Designing for Experience — A Requirements Framework for Enrolment Based and Public Facing E-Government Services

Christopher Charles Porter

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Computer Science

University College London

February 3, 2015

I, Christopher Charles Porter, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

To Christine

Abstract

User-centricity is a pre-requisite for a truly transformational e-government strategy. This goes beyond visual design and appeal, and ties down to a rudimentary measure of how far people are willing to go to enrol for and use e-government services. Enrolment can have a serious impact on the success of online government services. Different services require different levels of identity assurance, and different enrolment processes are put in place to deliver them. But from the citizen's perspective these processes often require a disproportionate amount of effort, producing hurdles that affect user acceptance and ultimately service adoption. When enrolling to high-effort services is not mandatory, take-up is low; when it is compulsory, it causes resentment, and neither is desirable. Despite existing work on the impact of security and identity processes on end users there has been little work on how these contributions could be operationalised and adopted by practitioners and policy makers as part of the requirements development process. Research in HCI provides techniques to help practitioners design systems that are within general human capabilities, however such techniques are too generic to approximate use-time behaviour across user groups and within different contexts of use.

This thesis proposes Calibrated Personas, a user modelling technique that accumulates knowledge on user behaviour to model and fine-tune tolerance levels for workload and its impact on e-government service adoption (1) across user groups, (2) e-service types and (3) contexts of use. A user group calibration protocol was devised to facilitate data collection and model generation for user behaviour in enrolment-specific use cases. These models are in turn used to approximate user reactions towards design alternatives, reducing the gap between design-time knowledge (upon which decisions are made) and use-time knowledge. To facilitate this activity this work presents *Sentire* (‘*to listen*’), a requirements and design framework that combines industry-strength practices with user feedback simulations (referred to as UX-analytics). These simulations in turn inform the requirements development process with actionable feedback as part of an iterative design process. This thesis considers tool support for *Sentire* as central to the investigation in order to facilitate adoption by practitioners and to encourage knowledge sharing and re-use within the e-government domain. For this reason, an online collaborative computer-aided software engineering (CASE) tool was developed and evaluated throughout the various real-world interventions carried out for this thesis. *Sentire* was applied to two new national e-services and also in the evaluation of an existing one. User-studies and expert evaluation were instrumental to the evolution and validation of the main contributions and deliverables arising from this thesis.

Acknowledgements

Thanks to everyone who contributed towards this thesis. My supervisors, M. Angela Sasse and Emmanuel Letier, my colleagues at the Information Security Research Group and support staff at Malet Place. I would also like to thank my family for their continuous support as well as everyone at the Faculty of ICT at the University of Malta. Finally, I would like to thank my wife Christine for her love, patience, support and encouragement throughout the past five years.

To all, thank you.

Vignette – A flashback

Back in 2008 I was on the bus on my way to work. I noticed an elderly woman who happened to sit in the seat in front of me reading an electronic identity enrolment agreement form given to her by the newly established e-ID office in Valletta (Malta) – she must have just completed the first part of the enrolment process. This form provided detailed instructions on how to activate her account, involving an activation PIN which was to be sent separately by post together with a temporary password printed on the form, with which she could create a new password (valid for 90 days). She seemed confused and hastily folded the form and placed it back into her handbag, sighed and looked out the window at the 9am traffic – with a distant gaze. Then it dawned on me. We are too pre-occupied with technical sophistication and legalities – leaving many people behind along the way. I am convinced that this lady felt helpless, and maybe, incompetent. I felt sorry for her.

I used to work in a software house that was subcontracted to build a significant portion of the national identity management system (NIDMS), and have seen these forms in test runs, but never thought much of them. There was a lot of energy from the project team who used to work closely with the technology vendor to find workarounds on issues specific to identity federation, certificate issuing/revocation processes and single sign-on (SSO) amongst others. Technologically this was a huge project, involving a large number of stakeholders, including policy makers, law makers, international technology vendors and the government IT agency which was responsible for operational issues (i.e., data-centre management and security). The government was working on tight political deadlines and NIDMS was expected to be delivered in a short time-frame.

Following the incident on the bus I was convinced that we can do much better than that, even though technically, the current solution was practically airtight. But what can be done?

This was the starting point.

Contents

1 Introduction **23**

1.1 Thesis Motivation . . . 23

1.1.1 Real-world problem . . . 24

1.1.1.1 Why bother with e-government services? . . . 26

1.1.2 Science problem . . . 27

1.1.3 Research gap . . . 28

1.2 Research Question . . . 29

1.3 Contributions to the Interface Between Requirements Engineering and HCI(Sec) . . . 30

1.4 Thesis Scope . . . 32

1.5 Publications and Awards . . . 33

1.5.1 Publications . . . 33

1.5.2 Forthcoming publications . . . 35

1.5.3 Posters . . . 36

1.5.4 Multimedia . . . 36

1.5.5 Industry recognition . . . 36

1.6 Overview of studies in this thesis . . . 36

1.7 Thesis Structure . . . 38

2 Background **39**

2.1 Usability, Experience and Lived Experience . . . 39

2.1.1 User Interface and Usability . . . 39

2.1.2 User Experience – UX . . . 40

2.1.3 Users' Lived Experience – ULX . . . 41

2.2 The Interface Between RE and HCI . . . 41

2.2.1 Classical HCI considerations in RE . . . 41

2.2.2 RE frameworks for usable e-government . . . 46

2.2.3 User involvement in RE . . . 50

2.2.4 User-centred design techniques . . . 52

2.3 Enrolment Processes and User Behaviour . . . 55

2.3.1 Evidence from the field . . . 58

2.3.2 Evidence from literature . . . 59
2.3.3 Lessons from the private sector . . . 61
2.3.4 HCI-Sec and identity mechanisms . . . 62
2.3.5 Human factors in security – workload . . . 63
2.3.5.1 NASA-TLX . . . 65
2.4 Process-Related Threats to UX in E-Service Projects . . . 67
2.4.1 E-service project contracts and processes . . . 67
2.4.2 “Getting UX into the contract” — and the process . . . 68
2.5 People-Related Threats to Usability in E-Service Projects . . . 71
2.5.1 The disconnect between software engineering and HCI . . . 71
2.6 Glossary of Terms . . . 73
2.7 Conclusions . . . 77

3 Methodology 78
3.1 Research Methods in HCI, Requirements and Design . . . 78
3.1.1 Data collection techniques . . . 79
3.1.2 Qualitative data analysis and interpretation for design . . . 80
3.1.2.1 Grounded theory and thematic analysis . . . 80
3.1.2.2 Coding . . . 81
3.1.2.3 Practical issues with qualitative techniques . . . 83
3.1.3 Case study as a research method . . . 84
3.2 Methodology Adopted for this Thesis . . . 86
3.2.1 Exemplar study to test proposed design framework . . . 90
3.3 Summary . . . 91

4 Understanding User Attitudes Towards Enrolment to Build Behaviour Based Models 92
4.1 Identifying Design Factors within Enrolment Processes . . . 92
4.1.1 Aims . . . 92
4.1.2 Method . . . 93
4.1.2.1 Participants . . . 93
4.1.2.2 Process . . . 93
4.1.3 Study outcomes . . . 98
4.1.4 Abstracting low level details for modelling purposes . . . 107
4.2 Building User Group Behavioural Models for Reuse . . . 108
4.2.1 Modelling for prediction . . . 109
4.2.2 Regression models . . . 110
4.2.2.1 Linear regression . . . 110
4.2.2.2 Multiple regression . . . 112
4.2.2.3 Linear regression and Generalised Linear Models . . . 113

4.2.2.4 Logistic regression . . . 114
4.2.3 Modelling willingness to complete a task . . . 118
4.2.4 Modelling perceived workload . . . 119
4.3 User Group Calibration (UGC) . . . 120
4.3.1 Calibration process . . . 123
4.3.2 A note on perceived workload calibration . . . 129
4.3.3 User group calibration and the context of use . . . 131
4.3.4 Sampling for calibration . . . 133
4.4 User Group Knowledge Base . . . 134
4.5 Summary . . . 137

5 A Requirements and Design Framework for Enrolment Based Public Facing E-Services 139
5.1 From Deferred to Realtime Feedback on Design Decisions . . . 139
5.2 Government Wide E-Service Requirements and Design Strategy . . . 141
5.3 Adopting and Extending *Volere* . . . 142
5.4 *Sentire* . . . 143
5.5 Supporting Tools and Techniques . . . 156
5.6 Observations and Criticisms . . . 159
5.7 Tool Support . . . 160
5.7.1 Policy makers and CASE tools . . . 160
5.7.2 Iterative prototyping . . . 161
5.7.3 First generation . . . 161
5.7.4 Learning from and evolving the CASE tool . . . 162
5.7.4.1 User group calibration . . . 162
5.7.4.2 User group library . . . 162
5.7.4.3 *Sentire* workflow . . . 163
5.7.4.4 User feedback simulation engine . . . 163
5.7.4.5 Reporting . . . 165
5.7.4.6 Domain visualisation . . . 166
5.7.4.7 Collaboration . . . 166
5.8 Summary . . . 168

6 Case Study 1: Publishing a Tender for a National Employment Agency 169
6.1 Defining the Context . . . 169
6.2 Aims . . . 169
6.2.1 Research objectives . . . 169
6.2.2 Practical objectives . . . 170
6.3 Method . . . 170
6.3.1 User group calibration (UGC) . . . 171

6.3.2 Product use case annotation . . . 173
6.3.3 Simulating user feedback and revisiting PUCs and requirements . . . 175
6.4 Evaluation and Findings . . . 176
6.4.1 Theoretical evaluation of calibration models . . . 176
6.4.1.1 Willingness to complete task (WCT) model . . . 176
6.4.1.2 Perceived workload (PEW) model . . . 177
6.4.2 Evaluation of task completion predictions . . . 178
6.4.3 Framework contributions and modifications . . . 179
6.4.3.1 Redefining delay . . . 179
6.4.3.2 Collaborative workspace . . . 179
6.4.4 Plans for next study . . . 179
6.5 Summary . . . 180

7 Case Study 2: Non E-Government Service Evaluation with Undergraduate Students 182
7.1 Aims . . . 182
7.1.1 Research objectives . . . 182
7.2 Method . . . 183
7.2.1 Participants . . . 183
7.2.2 Process . . . 183
7.2.2.1 User group calibration (UGC) . . . 184
7.2.2.2 Product use case annotation . . . 186
7.2.2.3 Simulating user feedback . . . 186
7.3 Evaluation and Findings . . . 187
7.3.1 Theoretical evaluation of calibration models . . . 187
7.3.1.1 Willingness to complete task (WCT) model . . . 187
7.3.1.2 Perceived workload (PEW) model . . . 187
7.3.2 Evaluation of task completion predictions . . . 189
7.3.3 Focus group findings . . . 191
7.3.3.1 Theme 1: Competition, trust and service features . . . 191
7.3.3.2 Theme 2: Social influence . . . 191
7.3.3.3 Theme 3: Privacy in security tasks . . . 192
7.3.3.4 Theme 4: Workload . . . 192
7.3.3.5 Theme 5: Usage patterns and lifestyles . . . 193
7.3.3.6 Theme 6: Convenience . . . 193
7.3.4 Framework contributions and modifications . . . 194
7.3.4.1 Adding meta-data to quantitative behavioural models . . . 194
7.3.4.2 One-size does not fit all . . . 194
7.3.5 Plans for next study . . . 195
7.4 Summary . . . 195

