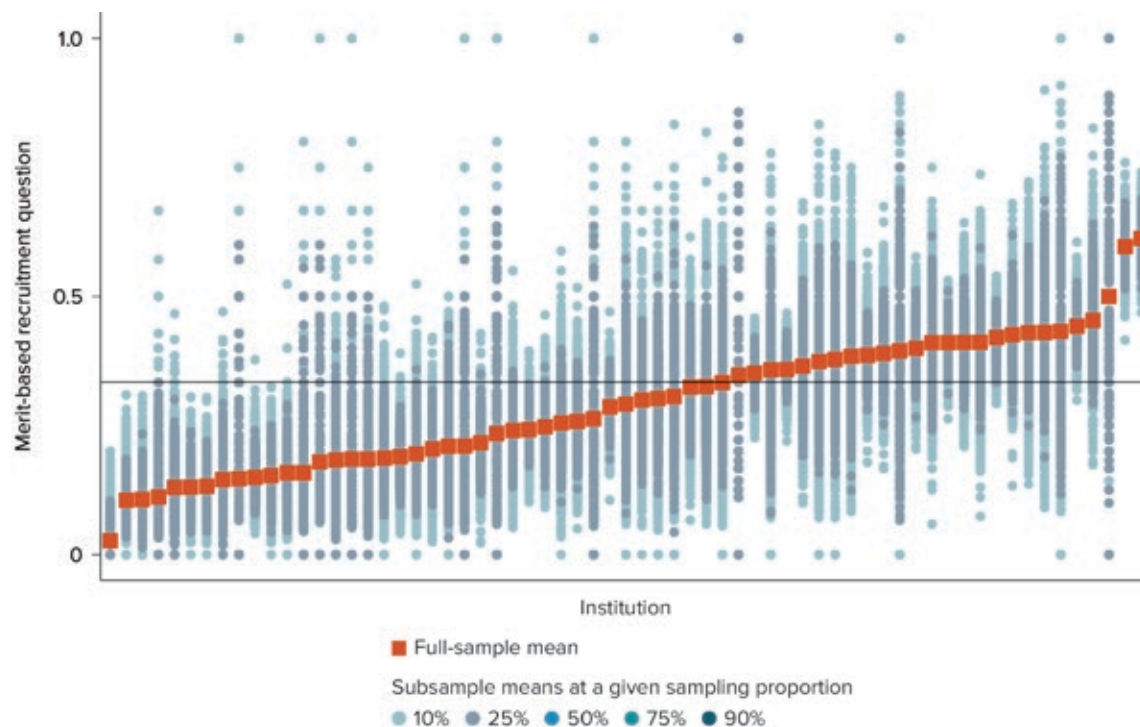
**FIGURE H.6** Chile: Performance Index



Source: Original figure for this publication.

Note: Institutional means are actual and simulated at different sampling proportions across 1,000 simulations.

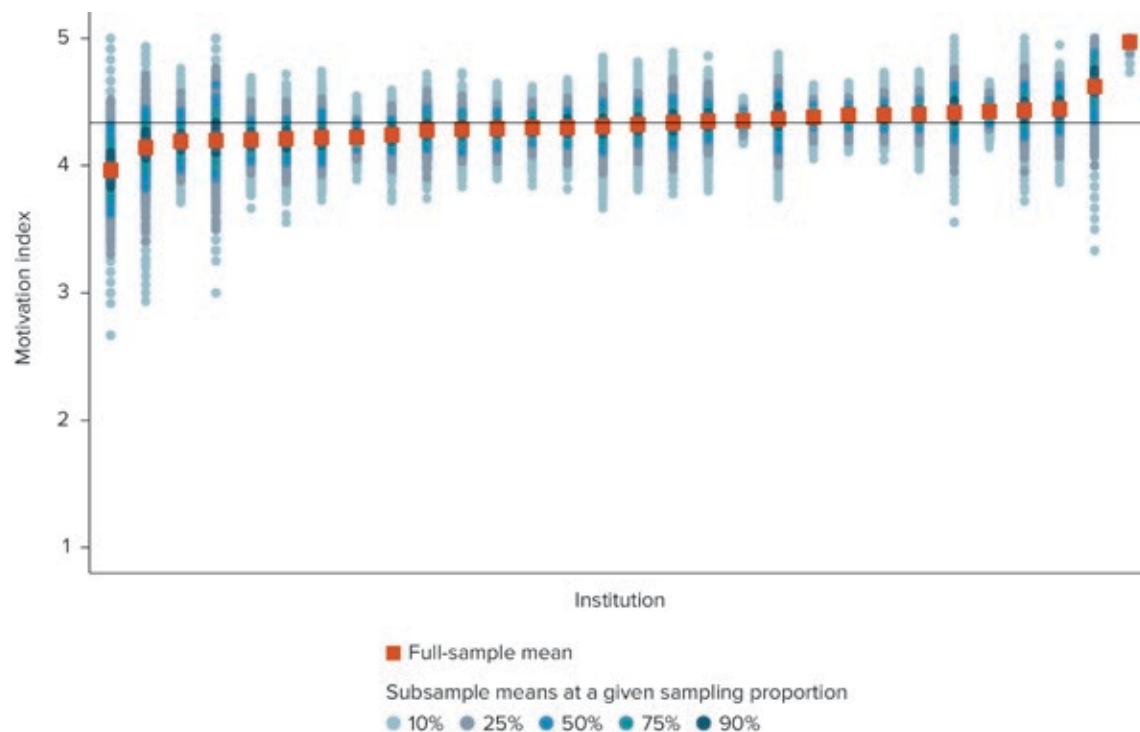
**FIGURE H.7** Chile: Recruitment Question



Source: Original figure for this publication.

Note: Institutional means are actual and simulated at different sampling proportions across 1,000 simulations.

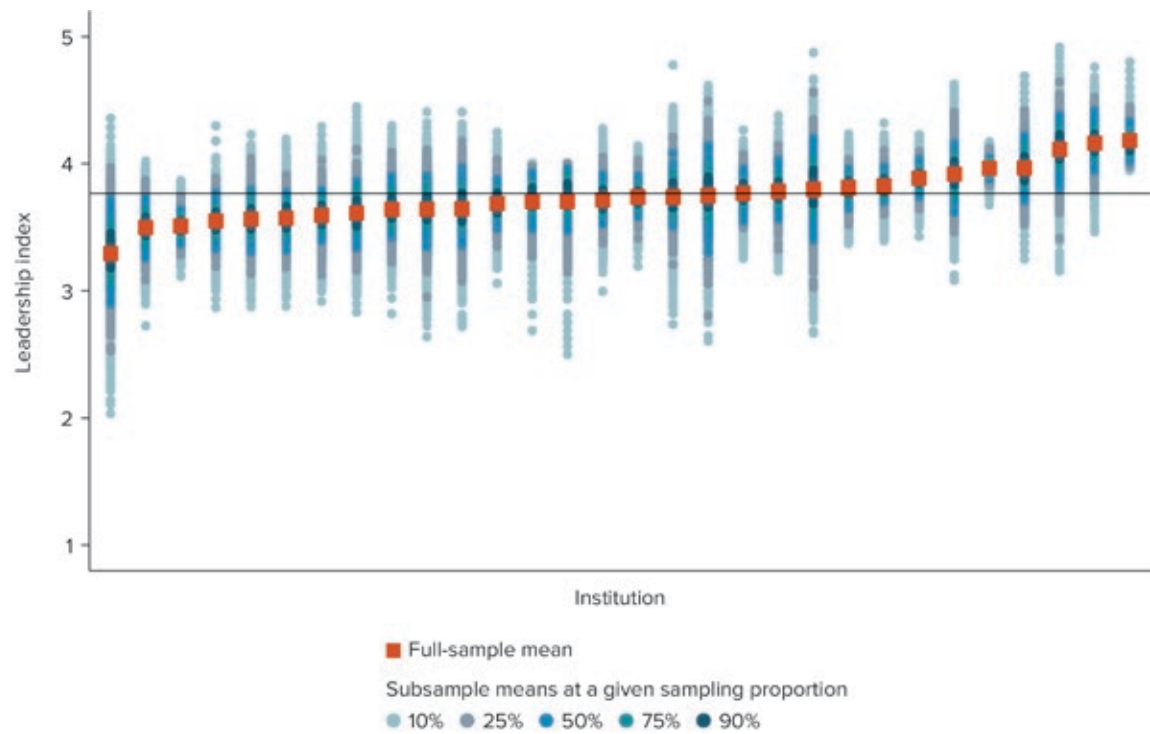
**FIGURE H.8** Romania: Motivation Index



Source: Original figure for this publication.

Note: Institutional means are actual and simulated at different sampling proportions across 1,000 simulations.

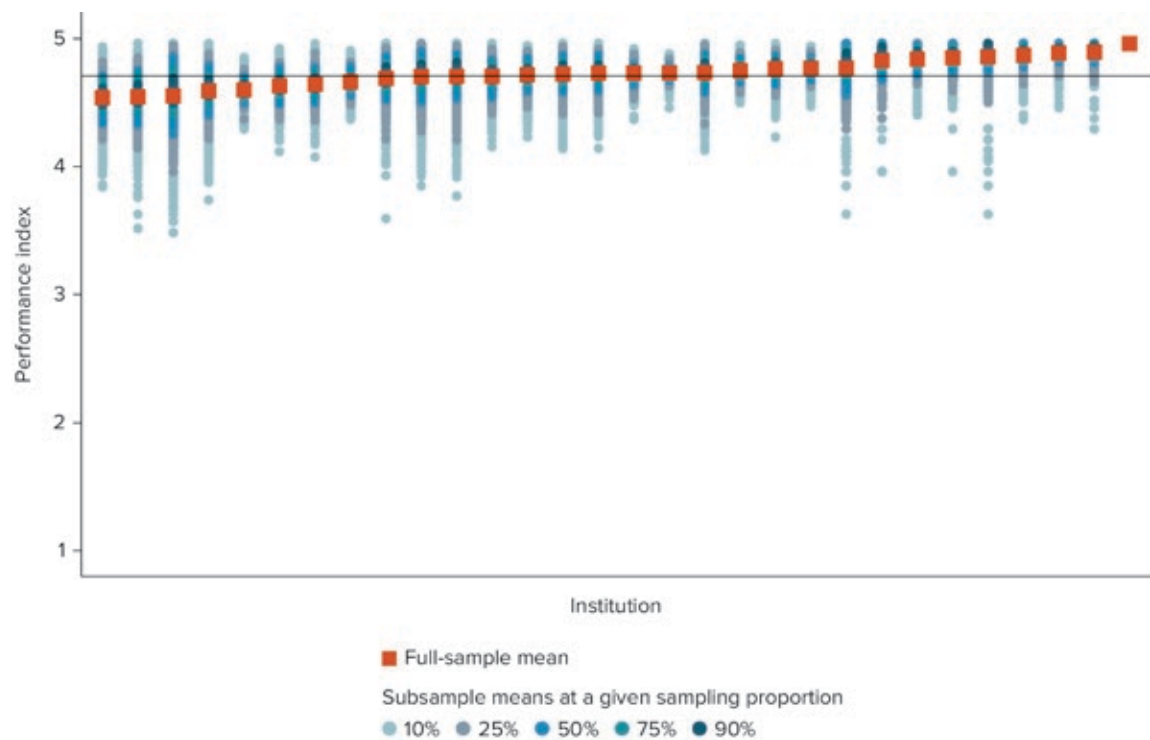
**FIGURE H.9** Romania: Leadership Index



Source: Original figure for this publication.

Note: Institutional means are actual and simulated at different sampling proportions across 1,000 simulations.

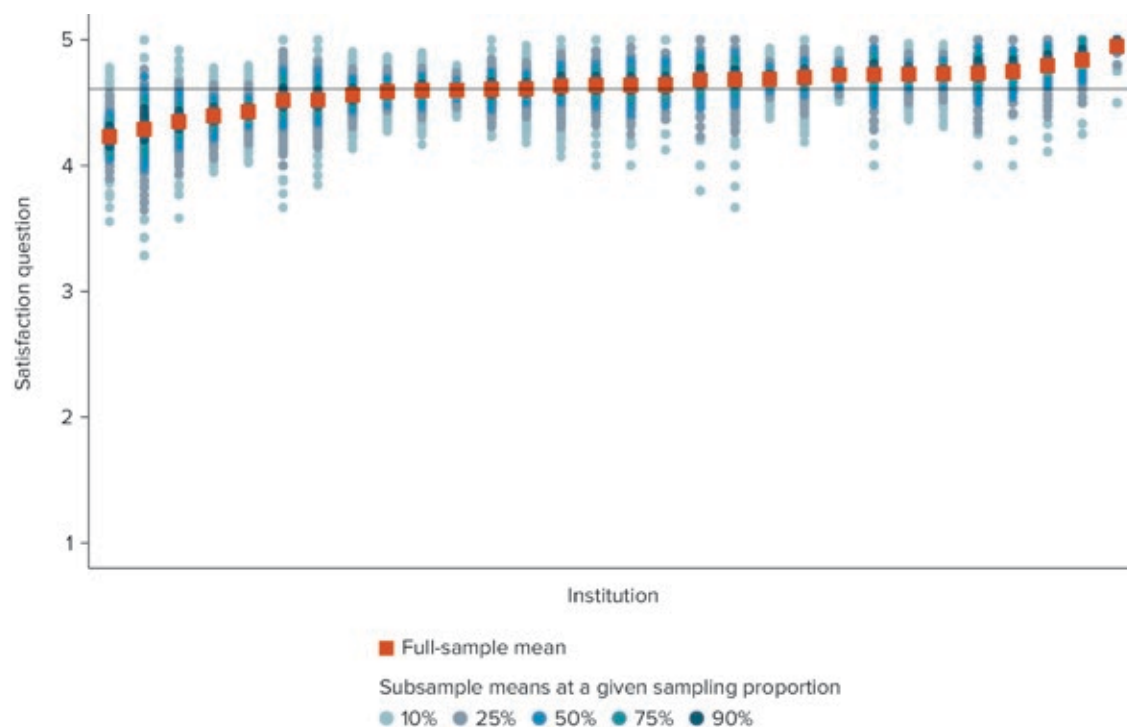
**FIGURE H.10** Romania: Performance Index



Source: Original figure for this publication.

Note: Institutional means are actual and simulated at different sampling proportions across 1,000 simulations.

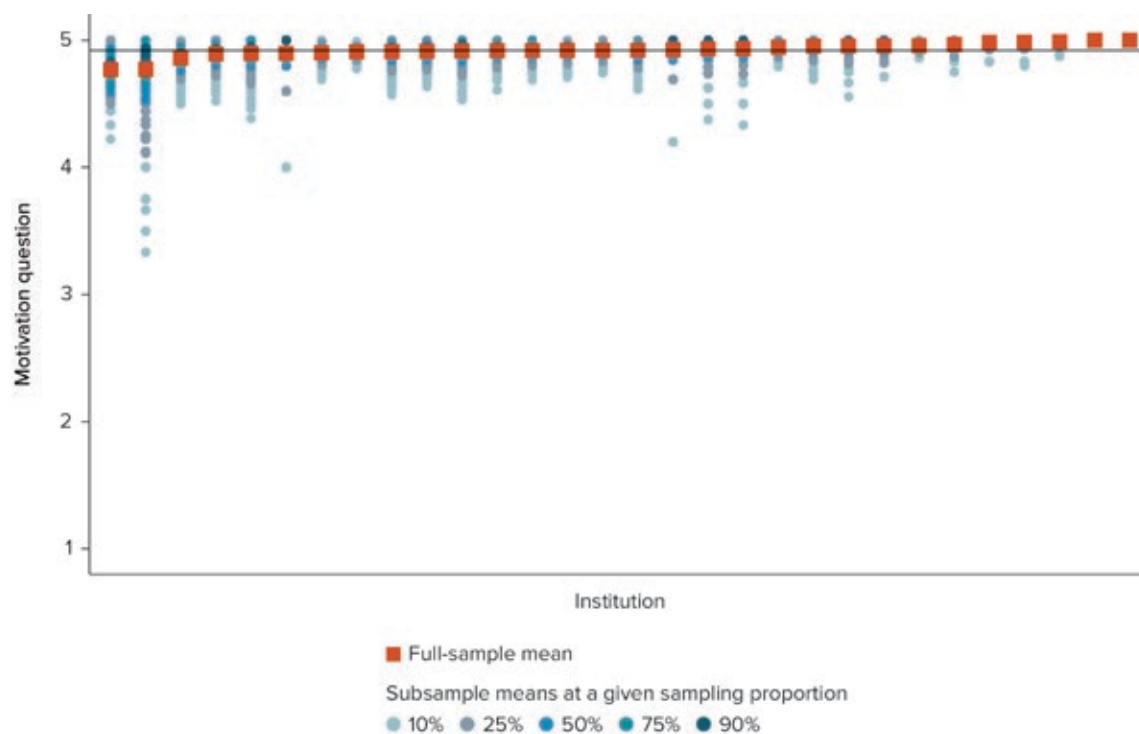
**FIGURE H.11** Romania: Satisfaction Question



Source: Original figure for this publication.

Note: Institutional means are actual and simulated at different sampling proportions across 1,000 simulations.

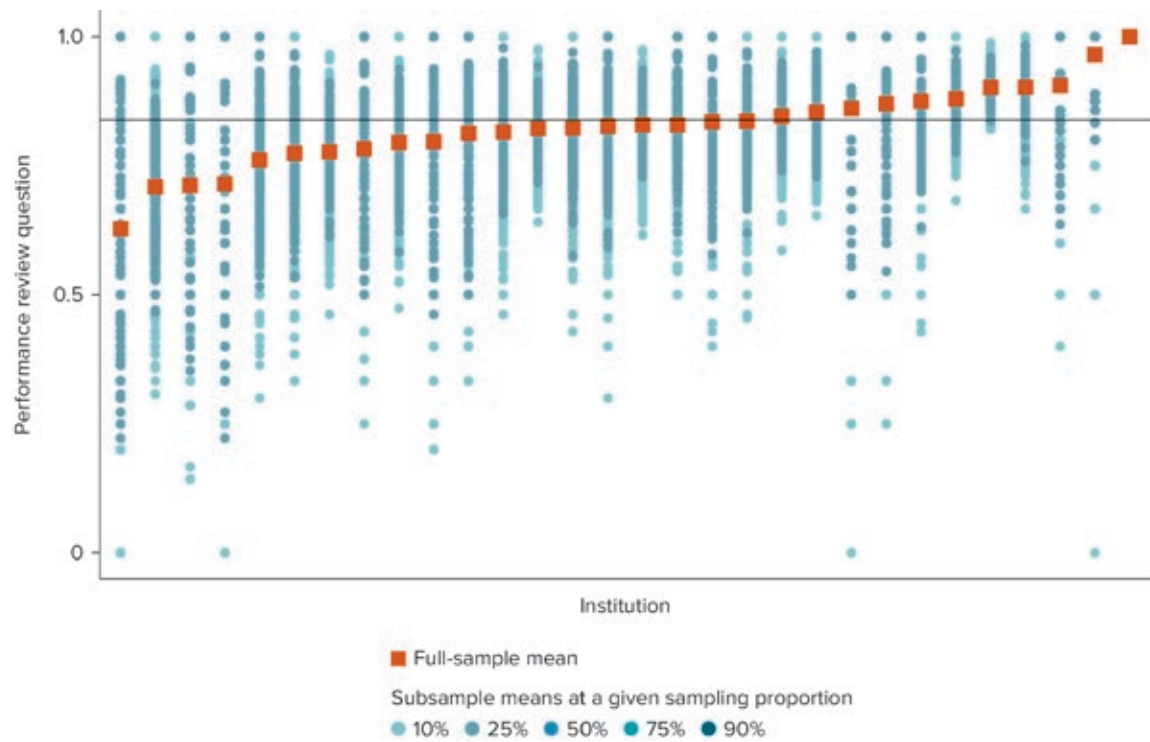
**FIGURE H.12** Romania: Motivation Question



Source: Original figure for this publication.