8 Case Study 3: Assessing NASA-TLX With Younger Users — Evaluating a Compulsory E-Service for Digital Natives 196

8.1 Defining the Context . . . 196
8.2 Aims . . . 197
8.3 Method . . . 197
8.3.1 Participants . . . 197
8.3.2 Process . . . 197
8.3.3 Sequential overview of study activities . . . 198
8.3.3.1 Online questionnaire . . . 198
8.3.3.2 Processing quantitative questionnaire data . . . 198
8.3.3.3 Processing qualitative questionnaire data . . . 199
8.3.3.4 Follow-up workshops . . . 199
8.3.3.5 Analysing and synthesising results . . . 199
8.4 Results . . . 199
8.4.1 Online questionnaire . . . 199
8.4.1.1 Participants . . . 199
8.4.1.2 Students' reported experience . . . 200
8.4.1.3 Students' reported workload . . . 201
8.4.2 Follow-up workshops . . . 204
8.4.2.1 Participants . . . 204
8.4.2.2 Perceived workload by consensus . . . 204
8.4.2.3 Sensitivity of NASA-TLX . . . 205
8.5 Discussion . . . 207
8.5.1 Digital natives and NASA-TLX . . . 207
8.5.1.1 Workload manifests itself in different ways . . . 207
8.5.1.2 Demystifying workload dimensions . . . 207
8.5.1.3 Keep out of reach of digital natives? . . . 208
8.5.2 NASA-TLX, e-government enrolment and digital natives — does it really work? 209
8.6 Recommendations . . . 212
8.6.1 Encourage secure behaviour . . . 212
8.6.1.1 Measure . . . 212
8.6.1.2 Simplify and guide . . . 212
8.6.1.3 Integrate and defer . . . 213
8.6.1.4 Communicate utility and value . . . 214
8.6.2 Modifying NASA-TLX for use in enrolment . . . 214
8.6.3 Privacy and frustration . . . 215
8.7 Conclusions . . . 215
8.8 Process Evaluation . . . 216

8.8.1 Framework contributions and modifications . . . 216
8.8.2 Plans for next study . . . 216
8.9 Summary . . . 216

9 Case Study 4: Building a National Consumer Affairs E-Service 217
9.1 Defining the Context . . . 217
9.2 Aims . . . 217
9.2.1 Research objectives . . . 217
9.2.2 Practical objectives . . . 218
9.3 Method . . . 218
9.4 Evaluation and Findings . . . 231
9.4.1 Theoretical evaluation . . . 231
9.4.1.1 Persona evolution through calibration . . . 231
9.4.1.2 Partial user models . . . 232
9.4.1.3 Indicative can be as good as precise . . . 232
9.4.1.4 Contextual feedback is possible with in-context calibration . . . 232
9.4.2 Evaluation of task completion predictions . . . 233
9.4.2.1 Participants . . . 233
9.4.2.2 Results . . . 234
9.4.3 Practical modifications to the *Sentire* CASE tool . . . 235
9.5 Summary . . . 237

10 Conclusions 238
10.1 Contributions . . . 241
10.1.1 Calibrated Persona – a technique to model and predict user reactions to and perceptions of e-service enrolment processes . . . 241
10.1.2 *Sentire* – a requirements framework based on simulated user feedback . . . 242
10.1.3 Collaborative tool support for *Sentire* . . . 243
10.1.4 Other contributions . . . 243
10.1.4.1 A study on user attitudes towards enrolment processes . . . 243
10.1.4.2 Assessment of NASA-TLX's sensitivity for enrolment-specific perceived workload on younger audiences . . . 244
10.1.4.3 Testable fit-criteria for experience related (non-functional) requirements 244
10.1.4.4 User-group knowledge base for reuse across government projects . . . 245
10.2 Critical Reflection . . . 245
10.2.1 Empirical validation limitations . . . 245
10.2.1.1 Evaluating quantitative results in case studies . . . 245
10.2.1.2 Evaluating qualitative studies . . . 246
10.2.2 Limitations of the calibration process . . . 246

10.2.3 Final remark on *Sentire* . . . 247
10.3 Future Work . . . 247
10.3.1 Adoption of *Sentire* for other critical e-service design aspects . . . 247
10.3.2 Simplifying simulated feedback . . . 248
10.3.3 Collaboration with statistical sciences research groups . . . 248
10.3.4 Calibrating non-technical users . . . 250
10.3.5 Remote and large-scale calibration . . . 251
10.3.6 Calibrated user group marketplace and cold start issues . . . 251
10.4 Expert Evaluation . . . 252

Appendices **253**

A Research Artefacts **253**
A.1 Initial Theory – Interview Guide . . . 253
A.1.1 Experiences in registration and authentication processes . . . 253

B Colophon **256**

Bibliography **257**

List of Figures

1.1 Personal observations on general design problems and their potential impact on product rework, UX and ULX (Users' Lived eXperience – see Section 2.6). 25
1.2 Calibrated Personas – this thesis' main contribution to HCI research and practice. This technique helps to reduce the gap between design-time knowledge on user behaviour and actual use-time data for specific systems, contexts and user groups 31
1.3 Thesis map – list of chapters and their relationship with the sub-research questions, contributions and resulting publications 38

2.1 Citizen-centric RE for e-government projects [166] 47
2.2 Finding the right balance for user involvement during the requirements development process – building on Boehm and Papaccio's theory on the cost of late corrections [16] . 51
2.3 Keystroke Level Modelling sequence for a manual login process (taken from NIST-IR7983 [158]) 69

3.1 The overarching multiple-case methodology used for this thesis 89

4.1 Initial coding using Atlas.ti 94
4.2 Emergence of related codes through axial coding using Atlas.ti 94
4.3 Thematic analysis – codebook for the first round of interviews visualised as a code cloud (size of text indicates frequency of occurrence) 95
4.4 Coding progress shows that most of the codes emerged by the third interview for the first set of interviews 96
4.5 Coding progress shows that most of the codes also emerged by the third interview for the second set of interviews 97
4.6 Theme discovery through code maps (first coding round) – code family for delays 98
4.7 Theme discovery through code maps (second coding round) – code family for delays . . 99
4.8 Theme discovery through code maps (first coding round) – code family for interruptions 99
4.9 Theme discovery through code maps (second coding round) – code family for interruptions 100
4.10 Theme discovery through code maps (first coding round) – code family for items to recall in enrolment form 100
4.11 Theme discovery through code maps (second coding round) – code family for items to recall in enrolment form 100

4.12 Theme discovery through code maps (first coding round) – code family for items to generate . . . 101
4.13 Theme discovery through code maps (second coding round) – code family for items to generate . . . 101
4.14 Theme discovery through code maps (first coding round) – code family for type of service 102
4.15 Theme discovery through code maps (second coding round) – code family for type of service . . . 102
4.16 A linear regression model for medical measurements on a group of students (*weight* vs *height*). Residuals are marked as vertical lines between actual data points (dots) and the model (line) (Image source: http://goo.gl/Cb0CJ3) . . . 111
4.17 A multiple regression model with two predictors (Image source: http://goo.gl/t0UVfD) . 113
4.18 Outcome variable following a Poisson distribution (fictitious example) . . . 114
4.19 Surveying enrolment pages (*Jobsearch.gov.au*) . . . 120
4.20 Calibration process overview . . . 123
4.21 One of the nine tasks presented during calibration . . . 123
4.22 Another task presented during calibration . . . 124
4.23 Participants are presented with notifications whenever interruptions or delays are present 124
4.24 Calibration task evaluation form . . . 125
4.25 Task completion parameter estimates for the *young urban professionals (30–40)* user group 128
4.26 Perceived workload parameter estimates for the *young urban professionals (30–40)* user group . . . 128
4.27 The NASA-TLX evaluation section presented in the UGC evaluation sheet following each task . . . 129
4.28 Each participant needs to complete a pairwise comparison of all the workload scales in order to generate a weighted workload measure, thus reducing between-rater variability . 130
4.29 Weighting for *Physical Demand* and *Temporal Demand* across three user groups . . . 131
4.30 Weighting for *Performance* and *Mental Demand* across three user groups . . . 132
4.31 Weighting for the various NASA-TLX workload dimensions generated from the user group calibration exercise conducted with *55+ year old computer newbies*. These are then used to generate warnings on design aspects that can have a stronger negative impact on this particular group of users . . . 136
4.32 Weighting for the various NASA-TLX workload dimensions generated from the user group calibration exercise conducted with *young urban professionals (30–40 years old)*. These are then used to generate warnings on design aspects that can have a stronger negative impact on this particular group of users . . . 137

5.1 Generic SDLC for e-government services . . . 140
5.2 Amended SDLC workflow – introducing requirements and design-time testing for issues related to Critical Design Decisions (CDD) . . . 140

5.3 *Sentire*'s extensions to the *Volere* process . . . 142
5.4 *Sentire*, a requirements framework that listens to personas – extending the *Volere* process. *Sentire*-specific steps are labelled from A to E and are drawn in a lighter shade. . . . 144
5.5 The *Volere* requirements shell or snow card . . . 147
5.6 Enrolment-specific use case annotation . . . 148
5.7 Process to uncover design factors for critical design aspects . . . 148
5.8 *Sentire* – generating simulated user feedback based on Calibrated Personas and annotated use case scenarios . . . 150
5.9 *Sentire* – simulating user feedback on the willingness to complete the task (for each user group) . . . 151
5.10 *Sentire* – simulating user feedback on perceived workload (for each user group) . . . 151
5.11 *Sentire* – dashboard for simulated user feedback generated via a custom built CASE tool. The pie charts on the left denote each user groups' predicted willingness to complete the current product use case while the histogram (right) represents perceived enrolment workload (the various user groups are represented by Calibrated Personas) . . . 152
5.12 A requirements snow card indicating precise fit-criteria for a usability and humanity requirement based on the business owner's experience, insights and previous cost-benefit assessments. These will in turn inform an iterative product use case design process guided by verifiable base-conditions indicating when an acceptable design has been reached (measured through simulated user feedback) . . . 153
5.13 *Sentire* – user group library implemented within the CASE tool . . . 154
5.14 *Sentire* – assigning regression coefficients (for the *WCT* and *PEW* models) to a user group 155
5.15 Ian Alexander's stakeholder taxonomy and the onion model . . . 156
5.16 Artefacts library in *Sentire*'s CASE tool . . . 158
5.17 Card sorting can be used to inform the e-service's information architecture . . . 159
5.18 Conceptual view of *Sentire*'s architecture . . . 162
5.19 *Sentire* – presenting calibration tasks randomly to avoid the perception of incrementing workload . . . 163
5.20 *Sentire* – Calibrated Personas (highlighted sections indicate the calibrated user group with which each project persona is associated – see Figure 5.21 for a user group example) 164
5.21 *Sentire* – user group dialog . . . 165
5.22 *Sentire* – project landing page . . . 165
5.23 *Sentire* – use case annotation (average time is not being used in the calculations). Type of Service (*ToS*) is specified at product use case level (rather than at scenario level) . . . 166
5.24 *Sentire* – project maps allowing for the visualisation of entities and their dependencies (from different perspectives). Various de-cluttering techniques are provided, especially for larger projects with hundreds of interlinked entities (e.g., highlighting of smaller project branches and dragging of overlapping items) . . . 167

6.1 Plan of action for the HRIU case study – highlighted steps are discussed in this chapter . 170
6.2 Facilitated workshop at ETC 171
6.3 Human resource manager persona. This persona is a member of the *young urban professionals (30–40)* user group 172
6.4 An HR manager participating in a user group calibration exercise. One-to-one in-context calibration was selected and calibration was conducted at the participants' premises . . . 172
6.5 Complete results for the task completion (*WCT*) regression coefficients generated for the *young urban professionals (30–40)* user group 173
6.6 Complete results for the perceived workload (*PEW*) regression coefficients generated for the *young urban professionals (30–40)* user group 173
6.7 Login page at www.etc.gov.mt – both e-ID and ETC-specific enrolment processes are provided both of which provide the same level of access and identity assurance 176
6.8 *Young urban professionals' (30–40) PEW* model – normality test for workload (visually right skewed – denoting a non-normal distribution) 177
6.9 Feedback from actual users and the predictions generated via *Sentire* for Scenario 1 . . . 179
6.10 Feedback from actual users (excluding missing values) and the predictions generated via *Sentire* for Scenario 2. If missing values are considered as 'non-adopters' the user evaluation figure would go down to 88%. 179
6.11 *Google Docs* was used for the first case study. It was convenient for document co-editing however the lack of resource linking capabilities provided a challenge especially when the number of use cases and associated requirements grew. 180

7.1 Plan of action for the *undergraduate students (non-IT)* user group case study 183
7.2 University student persona (retrieved from *Sentire*'s persona library). This assumption (hypothesis) persona is a member of the *undergraduate students (non-IT)* user group . . 184
7.3 An undergraduate student participating in a user group calibration exercise. This was carried out in group within a lab environment. 184
7.4 Complete results for the task completion (*WCT*) regression coefficients generated for the *undergraduate students (non-IT)* user group 185
7.5 Complete results for the perceived workload (*PEW*) regression coefficients generated for the *undergraduate students (non-IT)* user group 185
7.6 Undergraduate students' (non-IT) *PEW* model – normality test for workload (visually right skewed, denoting a non-normal distribution) 188
7.7 Feedback from actual users and the predictions generated via *Sentire*. This chart shows the percentage of undergraduate students (non-IT) who would be willing to enrol on *Blog.com* 190
7.8 Feedback from actual users and the predictions generated via *Sentire*. This chart shows the percentage of undergraduate students (non-IT) who would be willing to enrol on *WordPress.com* 190

7.9 Feedback from actual users and the predictions generated via *Sentire*. This chart shows the percentage of undergraduate students (non-IT) who would be willing to enrol on *LiveJournal.com* 190

8.1 Mean weighted workload (*MWW*) for e-service users (online) and for those who adopted the offline exam registration process (at the exams registration department) 202
8.2 Adjusted rating for e-service users who already owned an e-ID (adjust rating = weighting x raw rating) 203
8.3 Adjusted rating for e-service users who had to enrol for an e-ID (adjust rating = weight x raw rating) 203
8.4 Participants had to agree on the level of perceived enrolment-specific workload (from personal experience) for several online services 204
8.5 Perceived enrolment-specific workload for the most common online services 205
8.6 This chart shows the overall mean workload for the three tasks listed in Table 8.8 210

9.1 Consumer Advice Portal project workflow – *T* labels represent sequential tasks while *P* labels indicate parallel tasks 218
9.2 Business use case screen in *Sentire*'s CASE tool 220
9.3 Persona library in *Sentire*'s CASE tool 220
9.4 Facilitated in-context calibration session with a participant 222
9.5 Creation of a new persona hypothesis to reflect an emerging user archetype. Persona posters (shown here) were used during project meetings. 222
9.6 Complete results for the task completion (*WCT*) regression coefficients generated for the *confident newbies (55+)* user group 224
9.7 Complete results for the perceived workload (*PEW*) regression coefficients generated for the *confident newbies (55+)* user group 224
9.8 Persona posters were highly visible during meetings 225
9.9 Card sorting exercise following an initial iteration of product use case designs 226
9.10 Second stage of the card-sorting exercise was to determine user journeys 227
9.11 Simulated feedback for the different user groups represented by the various Calibrated Personas used throughout the design process. The feedback shown above was generated using *Sentire*'s CASE tool on the '*enrol with an MCCAA account*' product use case (Inset: *Sentire* workshop participants) 229
9.12 Atomic requirements were specified using a modified version of *Volere*'s requirements Snow Card template 230
9.13 Prototyping the e-service based on the initial information architecture session 230
9.14 An eye-tracking session participant 231
9.15 A conceptual representation of alerts shown when simulations are generated via partial user models 232

9.16 *Undergraduate students (18–25)* – feedback from actual users and predictions generated via *Sentire* for the willingness to adopt the e-service and complete the primary task online across the four enrolment scenarios 235
9.17 *Young urban professionals (30–40)* – feedback from actual users and predictions generated via *Sentire* for the willingness to adopt the e-service and complete the primary task online across the four enrolment scenarios 236
9.18 *Confident newbies (55+)* – feedback from actual users and predictions generated via *Sentire* for the willingness to adopt the e-service and complete the primary task online across the four enrolment scenarios 236

10.1 Calibrated Personas – embedding user behavioural models (generated via UGC sessions) within project personas. This adds a 'voice' to traditional personas via simulated user feedback. 241
10.2 *Sentire* – embedding Calibrated Personas within *Volere*'s Quality Gateway. This introduces user feedback simulations to the requirements development process. 242
10.3 Calibrated Personas can be adopted to model user behaviour for other critical e-service design aspects 248
10.4 Providing coarser grained results (right) might be more effective for practitioners (as observed during the final case study), whereby the original reporting format (left) might create a perception of more work – thus discouraging adoption or proper use of the tool (*Note: PEW = Perceived Enrolment Workload*) 249

List of Tables

1.1 Main contributions arising from this thesis . . . 32
1.2 Minor contributions . . . 32
1.3 Outline of studies and interventions carried out for this thesis – grouped by case study . . 37

2.1 A persona characteristic argument example (adapted from [55]) . . . 54
2.2 Levels of Identity Assurance (LoIA) compiled from recommendations published by the UK [65], US [61] and Canadian [121] governments . . . 57
2.3 Example of a tally count resulting from a pairwise comparison exercise of the six workload scales . . . 66
2.4 NASA-TLX data for one participant . . . 66

4.1 Enrolment-specific design factors, operationalised for modelling purposes . . . 105
4.2 Parameter estimates for weight (outcome variable) and height (predictor) data . . . 111
4.3 Parameter estimates for *weight* (outcome variable) and *height* and *age* (predictors) . . . 112
4.4 Example parameter estimates for the *willingness to complete task* outcome variable . . . 117
4.5 Dependent variable encoding (*willingness to complete task*) . . . 117
4.6 A small sample of actual observations together with their respective modelled outcomes (*expected*). Workings for values in bold are shown in equations 4.10 and 4.11 . . . 118
4.7 Set of nine enrolment tasks generalised from a survey of commonly found design configurations across various e-services (from low to high workload and assurance levels) . 120
4.8 The set of nine enrolment tasks were modified to include multiple delay intensities . . . 121
4.9 Examples of real-world e-services adopting enrolment processes similar to the ones presented in Table 4.8 . . . 122
4.10 Sample output from a calibration exercise . . . 126
4.11 Regression coefficients generated for the *young urban professionals (30–40)* user group. These coefficients explain the user group's reactions to the various enrolment-related design factors . . . 127
4.12 Multiple project personas linked with different user groups, distinguished by some common factor(s) . . . 134

5.1 Project blastoff (deliverables) . . . 145
5.2 Trawling for requirements (deliverables) . . . 146

5.3 Writing and prototyping requirements (deliverables) . . . 147
5.4 Use case annotation (deliverables) . . . 148
5.5 Quality gateway (deliverables) . . . 149
5.6 User feedback simulation (deliverables) . . . 149
5.7 Requirements specification (deliverables) . . . 153
5.8 Reuse library (deliverables) . . . 154

6.1 Regression coefficients generated for the *young urban professionals (30–40)* user group. These coefficients explain the user group's reactions to specific enrolment-related design factors . . . 172
6.2 Scenarios identified for the *Receive Engagement Form Online* Product Use Case. Scenarios vary from each other since they have different security tasks . . . 174
6.3 Step 1 (first scenario): Employer enrols for an e-ID by visiting a Registration Authority (PIN is received by post) . . . 175
6.4 Step 1 (second scenario): Employer enrols for an account on ETC's portal (by submitting additional details for manual verification) . . . 175
6.5 Predictions for perceived workload and willingness to use the e-service based on Maryanne Jones (associated with the *young urban professionals (30–40)* user group) . . . 175
6.6 *Young urban professionals' (30–40) WCT* model – testing fitness to the data . . . 177
6.7 *Young urban professionals' (30–40) PEW* model – tests of normality . . . 177
6.8 *Young urban professionals' (30–40) PEW* model – test of model effects . . . 178
6.9 *Young urban professionals' (30–40) PEW* model – testing goodness of fit . . . 178
6.10 *Young urban professionals' (30–40) PEW* model – omnibus test . . . 178
6.11 Set of nine enrolment pages were modified to include multiple delay intensities . . . 180

7.1 Regression coefficients for the *undergraduate students (non-IT)* user group. These coefficients explain the user group's reactions to specific enrolment-related design factors . . . 185
7.2 *Blog.com*'s enrolment process . . . 186
7.3 *WordPress.com*'s enrolment process . . . 186
7.4 *LiveJournal.com*'s enrolment process . . . 186
7.5 Predictions for perceived workload and the willingness to complete the task, generated for the above use cases and based on Jane Smith's behavioural models (derived from the *undergraduate students (non-IT)* user group) . . . 186
7.6 *Undergraduate students' (non-IT) WCT* model – testing fitness to the data . . . 187
7.7 *Undergraduate students' (non-IT) WCT* model – testing goodness of fit . . . 187
7.8 *Undergraduate students' (non-IT) WCT* model – likelihood ratio tests . . . 187
7.9 *Undergraduate students' (non-IT) PEW* model – tests of normality . . . 187
7.10 *Undergraduate students' (non-IT) PEW* model – test of model effects . . . 188
7.11 *Undergraduate students' (non-IT) PEW* model – testing goodness of fit . . . 188