Note: Institutional means are actual and simulated at different sampling proportions across 1,000 simulations.

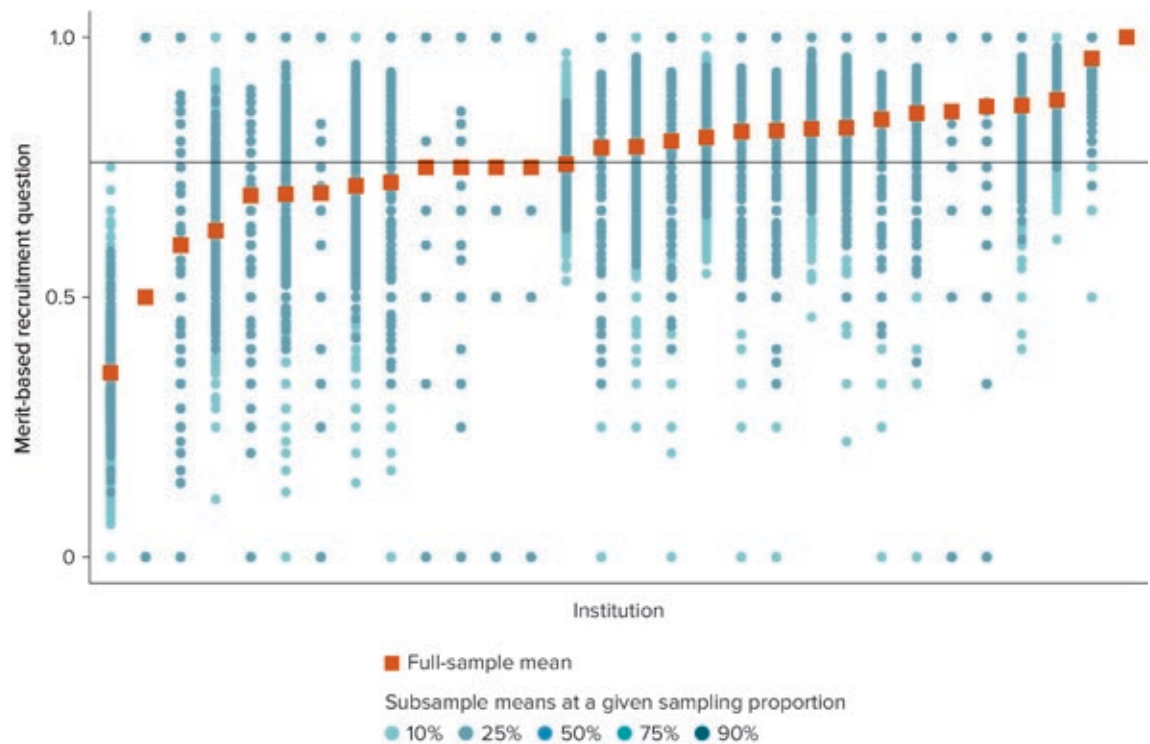
**FIGURE H.13** Romania: Performance Review Question



Source: Original figure for this publication.

Note: Institutional means are actual and simulated at different sampling proportions across 1,000 simulations.

**FIGURE H.14** Romania: Recruitment Question



Source: Original figure for this publication.

Note: Institutional means are actual and simulated at different sampling proportions across 1,000 simulations.



# APPENDIX I

## Further Details of Survey Questions

### Chapter 21 Appendix

**TABLE I.1** Survey Question Phrasing

Survey country	Variable	Phrasing	Original scale	Rescaled	Nonmissing
Chile	Job satisfaction	Estoy satisfecho/a con mi trabajo. (Translation: I am satisfied with my job.)	1 (strongly agree) to 5 (strongly disagree)	Reversed (so that higher values indicate stronger agreement)	10,926 (92.2%)
	Pay satisfaction	Estoy satisfecho/a con mi remuneración. (Translation: I'm satisfied with my pay.)	1 (strongly agree) to 5 (strongly disagree)		11,082 (93.6%)
	Motivation	Doy mi mejor esfuerzo para cumplir con mi trabajo, independiente de las dificultades. (Translation: I put my best effort to perform my work, independent of difficulties.)	1 (strongly agree) to 5 (strongly disagree)		10,955 (92.5%)
	Leadership 1 – Trust	Mi superior directo es confiable. (Translation: My direct superior is trustworthy.)	1 (strongly agree) to 5 (strongly disagree)		10,605 (89.5%)
	Leadership 2 – Generates motivation	Mi superior directo transmite y genera entusiasmo sobre la visión y misión de nuestro servicio. (Translation: My direct superior transmits and generates enthusiasm about the vision and mission of our service.)	1 (strongly agree) to 5 (strongly disagree)		10,675 (90.1)
	Performance incentives	¿Una evaluación positiva de mi desempeño podría ayudarme a obtener un ascenso? (Translation: A positive evaluation of my performance can help me to get promoted to a better position.)	1 (strongly agree) to 5 (strongly disagree)		9,303 (78.5%)
	Goal clarity	Tengo una comprensión clara de la misión y los objetivos de mi servicio. (Translation: I have a clear understanding of the mission and the objectives of my organization.)	1 (strongly agree) to 5 (strongly disagree)		10,973 (92.6%)
	Task clarity	Tengo una comprensión clara de cómo mi trabajo contribuye a la misión y los objetivos de mi servicio. (Translation: I have a clear understanding of how my work contributes to the mission and the objectives of my organization.)	1 (strongly agree) to 5 (strongly disagree)		10,978 (92.7%)

(continues on next page)



**TABLE I.1 Survey Question Phrasing (continued)**

Survey country	Variable	Phrasing	Original scale	Rescaled	Nonmissing
China	Job satisfaction	Are you generally satisfied with your job?	1 (very dissatisfied) to 5 (very satisfied)	—	2,477 (98.5%)
	Pay satisfaction	—	—	—	—
	Motivation	—	—	—	—
	Leadership 1 – Trust	—	—	—	—
	Leadership 2 – Generates motivation	—	—	—	—
	Performance incentives	Are you satisfied that your colleague's promotion is based on merit?	1 (strongly disagree) to 5 (strongly agree)	—	2,473 (98.4%)
	Goal clarity	—	—	—	—
	Task clarity	—	—	—	—
Colombia	Job satisfaction	Estoy satisfecho con mi trabajo. (Translation: I am satisfied with my work.)	—	—	9,693 (49.5%)
	Pay satisfaction	—	—	—	—
	Motivation	Doy mi mejor esfuerzo para cumplir con mi trabajo, sin importa las dificultades que existen. (Translation: I put my best effort to complete my work, independent of difficulties.)	—	—	9,710 (49.6%)
	Leadership 1 – Trust	—	—	—	—
	Leadership 2 – Generates motivation	—	—	—	—
	Performance incentives	—	—	—	—
	Goal clarity	—	—	—	—
	Task clarity	Tengo una comprensión clara de lo que se espera de mi cuando teletrabajo o realizo trabajo en casa. (Translation: I have a clear understanding of what is expected of me when teleworking or working from home.)	—	—	17,595 (89.9%)

*(continues on next page)*

**TABLE I.1 Survey Question Phrasing (continued)**

Survey country	Variable	Phrasing	Original scale	Rescaled	Nonmissing
Ethiopia	Job satisfaction	To what extent would you say you are satisfied with your experience of the civil service?	1 (very dissatisfied) to 4 (very satisfied)	Scale stretched: '3' recoded as '4' and '4' recoded as '5' to compare with the other surveys more easily, since here the original scale did not have a neutral option ('3').	1,117 (64%)
	Pay satisfaction	To what extent would you say you are satisfied with your salary?	1 (very dissatisfied) to 4 (very satisfied)	Scale stretched: '3' recoded as '4' and '4' recoded as '5' to compare with the other surveys more easily, since here the original scale did not have a neutral option ('3').	1,125 (64.4%)
	Motivation	Imagine that when you started your motivation was 100. What number would you say your motivation was now relative to that?	Continents scale	—	—
	Leadership 1 – Trust	—	—	—	—
	Leadership 2 – Generates motivation	—	—	—	—
	Performance incentives	On a scale of 1 to 5, how confident are you that you will get promoted if you perform your job well?	1 (very unconfident) to 5 (very confident)	—	—

(continues on next page)

**TABLE I.1 Survey Question Phrasing (continued)**

Survey country	Variable	Phrasing	Original scale	Rescaled	Nonmissing
Ethiopia (continued)	Goal clarity	Does your directorate have a clear set of targets derived from the organization's goals and objectives? Are they used to determine your work schedule?	<p>1 (The directorate does not have defined targets.)</p> <p>2 (The directorate has loosely defined targets, but there is no real connection between them and the tasks assigned to the staff. Midlevel staff have no real understanding of the targets.)</p> <p>3 (Targets are assigned to the directorate, as well as to the manager and employee levels, and these are generally well understood by midlevel staff. However, the tasks assigned to staff are not always related to those targets.)</p> <p>4 (Targets are clearly defined for the directorate and manager-level staff and are well understood by the midlevel staff. Tasks are typically closely related to the targets, although the connection is not always immediately obvious.)</p> <p>5 (Targets are clearly defined for the directorate, manager, and employee levels, and are well understood by all staff. All tasks are directly derived from the targets, which are regularly reviewed to ensure they remain on track.)</p>	—	1,121 (64.2%)
	Task clarity	When you arrive at work each day, do you and your colleagues know what their individual roles and responsibilities are in achieving the organization's goals?	<p>1 (Staff do not know what their roles and responsibilities are.)</p> <p>2 (Some staff have some idea of their roles and responsibilities are; it depends on what is going on in their organization at that time.)</p> <p>3 (Staff have a good idea of their roles and responsibilities, but it is not always clear how they contribute to their organization's goals.)</p> <p>4 (Generally, staff have a good understanding of their roles and responsibilities and how these contribute to the goals of their organization.)</p> <p>5 (Staff have a very good understanding of their roles and responsibilities. Their own roles and goals are clearly interconnected to those of their organization.)</p>	—	368 (21.1%)

(continues on next page)

**TABLE I.1 Survey Question Phrasing (continued)**

Survey country	Variable	Phrasing	Original scale	Rescaled	Nonmissing
Ghana	Job satisfaction	—	—	—	—
	Pay satisfaction	My salary is very satisfactory.	1 (strongly disagree) to 5 (strongly agree)	—	2,632 (99.9%)
	Motivation	I would feel an obligation to take time from my personal schedule to generate ideas/solutions for the organization if it is needed.	1 (strongly disagree) to 5 (strongly agree)	—	1,103 (41.9%)
	Leadership 1 – Trust	—	—	—	—
	Leadership 2 – Generates motivation	—	—	—	1,384 (52.5%)
	Performance incentives	On a scale of 1 to 5, how confident are you that you will keep your position and get promoted if you perform your job well?	1 (very unconfident) to 5 (very confident)	—	1,276 (48.4%)
	Goal clarity	Does your division have a clear set of targets derived from the organization's goals and objectives? Are they used to determine your work schedule?	1 (The division's targets are very loosely defined or not defined at all; if they exist, they are rarely used to determine our work schedule and our activities are based on ad hoc directives from senior management.) 1.522.53 (Targets are defined for the division and its individual officers (managers and staff). However, their use is relatively ad hoc and many of the division's activities do not relate to those targets.) 3.544.55 (Targets are defined for the division and individuals (managers and staff) and they provide a clear guide to the division and its staff as to what, the division should do. They are frequently discussed and used to benchmark performance.)	—	1,503 (57%)
	Task clarity	When you arrive at work each day, do you and your team know what your individual roles and responsibilities are in achieving the organization's goals?	1 (No. There is a general level of confusion as to what the division is trying to achieve on a daily basis and what individual's roles are toward those goals.) 1.522.53 (To some extent, or at least on some days. The division's main goals and individual's roles to achieve them are relatively clear, but it is sometimes difficult to see how current activities are moving us towards those.) 3.544.55 (Yes. It is always clear to the body of staff what the division is aiming to achieve with the day's activities and what individual's roles and responsibilities are toward that.)	—	1,510 (57.3%)

*(continues on next page)*

**TABLE I.1 Survey Question Phrasing (continued)**

Survey country	Variable	Phrasing	Original scale	Rescaled	Nonmissing
Guatemala	Job satisfaction	—	—	—	—
	Pay satisfaction	Estoy satisfecho con mi salario u honorarios, incluyendo otros complementos salariales. (Translation: I'm satisfied with my salary or fees including other salary supplements.)	1 (strongly disagree) to 5 (strongly agree)	—	1,138 (98.9%)
	Motivation	—	—	—	—
	Leadership 1 – Trust	¿Con qué frecuencia realiza las siguientes acciones su superior directo: cumple sus promesas y compromisos? (Translation: How frequently does your immediate supervisor realize the following actions: keeps his promises and commitments?)	1 (never) to 5 (very frequently/always)	—	579 (98.9%)
	Leadership 2 – Generates motivation	¿Con qué frecuencia realiza las siguientes acciones su superior directo: enuncia y genera entusiasmo por la visión y misión de mi institución? (Translation: How frequently does your immediate supervisor realize the following actions: Enunciates and generates enthusiasm for the vision and mission of my institution?)	1 (never) to 5 (very frequently/always)	—	585 (50.3%)
	Performance incentives	En una escala del 1 al 5, ¿Cuál confiado está de que será promovido en el futuro si es que desempeña bien su trabajo? (Translation: On a scale from 1 to 5, how confident are you that in the future you will be promoted if you perform your job well?)	1 (never) to 5 (very frequently/always)	—	574 (49.9%)
	Goal clarity	¿Se utilizan los objetivos para determinar su agenda de trabajo? (Translation: Are the objectives (of the institution) used to determine your work agenda?)	1 (The institution does not have defined objectives.) 2 (The institution has vaguely defined objectives, but there is no real connection between them and the tasks assigned to the staff.) 3 (The institution has well-defined objectives; however, the tasks assigned to the staff are not always related to those objectives.) 4 (The tasks are often related to the objectives, although the connection is not always obvious.) 5 (The tasks are derived directly from the objectives, which are periodically revised to ensure they stay on the right track.)	—	748 (65%)

(continues on next page)

**TABLE I.1 Survey Question Phrasing (continued)**

Survey country	Variable	Phrasing	Original scale	Rescaled	Nonmissing
Guatemala (continued)	Task clarity	Cuando llegan al trabajo todos los días, ¿Sabén Ud. y sus colegas cuáles son sus funciones y responsabilidades individuales? (Translation: When you arrive at work every day, do you and your colleagues know what your functions and individual responsibilities are?)	1 (The staff do not know what are their functions and responsibilities.) 2 (The staff have some idea about their functions and responsibilities within their work team.) 3 (The staff know well their functions within their work team.) 4 (The staff know well their functions within their unit.) 5 (The staff know perfectly what their functions and responsibilities are within their unit and their own institution.)	—	747 (64.9%)
Liberia	Job satisfaction	To what extent would you say you are satisfied with your experience of the civil service?	1 (very dissatisfied) to 4 (very satisfied)	Scale stretched: '3' recoded as '4' and '4' recoded as '5' to compare with the other surveys more easily, since here the original scale did not have a neutral option ('3').	2,651 (98.6%)
	Pay satisfaction	How satisfied are you with your total income?	1 (very dissatisfied) to 4 (very satisfied)	Scale stretched: '3' recoded as '4' and '4' recoded as '5' to compare with the other surveys more easily, since here the original scale did not have a neutral option ('3').	2,670 (99.3%)
	Motivation	How motivated are you to work as a civil servant today?	0 (not motivated at all) to 10 (extremely motivated)	—	2,687 (99.9%)
	Leadership 1 – Trust	How much do you trust each of the following types of people: supervisors in your unit?	(1 not at all) to 4 (I trust them a lot)	Scale stretched: '3' recoded as '4' and '4' recoded as '5' to compare with the other surveys more easily, since here the original scale did not have a neutral option ('3').	839 (31.2%)

*(continues on next page)*



**TABLE I.1 Survey Question Phrasing (continued)**

Survey country	Variable	Phrasing	Original scale	Rescaled	Nonmissing
Liberia (continued)	Leadership 2 – Generates motivation	—	—	—	—
	Performance incentives	How confident are you that you will get promoted if you perform your job well?	1 (not confident at all; it will not happen) to 4 (very confident; you are certain it will happen)	Scale stretched: '3' recoded as '4' and '4' recoded as '5' to compare with the other surveys more easily, since here the original scale did not have a neutral option ('3').	486 (18.1%)
	Goal clarity	Does your unit have clearly defined targets? Are they used to determine the unit's work schedule?	1 (No. The unit does NOT have defined targets.)  2 (Partially. The unit has loosely defined targets, but the staff do not understand these and they are not used to inform what tasks staff do.)  3 (Partially. The unit has loosely defined targets, but the staff do not understand these and they are not used to inform what tasks staff do.)  4 (Yes, targets are clearly defined for the unit and the managers, and are well understood by the midlevel staff. Tasks are typically closely related to the targets, BUT the connection is not always clear.)  5 (Yes, there is a clear set of targets defined for the unit, which are well understood by all staff. These targets inform all tasks.)	—	1,407 (52.3%)

(continues on next page)