7.12 *Undergraduate students' (non-IT) PEW* model – omnibus test 189

8.1 Workload dimension weighting by students who used the e-service and who already owned an e-ID 202
8.2 Workload dimension weighting by students who used the e-service but had to enrol for an e-ID 202
8.3 Various services' enrolment processes, their design factors and consensus based perceived workload 205
8.4 Workload dimension weighting by students following the final pairwise comparison . . . 206
8.5 Median value for the mean weighted workload (*MWW*) score across all participants for the nine fictitious enrolment processes presented during the user group calibration (UGC) exercise 206
8.6 Workload dimension weighting (median) varied when students were supervised as opposed to unsupervised responses (i.e., no immediate help was available) 208
8.7 Tests to determine whether there is a statistically significant difference between an Unsupervised and a Supervised TLX weighting exercise (i.e., pairwise comparison) 209
8.8 This table shows three different tasks from the user group calibration exercise denoting the participants' perceived mean weighted workload (*MWW*) 210
8.9 Contrasting perceived enrolment workload (*PEW*) derived by consensus from actual enrolment processes with TLX-based Mean Weighted Workload (*MWW*) values for similar, but fictitious tasks 210
8.10 Tests to determine whether there is a statistically significant difference between reported workload levels for subsequent incrementally (theoretical) demanding tasks 211
8.11 This table shows the set of nine calibration tasks together with their respective median *MWW* values alongside the median *RTLX* values 212

9.1 This table shows regression coefficients generated for the *confident newbies (55+)* user group (represented by Mary Piscopo) 223
9.2 *Confident newbies' (55+) WCT* model – testing fitness to the data 223
9.3 *Confident newbies' (55+) WCT* model – likelihood ratio tests 224
9.4 CAP – enrolment process alternative 1 233
9.5 CAP – enrolment process alternative 2 233
9.6 CAP – enrolment process alternative 3 234
9.7 CAP – enrolment process alternative 4 234
9.8 Predictions for perceived workload and the willingness to complete the task, generated for the four enrolment scenarios and user groups 234
9.9 Actual user feedback (on *WCT*) for the four scenarios. Respondents are grouped based on demographic similarities to project personas 234

10.1 Revisiting the sub-research questions 240

Chapter 1

Introduction

This chapter outlines the real-world as well as the underlying science problems motivating this thesis. A list of associated research efforts introduce the relevance of these problems while helping to highlight the research gaps that this thesis tackles. The research question serves as the opening statement and sets the tone and direction of this work. The rest of the chapter gives an overview of the work undertaken, the research strategy adopted and the structure used for this thesis.

1.1 Thesis Motivation

Before designing for users it is crucial that their relationships with the social and physical environments as well as the product that is being developed are clearly defined [38]. Without user models one runs the risk of relying on assumptions based on subjective and unstructured information, widening the gap between the known and actual design outcomes (unknown). This gap is materialised as a set of challenges practitioners face when taking the creative leap – from reasoning about a system to actually taking design decisions.

"*Just as physicists have created models of the atom based on observed data and intuitive synthesis of the patterns in their data, so must designers create models of users based on observed behaviours and intuitive synthesis of the patterns in the data. Only after we formalise such patterns can we hope to systematically construct patterns of interaction that smoothly match the behaviour patterns, mental models, and goals of users. Personas provide this formalization.*" Alan Cooper [38].

Alan Cooper also maintains that systems reflect company culture and values, and each system tells a story "*like body language in the way it reveals your inner personality to a patient observer*" – [37].

Jokela and Buie [84] state that Government systems "*continue to demonstrate poor usability*" while providing a "*less-than-satisfying UX* [User eXperience]". Nonetheless a new wave of systems thinking is leaving an impact on how government systems are perceived and built and this is evidenced through the recruitment of industry experts to assist in building user-centric systems. UX is gaining more recognition in this domain. An example of this trend can be seen through the UK's Government Digital Services' (GDS) move towards user-centric thought and design processes whereby a set of ten design principles have been proposed and published for all agencies to adopt. GDS have also adopted a strong philosophical stance in that usability, accessibility and overall experience should be implicitly built within their

services, as opposed to being treated as an afterthought. In a recent statement GDS announced that it removed the explicit accessibility statement for a reason: "*... we realised that if we wanted accessibility to be a basic part of GOV.UK, we had to treat it like all the other basic parts. We didn't have statements for creative design, technical merit, or user friendliness, so why single out accessibility?*" [84]. Another example is the US's General Services Administration's *DigitalGov UX Program* (formerly known as 'GSA's *First Fridays Testing Program*' [24]). This program provides UX-related training for federal employees as well as support for usability testing initiatives. Usability testing starter kits are also provided to kick-start UX testing procedures[1].

Although usability and UX are seeing their way in government projects, in a study conducted across 38 project contracts for public software systems [170] the authors noticed that in most cases statements about usability were vague and focused on precise design or process features (e.g., *software must be tested for usability and screen must have an exit button*). Jokela and Buie [84] argue that these are not proper usability requirements since they lack verifiability, validity and comprehensiveness. Subsequently these were compared to wish lists or strategic desires that may or may not support the user's primary task. Policy makers were also reported to state that in order to ensure usability they generally ask end-user representatives to provide a subjective rating on system demonstrations given by the contractors. Another issue the authors highlight is the distribution of responsibility towards UX. Who should be accountable for negative UX? Should this be tackled as part of an overall design process?

Negative user experience in the government sector could have damaging effects on the service provider's goodwill in both the short and long-term. Citizens will not benefit from added convenience (as projected by the service provider) especially when design issues are severe, possibly resulting in non-adoption, frustration and resentment [127]. Unlike in the commercial space, there are generally no competing providers for government services within a national context, and policy makers may be tempted to push their digital agenda by introducing compulsory e-services with no alternative service provision channels (see Chapter 8). With compulsion, e-service adoption may superficially improve however this may on the other hand affect the users' lived experience by introducing feelings of resentment towards the service, its provider and possibly towards the e-government policy strategy in general. This thesis does not propose methods by which one can measure or predict resentment, however through the use of multi-dimensional workload rating techniques (see Section 2.3.5.1) together with the modelling techniques presented in Section 4.2, one can obtain a good indication of conditions that may lead to resentment (e.g., a situation wherein uptake is predicted to be good while at the same time indicating high levels of frustration and perceived effort across different user groups).

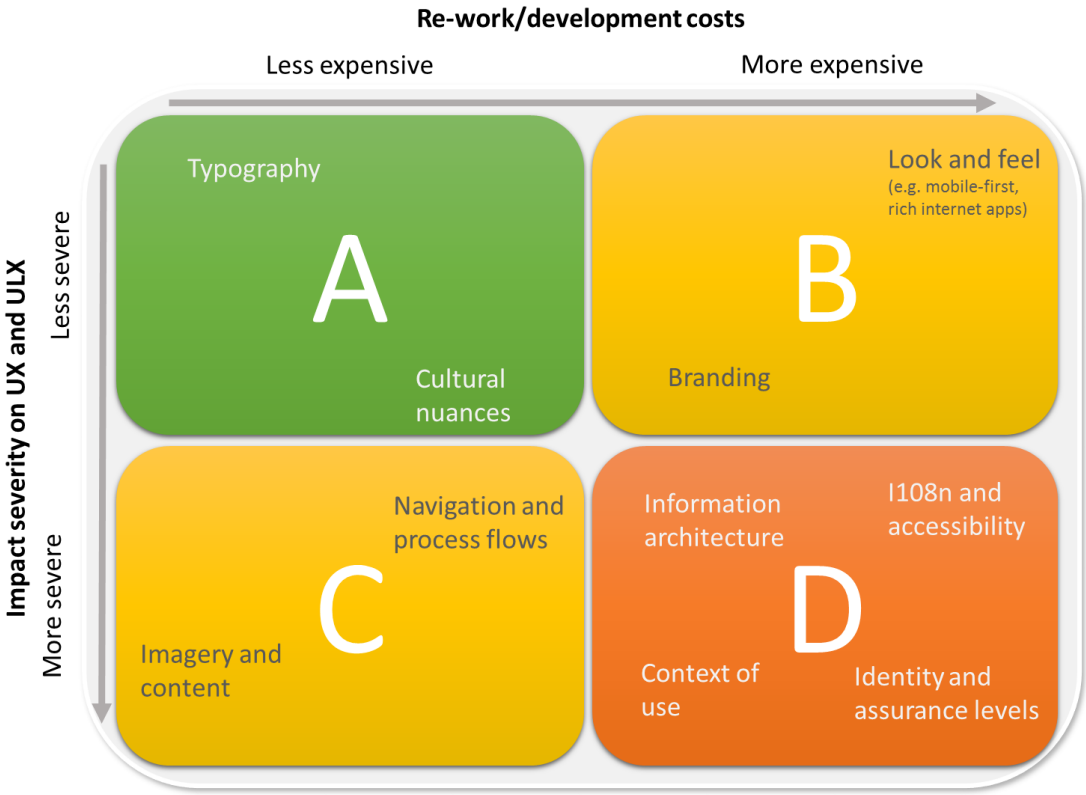
1.1.1 Real-world problem

Certain design decisions have greater implications than others, whereby some are more expensive to rectify at a later stage while others are more forgiving, assuming adherence to architectural best practices. An undetected issue within the requirements could result in very expensive fixes if this finds itself in the final product – the further downstream the issue is detected, the more expensive it is to rectify – and all

[1]Government Services Administration, DigitalGov User Experience Program, http://www.usability.gov/how-to-and-tools/guidance/gsa-first-fridays-program.html, (accessed February 2015)

of this depends on the initial requirements specification. The earlier a requirements error is detected, the cheaper it is to fix [138]. Figure 1.1 provides a matrix representing the possible levels of impact (on users and project resources) arising from issues in specific areas of design. This representation is based entirely on personal experiences and observations. Four quadrants (A, B, C and D) denote the different levels of re-development costs as well as impact severity on the end-users' experience that can be caused (directly or indirectly) by mistakes in the various design elements shown within the matrix.

Figure 1.1: Personal observations on general design problems and their potential impact on product rework, UX and ULX (Users' Lived eXperience – see Section 2.6).



Although this work focuses on design issues similar to those in quadrant D, it is believed that designing for the user experience presents a complex set of problems and requires close collaboration between researchers and industry professionals, informing one another to build a steady momentum towards achieving a basic set of constructs that define a well rounded understanding of the domain.

Usability has been taken seriously by many bodies, including the International Organisation for Standardization (ISO), the British Standards Institute (BSI), and by several governments around the world. This led to the establishment of international guidelines that help designers and developers specify and measure usability in terms of effectiveness, efficiency and satisfaction. Molich and Dumas (in [84]) have established that even though international standards exist, the professional experience of the entity conducting the usability tests will influence the final outcome and results. This shows that this kind of testing is not as robust and scientific as one would believe. Requirements have to be specific enough to guide the testing process with measurable and verifiable fit-criteria. Jokela and Buie go a step further and

state that usability testing does not guarantee good usability. Further to this, user experience cannot be specified or measured unless rigorous, albeit, expensive monitoring is carried out throughout the project's lifecycle. It is very difficult to specify experience-related requirements in a measurable manner and vague and subjective statements are generally used in requirements documents. In principle they all lead to the holy grail of UX: "*the system must make users happy*". Good user experience for public facing e-services, using any form or iteration of the above statement, is extremely difficult to measure, and therefore, guarantee. Through his book, '*Quantifying the User Experience*' [146] as well as several publications, Jeff Sauro has taken this challenge a step further by suggesting the use of statistical techniques to make sense of this fuzzy and subjective problem. Sauro aims to quantify the user experience through statistical analysis of human behaviour in order to ultimately "*provide meaning through measurement*"[2].

1.1.1.1 Why bother with e-government services?

By surveying call for tenders across European states (and beyond) the author observed that e-government services are commissioned (or revisited) on a regular basis. This was especially the case for narrow scoped and public facing transactional services. Examples of such projects include the (re)development of official tourism portals for regions or cities, business directories, pollution reporting systems, self-service portals (e.g., housing and benefits) and so forth. In private communication with the Government Enterprise Architect at the government IT agency in Malta, the smallest state in Europe, he reported that there are over 1,500 documented transactional government services offered by the Maltese government, most of which are still provided manually using paper-based forms. Driven by the need to reduce costs, improve service quality and compliance with legal directives, government entities have been pushed to develop an e-service strategy. This also implies that existing and new services will need to be maintained, updated and replaced in the future, making a case for the need to investigate and propose better design processes for e-government services. Unfortunately, as in Malta's case, internal politics can lead to sub-optimal decision making, such as the push for the adoption of a national e-ID scheme that imposes strict assurance by default at first interaction[3], thus creating a high barrier for adoption and use.

Using public funds, service providers must deliver *useful* services as efficiently as possible – which funds could otherwise be used to tackle other fundamental issues of national importance. Bad use of tax-payers' money signals lack of competence which may in turn undermine trust in government. The central difference between the public and private sector is that in the former the tax-payer (user) cannot simply 'walk away' if a service is unsatisfactory – potentially causing wide-spread resentment. Accountability is important, however decision makers must also be supported with appropriate tools and expertise to formulate and implement a positive, useful and cost-effective e-government strategy.

By following the development of e-government projects throughout the years, the author noted that e-government projects generally suffer from a set of specific threats, including: heavy political influences and lobbying (e.g., favouring one technology platform or vendor in the calls for tender), a wide skills-gap across the several project stakeholders (e.g., decision makers have to trust the advice given by their

[2] Jeff Sauro (2014), Measuring Usability LLCs about pages, http://www.measuringusability.com/about.php, (accessed December 2013)

[3] The e-ID enrolment process can be found at https://mygov.mt/PORTAL/webforms/howdoigetaccesstomygov.aspx, (accessed 6th January 2013)

consultants) and also unforgiving monitoring by pressure groups and government opposition, adding pressure on, and tension within the team. A scientific, systematic, simple, transparent and repeatable process that builds upon cross-project and cross-agency cumulative knowledge is a must, denoting care in the use of public funds. Knowledge management and reuse during product development is important for commercial entities to survive, acknowledging the risks of brain drain due to staff turnover (e.g., maintaining a component library within a computer-aided development environment, using company-wide knowledge bases to store and disseminate process experiences, and so forth). The author believes that this is also important in a government context especially when continuity is fragmented with several government bodies (e.g., ministries, authorities and agencies) contracting different developers to build, deploy and sometimes maintain or extend existing systems.

1.1.2 Science problem

Seffah et al. [147] argue that a large percentage of software maintenance costs are associated with user-specific issues (e.g., usability and accessibility). The authors argue that there is a methodological gap in the way interactive software is built, lacking explicit, systematic and empirical ways to specify, test and validate usability requirements across the entire development process [147]. This problem is amplified even further thanks to a *people gap*, in culture and skills, between software developers and behavioural science practitioners.

Humans generally seek the path of least resistance [26] and have a finite and expendable budget of compliance towards security measures [12]. For instance, when withdrawing cash, the user has to prove that she is the rightful owner of the bank card by putting in a unique number (PIN). This extra effort to complete the transaction seems to be acceptable within this particular context. However any additional identification or security measure (e.g., voice recognition) will consume more of this *compliance budget* (see Section 2.3), causing additional friction which may in turn lead to non-compliance or task abandonment. This problem is more evident at the point of entry for e-government services, whereby citizens may be discouraged from signing-up due to a cumbersome enrolment process. Pfleeger, Sasse and Caputo [126] suggest that a usable security mechanism is not merely a mechanism that users are capable of using, but one that aligns well to the performance requirements of the task at hand, is in line with users' goals and considers the context of use and its inherent physical and social constraints. This work is not about usability of specific security mechanisms (e.g., passwords), but on the level of perceived workload (i.e., hurdles) their adoption generates. This perception can vary from one user group to another and may also vary depending on the context of use and type of service being considered – adding multiple levels of complexity to the inquiry. These hurdles can ultimately impede or discourage users from gaining access to e-government services. There is a continuous, generally sub-conscious, cost-benefit exercise when people are faced with tasks that do not contribute directly to the primary task or goal (e.g., enrolment). Pfleeger, Sasse and Caputo [126] refer to most security mechanisms as 'gatekeepers', whereby if such mechanisms are not usable they may ultimately undermine service accessibility. Beautement et al. state that "*additional authentication hurdles cause delay in accessing systems and cause frustration at having to repeat a task*" [12]. This issue affects all types of users including highly technical users such as

administrators and software developers who "*often struggle to keep up with the increased complexity and workload created by security mechanisms*" [12]. Zurko and Simon also state that the username/password authentication mechanism, "*the most widely used*" mechanism around, is "*unsuitable for providing both easy-to-use and effective security to most end users*" [180].

Beautement et al. [12] postulate that the "*most direct*" way to make a positive impact on the cost-benefit perception is to reduce workload, both mental and physical. Workload is the currency used in this trade-off between the costs generated by an identity or security related task and the benefits obtained thereafter. Reducing friction caused by security tasks on business processes is what "*well designed security*" is all about. Pfleeger and Caputo [125] refer to heuristics, a term used in behavioural sciences to explain the natural (or learned) human reaction to reduce cognitive load in a given situation.

Theoretically this is what designers need to do, however the main research problem here is the lack of quantitative knowledge on the user's elasticity to identity-related tasks (1) in specific situations and (2) across different user groups. This lack of knowledge leads researchers and practitioners to throw guesses at what might be acceptable and what might cause excessive friction on the user (which might result in service abandonment or reversion to traditional channels or practices). Beautement et al. [12] argue that data on "*hotspots of high* [security] *friction*" is essential and "*without this data, or with only poor approximations there-of,* [any decision support tool] *would be useless, and possibly even damaging if it leads to poor decision making*".

There seems to be no practical and systematic way to tackle this problem. The author believes that the requirements stage is the point at which this issue should be handled. Failing to do so may result in expensive rework that could possibly be avoided with some additional effort at the requirements development stage [138]. This leads the author to believe that an early detection mechanism for errors in requirements is desirable.

Requiring excessive identity assurance (with respect to the actual and perceived risk levels) can have serious consequences on the e-service's success, however it is very challenging to assess and select an appropriate identity assurance process systematically for use in specific contexts and with different groups of users. A balance needs to be found between what users are willing to accept (and cope with) and the identity assurance requirements for a given context. This thesis attempts to understand the underlying factors that might lead to improvements in the requirements development process. Tool support is also essential to allow non-technical people to design secure and yet acceptable enrolment-based e-services.

1.1.3 Research gap

In her work Renaud [133] lists a number of design principles for security mechanisms in e-government services. The first two are: "*carry out a formal threat analysis for the protected assets*" and "*put as small a burden as possible on the citizen*". Following these two principles it can be deduced that security levels should reflect the risk levels involved while minimising the burden on users – the main challenge identified here, which is also a gap in knowledge, is the operationalisation of the relationship between *risk levels*, *identity processes* and *citizen workload*. How can one predict the point at which additional

security measures (identity assurance) will surpass the users' acceptability threshold? Will this vary across different service contexts? And will this perception vary across user groups? Assurance levels that surpass the acceptability threshold may result in (1) *circumvention of security measures* and (2) *service abandonment or reversion to traditional channels.* On the other hand low assurance levels may increase the risk for both users and service providers – including attacks on protected assets as well as identity theft. Identity mechanisms such as enrolment processes are considered to be gatekeepers to the fulfilment of user goals and the conveyors of first impressions, thus central to the design of positive user experiences.

National laws, international standards and entities exist to define techniques and tools to specify, measure and validate usability, however this is not the case for user experience. The field is not mature enough [84] and there are no generally accepted techniques or UX parameters to specify, measure and validate systems. Checklists exists to verify usability factors nonetheless it is very difficult to measure the users' experience in a systematic and scientific way (e.g., level of frustration, resentment or willingness to complete a task). Further to this, e-government service designers need to find a balance between appropriate identity assurances (given the value of assets being protected) and a positive user experience while designing enrolment-based services guided by a systematic and usable requirements development process. Herley and Oorschot [72] argue that one of the research directions required is that of finding better means to identify actual requirements in support of better identification of best-fit mechanisms for specific scenarios.

This is the research gap being addressed in this thesis. In order to better understand the problem it was decided to focus on one specific, yet critical and common e-service element: *the enrolment process*. The complexity of issues surrounding the UX domain is acknowledged, however the author believes that enrolment is the first and major hurdle in most online activities and given that users are goal-driven [144, 12, 128] it is strongly believed that this design element can potentially make or break an e-government service. The existence of this risk has been evidenced in both the commercial and public sectors. A detailed discussion can be found in Sections 2.3.1, 2.3.2 and 2.3.3.

Seffah and Metzker [148] also iterate the need for tools to "*support developers in acquiring and sharing [user-centred design] and software engineering best practices*" and to analyse and visualise the large amount of observational data collected during usability studies [148]. Section 5.7 discusses the need for tool support and the integration of usability and experience design practices into the requirements development process.

1.2 Research Question

This thesis proposes the following question: **How can user behavioural modelling support the requirements process to encourage takeup in enrolment based and public facing e-government services?** Having a strong exploratory component this question is further dissected into a series of sub-questions that drive the inquiry:

SRQ1 What is the relationship between different levels of enrolment-specific friction and the adoption,

security and cost of e-government services?

SRQ2 Which enrolment-specific design factors contribute to friction?

SRQ3 How can user behaviour be modelled to simulate reactions towards friction in new enrolment processes – given different e-service contexts and across a rich diversity of user groups?

SRQ4 How can UX simulations help non-HCI practitioners design better e-services?

An overview of how these research questions are addressed within this thesis is provided in Section 1.7.