**TABLE I.1 Survey Question Phrasing (continued)**

Survey country	Variable	Phrasing	Original scale	Rescaled	Nonmissing
Liberia (continued)	Task clarity	When arriving at work every day, do staff in the unit know what their individual roles and responsibilities are in achieving the unit's goals?	<p>1 (No. Staff do not know what their roles and responsibilities are.)</p> <p>2 (Some staff have some idea of their roles and responsibilities. It depends on what is going on in the name of the ministry, agency, commission [MAC]] at that time.)</p> <p>3 (Yes, staff generally have a good idea of their roles and responsibilities, but it is not always clear how they contribute to the MAC goals.)</p> <p>4 (Staff have a good understanding of their roles and responsibilities and how these contribute to their organization's goals.)</p> <p>5 (Staff have a very good understanding of their roles and responsibilities and how these contribute to their organization's goals.)</p>	—	1,410 (52.4%)
Philippines	Job satisfaction	—	—	—	—
	Pay satisfaction	You are satisfied with the pay you receive for your work.	1 (strongly disagree) to 5 (strongly agree)	—	1,768 (100%)
	Motivation	—	1 to 100	—	—
	Leadership 1 – Trust	—	—	—	—
	Leadership 2 – Generates motivation	—	—	—	—
	Performance incentives	—	—	—	—
	Goal clarity	Management and staff in your bureau are working together to set clear achievable targets for your bureau.	1 (strongly disagree) to 5 (strongly agree)	—	1,766 (99.9%)
	Task clarity	—	—	—	—
Romania	Job satisfaction	Overall, I am satisfied with my job.	1 (strongly disagree) to 5 (strongly agree)	—	2,716 (98%)
	Pay satisfaction	I am satisfied with my salary.	1 (strongly disagree) to 5 (strongly agree)	—	2,690 (97%)
	Motivation	I put forth my best effort to get my job done regardless of any difficulties OR I am willing to do extra work for my job that isn't really expected of me.	1 (strongly disagree) to 5 (strongly agree)	—	2,726 (98.3%)
	Leadership 1 – Trust	Can be trusted to carry out promises and commitments. (face-to-face mode only)	1 (never) to 5 (very frequently, if not always)	—	1,624 (58.6%)

*(continues on next page)*

**TABLE I.1 Survey Question Phrasing (continued)**

Survey country	Variable	Phrasing	Original scale	Rescaled	Nonmissing
Romania (continued)	Leadership 2 – Generates motivation	Communicates effectively the institution's vision and mission to employees. (face-to-face mode only)	1 (never) to 5 (very frequently, if not always)	—	1,367 (60.1%)
	Performance incentives	On a scale of 1 to 5, how confident are you that you will get promoted to the next professional grade if you perform your job well? (Note: Divided into individual vs. organizational-level modules.)	1 (very unconfident) to 5 (very confident) 6 – I cannot get promoted, because I am at the highest professional grade. 7 – I am not interested in getting promoted.	Responses 6 and 7 recoded to NA.	<512 (22.1%)
	Goal clarity	I have a good understanding of my institution's goals.	1 (strongly disagree) to 5 (strongly agree)	—	2,707 (97.7%)
	Task clarity	When I arrive at work each day, I know what my individual roles and responsibilities are in achieving the institution's goals.	1 (strongly disagree) to 5 (strongly agree)	—	2,723 (98.2%)
United States	Job satisfaction	Considering everything, how satisfied are you with your job?	1 (very dissatisfied) to 5 (very satisfied)	—	573,255 (95%)
	Pay satisfaction	Considering everything, how satisfied are you with your pay?	1 (very dissatisfied) to 5 (very satisfied)	—	572,853 (94.9%)
	Motivation	When needed I am willing to put in the extra effort to get a job done.	1 (strongly disagree) to 5 (strongly agree)	—	601,274 (99.6%)
	Leadership 1 – Trust	I have trust and confidence in my supervisor.	1 (strongly disagree) to 5 (strongly agree)	—	582,758 (96.5%)
	Leadership 2 – Generates motivation	In my organization, senior leaders generate high levels of motivation and commitment in the workforce.	1 (strongly disagree) to 5 (strongly agree)	—	565,650 (93.7%)
	Performance incentives	Awards in my work unit depend on how well employees perform their jobs.	1 (strongly disagree) to 5 (strongly agree)	—	558,198 (92.5%)
	Goal clarity	Managers communicate the goals of the organization.	1 (strongly disagree) to 5 (strongly agree)	—	569,466 (94.3%)
	Task clarity	I know how my work relates to the agency's goals.	1 (strongly disagree) to 5 (strongly agree)	—	598,601 (99.2%)

Source: Original table for this publication.

Note: The table displays information on all surveys and variables used in the analyses in chapter 21. It specifies the name of each variable, original phrasing of a relevant survey question, original response scale, details of any rescaling operations that were performed to facilitate comparisons, and also statistics (absolute and relative number) on nonmissing responses for a given variable. For each survey, the same set of variables is displayed, even if some of there were not present in a given setting, to facilitate comparisons. — = not applicable.

**TABLE I.2 Robustness Check: ANOVAs and Nested Model with Outliers**

Variable	Models	Sum of squares	RSS	df	F-statistic	Pr (>F)	R <sup>2</sup>	Adj. R <sup>2</sup>
Job satisfaction	Demographics	—	8,458.04	—	—	—	0.08	0.08
	Demographics + country	107.89	8,350.15	3.00	68.77	0.00	0.10	0.10
	Demographics + country + unit	169.93	8,180.22	102.00	3.19	0.00	0.11	0.11
	Demographics + country + unit + subunit	362.60	7,817.62	504.00	1.38	0.00	0.15	0.12
	Nested RE model	—	—	—	—	—	0.15	—
Pay satisfaction	Demographics	—	13,230.41	—	—	—	0.11	0.11
	Demographics + country	3,780.90	9,449.51	3.00	1,129.26	0.00	0.36	0.36
	Demographics + country + unit	1,103.07	8,346.44	111.00	8.90	0.00	0.44	0.43
	Demographics + country + unit + subunit	963.81	7,382.63	347.00	2.49	0.00	0.50	0.47
	Nested RE model	—	—	—	—	—	0.45	—
Motivation	Demographics	—	592.31	—	—	—	0.02	0.01
	Demographics + country	38.00	554.30	1.00	233.32	0.00	0.08	0.08
	Demographics + country + unit	12.09	542.21	18.00	4.12	0.00	0.10	0.09
	Demographics + country + unit + subunit	18.54	523.67	82.00	1.39	0.01	0.13	0.10
	Nested RE model	—	—	—	—	—	0.24	—

Source: Original table for this publication.

Note: *Demographics* refers to the inclusion of the following set of variables: respondent's gender and tenure in public service, as well as age (all except for the United States) and managerial status (for Chile, Colombia, Ghana, Guatemala, and United States). Country, unit, and subunit enter the regression as dummy variables, taking the value 1 if the respondent is in the corresponding country, unit, or subunit, respectively. The first four lines for each variable summarize test statistics for analyses of variance (ANOVAs) and how the model fit compares to the next more complex model. The first line refers to a model that only includes demographic predictor variables. The second one adds country dummies, the third adds unit-level dummies, and the fourth, subunit-level dummies. The *F*-test for each model indicates whether it has a better fit than the simpler model specified above. Models with lower residual sums of squares (RSS) and higher (adjusted) *R*<sup>2</sup> explain a larger proportion of the variance. The last line for each variable reports the model fit for a nested model that nests subunits into units and units into countries. If the adjusted *R*<sup>2</sup> of the nested model is larger than that in the lines reported above it, it indicates that the nested model is a better fit. df = degrees of freedom; Pr = probability. — = not applicable.

**TABLE I.3 Compare Models, Full Data Set versus List-Wise Deletion: ANOVAs,  $R^2$** 

Variable	Modification	Model	Residual df	RSS	df	Sum of squares	F-statistic	Pr	$R^2$	Adj. $R^2$
Job satisfaction	List-wise deletion	Demographics	24,715	19,382.56	–	–	–	–	0.03	0.03
		Demographics + country	24,712	17,154.54	3	2,228.03	11,41.51	0.00	0.14	0.14
		Demographics + country + unit	24,469	16,539.42	243	615.12	3.89	0.00	0.17	0.17
		Demographics + country + unit + subunit	22,780	14,820.82	1,689	1,718.60	1.56	0.00	0.26	0.20
		Nested RE model	–	–	–	–	–	–	0.44	–
	Box-Cox transformation	Demographics	616,400	984,754.94	–	–	–	–	0.01	0.01
		Demographics + country	616,394	969,112.64	6	15,642.30	1,704.36	0.00	0.03	0.03
		Demographics + country + unit	616,091	953,859.15	303	15,253.48	32.91	0.00	0.04	0.04
		Demographics + country + unit + subunit	614,047	939,269.63	2,044	14,589.52	4.67	0.00	0.06	0.05
		Nested RE model	–	–	–	–	–	–	0.24	–
Pay satisfaction	List-wise deletion	Demographics	18,321	35,185.68	–	–	–	–	0.03	0.03
		Demographics + country	18,317	23,714.11	4	11,471.58	2,418.44	0.00	0.35	0.35
		Demographics + country + unit	18,124	22,225.93	193	1,488.18	6.50	0.00	0.39	0.38
		Demographics + country + unit + subunit	16,665	19,762.08	1,459	2,463.86	1.42	0.00	0.45	0.40
		Nested RE model	–	–	–	–	–	–	0.49	–
	Box-Cox transformation	Demographics	609,406	736,959.36	–	–	–	–	0.01	0.01
		Demographics + country	609,399	704,163.73	7	32,795.63	4,304.05	0.00	0.06	0.06
		Demographics + country + unit	609,130	691,504.17	269	12,659.56	43.23	0.00	0.07	0.07
		Demographics + country + unit + subunit	607,351	661,119.44	1,779	30,384.73	15.69	0.00	0.11	0.11
		Nested RE model	–	–	–	–	–	–	0.40	–

*(continues on next page)*

**TABLE I.3 Compare Models, Full Data Set versus List-Wise Deletion: ANOVAs,  $R^2$  (continued)**

Variable	Modification	Model	Residual df	RSS	df	Sum of squares	F-statistic	Pr	$R^2$	Adj. $R^2$
Motivation	List-wise deletion	Demographics	24,767	8,859.33	–	–	–	–	0.01	0.01
		Demographics + country	24,764	8,594.48	3	264.84	262.76	0.00	0.04	0.04
		Demographics + country + unit	24,542	8,423.15	222	171.33	2.30	0.00	0.06	0.05
		Demographics + country + unit + subunit	22,840	7,673.73	1,702	749.42	1.31	0.00	0.14	0.07
		Nested RE model	–	–	–	–	–	–	0.11	–
	Box-Cox transformation	Demographics	642,380	848,654.29	–	–	–	–	0.01	0.01
		Demographics + country	642,375	840,307.36	5	8,346.93	1,292.33	0.00	0.02	0.02
		Demographics + country + unit	642,093	834,928.20	282	5,379.16	14.77	0.00	0.02	0.02
		Demographics + country + unit + subunit	640,062	826,810.76	2,031	8,117.44	3.09	0.00	0.03	0.03
		Nested RE model	–	–	–	–	–	–	0.44	–
Leadership trust	List-wise deletion	Demographics	11,516	14,890.51	–	–	–	–	0.00	0.00
		Demographics + country	11,514	147,44.67	2	145.83	60.74	0.00	0.01	0.01
		Demographics + country + unit	11,456	143,46.60	58	398.07	5.72	0.00	0.04	0.04
		Demographics + country + unit + subunit	10,750	12,905.69	706	1,440.90	1.70	0.00	0.14	0.08
		Nested RE model	–	–	–	–	–	–	0.10	–
	Box-Cox transformation	Demographics	608,171	1,113,578.1	–	–	–	–	0.01	0.01
		Demographics + country	608,167	1,112,483.9	4	1,094.22	152.95	0.00	0.01	0.01
		Demographics + country + unit	608,033	1,100,294.6	134	12,189.28	50.86	0.00	0.02	0.02
		Demographics + country + unit + subunit	607,044	1,085,723.0	989	14,571.60	8.24	0.00	0.03	0.03
		Nested RE model	–	–	–	–	–	–	0.05	–
Leadership motivation	List-wise deletion	Demographics	13,409	19,241.96	–	–	–	–	0.00	0.00
		Demographics + country	13,406	18,281.92	3	960.04	251.31	0.00	0.05	0.05
		Demographics + country + unit	13,291	17,604.61	115	677.31	4.63	0.00	0.09	0.08
		Demographics + country + unit + subunit	12,218	15,558.49	1,073	2,046.12	1.50	0.00	0.19	0.11
		Nested RE model	–	–	–	–	–	–	0.14	–

(continues on next page)



**TABLE I.3 Compare Models, Full Data Set versus List-Wise Deletion: ANOVAs,  $R^2$  (continued)**

Variable	Modification	Model	Residual df	RSS	df	Sum of squares	F-statistic	Pr	$R^2$	Adj. $R^2$
Leadership motivation (continued)	Box-Cox transformation	Demographics	591,842	969,877.90	–	–	–	–	0.01	0.01
		Demographics + country	591,838	964,909.59	4	4,968.31	800.30	0.00	0.01	0.01
		Demographics + country + unit	591,679	937,603.61	159	27,305.99	110.65	0.00	0.04	0.04
		Demographics + country + unit + subunit	590,411	916,324.12	1,268	21,279.49	10.81	0.00	0.06	0.06
		Nested RE model	–	–	–	–	–	–	0.14	–
Meritocratic promotion	List-wise deletion	Demographics	12,340	28,368.88	–	–	–	–	0.01	0.01
		Demographics + country	12,336	23,104.74	4	5,264.13	755.86	0.00	0.19	0.19
		Demographics + country + unit	12,144	22,008.51	192	1,096.24	3.28	0.00	0.23	0.22
		Demographics + country + unit + subunit	10,818	18,835.24	1,326	3,173.27	1.37	0.00	0.34	0.25
		Nested RE model	–	–	–	–	–	–	0.30	–
	Box-Cox transformation	Demographics	585,446	887,439.70	–	–	–	–	0.03	0.03
		Demographics + country	585,439	879,234.12	7	8,205.58	812.63	0.00	0.03	0.03
		Demographics + country + unit	585,171	860,601.37	268	18,632.75	48.20	0.00	0.06	0.05
		Demographics + country + unit + subunit	583,544	841,764.24	1,627	18,837.13	8.03	0.00	0.08	0.07
		Nested RE model	–	–	–	–	–	–	0.27	–
Goal clarity	List-wise deletion	Demographics	15,469	10,188.80	–	–	–	–	0.03	0.03
		Demographics + country	15,465	9,016.04	4	1,172.76	555.88	0.00	0.14	0.14
		Demographics + country + unit	15,272	8,473.67	193	542.37	5.33	0.00	0.19	0.18
		Demographics + country + unit + subunit	13,927	7,345.56	1,345	1,128.11	1.59	0.00	0.30	0.22
		Nested RE model	–	–	–	–	–	–	0.50	–
	Box-Cox transformation	Demographics	601,764	989,415.21	–	–	–	–	0.01	0.01
		Demographics + country	601,757	966,805.58	7	22,609.63	2,084.26	0.00	0.03	0.03
		Demographics + country + unit	601,488	947,395.70	269	19,409.88	46.56	0.00	0.05	0.05
		Demographics + country + unit + subunit	599,824	929,537.61	1,664	17,858.08	6.93	0.00	0.07	0.07
		Nested RE model	–	–	–	–	–	–	0.23	–

(continues on next page)

**TABLE I.3 Compare Models, Full Data Set versus List-Wise Deletion: ANOVAs,  $R^2$  (continued)**

Variable	Modification	Model	Residual df	RSS	df	Sum of squares	F-statistic	Pr	$R^2$	Adj. $R^2$
Task clarity	List-wise deletion	Demographics	35,804	24,228.95	—	—	—	—	0.01	0.01
		Demographics + country	35,799	22,332.46	5	1,896.50	632.98	0.00	0.08	0.08
		Demographics + country + unit	35,482	21,651.81	317	680.65	3.58	0.00	0.11	0.10
		Demographics + country + unit + subunit	32,996	19,772.23	2,486	1,879.58	1.26	0.00	0.19	0.12
		Nested RE model	—	—	—	—	—	—	0.45	—
	Box-Cox transformation	Demographics	649,539	991,216.84	—	—	—	—	0.01	0.01
		Demographics + country	649,532	981,092.91	7	10,123.93	977.97	0.00	0.02	0.02
		Demographics + country + unit	649,155	969,147.51	377	11,945.40	21.43	0.00	0.03	0.03
		Demographics + country + unit + subunit	646,343	955,851.08	2,812	13,296.43	3.20	0.00	0.05	0.04
		Nested RE model	—	—	—	—	—	—	0.34	—

Source: Original table for this publication.

Note: For each dependent variable, two sets of models are presented. *List-wise deletion* rows display models where instead of imputing missing values (as in table I.2) all rows with missing values for any of the variables included in a model are removed. In turn, *Box-Cox* rows show results after adjusting the dependent variable for nonnormal distribution, using Box-Cox transformation. The first four lines for each variable summarize test statistics for analyses of variance (ANOVAs) and how the model fit compares to the next more complex model. The first row for a given dependent variable-modification combination always refers to a model that only includes demographic predictor variables. Those include the following set of variables: respondent's gender and tenure in public service, as well as age (present in all surveys except for the United States) and managerial status (for Chile, Colombia, Ghana, Guatemala, and United States). Rows two through four progressively add country-, unit-, and subunit-level dummies to the model. The *F*-test for each model indicates whether it has a better fit than the simpler model specified above. Models with lower residual sums of squares (RSS) and higher (adjusted)  $R^2$  explain a larger proportion of the variance. The last line for each variable reports the model fit for a nested model, which nests subunits into units and units into countries. If the  $R^2$  of the nested model is larger than that in the lines reported above it, it indicates that the nested model is a better fit. df = degrees of freedom; Pr = probability; RE = residual error. — = not applicable.