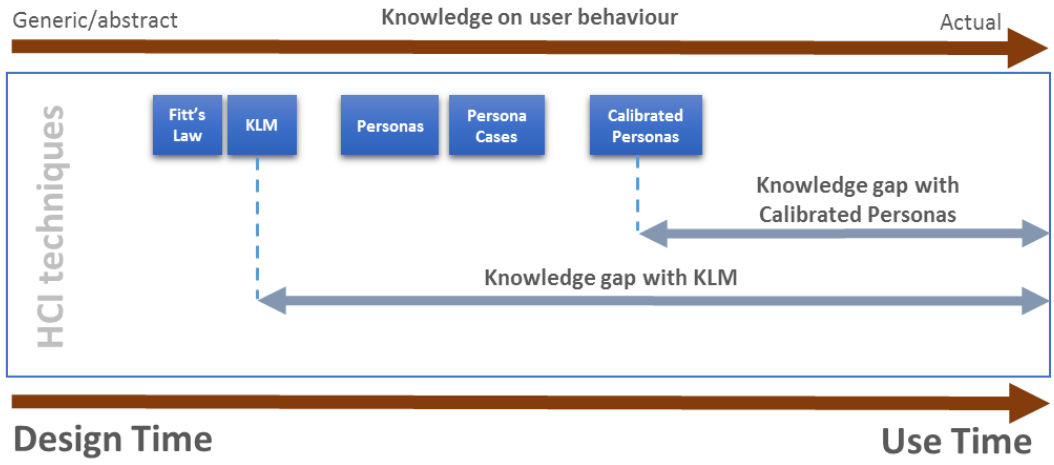
1.3 Contributions to the Interface Between Requirements Engineering and HCI(Sec)

Human capabilities and limitations provide natural upper and lower boundaries limiting the extent of possibilities for human-computer interaction (HCI) research and practice. Empirical evidence on human capabilities and limitations as well as modelling techniques have been adopted and also produced by the HCI community, in particular to the field of secure interaction design and usable security (HCISec). These cover aspects such as users' sensory capabilities, motor and cognitive capabilities and limitations as well as user perceptions, beliefs and motivations [21, 144, 143, 158, 29, 175, 131]. Several theories, models and frameworks have been imported into HCI from various disciplines, such as the cognitive, social and organisational sciences and these are discussed in some depth by Sharp, Rogers and Preece in [139]. By ignoring established HCI theory, practitioners run the risk of specifying requirements or designing tasks that are beyond basic human capabilities (e.g., enforcing a password policy requiring users to generate a complex 16-character non-dictionary password). Existing HCI theories and techniques can help designers observe general design limitations, however they stop short from approximating (predicting) user behaviour with respect to the service under consideration, its context of use and its potentially heterogeneous user base. Techniques such as KLM (Keystroke Level Model) [29] and associated tools [87] can help designers predict workload for tasks such as online form submissions, factoring in time demands for mental and motor operations based on a set of parameters reflecting general user capabilities (e.g., good typist vs average or poor typist). Although such insights are helpful, these can be said to be operating on abstract generalisations of users, user behaviour and capabilities, while ignoring the type of service and its context of use. This thesis acknowledges the complexity of the e-government domain which includes different categories of e-services, heterogeneous user groups with different capabilities all of whom operate within and across different contexts of use. For this reason a finer-grained user modelling technique is presented, building upon general human capabilities while fine-tuning (and progressively accumulating knowledge about) known capability boundaries and user groups' tolerance-levels towards workload as exhibited within different contexts. This modelling technique aims to better reflect reality at design-time and is embedded within the persona construct – referred to as Calibrated Personas (see Chapter 4). Building on Cooper's personas [35, 38] a computational aspect is added through embedded user behavioural models which are in turn used to simulate user feedback on different aspects of a

proposed use case – this reporting technique is referred to as UX-analytics. Calibrated Personas acknowledge the heterogeneity of user behaviour through the inclusion of aspects such as personal experiences, levels of confidence, tolerance towards workload and daily routines within a systematic calibration process – referred to as User Group Calibration. Various empirical studies have been conducted and have shown that Calibrated Personas are sensitive enough to provide design-time predictions (i.e., simulated feedback) on different user groups' reactions towards design decisions related to enrolment within public facing e-government services. These simulations have been shown to influence design decisions taken by e-service project teams during the requirements development process (see Chapters 6 and 9).

Existing user centred design techniques can provide some indications on how users may react to specific design decisions, however a wide gap exists between design-time knowledge on user behaviour (provided by such techniques) and actual use-time data (see Figure 1.2). Calibrated Personas close this gap even further through simulated user feedback – for specific user groups, service types and contexts of use – minimising the 'unknown' during decision making, specifically on critical aspects of e-service design. Calibrated Personas' underlying behavioural models are progressively refined through knowledge accumulation across consecutive projects. This thesis presents Calibrated Personas as the main contribution to the field of HCI.

Figure 1.2: Calibrated Personas – this thesis' main contribution to HCI research and practice. This technique helps to reduce the gap between design-time knowledge on user behaviour and actual use-time data for specific systems, contexts and user groups



In order to support practitioners, this thesis presents *Sentire* – a requirements framework that builds upon and extends *Volere* (a widely adopted industry strength requirements development process [138]). *Sentire* extends *Volere*'s Quality Gateway with Calibrated Personas which in turn allow for user feedback simulations as part of the requirements quality assurance stage. Simulated user feedback introduces the opportunity to specify *quantitative and thus testable fit-criteria for user experience requirements* within the requirements development process through measurable base-conditions informing an iterative design process for enrolment-based product use cases. *Sentire* is presented as a primary contribution to the field of software engineering, and requirements engineering (RE) in particular. Tool support is also presented

for *Sentire* as a contribution to practice.

Table 1.1 lists the major contributions arising from this thesis, while Table 1.2 provides an outline of secondary (minor) contributions produced throughout the various studies. The reader is also directed to specific sections within this thesis that support each contribution.

Table 1.1: Main contributions arising from this thesis

		Contributing to		
	Major contributions	**Research**	**Practice**	**Thesis section(s)**
C1	*Calibrated Personas* – a technique to model and predict user reactions to and perceptions of e-service enrolment processes	✓ HCI(Sec)		4.2.3, 4.2.4
C2	*Sentire* – a requirements framework based on simulated user feedback (using Calibrated Personas)	✓ HCI/RE	✓	5
C3	Collaborative tool support for *Sentire*		✓	5.7

Table 1.2: Minor contributions

		Contributing to		
	Minor contributions	**Research**	**Practice**	**Thesis section(s)**
C4	A study on user attitudes towards enrolment processes	✓ HCISec	✓	4
C5	Assessment of NASA-TLX's sensitivity for enrolment-specific perceived workload on younger audiences	✓ HCISec		8
C6	Testable fit-criteria for experience related (non-functional) requirements		✓	5.3
C7	User group knowledge base for reuse across government projects		✓	4.4, 6, 7, 8, 9

1.4 Thesis Scope

Calibrated Personas and *Sentire* are limited to specific facets of the larger user experience conundrum within public facing and enrolment based e-services. Other authors are tackling various isolated aspects of this bigger problem (see Chapter 2.3.4). This thesis considers overzealous enrolment processes as a major contributor to task abandonment during first-time interaction with e-government services. Over-protection may not be detected due to a disconnect between the requirements development process and actionable early-stage user feedback. This thesis tackles issues of design from a user's lived experience perspective (e.g., feeling discouraged from using online services), which may in turn have long term effects on the e-service and possibly on its provider (e.g., resentment towards policy makers). This may be caused by high-friction processes and the various impacts these might have on the different user groups. Several authors have been studying more granular aspects of user experience, including form design and typography [154], information architecture [112], psychology of page layout and wording [171], and content strategies (Bailie in [24]) amongst others. The author believes that user experience is a multi-faceted science involving various disciplines and it would be an egregious shortcoming to

consider user experience as a purely visual science bound to the human-system interaction space and time.

This thesis outlines a method by which one may build quantitative models to explain users' reactions towards critical design decisions (see Chapter 4) – in particular, decisions related to enrolment processes. These decisions may introduce friction within the primary task independently from small-UX decisions (e.g., visual appeal). Based on this argument, one cannot guarantee a high task-completion rate even if the most appealing user interface (UI) is produced, one which is easy to use and follow. This is especially true when such UI would be interfacing a high friction underlying process (or workflow) that would have an impact on the users' lived experience (ULX), beyond the interaction space and time.

It would be a mistake to claim that Calibrated Personas and their use within *Sentire* can precisely predict the success or failure of an e-service, mainly because many other influencing factors exist (technical and otherwise) that would need to be taken into consideration. Nonetheless these techniques provide a systematic approach to inform decision making on critical aspects of e-service design as part of the requirements development process through measurable and comparable feedback. *Sentire* promotes the idea of UX-analytics as part of the requirements development process while creating active awareness that every design decision will have an impact on different user groups. Calibrated Personas are used to generate user-group specific feedback at design-time, enabling project teams to fine tune their designs according to their target audience. This also allows project teams to balance their own goals (e.g., identity assurance and service adoption) with expected user response (e.g., the willingness to accept a specific level of workload to obtain marginal short or long term benefits) – see Section 5.4.

1.5 Publications and Awards

1.5.1 Publications

Except for P6, the author was the lead researcher for all of the publications listed below (including authorship). Figure 1.3 provides a map denoting how each publication contributes to individual chapters presented in this thesis.

P1 **Porter, C**. (2011). Privacy and usability in SMS-based G2B/B2G m-Government: STK and SMS: Balancing privacy and usability. EUROCON2011 (IEEE). Lisbon, Portugal.

Abstract: The provision of sensitive information over SMS has been held back due to the inherent privacy problems of SMS. Sending messages as plain-text carries multiple risks. SMS encryption is one solution to this problem. However software written for specific devices might affect the User Experience (UX) while impacting existing investments. Also, using the phone's memory as a data store provides no guarantees against intrusion attacks. A standards-based STK (SIM Toolkit) application has been adopted in order to strike a balance between usability and mobility. An 8-bit microcontroller was used as the main platform for this security application, implementing Twofish symmetric encryption to enhance privacy and confidentiality. Creating synergy between Security and Usability is a challenge, and this paper discusses the role of STK and SMS in G2B m-Government.

Story: This paper represents the initial direction taken by this thesis. It builds on previous work with

the aim to expand and explore the contribution from an HCI perspective. With further reading and discussions with other researchers it was becoming clearer that the problem may not necessarily lie with technological constraints but with a clear disconnect between the engineering disciplines and behaviour sciences (in research and practice). This realisation repositioned the thesis's scope to explore how this gap could be reduced, if at all.

P2 **Porter, C**., Sasse M., A., & Letier, E. (2012). Designing acceptable user registration processes for e-services. Proceedings of HCI 2012 – The 26th BCS Conference on Human-Computer Interaction, Birmingham, UK

Abstract: User registration can have a serious impact on the success of online government services. Different services require different levels of identity assurance, and different registration processes are put in place to deliver them. But from the citizen's perspective, these processes often require a disproportionate amount of effort, which reduces users' acceptance. Typically, when sign-up to high-effort services is not mandatory, take-up is low; when it is compulsory, it causes resentment, and neither is desirable. Designers of services requiring registration currently have no way of assessing likely user acceptance at design-time. We are introducing a tool that allows system designers to identify the impact of registration processes on different groups of users, in terms of workload and friction. Personas have been successfully applied to assist security designers, and we extend the concept with statistical properties, and introduce the User Group Calibration (UGC) exercise to calibrate the different personas for sensitivity to specific identity-related elements.

P3 **Porter, C**; Sasse M., A., & Letier, E. (2013). Giving a voice to personas in the design of e-government identity processes. From Research to Design: Challenges of Qualitative Data Representation and Interpretation in HCI – BCS HCI 2013, Uxbridge, UK

Abstract: Identity processes, such as enrolment and authentication, can have a negative impact on the user's experience. By using personas designers get a better understanding of the end user during the design process. Personas represent a user archetype to assist in the development of [digital] products. However this technique involves a measure of subjective interpretation. Following a qualitative empirical exercise we extend the persona concept to include statistical capabilities in order to inform the decision making process through measurable and comparable feedback. This feedback indicates how acceptable an identity mechanism is for a specific group of users. For this purpose we propose Calibrated Personas, an extension of the persona design tool that encapsulates the necessary regression coefficients which can help us predict perceived workload and users' willingness to complete a task given specific design decisions.

P4 **Porter, C**; Letier, E. & Sasse M., A. (2014). Building a national e-service using *Sentire*: Experience report on the use of *Sentire*: a *Volere*-based requirements framework driven by Calibrated Personas and simulated user feedback, RE14, Karlskrona, Sweden

Abstract: User experience (UX) is difficult to quantify and thus more challenging to require and guarantee. It is also difficult to gauge the potential impact on users' lived experience, especially at the earlier

stages of the development life cycle, particularly before high-fidelity prototypes are developed. We believe that the enrolment process is a major hurdle for e-government service adoption and badly designed processes might result in negative repercussions for both the policy maker and the different user groups involved; non-adoption and resentment are two risks that may result in low return on investment (ROI), lost political goodwill and ultimately a negative lived experience for citizens. Identity assurance requirements need to balance out the real value of the assets being secured (risk) with the user groups' acceptance thresholds (based on a continuous cost-benefit exercise factoring in cognitive and physical workload). Sentire is a persona-centric requirements framework built on and extending the Volere requirements process with UX-analytics, reusable user behavioural models and simulated user feedback through Calibrated Personas. In this paper we present a story on how Sentire was adopted in the development of a national public facing e-service. Daily journaling was used throughout the project and a custom built cloud-based CASE tool was used to manage the whole process. This paper outlines our experiences and lessons learnt.

1.5.2 Forthcoming publications

P5 **Porter, C**; Sasse M., A & Letier, E. (2014). NASA-TLX – from Aircraft to E-government: A Study on Enrolment Processes and Related Workload in a Compulsory E-service for Digital Natives, (Manuscript in preparation for submission at the Government Information Quarterly (GIQ) Journal)

Abstract: In 2013 Malta launched a new e-service for students aged 16–18 who were applying for their A-Level exams. Adoption was compulsory and students also needed to enrol for a national e-ID to gain access to the service. This was a one-off opportunity whereby theory could be applied on an actual nation-wide case study. First we seek to explain the impact of perceived workload on the students' lived experience while substantiating these measurements through qualitative insights. The second goal is to present new insights on the applicability and sensitivity of NASA-TLX as a multidimensional and subjective workload measurement technique within this context and with this particular group of users. We found a fair level of statistical significance in workload sensitivity however there are specific exceptions and caveats in adopting this tool. For this reason we propose some modifications to the TLX exercise and reassess the validity and need of the steps involved. We then conclude this paper with a set of actionable recommendations for both practitioners and researchers. This study focuses on digital natives – people who have grown up with, and are highly accustomed to digital technology [130]. This user group is contrasted with digital immigrants – people who have adopted technology later on in life by choice or necessity.

Story: Part of this paper was initially submitted at the ACM Digital Identity Management Workshop following the completion of the third case study (see Chapter 8), but was rejected. This motivated the author to expand this study to include a follow up exercise to test the sensitivity of NASA-TLX, discussed in Section 8.4.2.

P6 Brown, N; Makri, S; Oussama, M; **Porter, C**; Stockman, T; (2014). From qualitative data to design. (Manuscript in preparation for submission at the Interacting with Computers (IwC) Journal)

Author's contribution: This paper presents a discussion on qualitative research techniques and the challenges HCI researchers face when moving from research to actual product design. The paper also presents a set of practical recommendations on the use of well known qualitative methods. The author discusses the appropriateness of mixed-method techniques in qualitative HCI research. Only the author's contribution is reported in this thesis (see Section 3.1).

1.5.3 Posters

P7 **Porter, C**. (2013). SENTIRE / designing human-centric identity systems. Poster session presented at the Information Assurance Advisory Council (IAAC) Symposium, London, 2013

1.5.4 Multimedia

A five-minute explainer video was created for dissemination purposes. This provides a high-level introduction to the techniques presented in this thesis together with an overview of the main deliverables. The target audience includes potential industry partners, research collaborators and end-users (i.e., government agencies). The explainer video can be accessed through this link: `http://vimeo.com/85960146`.

1.5.5 Industry recognition

Information Assurance Advisory Council (IAAC) – *Sentire* was presented at the UK's Information Assurance Advisory Council (IAAC) during the 2013 annual symposium at the BT Centre Auditorium in London. It was selected for the best poster award by an independent judging panel based on innovation, content and presentation.

International Design for Experience Awards – *Sentire* was shortlisted as a semi-finalist at the 2013 International Design for Experience Awards. This awards programme was developed by a community of experience design professionals and academics. Tobii, Atlassian and Citrix were amongst the winners of the 2013 edition. *Sentire* is featured on the design for experience awards site[4] and will also feature in UX Magazine.

1.6 Overview of studies in this thesis

Due to its exploratory nature, all of the case studies conducted for this thesis have been planned and carried out incrementally, whereby questions raised during one study informed the selection of and planning for the subsequent study. Each case study may contain a real-world intervention as well as one or more sub-studies as required to satisfy its overall aims. In turn, the entire set of interventions and sub-studies contribute towards the research question driving this thesis. Table 1.3 provides an outline of studies and interventions conducted for this thesis.

[4]UX Magazine – Design for Experience Awards Gallery, http://awards.designforexperience.com/gallery/2013/innovative-technique-or-tool/university-college-london, (accessed June 2014)

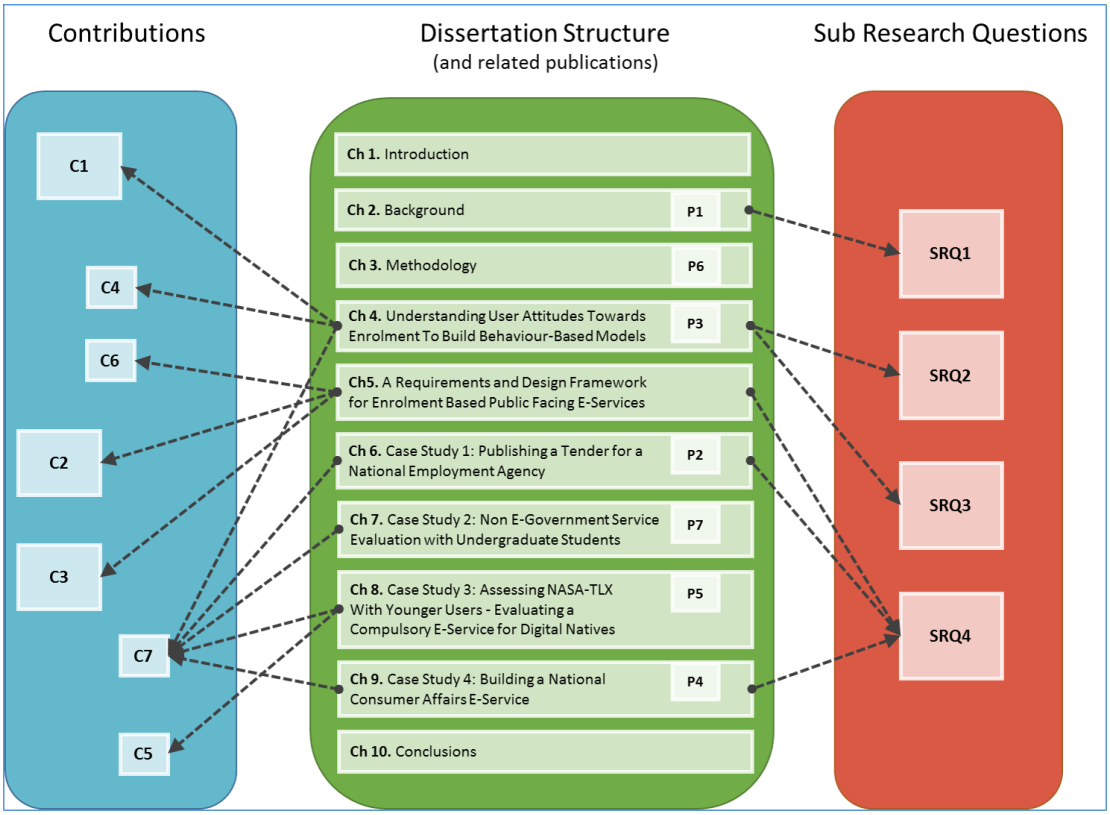
Table 1.3: Outline of studies and interventions carried out for this thesis – grouped by case study

	Initial theory formulation	**Case 1. HRIU**		**Case 2. Undergraduate students**			**Case 3. Digital Natives**		**Case 4. CAP**	
	Study 1	**Intervention**	**Study 2**	**Study 3**	**Study 4**	**Study 5**	**Study 6**	**Study 7**	**Intervention**	**Study 8**
Section	**4**	**6**	**6.4.2**	**7**	**7.3.2**	**7.3.3**	**8**	**8.4.2**	**9**	**9.4.2**
Aim(s)	1. Identify enrolment-process related design factors that have a negative impact on user experience (based on feedback from regular internet users) 2. Formalise a user behaviour modelling technique	1. First real-world intervention using *Sentire* and Calibrated Personas to draft a requirements document for a new e-service commissioned by the National Employment Agency 2. Revise and refine proposed techniques	User evaluation to assess *Sentire*'s predictions	1. Test the sensitivity of the User Group Calibration process (UGC) with another group of users 2. Revise and refine proposed techniques	User evaluation to assess *Sentire*'s predictions	Construct further insights on this user group's general attitudes towards enrolment, the main causes of frustration and factors that influence their decision making process	Study the impact of a compulsory e-service on the users' lived experience through qualitative experience reports and NASA-TLX workload ratings (Service: a compulsory online registration and payment e-service for the 2013 A-level exam sessions)	1. Assess the effectiveness and applicability of NASA-TLX with younger participants 2. Test the sensitivity of NASA-TLX for enrolment related workload 3. Revise and refine the User Group Calibration (UGC) process	1. Field evaluation of the latest iteration of *Sentire* and its associated tools and techniques on a new national e-service (Consumer Affairs Portal) for the Malta Competition and Consumer Affairs authority (MCCA)	User evaluation to assess *Sentire*'s predictions
Method(s)	Semi-structured interviews	Real-world intervention	Online questionnaire	Lab-based workshop	Online questionnaire	Focus groups	Online questionnaire	Lab-based workshops	Real-world intervention	Online questionnaire
Target user group(s)	Regular internet users	Young urban professionals (30-40 years old)		Undergraduate students (18-21 years old) majoring in any non-IT related field of study			A-Level students (16-18 years old)		*55+ technology newbies* as well as previously calibrated user groups	

1.7 Thesis Structure

Figure 1.3 outlines the structure used for this thesis including research questions, contributions and resulting publications linked to their corresponding section(s).