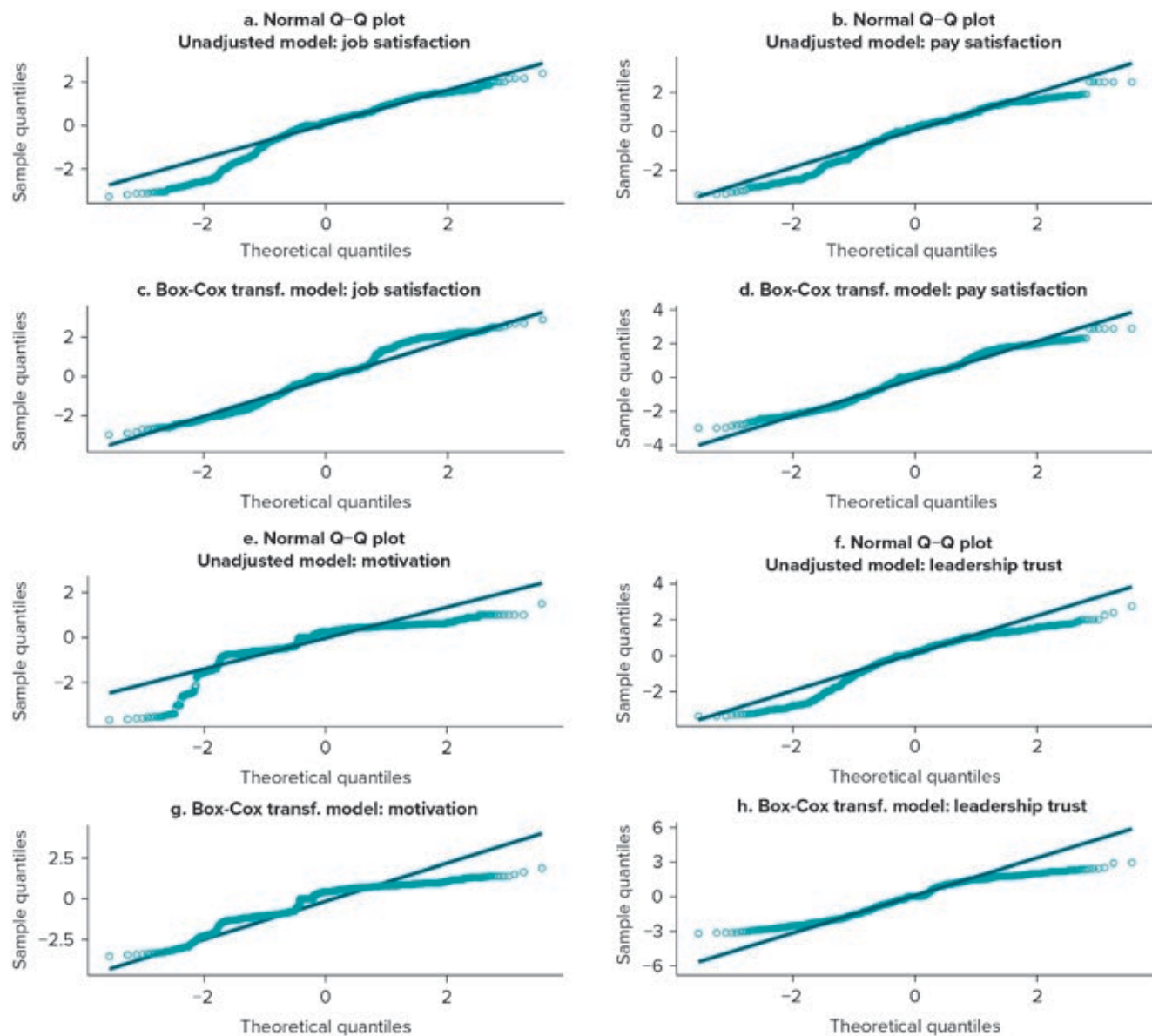
**TABLE I.4 Regression Diagnostic Statistics**

Variable	Outliers (N)	$R^2$ no outliers	Outliers Cook's	Lambda	$R^2$ Cook's	$R^2$ Box- Cox	Adj. $R^2$ no outliers	Adj. $R^2$ Cook's	Adj. $R^2$ Box- Cox	RE no outliers	RE Cook's	RE Box-Cox
Job satisfaction	1,845	0.23	1,287	2.7	0.386	0.345	0.186	0.357	0.311	0.534	0.703	1.038
Pay satisfaction	0	0.536	981	0.2	0.64	0.541	0.503	0.615	0.508	1.018	0.896	1.051
Motivation	791	0.278	1,057	3.5	0.449	0.328	0.237	0.42	0.292	0.52	0.595	1.1
Leadership trust	493	0.285	297	2.2	0.241	0.202	0.192	0.161	0.112	0.601	0.898	1.182
Leadership motivation	0	0.262	377	2.2	0.338	0.281	0.172	0.258	0.193	0.983	0.821	1.127
Meritocratic promotion	0	0.406	692	1.7	0.521	0.418	0.326	0.457	0.34	1.1	0.928	1.143
Goal clarity	597	0.407	635	3.2	0.455	0.427	0.348	0.401	0.37	0.579	0.68	1.105
Task clarity	1,190	0.19	1,400	3.2	0.19	0.194	0.146	0.149	0.15	0.598	0.727	1.192

Source: Original table for this publication.

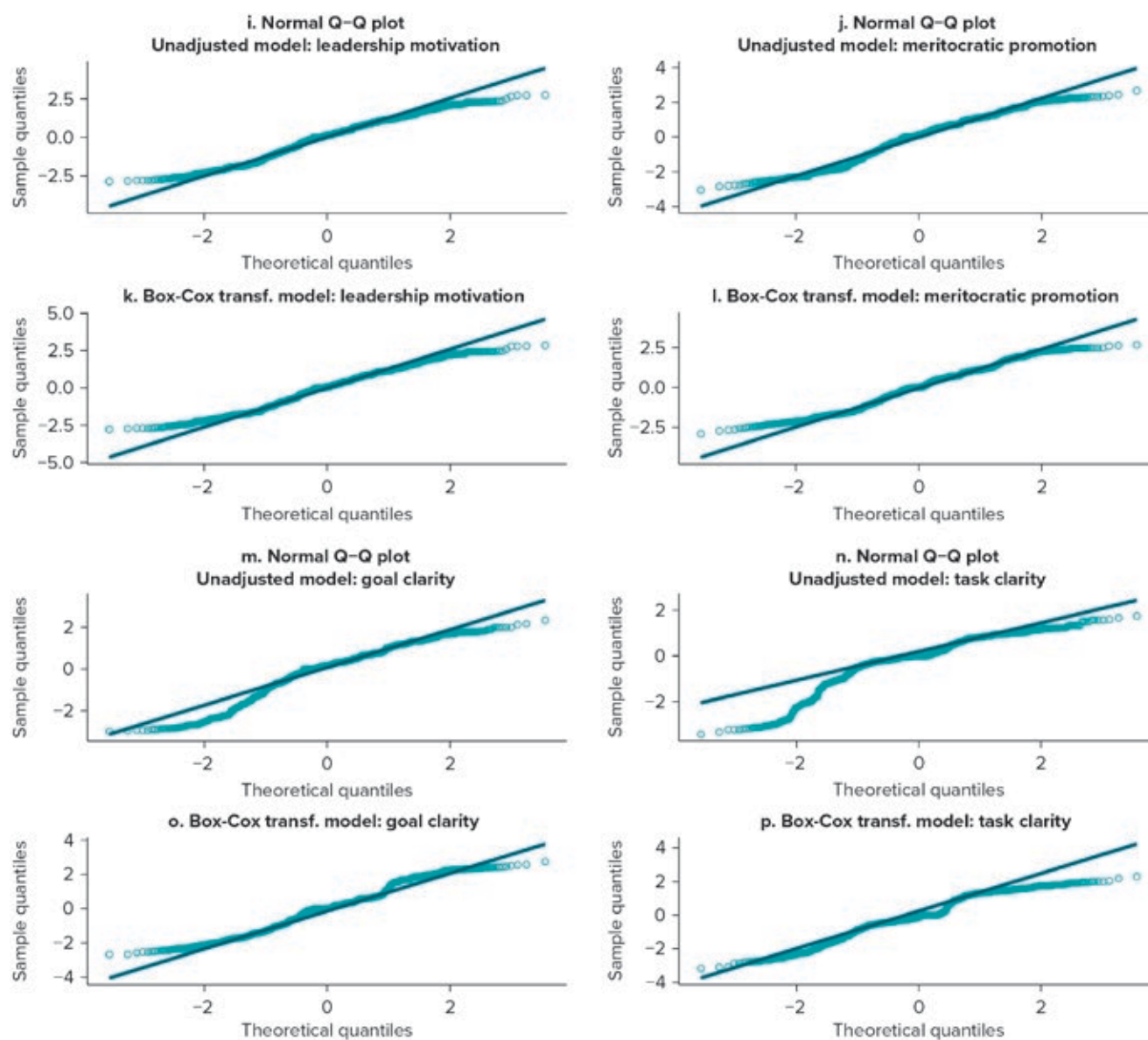
Note: The table summarizes several regression diagnostics aimed at showing the prevalence of outliers and the adequacy of transformations. The second column summarizes the number of outliers for each variable. The third, the  $R^2$  (a measure of the proportion of variance explained) for a model run on data with all outliers removed. The fourth column reports the number of outliers as determined by Cook's distance. Cook's distance uses both leverage and the residual error term to calculate the degree of "outlierness" for each point. The median point as in the analyses here is used as a cut-off point. The fifth column reports the lambda for the Box-Cox transformation. The lambda is a transformation factor that is nonzero and nonnegative. The next five columns report the  $R^2$  for models using transformations to address the presence of outliers. The  $R^2$  is larger than that reported in the third column if the transformation improved the model fit. The last four columns report the residual error for each of the models after transformation. The models with the smallest residual error (RE) are the best fit.

**FIGURE I.1** Residuals before and after Box-Cox Transformation



(continues on next page)

**FIGURE I.1** Residuals before and after Box-Cox Transformation (*continued*)



## APPENDIX J

# Framework for Coding the Complexity and Sensitivity of Questions in Public Administration Surveys

## Chapter 22 Appendix

### SUMMARY OF THE TOOLKIT

This toolkit presents a framework for manual coding of complexity and sensitivity of survey questions that is utilized in chapter 22 of *The Government Analytics Handbook* for analyzing item nonresponse across three public officials surveys (Guatemala, Romania, United States). The framework is based on past research regarding the psychological, linguistic, and social considerations that might cause survey respondents to be unable to answer a particular question (complexity) or be unwilling to do so (sensitivity). Overall question complexity is assessed by averaging scores across 10 subindicators, each scored on a scale from 0 to 2. The subindicators aim to measure the mental effort required from respondents in the process of comprehending the question, retrieving and integrating the relevant information, and then translating it to an answer. Sensitivity is measured in the same manner across four subindicators, which are concerned with the invasion of the privacy, social-emotional and formal threats of disclosure, as well as the interaction between the two. By presenting the rationale and the coding framework for each of the subindicators, this toolkit can act as a manual for researchers who aim to measure survey question complexity and sensitivity, both in public officials surveys, as well as general population surveys more broadly. It might also provide a reference for survey designers, who would like to minimize the complexity and sensitivity of their questions, thus limiting the prevalence of item nonresponse.



## BACKGROUND

The framework has been designed to assess the level of complexity and sensitivity of survey questions.<sup>1</sup> Each question is assessed on a scale from 0 to 2, with 0 being the least complex/sensitive and 2 most complex/sensitive, across a range of subindicators whose arithmetic average determines overall complexity and sensitivity of a question.

The **10 complexity subindicators**, which are largely based on Krosnick and Presser (2009), assess the difficulty of the four-stage mental process involved in answering a survey question:

- **Question comprehension** refers to the ease of understanding a question based on its linguistics, presence of reference frame, and the number of subquestions being asked.
- **Information retrieval** refers to the ease of retrieving the information the question is asking, which depends on the respondent's familiarity with the subject, the time frame, and specificity of information required.
- **Information integration** refers to the extent to which retrieved information must be integrated and evaluated to formulate an answer.
- **Translation of information to answer** refers to the ease of mapping the formulated answer into the predefined answer options.

The **four sensitivity subindicators**, which are largely drawn from Helmke and Levitsky (2006) and Tourangeau and Yan (2007), identify the sources of sensitivity in a question, which comes from the formal and informal institutions (the latter also includes invasion of privacy) that regulate human interaction in the society and the interaction between these two institutions. Consequently, a question's sensitivity level is assessed based on four different subindicators:

- **Invasion of privacy** refers to the extent to which respondent is asked to discuss "taboo" or private topics that may be inappropriate in everyday conversation.
- **Social-emotional threat of disclosure** (informal sensitivity) refers to the extent to which the respondent is concerned with the social and/or emotional consequences of a truthful answer should the information become known to a third party.
- **Threat of formal sanctions** (formal sensitivity) refers to the extent to which the respondent is concerned with the legal and/or formal consequences of a truthful answer should the information become known to a third party.
- **Relationship between formal and informal sensitivity** refers to the likelihood that a behavior/attitude may cause a threat of both social-emotional disclosure and formal sanctions.

The framework attempts to produce a standardized set of rules that enable an objective assessment of complexity and sensitivity of public administration surveys. For each survey item, the assessor can obtain the average score of both complexity and sensitivity, which is based on the arithmetic mean of all the assigned scores for all components that affect complexity and sensitivity.

The following section discusses each subindicator of complexity and sensitivity in turn, explaining its scope, rationale for inclusion, and an applied example of how to use it for coding a survey question, based on questions included in the Romania Public Administration Survey in 2019.

## COMPLEXITY

### Question comprehension

---

**Component name:** Length and complexity of syntax

---

**Question:** How long is a question and how complicated is the syntax used in the question?

**Coding:** Categorical; {0,1,2}

**Scale:**

0 (short question with easy syntax)

1 (short question with difficult syntax or long question with easy syntax)

2 (long question with difficult syntax)

**Description:** This component has two dimensions. It aims to gauge the length of a question and the difficulty level of the syntax used in the question. For the length dimension, first, count the number of characters (including spaces) for every question in the survey. Afterwards, find out both the maximum ( $y$ ) and minimum number of characters ( $x$ ) for a question in the survey.

Let  $D = \frac{(y+1-x)}{2}$ . Questions with number of characters ( $n$ ) between  $x < n < x + D$  are considered as short and questions with  $n$  between  $x + D \leq n < x + 2D$  are considered as long. Thereby, whether a question is short or long is determined by its length relative to the other questions in the questionnaire. Moreover, these definitions of “short” and “long” are also flexible since the values of  $y$  and  $x$  vary by survey.

For syntax complexity, simple sentences are considered as simple syntax, while compound, complex, and complex-compound sentences are considered as complicated syntax.

For the calculation of the final score for this component, first, assign a score of 0 for a short question and 1 for a long question. Similarly, assign a score of 0 for a question with simple syntax and 1 for a question with complicated syntax. Then, to obtain the final score, one can just add these two scores. A short question with easy syntax will get a total score of 0. A short question with complicated syntax or a long question with easy syntax will get a total score of 1. Lastly, a long question with difficult syntax will get a final score of 2. Note that this formula assumes equal weighting of length and complexity of syntax.

**Example:** PEM.1.2 in 2019 Romania Civil Servant Survey: “Have your objectives and performance objectives been set and discussed with you before your last performance evaluation?” In this survey question,  $y$  is equal to 437 characters,  $x$  is 19 characters, and thus  $D$  is equal to 210 (rounded up). Therefore, a question with characters between  $19 < n < 229$  is short and a question with characters between  $229 \leq n < 438$  is long. As PEM.1.2 has 120 characters, this question is considered to be short; therefore, we assign a score of 0. For syntax complexity, the question can be considered as a complex sentence, as it has both independent (“Have your objectives and performance objectives been set and discussed with you”) and dependent clauses (“your last performance evaluation”). Thus, we assign a score of 1 for the complexity of this question. Through the addition of these two scores, the question is coded as 1.

---

**Component name:** Vagueness in wording

---

**Question:** To what extent does the question asked contain vague words that leads to ambiguity?

**Coding:** Continuous; {0,1,2}

**Scale:**

0 (No vague words in the predetermined list are used, and the question is written with absolute clarity and specificity.)

1 (Some vague words are used, so that the meaning of a question is not described clearly, but there is no or little ambiguity.)

2 (Vague words in the question induce substantial vagueness, which might result in survey breakoff or further leads to great ambiguity.)

**Description:** This component aims to capture the extent to which the language used in a question is vague. *Vague* is defined as unclear, imprecise, and possibly but not necessarily ambiguous (Edwards et al. 1997) or open to interpretation. Common terms such as “good” and “well” are predetermined in a list of vague words by the assessor(s). A score of 0 is assigned to a question with no common term, while a 1 or 2 is assigned to a question containing one (or more) of such common terms, unless the question explicitly provides additional information that defines the common term(s) (see example below).

When a question is vague and the corresponding vagueness leads to significant ambiguity, arguably the question is more difficult to understand than vague but unambiguous questions. This is the rationale behind the score of 1 and 2. For score 1, although vagueness is present, respondents are able to finish the survey. For score 2, it is likely that this vagueness leads to confusion and frustration, which might prevent respondents from completing the survey (Edwards et al. 1997).

**Example:** REC.2.13.e(5) in 2019 Romania Civil Servant Survey: “To what extent do you agree or disagree with the following statements: The competition was fair?” “Fair” is a vague word and fairness is a vague concept. However, the question is unambiguous. Hence, this question is coded as 1.

Let us now consider two additional examples: REC.2.19.e(5): “To what extent do you agree or disagree with the following statements: Jobs in my institution are assigned based on the results of a formal selection process?” and REC.2.7: “Which of the following assessment methods were used in the selection process for your current position? Select all that apply.” The phrase “selection process,” which appears in both questions, can be a vague term. However, in REC.2.19, the adjective “formal” is used to describe “selection process,” which makes it less vague. Thus, the vagueness score for REC.2.19 should be lower than REC.2.7.

---

**Component name:** Presence of reference frame

---

**Question:** Are the necessary reference frames included in the question?

**Coding:** Categorical; {0,1,2}

**Scale:**

0 (Either there is no need for any reference frames in the question or the necessary reference frame(s) are clearly stated and an average respondent is not asking for further clarifications.)

1 (The necessary reference frame(s) are provided, but they are not well specified, which leads to an average respondent asking for further clarifications.)

2 (The necessary reference frame(s) are missing, and a respondent’s typical reaction is asking “compared to what?” or “compared to whom?”)

**Description:** This component reflects the extent to which well-defined reference frames are present in a question whenever necessary so that respondents understand the question in the way the questionnaire maker intended and answers from different respondents to a question are comparable (Ekinci 2015).

**Example:** REC.2.19.f in 2019 Romania Civil Servant Survey: “Preferred candidates are handed copies of exams before the actual exam takes place.” Here, “preferred” acts as a necessary reference frame, as it indicates that not all candidates obtain the exam before it is conducted. However, this is still not well specified,

as it is unclear which candidates are “preferred” by the institution, highlighting the need for further clarifications. As a result, this question is coded as 1.

---

**Component name:** Number of subquestions

---

**Question:** How many subquestions are there in a question block?

**Coding:** Continuous; {0,1,2}

**Scale:**

0 (There is no or one subquestion.)

1 (There are two or three subquestions.)

2 (There are four or more than four subquestions.)

**Description:** This component aims to count the number of subquestions embedded in the question block to which a question belongs. A question block with two subquestions will be assigned a score of 1, and this will also apply to both to subquestions (a) and (b).

**Example:** REC.3.1. in 2019 Romania Civil Servant Survey has two subquestions: REC.3.1.(a)(1): “Please indicate the number of years ago when you were promoted in class.” and REC.3.2.(b)(2): “Please indicate the number of years ago when you were promoted in grade.” Therefore, REC.3.1 (a)(1) and REC.3.1 (b)(2) both are coded as 2.

However, note that for the other components, each of these subquestions should be treated as a separate question, and therefore should be coded independently. For instance, assessors should rate vagueness in wording separately for REC.3.1.(a)(1) and REC.3.1 (b)(2) instead of giving one rating for REC.3.1.

## Information retrieval

---

**Component name:** Familiarity with the subject

---

**Question:** Are the respondents familiar with the subject of the question?

**Coding:** Categorical; {0,1,2}

**Scale:**

0 (The question relates to personal information about the respondent, such as their age, gender, or level of education; or if it relates to the respondent’s friends, immediate supervisors, subordinates, coworkers, or team.)

1 (The question relates to the respondent’s institution or organization as a whole.)

2 (The question relates to information beyond the scope of the respondent’s organization, such as other institutions or organizations, the public sector, or private sector as a whole.)

**Description:** This component reflects the extent to which respondents are knowledgeable on the subject of a question. It is assumed that a respondent is more familiar with subjects relating to her/himself or someone who is close to him/her.

Although the description is quite restrictive since there are other ways to assess familiarity—for instance, by using technical terms or jargon (Krosnick and Presser 2009)—it is hard to incorporate all aspects of familiarity in one component. Moreover, the chosen description can be argued to be the most appropriate for a public administration survey, due to its tendency to ask questions surrounding civil servants’ work environment, including coworkers and the institution as a whole.

**Example:** DWH.1.4 in 2019 Romania Civil Servant Survey: “What is your job title?” The respondent should be familiar with the question, as job title is regarded as the respondent’s personal information. Hence, the question is coded as 0.

---

**Component name:** Difficulty in recalling information

---

**Question:** To what extent is the relevant information difficult to be recalled?

**Coding:** Categorical; {0, 1, 2}

**Scale:**

0 (The question requires the respondent to recall general information in the present, e.g., including but not limited to basic demographic questions, facts about oneself, binary questions.)

1 (The question requires the respondent to recall general information in the past OR highly specific information in the present. Specific information includes but is not limited to quantitative questions, except age, asking about the year of occurrence, and opinion-based question, which is usually accompanied with a Likert scale.)

2 (The question requires the respondent to recall highly specific information in the past.)

**Description:** This component tests for the extent to which respondents are required to remember information based on the question’s level of specificity and time frame of interest (past/present). Willis and Lessler (1999) argued that the respondent might not remember a very specific information (e.g., exact count) or if the event had occurred in a distant past. This is the basis of why it is assumed that specific or past-related information is more difficult to be recalled by an average respondent. Therefore, on the scale from 0 to 2, a score of 0 is assigned for more general and present-related information, while a score of 1 is coded for a question asking specific information about the present or general information about the past. Meanwhile, the score 2 is assigned for most extreme cases of question relating to specific and distant information in the past.

**Example:** REC.2.12 in 2019 Romania Civil Servant Survey: “Which of the following factors were important for getting your current job in the public administration?” This question pertains to the past and asks for a specific level of information, as it requires the respondent to consider the level of importance of the following factors for getting a job: academic qualifications, job-specific skills, knowing someone with political links, having personal connections, etc. Therefore, this question is coded as 2.

## Information integration

---

**Component name:** Computational intensity

---

**Question:** To what extent is computation required to be made for a respondent to come up with an answer?

**Coding:** Categorical; {0, 1, 2}

**Scale:**

0 (The question does not require any computation to be made.)

1 (The question requires the respondent to do simple arithmetic calculation, which includes addition or subtraction.)

2 (The question requires the respondent to perform more complex arithmetic calculation, which includes multiplication or division.)

**Description:** This component tests for the extent to which basic arithmetic computations (addition, subtraction, multiplication, and division) are required to reach an answer. The more advanced the required computation is, the higher the chance that this overloads the respondent's working memory. It is assumed that addition and subtraction are less complex than multiplication and division; therefore, if a question requires the respondent to execute only the former, it will be coded as 1, and for the latter, a 2. The basis of this assumption is Willis and Lessler (1999), which discussed an example question asking about the proportion of time spent in some activities. They argued that asking the number of hours (only involving addition and subtraction) burdens the respondent less compared to asking proportions of time (which involves division).

**Example:** DWH.1.8 in 2019 Romania Civil Servant Survey: "How many years have you been in your current institution?" This question requires a respondent to calculate his/her length of service in the current institution by subtracting his/her starting year from the current year. Therefore, this question is coded as 1.

---

**Component name:** Scope of information

---

**Question:** To what extent is the relevant information based on the respondent's personal experience?

**Coding:** Categorical; {0, 1, 2}

**Scale:**

0 (The question requires the respondent to draw on only his/her personal experience)

1 (The question requires the respondent to draw equally on his/her personal experience and some additional information that the respondent is not heavily exposed to.)

2 (The question requires the respondent to only integrate information beyond his/her personal experience and the respondent is not exposed to this experience in his/her job.)

**Description:** This component tests for the extent to which answers are derived from information beyond the personal experience of the respondent. When the respondent can answer the question asked by only referring to his/her personal experience, then the question is the least complex. However, if it requires the respondent to integrate his/her personal experience with something that he/she is not familiar with, then this increases cognitive burden and the complexity of the question, and even more if the answer is mainly derived from something beyond the respondent's personal experience.

**Example:** REC.3.14 in 2019 Romania Civil Servant Survey: "Which of the following factors help employees get a promotion in your institution?" This question requires the respondent to reflect both on his/her personal experience and from other employees' (in which respondent is not heavily exposed to) past experience in getting a promotion. Therefore, this question is coded as 1.

## Translation of Information to Answer

---

**Component name:** Mismatch between categories and questions

---

**Question:** To what extent does the given response options match the possible answers to the question?



**Coding:** Categorical; {0, 1, 2}

**Scale:**

0 (The answer categories are completely matched with the potential answers to the question, for instance: (1) For a straightforward yes/no question: a binary scale is assigned. (2) For an opinion-based question: a Likert scale is assigned. (3) For a complex factual question (with many possible answers): a list of mutually exclusive and collectively exhaustive items is provided. (4) For an individual-specific question (e.g., job title, age): “record respond” or “record number” is offered.)

1 (The answer categories work with the question but are not perfectly suited for it, i.e., moderately mismatched. For example, a question that asks “I enjoy working at my organization” and has answer categories of only “Yes” or “No.” The binary answer options do make sense in the context of the question, but they limit the possible answers that someone might give to the question, as someone might only “somewhat enjoy” working at their organization, and would want to answer “somewhat agree” if the option is given.)

2 (The answer categories are severely mismatched to the question. For instance: (1) For a question with nuanced answers: a binary scale is assigned. (2) For a straightforward yes/no question: various response options are assigned (also includes a yes/no answer with some degree of explanation). (3) For questions that ask a time frame: the answer options overlap with each other.)

**Description:** This component tests for the extent to which the available answer options match the true answer to the question. Moreover, it also assesses whether there are other response options that do not necessarily belong to that particular question. Willis and Lessler (1999) reasoned that if the answer options do not match the question asked, the respondent will have more difficulty (and be confused) when mapping their true response to the given categories. This increases cognitive burden and thus makes the question more complex to be answered. An example of mismatch includes overlapping responses, for instance when “within the past 12 months” and “within the past 5 years” are offered for a time-frame-related question. The latter should instead be framed as “between 1 and 5 years ago.” Another example includes if the answer options to a straightforward yes/no question offer some explanation to the answer (e.g., Yes, because...). This is a mismatch, as the reason for the yes/no answer should instead be asked in a follow-up question.

**Example:** AWE.1.1 in 2019 Romania Civil Servant Survey: “How many more years do you intend to work in the public administration?” The potential responses are: (1) 1–2 years; (2) Maximum 5 years; (3) 10 years or more; (4) Rest of my career; (900) Don’t know; (998) Refused to answer. Answer options (1) and (2) overlap with one another, as if someone wishes to work for 2 additional years, this answer can be mapped into option (1) or (2). Option (2) should instead be phrased “3–5 years” instead of “maximum 5 years.” A similar argument applies for options (3) and (4), where there exists a mismatch. Therefore, this question is coded as 2.

---

**Component name:** Number of answers needed to answer the question (burden for answering the questions)

---

**Question:** How many responses are required to be picked to fully answer the question?

**Coding:** Categorical; {0, 1, 2}

**Scale:**

0 (The question requires the respondent to pick one answer to the question.)

1 (The question requires the respondent to pick two to three answers to the question.)

2 (The question requires the respondent to pick four to five answers to the question.)

**Description:** This component tests for the extent to which respondents are required to pick more than one answer to the question. It is argued that if the respondent is required to pick more than one answer (e.g., select all that apply), the higher the cognitive burden will be, and thus the more complex the question is. The rationale behind this argument is derived from existing literature that suggests that matrix/list questions are associated with higher cognitive burden and breakoff rate (Peytchev 2009; Steinbrecher, Roßmann, and Blumenstiel 2014; Tourangeau, Conrad, and Couper 2013). When a question requires a respondent to select multiple questions, it amounts to asking the respondent to give a yes/no answer to a list of questions. For instance, REC.2.12 asks respondents to consider factors that were important for getting their current job in the public administration. By requiring respondents to select all that apply, the question is essentially asking the respondents to rate (1) whether academic qualifications are important; (2) whether previous work experience is important; (3) whether job-specific skills are important; (4) and so on.

**Example:** D.W.H.1.3 in 2019 Romania Civil Servant Survey: “What status do you have in the public administration?” It is indicated in the question that the respondent is only allowed to pick one option only to answer this question. Therefore, this question is coded as 0.

## Sensitivity

---

**Subindicator name:** Invasion of privacy

---

**Question:** To what extent the respondent is asked to discuss “taboo” or private topics that may be inappropriate in everyday conversation?

**Coding:** Categorical; {0, 1, 2}

**Scale:**

0 (The question probes a topic that is unlikely to lie within the private domain, e.g., an individual’s sex or gender.)

1 (The question probes a behavior/topic that is considered to lie within the boundary between public and private domains, e.g., political views.)

2 (The question probes a behavior that is considered to be completely within the personal domain or overly taboo, e.g., drug use.)

**Description:** This component tests whether the question is probing any taboo topics or any topic that the respondent might feel is inappropriate to probe in a public administration survey. Questions related to a respondent’s income or religion may fall into this category. This form of sensitivity is not related to whether the respondent answers the question truthfully.

**Example:** REC 3.1 in the 2019 Romania Civil Servant Survey asks, “Do you remember how many years ago you last advanced in your career in the public administration to a better job—be it a position of higher pay or greater responsibilities?” where the respondent is asked to indicate the number of years. The question is coded as 1 as some respondents might feel that this is information that should be privy to them and the question is indirectly tied to income.

---

**Subindicator name:** Social-emotional threat of disclosure

---

**Question:** To which degree may the respondent be concerned with the social and/or emotional consequences of a truthful answer should the information become known to a third party?

**Coding:** Categorical; {0, 1, 2}

**Scale:**

0 (The reported behavior will not cause censure and is unlikely to lead to some degree of social stigma and/or embarrassment in the community.)

1 (The reported behavior is unlikely to cause censure but will lead to at least some degree of social stigma and/or embarrassment in the community.)

2 (The reported behavior will lead to censure and a substantial degree of social stigma and/or embarrassment in the community.)

**Description:** This component tests the degree of embarrassment or censure within the respondent's community that a truthful answer will cause if the information is made publicly available or leaked.

**Example:** PEM 1.17c in the 2019 Romania Civil Servant Survey asks, "To what extent do you agree with the following statements on performance evaluations? I feel pressure to give some team members higher ratings than their work performance justifies." The question is coded as a 1, since, if this is true and the truthful answer was made publicly available, it could be embarrassing for the respondent.

---

**Subindicator name:** Threat of formal sanctions

---

**Question:** To which degree may the respondent be concerned with the legal and/or formal consequences of a truthful answer should the information become known to a third party?

**Coding:** Categorical; {0, 1, 2}

**Scale:**

0 (The reported behavior does not lead to any kind of legal action and or punishment.)

1 (The punishment for the reported behavior is somewhat severe. In general, these are behaviors that would be considered misdemeanors or infractions in most legal systems. These behaviors are usually punished with monetary fines or some disciplinary action in the workplace.)

2 (The punishment for this reported behavior is very severe. In general, these are behaviors that would be considered felonies in most legal systems and will include some jail time.)

**Description:** This component tests the potential legal consequences that the behavior probed by the question could have. This type of question is only sensitive if the respondent's truthful answer departs from legal behaviors defined by formal institutions and legal regulations.

**Example:** AWE 4.1a in the 2019 Romania Civil Servant Survey asks, "How frequently do employees in your institution undertake the following actions? Accepting gifts or money from citizens." This question is coded as a 2 since the respondent might have a legal duty to report the illegal behavior to the relevant authorities.

---

**Subindicator name:** Relationship between informal and formal sensitivity

---

**Question:** What is the likelihood that a behavior/attitude may cause a threat of both social-emotional disclosure and formal sanctions?

**Coding:** Categorical; {0, 1, 2}

**Scale:**

0 (The probed behavior/attitude violates neither formal nor informal sensitivity and does not trigger sanction from either source.)

1 (The probed behavior/attitude violates both informal and formal sensitivity but is still unlikely to trigger sanction from any of the sources.)

2 (The probed behavior/attitude violates both informal and formal sensitivity and triggers severe sanctions from both sources.)

**Description:** This component measures the likelihood that a behavior/attitude may cause a threat of both social-emotional disclosure and formal sanctions. This type of question is logically more sensitive than ones that violate one type of institution while conforming to another.

## BEST PRACTICES FOR MANUAL CODING

- Where there is more than one assessor in the project, it is advisable that all assessors code selected sample questions together in order to reach an agreement on practicalities that further enhances objectivity.
- For the remaining questions, the assessors also need to discuss the questions for which they have disagreement in scoring (if any) and reach a final consensus.
- Where the assessor identifies systematic challenges to consistent coding, the assessor should make a note, and this needs to be discussed during the consensus meeting with other assessors, and if needed, with the supervisor of the project.
- Where the assessor finds the coding description to be unclear, the assessor should use their judgement to the best of his/her abilities. This should be discussed during the consensus meeting (or the initial sit-together) to ensure a standardized practice can be conducted.

## NOTE

1. The terms “survey questions” and “survey items” are used interchangeably in this appendix.

## REFERENCES

- Edwards, J., M. Thomas, P. Rosenfeld, and S. Booth-Kewley. 1997. *How to Conduct Organizational Surveys: A Step-by-Step Guide*. Los Angeles: SAGE Publications.
- Ekinci, Y. 2015. *Designing Research Questionnaires for Business and Management Students*. Los Angeles: SAGE Publications.
- Helmke, G., and S. Levitsky. 2006. *Informal Institutions and Democracy*. Baltimore: Johns Hopkins University Press.
- Krosnick, J. A., and S. Presser. 2009. *Question and Questionnaire Design. Handbook of Survey Research*. 2nd ed. San Diego: Elsevier.
- Peytchev, A. 2009. “Survey Breakoff.” *Public Opinion Quarterly* 73 (1): 74–97.
- Steinbrecher, M., J. Roßmann, and J. Blumenstiel. 2014. “Why Do Respondents Break off Web Surveys and Does It Matter? Results from Four Follow-Up Surveys.” *International Journal of Public Opinion Research* 27 (2): 289–302.
- Tourangeau, R., F. Conrad, and M. Couper. 2013. *The Science of Web Surveys*. Oxford: Oxford University Press.
- Tourangeau, R. and T. Yan. 2007. “Sensitive Questions in Surveys.” *Psychological Bulletin* 133 (5): 859–83. <https://doi.org/10.1037/0033-2909.133.5.859>.
- Willis, G., and J. T. Lessler. 1999. *Question Appraisal System: QAS-99*. Rockville: Research Triangle Institute.

# APPENDIX K

## Referent Questions and Organizational and Commitment Measures

### Chapter 23 Appendix

#### APPENDIX K.1 LIST OF REFERENT QUESTIONS

##### Recruitment (Romania)

###### *Individual referent*

1. The skills and knowledge I was tested on during the recruitment process match the skills and knowledge I need to perform my job.
2. What I do in my current job on a daily basis matches the job description of the position I hold.

###### *Organizational referent*

1. The recruitment process in my institution tests the skills and knowledge staff needs to perform their jobs.
2. What staff in my institution are doing in their jobs on a daily basis matches their formal job descriptions of their positions.

##### Promotions (Romania)

###### *Individual referent*

1. Please indicate the extent to which you agree or disagree with the following statement: The promotion process in my institution is clear.
2. Please indicate the extent to which you agree or disagree with the following statement: The promotion process in my institution is fair.

###### *Organizational referent*

1. Please indicate the extent to which you agree or disagree with the following statement: The promotion process I have to go through in my institution is clear.
2. Please indicate the extent to which you agree or disagree with the following statement: The promotion process I have to go through in my institution is fair.

## Turnover (Romania)

### *Individual referent*

1. Please indicate the extent to which you agree or disagree with the following statement: I spend time searching for other jobs.
2. Please indicate the extent to which you agree or disagree with the following statement: I want to quit my job.
3. Please indicate the extent to which you agree or disagree with the following statement: I want to quit my institution.
4. Please indicate the extent to which you agree or disagree with the following statement: I want to quit the public service.
5. Please indicate the extent to which you agree or disagree with the following statement: I will quit my job if I have the chance to get another job.

### *Organizational referent*

1. Please indicate the extent to which you agree or disagree with the following statement: Staff in my institution spend time searching for other jobs.
2. Please indicate the extent to which you agree or disagree with the following statement: Staff in my institution often think about quitting their jobs.
3. Please indicate the extent to which you agree or disagree with the following statement: Staff in my institution often think about quitting the institution.
4. Please indicate the extent to which you agree or disagree with the following statement: Staff in my institution often think about quitting the public service.
5. Please indicate the extent to which you agree or disagree with the following statement: Staff in my institution will quit their jobs if they have the chance to get another job.

## Dismissals (Romania)

### *Individual referent*

1. To what extent do you agree or disagree with the following statement: It would be difficult in practice to dismiss me from the public administration.
2. To what extent do you agree or disagree with the following statement: It would be difficult in practice to transfer me to another job against my will.

### *Organizational referent*

1. To what extent do you agree or disagree with the following statement: In my institution, it would be difficult in practice to dismiss employees.
2. To what extent do you agree or disagree with the following statement: In my institution, it would be difficult in practice to transfer employees to another job against their will.

## Promotion (Guatemala)

### *Individual referent*

1. On a scale of 1 to 5, how confident are you that if you perform well in your job you will receive a promotion?



2. To what extent do you agree with the following statement: The promotion process I have to go through in my organization is clear and fair.

*Organizational referent*

1. On a scale of 1 to 5, how confident are the majority of personnel in your institution that they will be promoted in the future if they perform their jobs well? 1 is very unconfident and 5 is very confident.
2. Please indicate the extent to which you agree that the process of promotion in your institution is clear and fair.

### **Turnover (Guatemala)**

*Individual referent*

1. I often think about leaving my current job.
2. I often think about leaving my organization.
3. I often think about leaving the public service.

*Organizational referent*

1. Staff in my organization often think about leaving their jobs.
2. Staff in my organization often think about leaving the organization.
3. Staff in my organization often think about leaving the public service.

### **Dismissals (Guatemala)**

*Individual referent*

1. It would be difficult to dismiss me from the public service.
2. It would be difficult to transfer or rotate me to another position against my will.

*Organizational referent*

1. It is difficult to dismiss staff in my organization.
2. It is difficult to transfer or rotate staff in my organization.

### **Leadership (Guatemala)**

*Individual referent*

1. How often does your direct supervisor do the following?: Communicates and encourages enthusiasm about the mission and vision of your organization.
2. How often does your direct supervisor do the following?: Leads by example.
3. How often does your direct supervisor do the following?: Says things that makes his/her staff feel proud to be a part of the organization.
4. How often does your direct supervisor do the following?: Promotes communication and accountability in relation to ethical and unethical practices associated with their work.
5. How often does your direct supervisor do the following?: Communicates the ethical standards clearly to his/her staff.

6. How often does your direct supervisor do the following?: Fulfills his/her promises and commitments.
7. How often does your direct supervisor do the following?: Puts my interests above his/hers.
8. How often does your direct supervisor do the following?: Supports my professional development.
9. How often does your direct supervisor do the following?: Is concerned for my well-being.

#### *Organizational referent*

1. How often does your organization's management do the following?: Communicate and encourage enthusiasm about the mission and vision of the organization.
2. How often does your organization's management do the following?: Lead by example.
3. How often does your organization's management do the following?: Say things that makes his/her staff feel proud to be a part of the organization.
4. How often does your organization's management do the following?: Promote communication and accountability in relation to ethical and unethical practices associated with their work.
5. How often does your organization's management do the following?: Communicate the ethical standards clearly to their staff.
6. How often does your organization's management do the following?: Fulfill their promises and commitments.
7. How often does your organization's management do the following?: Puts the interests of their staff above their own.
8. How often does your organization's management do the following?: Support the professional development of their staff.
9. How often does your organization's management do the following?: Are concerned for their well-being of their staff.

**TABLE K.1 Organizational Commitment and Identification Using Other Measures**

	Model 1 (Dismissals, Belonging)	Model 2 (Dismissals, Pride)	Model 3 (Dismissals, Commitment)	Model 4 (Recruitment, Belonging)	Model 5 (Recruitment, Pride)	Model 6 (Recruitment, Commitment)	Model 7 (Turnover, Belonging)	Model 8 (Turnover, Pride)	Model 9 (Turnover, Commitment)
Organizational-Level	0.365*	0.321*	0.257†	−0.254***	−0.370***	−0.306***	0.670***	0.496***	0.481***
	(0.178)	(0.160)	(0.137)	(0.071)	(0.066)	(0.056)	(0.088)	(0.086)	(0.074)
Commitment	0.450	0.434	0.272	−0.094	−0.170	−0.174†	−0.297*	−0.681***	−0.465***
	(0.292)	(0.275)	(0.234)	(0.117)	(0.114)	(0.096)	(0.143)	(0.144)	(0.124)
Organizational-Level × Commitment	−0.204	−0.189	−0.134	0.219**	0.351***	0.281***	−0.177†	−0.003	0.021
	(0.185)	(0.172)	(0.146)	(0.074)	(0.071)	(0.059)	(0.092)	(0.092)	(0.079)
(Intercept)	3.175***	3.211***	3.375***	4.965***	5.013***	5.020***	1.028***	1.431***	1.223***
	(0.281)	(0.257)	(0.221)	(0.113)	(0.107)	(0.090)	(0.137)	(0.134)	(0.117)
N	2,642	2,979	2,930	2,877	3,249	3,190	2,521	2,860	2,814
R <sup>2</sup> (adj.)	0.005	0.004	0.003	0.018	0.041	0.032	0.184	0.178	0.154

Source: Original table for this publication.

Notes: Results from ordinary least squares models. † p<0.100, \* p<0.050, \*\* p<0.010, \*\*\* p<0.001.

# APPENDIX L

## Further Details of Surveys

### Chapter 24 Appendix

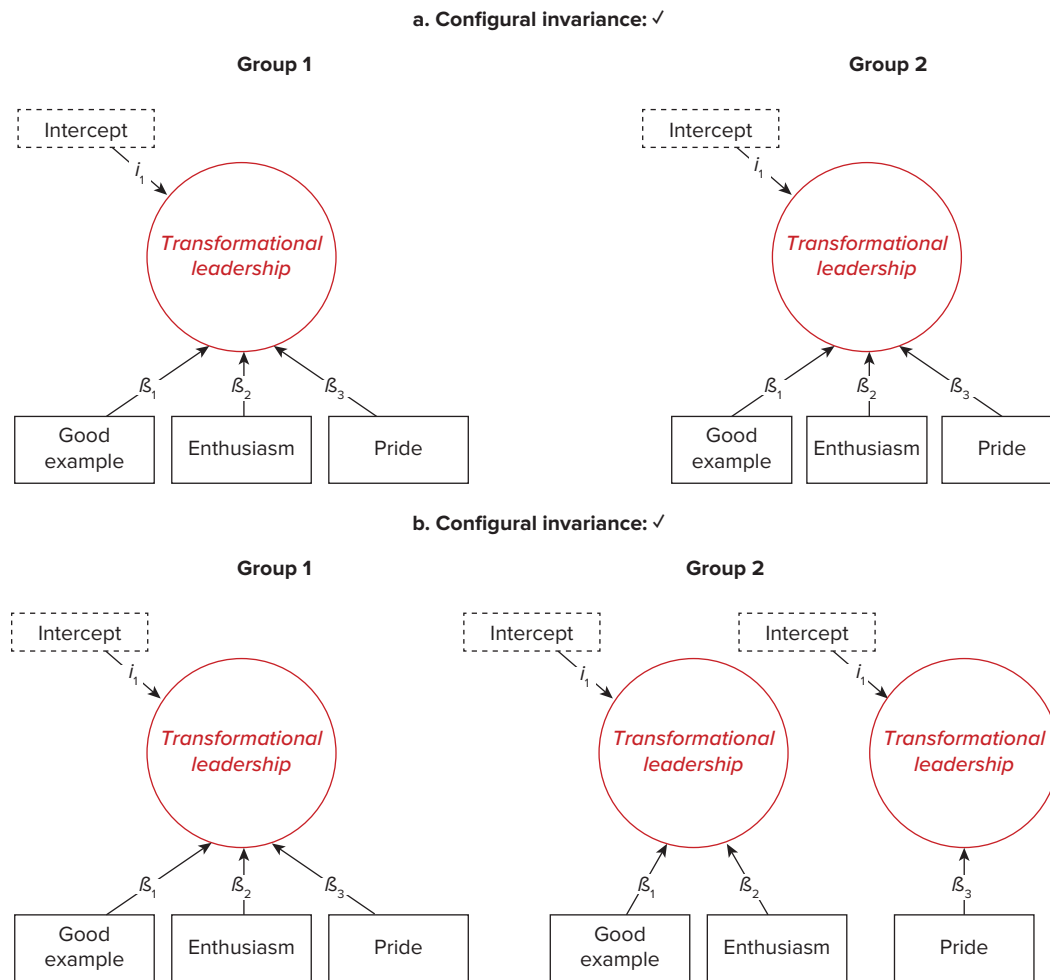
**TABLE L.1** Distribution of Respondents: Survey Sample vs. Civil Service Population

Country	Gender (% female)		Education (% university educated)		Managerial status (% managers)		Age (mean)	
	Survey	Population	Survey	Population	Survey	Population	Survey	Population
Albania	61.3	—	92.1	—	28.9	—	35.2	—
Bangladesh	21.4	18	53.4	—	21.6	27	38.5	—
Brazil	42.6	45	49.7	75	13.1	—	47.6	46
Chile	54.9	58	10.2	50	3.1	—	41.5	42
Estonia	69.3	56	53.4	61	17.8	—	43.3	43.3
Kosovo	41.7	—	46.8	—	25.8	—	43.0	—
Nepal	33.7	—	48.3	—	20.9	—	38.0	—

Source: Original table for this publication.

Note: Populationwide statistics were available only for selected demographics and or countries. For details, see Mikkelsen, Schuster, and Meyer-Sahling (2020).

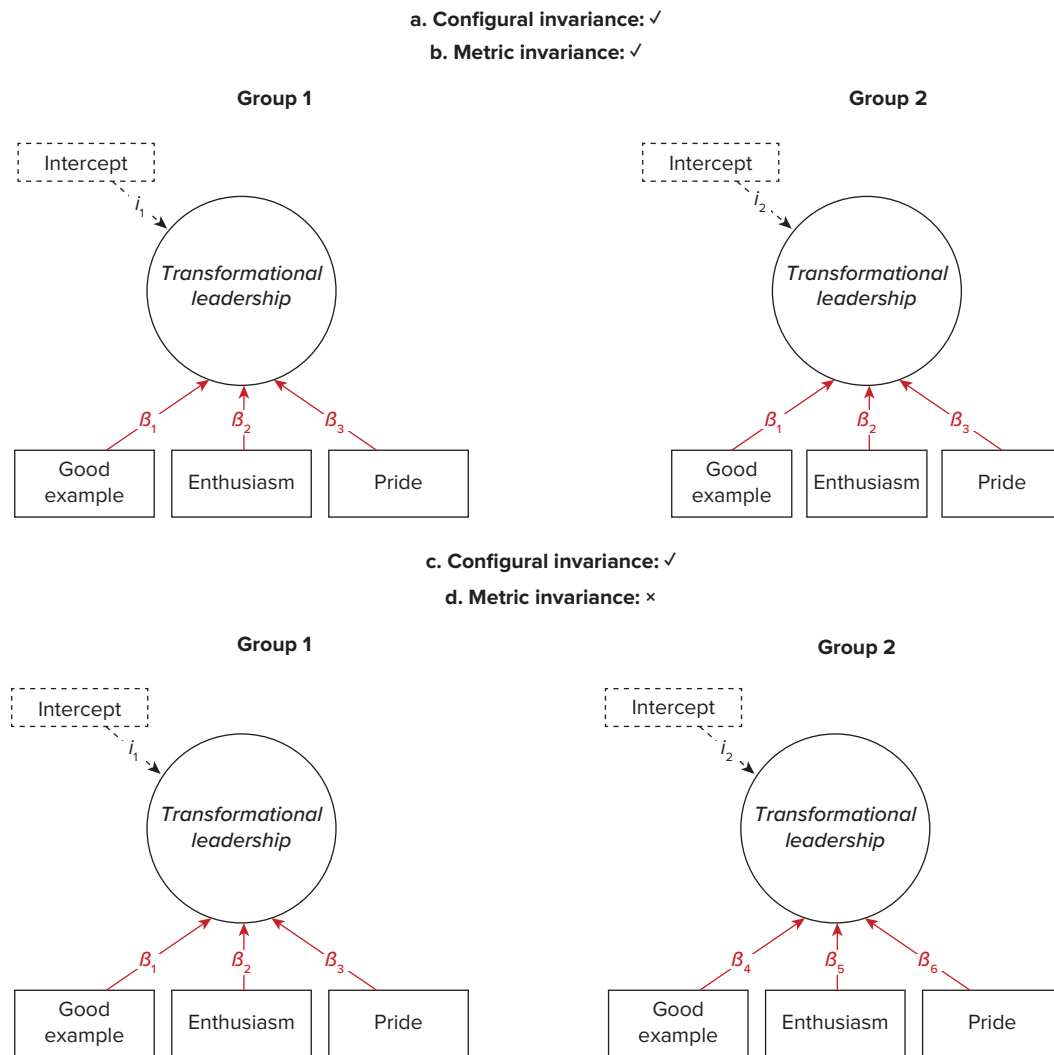
**FIGURE L.1 Concept of Configural Invariance Visualized**



Source: Original figure for this publication.

Note: The key metric of focus for configural invariance measurement is highlighted in red: structure of latent factors. Configural invariance means that the same model fits both groups. Lack of configural invariance in turn signifies that the observed variables measure different combinations of latent factors.

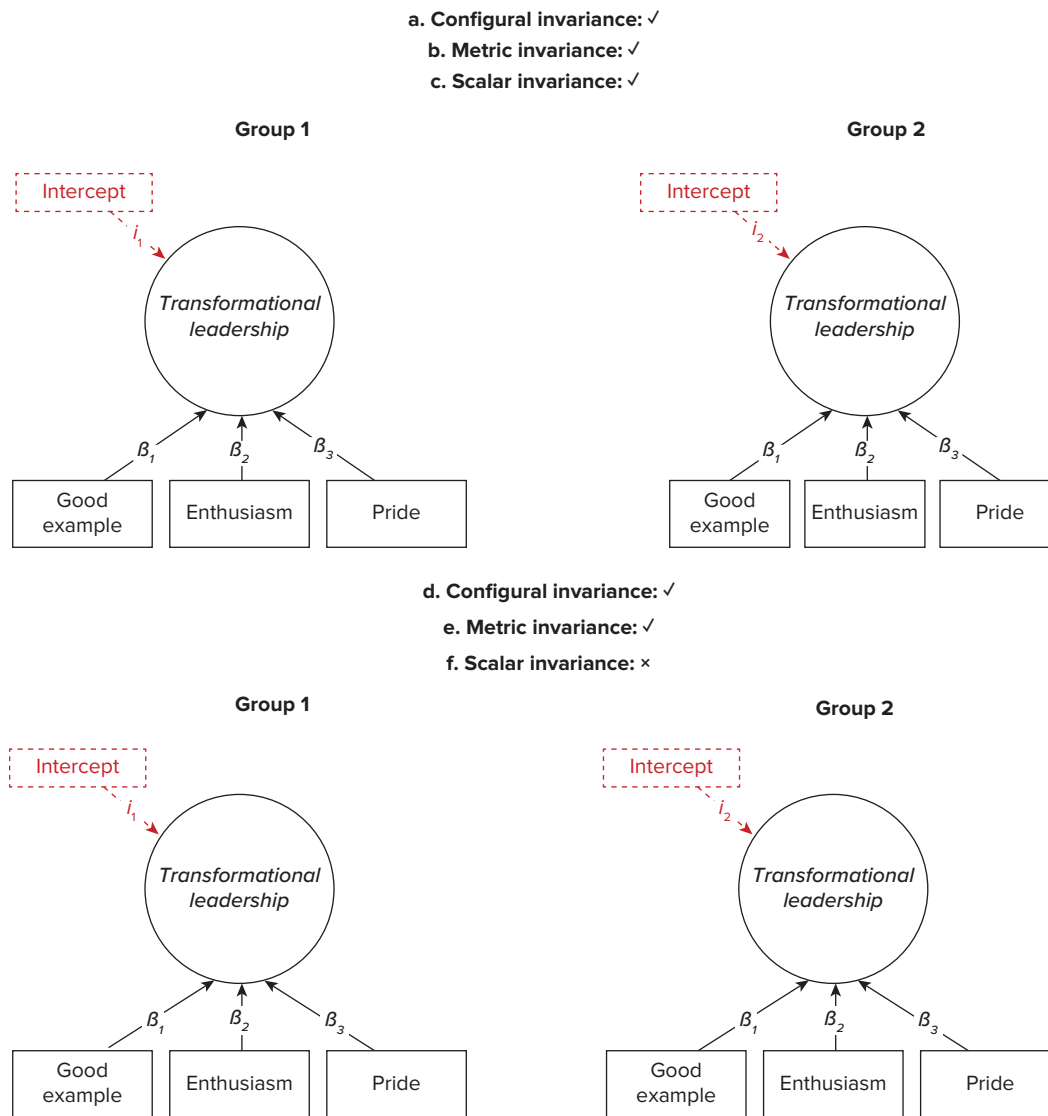
**FIGURE L.2 Concept of Metric Invariance Visualized**



Source: Original figure for this publication.

Note: The key metric of focus for metric invariance measurement is highlighted in red: factor loadings. Metric invariance means factor loadings are not statistically different from each other and therefore can be compared across groups.

**FIGURE L.3 Concept of Scalar Invariance Visualized**

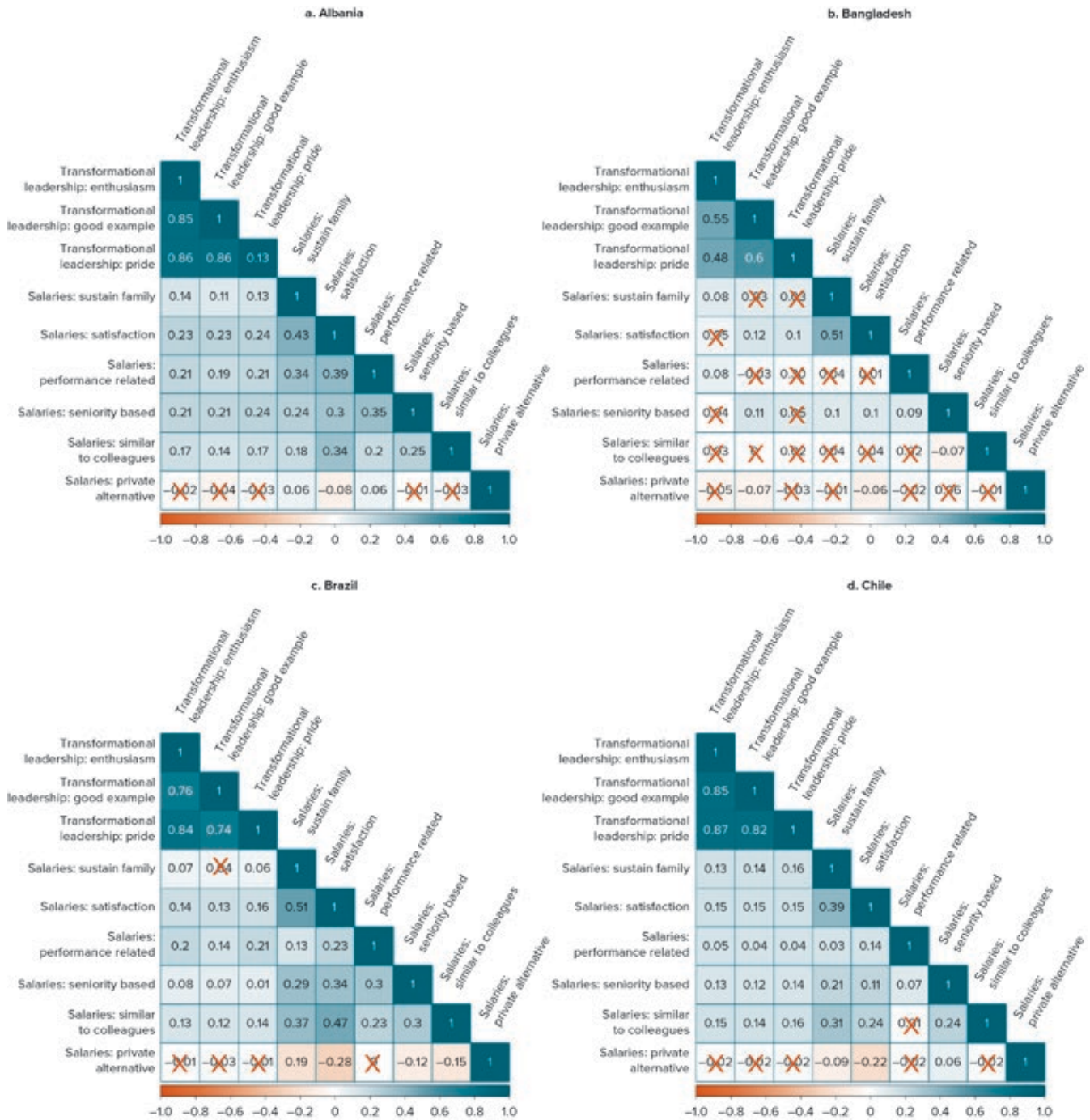


Source: Original figure for this publication.

Note: The key metric of focus for scalar invariance measurement is highlighted in red: value of intercepts of the latent factors. Scalar invariance means latent factor intercepts (means) are not statistically different from each other and therefore can be compared across groups.

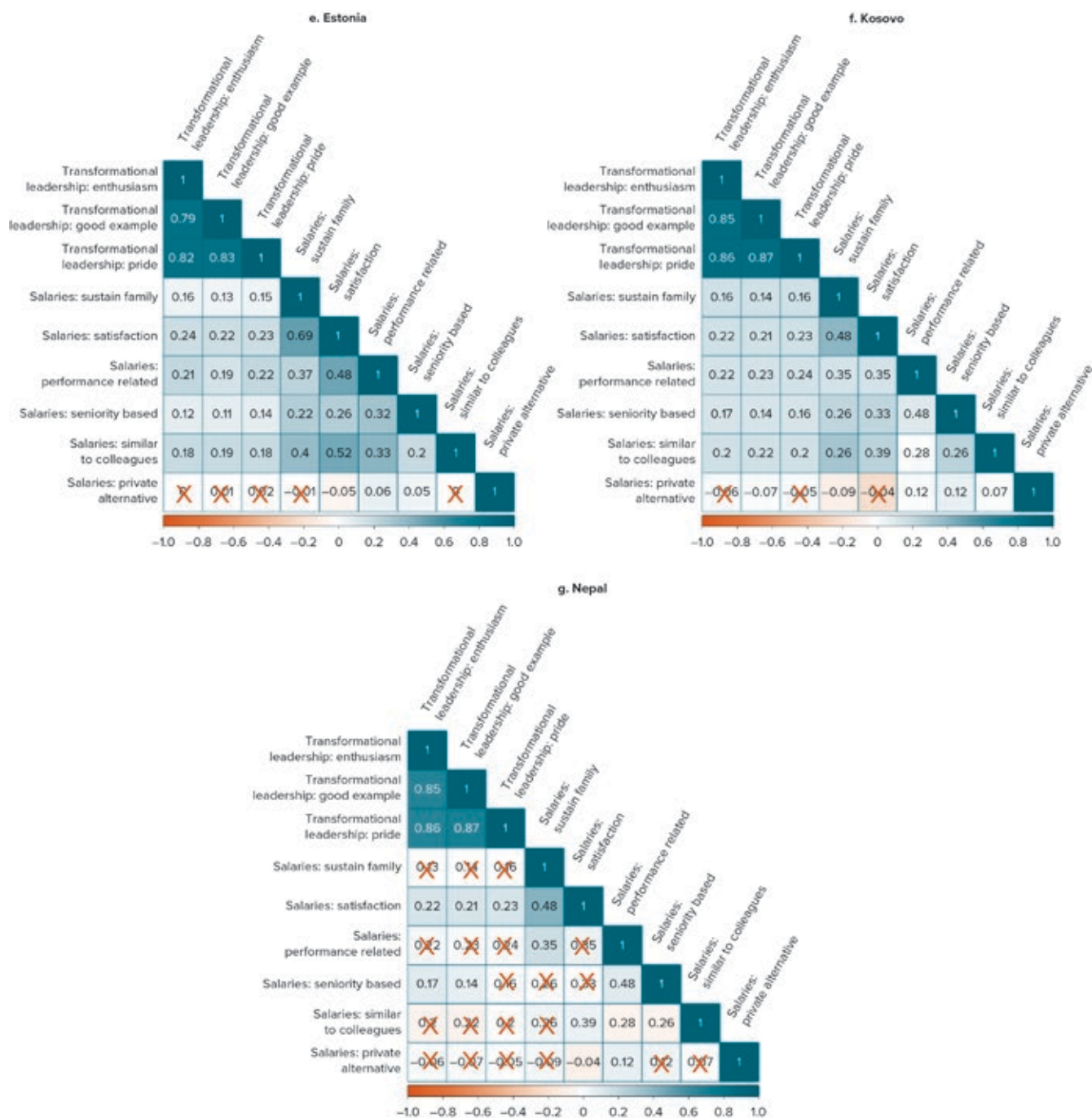


**FIGURE L.4** Correlations between Questions from the Transformational Leadership and Salaries Sections of the Global Survey of Public Servants Administered across Seven Countries



(continues on next page)

**FIGURE L.4** Correlations between Questions from the Transformational Leadership and Salaries Sections of the Global Survey of Public Servants Administered across Seven Countries (*continued*)



Source: Original figure for this publication.

Note: The correlograms demonstrate correlations among Global Survey of Public Servants questions using only complete observations. The exact phrasing of the questions was as follows:

1) Transformational Leadership: Enthusiasm: My direct superior articulates and generates enthusiasm for my organization's vision and mission; 2) Transformational leadership: Good example: My direct superior leads by setting a good example; 3) Transformational leadership: Pride: My direct superior says things that make employees proud to be part of this organization; 4) Salaries: Sustain family: I could sustain my household through my salary alone; 5) Salaries: Satisfaction: I am satisfied with my salary; 6) Salaries: Performance related: My work performance has had an influence on my salary in the civil service; 7) Salaries: Seniority based: My years of service in the civil service have had an influence on my salary; 8) Salaries: Similar to colleagues: I am paid at least as well as colleagues who have job responsibilities similar to me; 9) Salaries: Private alternative: It would be easy for me to find a job outside the public sector that pays better than my current job.



## REFERENCES

- Fukuyama, F., D. Rogger, Z. Hasnain, K. Bersch, D. Mistree, C. Schuster, K. Mikkelsen, K. Kay, and J. Meyer-Sahling. 2022. *Global Survey of Public Servants Indicators*. <https://www.globalsurveyofpublicservants.org>.
- Mikkelsen, K. S., C. Schuster, and J.-H. Meyer-Sahling. 2020. "A Cross-Cultural Basis for Public Service? Public Service Motivation Measurement Invariance in an Original Survey of 23,000 Public Servants in Ten Countries and Four World Regions." *International Public Management Journal* 24 (6): 739–61. doi:10.1080/10967494.2020.1809580.

# APPENDIX M

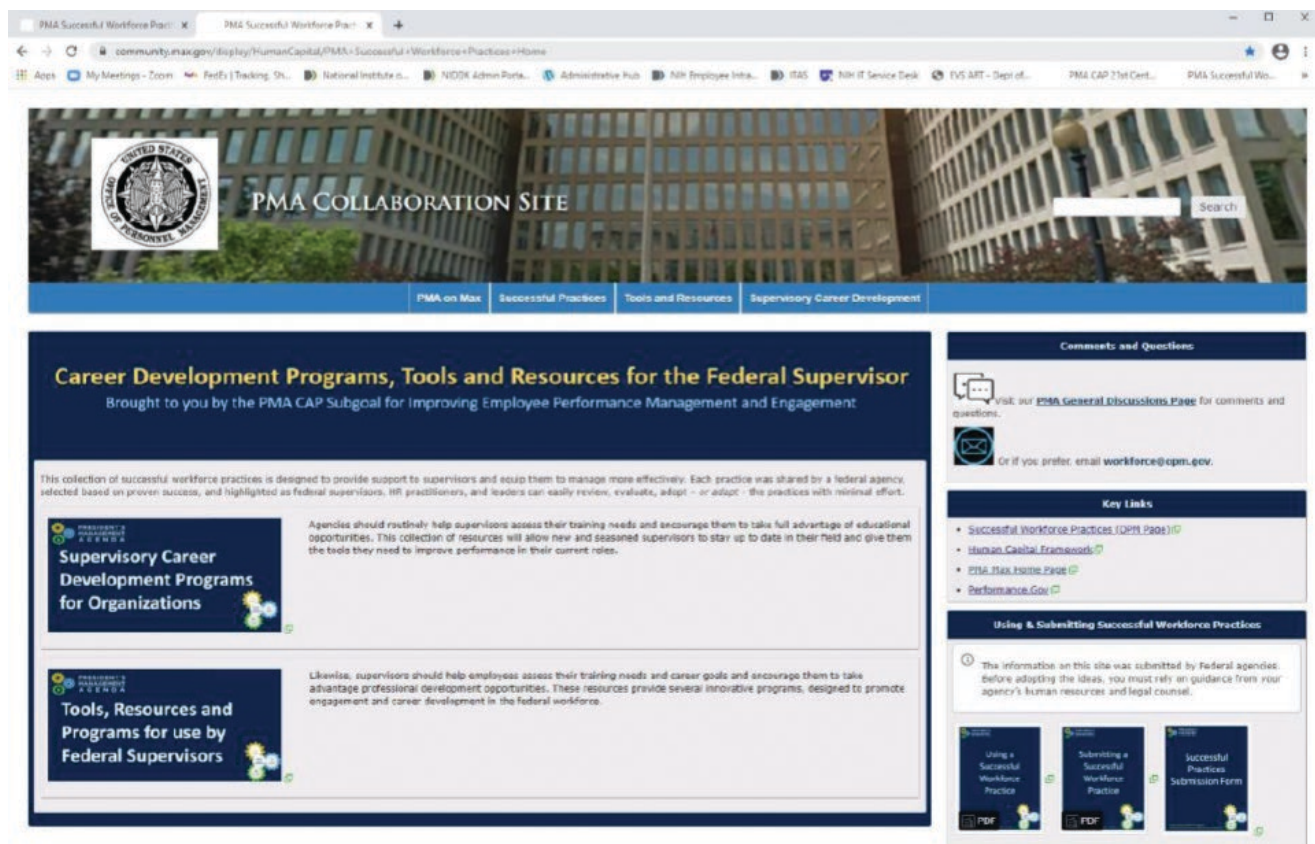
## US Federal Government Best Practices Resources

### Chapter 26 Appendix

#### APPENDIX M.1 PRESIDENT'S MANAGEMENT AGENDA BEST PRACTICES SHARING WEBSITE

The President's Management Agenda Developing a Workforce for the 21st Century Subcommittee for Improving Employee Performance Management and Engagement collected successful workforce practices from across the US federal government and created a platform to share them broadly. Screenshots from the platform are provided below.

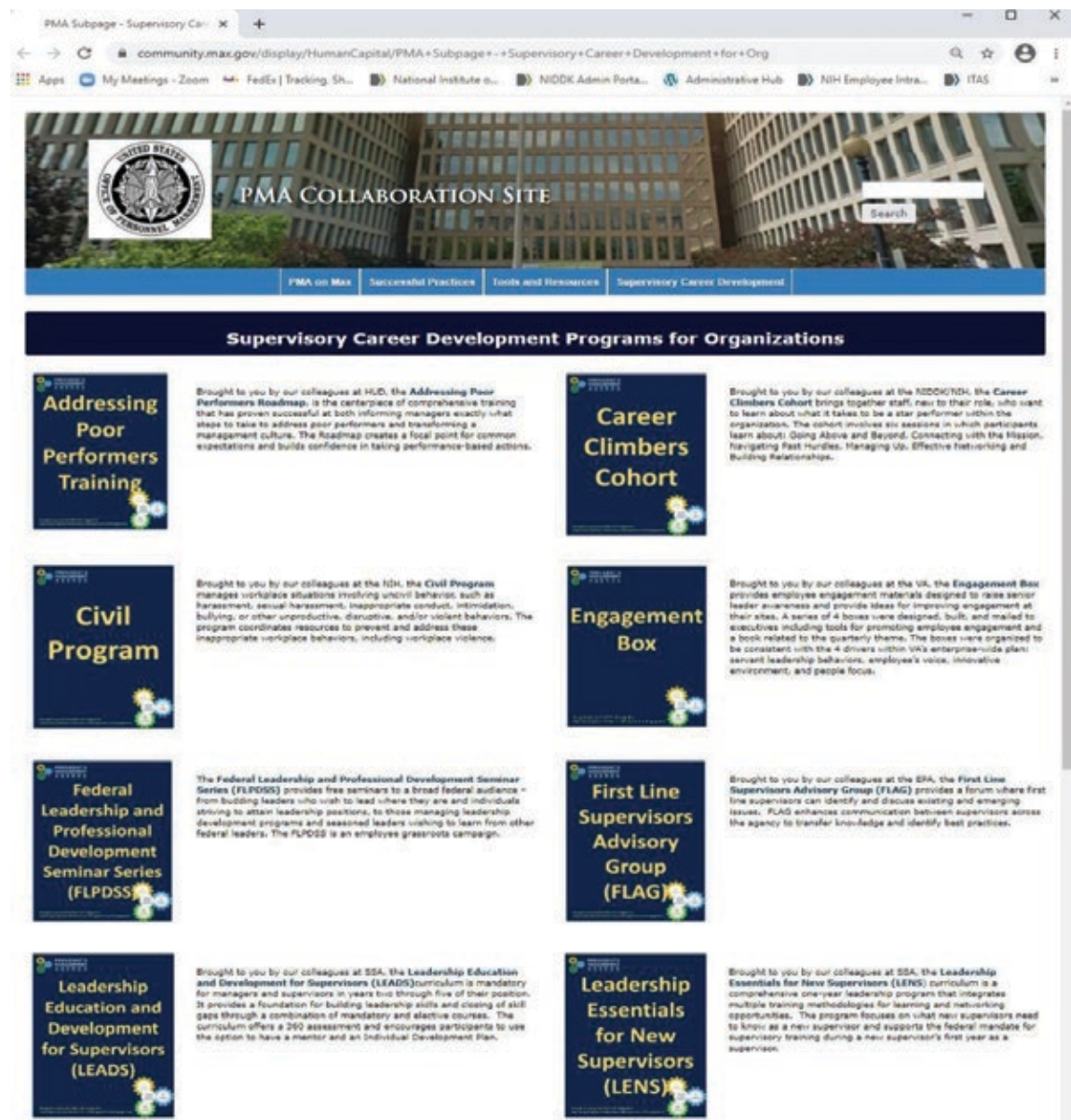
**FIGURE M.1** PMA Site on Career Development Programs



Source: Screenshot from President's Management Agenda (PMA) site.



**FIGURE M.2 PMA Site on Supervisory Career Development Programs**



Source: Screenshot from President's Management Agenda (PMA) site.

**FIGURE M.3 PMA Site on Career Climbers Cohort**



Source: Screenshot from President's Management Agenda (PMA) site.

## APPENDIX M.2 WORKFORCE POLICY BEST PRACTICE SHARING SITE

To advertise the existence of resources on best practice, US Office of Personnel Management (OPM) has set up their own workforce policy best practice sharing site that links to many of the resources from the President's Management Agenda (PMA), as well as others such as the Employee Viewpoint Survey Analysis and Results Tool (EVS ART) outlined in appendix M.3.



**FIGURE M.4 OPM Link to PMA Resources**

The screenshot displays the OPM.gov website interface. At the top, there is a navigation bar with links for 'ABOUT', 'POLICY', 'INSURANCE', 'RETIREMENT', 'SUITABILITY', 'AGENCY SERVICES', and 'NEWS'. Below this, a breadcrumb trail reads: 'OPM.gov Main > Policy > Human Capital Management > Successful Workforce Practices'.

The main content area is titled 'Policy, Data, Oversight' and 'HUMAN CAPITAL MANAGEMENT'. It includes a paragraph about federal leaders' passion for creating a positive culture and a link to the 'Successful Workforce Practices' page. Below this, a section titled 'Highlighted Successful Practices' lists four programs:

- Successful Leaders Program**: At the Environmental Protection Agency. The SLP is a highly successful program that provides systematic training and development to managers. A link to the 'Successful Leaders Program' on MAX.gov is provided.
- Career Climbers Cohort**: At Health and Human Services, National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The cohort brings together staff who are new to their role and want to learn about what it takes to be a star performer. A link to the 'Career Climbers Cohort' on MAX.gov is provided.
- Engagement Box Program**: From the Veterans Health Administration (VHA) National Center for Organization Development (NCOD). The VHA NCOD developed employee engagement materials designed to raise senior leader awareness and provide them with ideas for improving engagement at their sites. A link to the 'Engagement Box Program' on MAX.gov is provided.
- Employee Viewpoint Survey Analysis and Results Tool (EVS ART)**: At Health and Human Services, National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). EVS ART is an innovative tool that can save the government millions of dollars annually in time—while providing better, quicker analysis of FEVS results. A link to the 'Employee Viewpoint Survey Analysis and Results Tool (EVS ART)' on MAX.gov is provided.

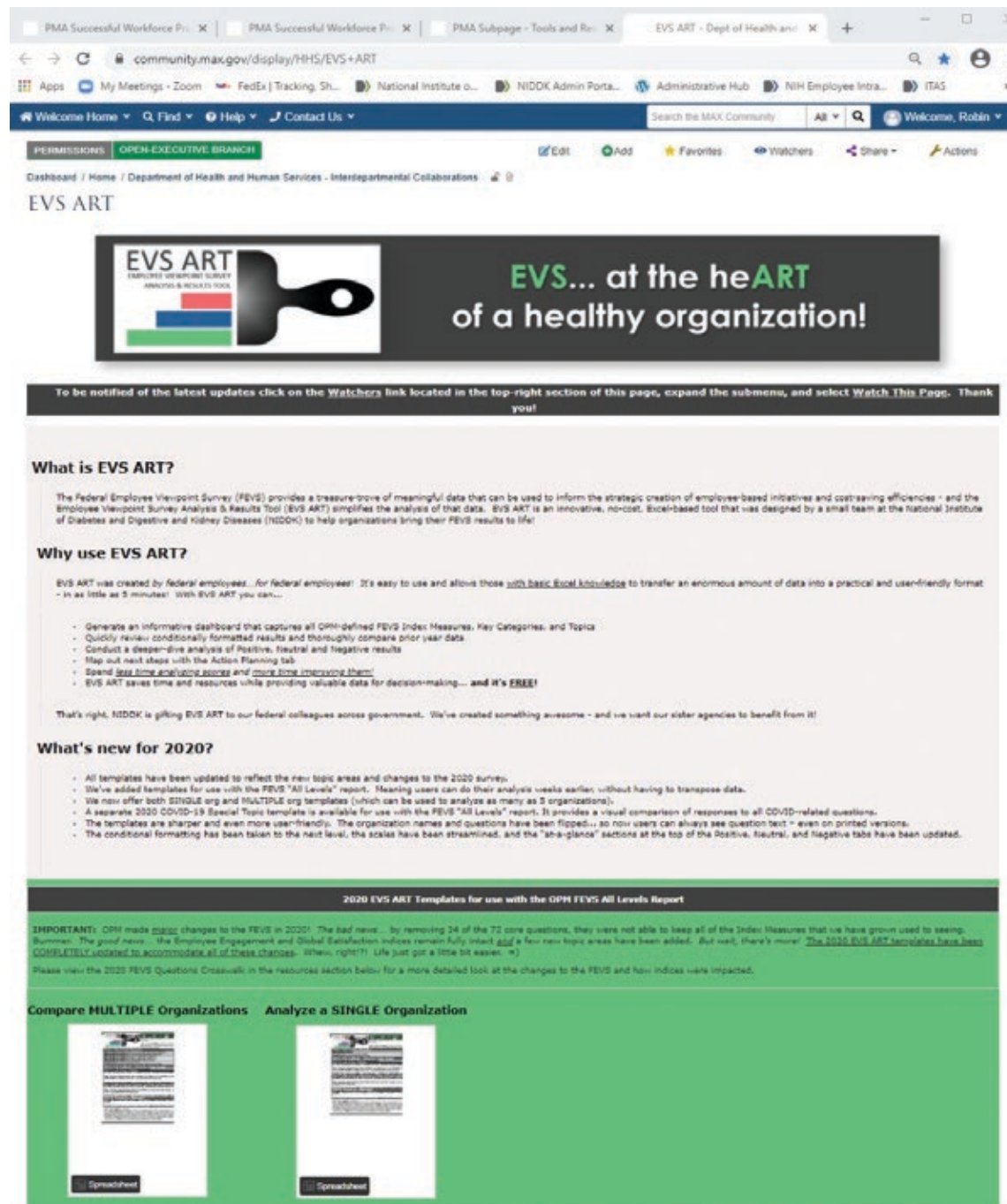
On the left side of the page, there is a sidebar with a list of links under the heading 'IN THIS SECTION'. The links include: Assessment & Selection, Classification & Qualifications, Data, Analysis & Documentation, Disability Employment, Diversity & Inclusion, Coronavirus Disease 2019, Employee Relations, Hiring Information, Human Capital Management, Closing Skills Gaps, Cybersecurity, Federal Workforce Priorities Report (FWPR), Strategic Foresight, Successful Workforce Practices (highlighted), Hiring Reform, HRStat, Human Capital Operating Plan (HCOP), Human Capital Reviews (HCR), Human Capital Framework, Labor-Management Relations, Oversight Activities, Pandemic Information, Pay & Leave, Performance Management, Senior Executive Service, Settlement Guidelines, Snow & Dismissal Procedures, Training & Development, Veterans Services, Work-Life, Workforce Restructuring, Policy FAQs, and Contact Policymakers. Below this, there is a section titled 'RESOURCES FOR' with links for New / Prospective Employees, Federal Employees, HR Professionals, and Managers.

Source: Screenshot from Office of Personnel Management (OPM) site.

## APPENDIX M.3 EMPLOYEE VIEWPOINT SURVEY ANALYSIS AND RESULTS TOOL

An Excel-based tool, the Employee Viewpoint Survey Analysis and Results Tool (EVS ART), has provided officials across government with a no cost, practical, and easy to use resource for analyzing the OPM Federal Employee Viewpoint Survey (FEVS) data sent to agencies by OPM. Screenshots from the platform are provided below.

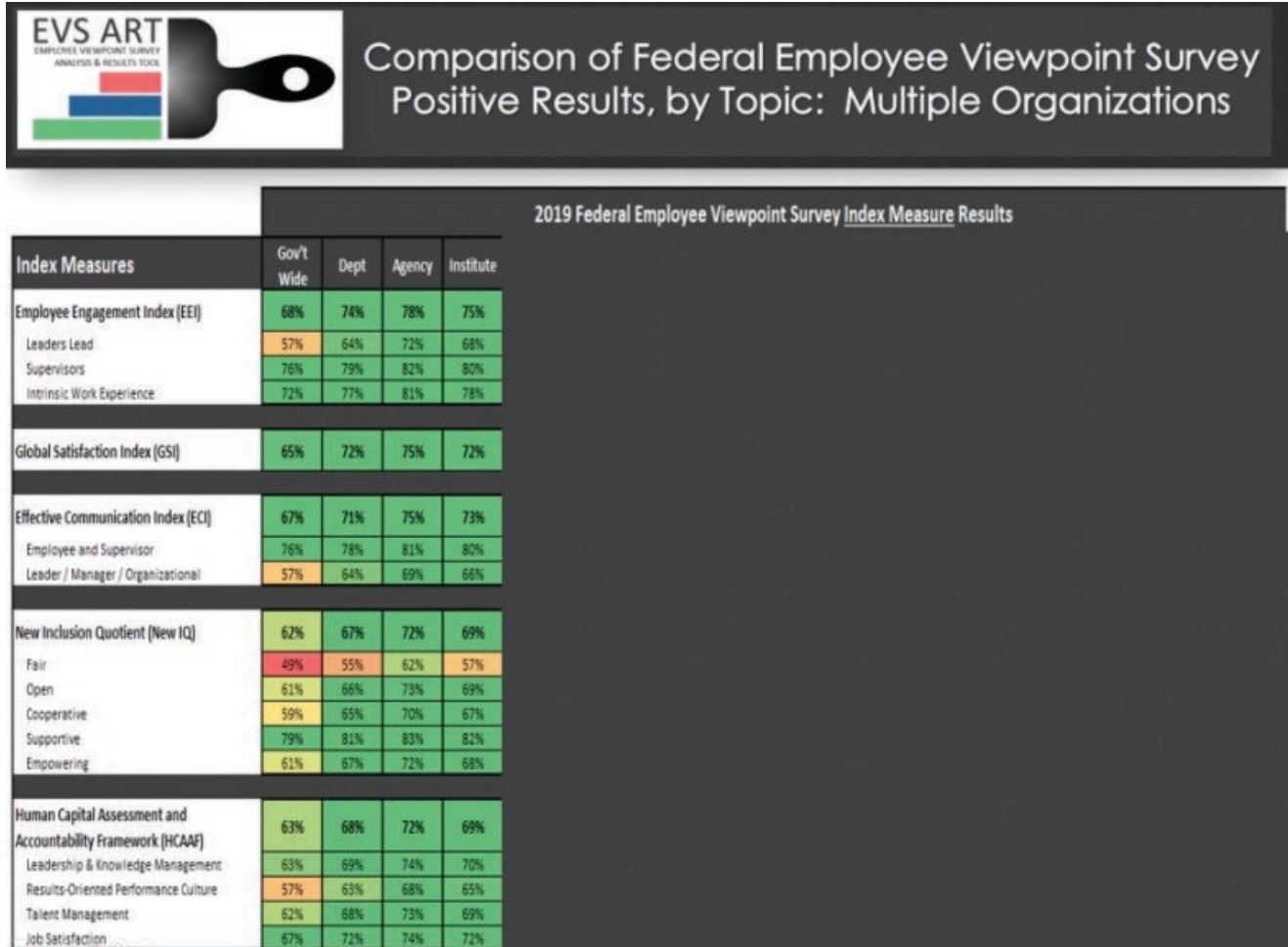
FIGURE M.5 EVS ART Online Dashboard



Source: Screenshot from Office of Personnel Management (OPM) site.

As the below screenshot shows, at the institute level, averages provide little indication of problem areas at specific organizations.

**FIGURE M.6** EVS ART Results Dashboard Showing Institute Level

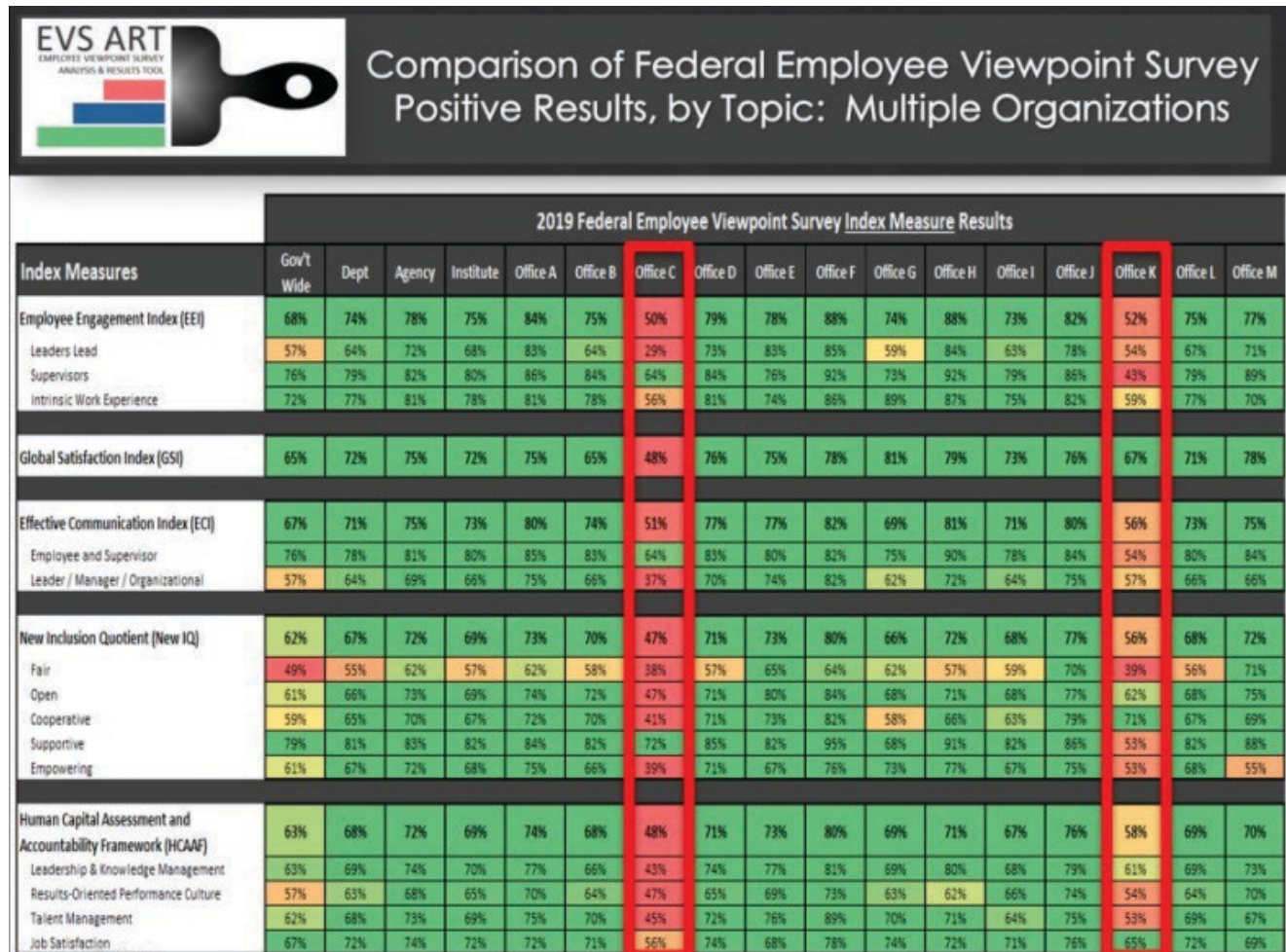


Source: Screenshot from Office of Personnel Management (OPM) site.



EVS ART makes a deeper dive to the office-level sample. Such a perspective indicates two problem offices otherwise hidden when rolled up to the parent organization.

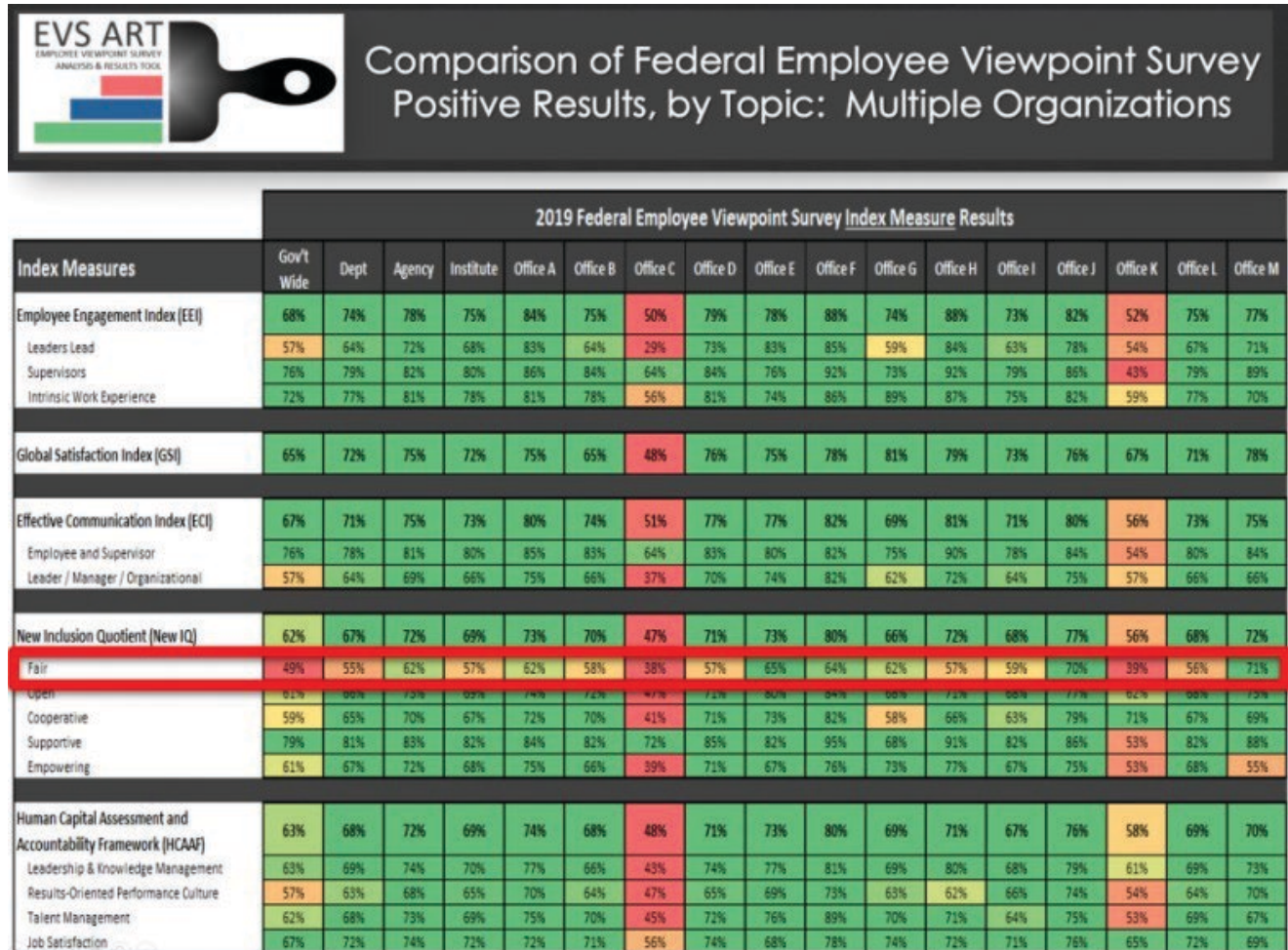
**FIGURE M.7** EVS ART Results Dashboard Showing Office Level, with a Focus on Organizations



Source: Screenshot from Office of Personnel Management (OPM) site.

A deeper dive into the data can also be done by topic. In the screenshot below, we see this highlight an opportunity for improvement across the institute in the fairness component of the survey.

**FIGURE M.8 EVS ART Results Dashboard Showing Office Level, with a Focus on Topics**



Source: Screenshot from Office of Personnel Management (OPM) site.





#### **ECO-AUDIT**

#### ***Environmental Benefits Statement***

The World Bank Group is committed to reducing its environmental footprint. In support of this commitment, we leverage electronic publishing options and print-on-demand technology, which is located in regional hubs worldwide. Together, these initiatives enable print runs to be lowered and shipping distances decreased, resulting in reduced paper consumption, chemical use, greenhouse gas emissions, and waste.

We follow the recommended standards for paper use set by the Green Press Initiative. The majority of our books are printed on Forest Stewardship Council (FSC)–certified paper, with nearly all containing 50–100 percent recycled content. The recycled fiber in our book paper is either unbleached or bleached using totally chlorine-free (TCF), processed chlorine-free (PCF), or enhanced elemental chlorine-free (EECF) processes.

More information about the Bank’s environmental philosophy can be found at <http://www.worldbank.org/corporateresponsibility>.



*The Government Analytics Handbook* presents frontier evidence and practitioner insights on how to leverage data to strengthen public administration. Covering a range of microdata sources—such as administrative data and public servant surveys—as well as tools and resources for undertaking the analytics, it transforms the ability of governments to take a data-informed approach to diagnose and improve how public organizations work.

Readers can order the book as a single volume in print or digital formats, or visit [worldbank.org/governmentanalytics](http://worldbank.org/governmentanalytics) for modular access and additional hands-on tools. The *Handbook* is a must-have for practitioners, policy makers, academics, and government agencies.

.....

“Governments have long been assessed using aggregate governance indicators, giving us little insight into their diversity and how they can practically be improved. This pioneering handbook shows how microdata can be used to give scholars and practitioners granular and real insights into how states work, and practical guidance on the process of state-building.”

— Francis Fukuyama, *Stanford University*, author of *State-Building: Governance and World Order in the 21st Century*

“*The Government Analytics Handbook* is the most comprehensive work on practically building government administration I have ever seen, helping practitioners to change public administration for the better.”

— Francisco Gaetani, *Special Secretary for State Transformation*, Government of Brazil

“The machinery of the state is central to a country’s prosperity. This handbook provides insights and methodological tools for creating a better shared understanding of the realities of a state, to support the redesign of institutions, and improve the quality of public administration.”

— James Robinson, *University of Chicago*, coauthor of *Why Nations Fail*



ISBN 978-1-4648-1957-5



SKU 211957