Figure 1.3: Thesis map – list of chapters and their relationship with the sub-research questions, contributions and resulting publications



Chapter 2

Background

This chapter starts by presenting a discussion on the interface between requirements engineering (RE) and human computer interaction (HCI), in both research and practice. This is then followed by a critical analysis of existing literature on requirements practices and user experience considerations adopted within the public sector. A discussion on the impact of enrolment processes on e-government service users and on the service provider itself is then provided, drawing on lessons learnt across various e-government initiatives, experiences shared by internet giants as well as existing research in usable security (HCI-Sec) and human factors (HF). This chapter concludes with a discussion on possible threats to e-service usability arising from both process and people-related issues within the public sector.

2.1 Usability, Experience and Lived Experience

2.1.1 User Interface and Usability

UI is the technology that allows humans and machines to interact. This interaction is broadly studied in the field of HCI (Human-Computer Interaction). Users control machines through hardware and software components (input) and in turn machines provide the necessary feedback (output) to allow users to decide on the next action to take. Interfaces vary according to the system and its context of use however the goal of any UI is to enable effective operation conducive to reaching the intended user goals while minimising effort, time and frustration. This calls for a whole set of disciplines, including accessibility, usability and user experience.

Usability goes beyond user interface design and includes the design of workflows required for users to achieve their goals effectively and efficiently in a specified context of use. Jakob Nielsen [114] describes usability as a quality attribute defined by several quality components, including (1) learnability (task accomplishment at first interaction), (2) efficiency (performance achievability), (3) memorability (re-establishment of efficiency in future interactions), (4) errors (measuring frequency, severity and recoverability) and (5) satisfaction (pleasantness in use and visual appeal). Usability is distinguished from utility in that the latter, although equally important, determines whether the system meets the user's needs – and if it is easy to use then it also becomes useful. Ease of use does not equate to utility and vice-versa, nonetheless similar techniques can be adopted to study both. Standards such as the ISO9241 series and the ISO10075 series provide system design principles for usability and ergonomics, including

mental workload [165]. These standards also provide guidelines on specifying the context of use and evaluating usability in terms of measures of user performance and satisfaction. User knowledge, goals and context of use have a huge impact on usability. If ignored they can all impact the effectiveness (task completion), efficiency (effort required) and satisfaction of users' interactions with a system.

2.1.2 User Experience – UX

UX is an overarching perspective on user interaction involving several disciplines and considering aspects such as emotions and impact on the users' psyche and well-being. Fredheim [60] also lists hedonic, aesthetic, affective and experiential aspects arising from UX, all of which are not easily measured [60]. A usable system with an elegant user interface may still have a negative impact on the user's experience. By way of an example, Jakob Nielsen [116] puts forth the same argument by suggesting that even if a film review website has a perfectly planned and executed user interface, poor UX may still exist if the underlying film database is limited in scope and breadth (e.g., excluding whole genres of films). ISO (International Standards Organisation) defines UX as "*a person's perceptions and responses that result from the use or anticipated use of a product, system or service*" [76].

UX is a broad, highly subjective and context-dependent domain [94] involving measures to make a product both useful and enjoyable [116] – to own, associate with and use. To add to its complexity UX (as a field) is perceived differently by different people with varying levels of experience [94] and more often than not it is used indiscriminately in the field of interaction design [60]. Nielsen [114] uses several terms to define UX including utility and usability (thus usefulness), elegance and simplicity. All in all Nielsen equates UX with the joy to own and use a product and to do this developers need to merge several disciplines, including engineering, marketing and design (spanning across interface, aesthetic and possibly tactual elements). A system may look good, be usable and is technically well-built, but it can still put its users in undesirable states of mind engaging feelings such as fear (e.g., security and privacy), stress (e.g., temporal or mental demand), resentment and mistrust. These feelings are vague and difficult to specify, measure and verify, especially at the requirements stage when the e-service is not yet built. At *usability.gov* the authors explain that UX best practices help to improve the quality of user's interaction with the product or service while improving its perceived value. At the same time UX is improved by ensuring usefulness, usability, desirability, clarity, findability, accessibility and credibility. This follows Peter Morville's recommendations in the UX Honeycomb [109]. UX should be supported with proper user-centred design processes and project management practices, as well as user research, usability evaluation, information architecture planning, UI design, interaction design, visual design, content strategy, accessibility evaluation and web analytics for user behaviour analysis. Dimmick [42] discusses the merits of *big UX* (or *big picture UX*) as opposed to *small UX* [42] while Saucken et al. [168] adopt the terms *macro* and *micro UX*. Although both are related to user interaction with systems, the main difference lies in the level of design abstraction. *Big picture* or *macro UX* looks at the strategic impact of a system on the users' mental model of the service and the users' perception towards its provider including expectations, trust and loyalty. On the other hand *small* or *micro UX* is concerned with the tactical design decisions that may have an immediate and direct impact on how users interact

with systems, possibly resulting in satisfaction and utility. Saucken et al. [168] define *micro UX* as "*design in detail concerning material, usability and interface*".

2.1.3 Users' Lived Experience – ULX

"*It is only by seeing technology as participating in felt experience that we understand the fullness of its potential*" [105]. McCarthy and Wright believe that the interface between technology and the social user has wider and deeper connotations on the user that go beyond product-specific disruptions bound to the immediate human-technology interaction space and time. A difficulty while interacting with technology may have an impact on users that goes beyond '*the interface*' – with repercussions extending into the users' daily lives and away from the technology itself. Not being able to complete the primary task may lead the user to doubt in her own capabilities or induce feelings of helplessness, anger and frustration. These experiences may consequently impact the users' life as a social being. When designing for the lived experience, one should consider aspects such as the user's personal experiences, culture and values [75]. Feelings of fear, pre-occupation and frustration may result when a system conflicts with personal beliefs, values or experiences. Rahaman and Sasse [131] provide several examples of how identity systems may affect the users' lived experience. The authors [131] explain how a new system introduced under the Poor Laws in 17th century England (requiring beggars to wear a bright coloured badge on their shoulder indicating their parish of origin) caused its 'users' to feel rejected. This was mainly due to the fact that wearers of such badges had their social status publicly exposed, introducing feelings of shame that go beyond the usability and effectiveness of the system itself.

2.2 The Interface Between RE and HCI

2.2.1 Classical HCI considerations in RE

Zave [179] defines requirements engineering as the branch of software engineering concerned "*with the real-world goals for, functions of, and constraints on software systems. It is also concerned with the relationship of these factors to precise specifications of software behaviour, and to their evolution over time and across software families*".

Karl Wiegers states that requirements engineering is a communication activity and not a purely technical activity [172]. This can be backed up by the fact that if communication barriers exist, and different people have different ideas of what the requirements actually are, problems will surely develop in the early phases of the project. Wiegers also explains that RE is the understanding [by all stakeholders] of what is intended to be built, before actually building it.

Jackson [77] argues that the term *requirements* is often used to specify functional requirements and in [78] it is stated that requirements are located in the environment and are to be segregated from the machine or product to be built. Specifications on the other hand are restricted forms of requirements, giving boundaries and rules which are to be adopted in order to implement an effective solution. Jackson divides systems into the environment and the machine, and through this he denotes the importance of formalisation techniques. This is due to the fact that most environments provide haphazard and widely heterogeneous variables, changing from one industry to the other and from organisation to the next. The

goal for the formalisation techniques as described by Jackson is to provide a "*faithful approximation of the informal reality*" to the customer.

Lamsweerde [88, 89] suggests a number of activities which are pertinent to RE, namely; domain analysis, elicitation, negotiation and agreement, specification, analysis of specification, documentation and evolution. These activities are iterative in nature in an effort to elaborate and manage a complete specification for a new system or product. Other authors, such as Goguen and Linde [64] denote that many scientists see RE as the phase in which scientific processes stop and chaos begins. Is there any order in the social world? For this reason, added importance is given to the social context wherein social order may not be readily visible and "*obvious*". This soft view of the social context elicits an important aspect of RE; the importance of having proper communication channels between all the stakeholders. On the same lines, Leveson [95] considers RE to pertain to the fields of cognitive psychology, systems theory and human-machine interaction. Her work concludes that it is of greater importance to understand the intention for which the system was or is to be designed [95]. Having set a psychological framework for, and having understood the reason why the system is to be developed, that is, knowing the goals which the system is to help achieve, clears the way for successful translation of the business' requirements into an operational technological solution.

One of the largest problems of evolutionary and long-term management of requirements is cost. Boehm and Papaccio [16] state that late corrections in software (already in operation) may cost 50 and up to 200 times more than if such corrections are carried out at the requirements stage. When this statement was written (in the 80's), technology was not as enabling as today's, while project management methodologies which were generally followed discouraged late amendments to the function points currently present and planned within the system. So the 200 times multiplier might be diluted down to a relatively lower figure due to technological and management advancements. Whichever figure that might be the problem persists – late changes will still be a burden, especially in highly competitive markets.

The determinant of success for any task or goal is the degree by which it satisfies the purpose or intention for which it was initially commissioned. Nuseibeh and Easterbrook [118] earmark software systems requirements engineering as the process for identifying this "*purpose*" (or "*intention*"). It is suggested that this process should take the form of a documented study including an analysis of all the stakeholders and their needs, while producing artefacts which are fit for analysis, communication (between all parties), and eventually translatable for implementation.

In 2002 at the RE conference held in Essen (Germany), RE was positioned as a "*well-recognized practice and research area*". The importance of "*skills, processes, methods, techniques and tools*" within RE was also given its due importance. Within the same definition, the issue of diversity in terms of application domains was denoted as an underlying challenge which practitioners face on a project-by-project basis.

Axel Van Lamsweerde outlines a number of requirement categories in [90], including:

Functional Requirements Define what the system shall do

Non-Functional Requirements Define constraints and parameters within which functional require-

ments should be satisfied

Quality Requirements Or quality attributes, adding the level of requirement satisfaction, defining integrity, availability, reliability and privacy amongst others.

Compliance Requirements Define requirement conformance to issues such as regulations, norms and cultures amongst others.

Architectural Requirements Architectural nature of the system, including software component distribution and installation requirements.

Development Requirements Requirements related to the way systems are to be developed, including requirements such as re-usability, portability and maintainability amongst others.

Castro, Kolp and Mylopoulos denote that non-functional requirements have not received enough attention, and are "*less well understood than other, less critical factors in software development*" [30]. Should solution design be based on intentional specifications or on operational ones? Intentions specify the goals of a system, while operational specifications explicitly specify the system operations, their rules and dependencies among others. van Lamsweerde stipulates that the two approaches are complementary [90], in which intentional specifications "*leave the operations realizing them implicit*" and operational specifications "*leave the intentions underlying them implicit*". The approach proposed is called goal operationalisation, in which leaf goals are mapped to operations. Each goal is assigned to a specific agent, whose responsibility is to execute the respective operations under specific conditions in order to achieve such goals. Goal operationalisation is discussed in [90].

Goal oriented requirements engineering techniques such as the i* Framework and KAOS are technically enabling when modelling complex scenarios. The author believes that the complexity in visual representations (even in the simplest forms) produces barriers for successful communication with non-technical stakeholders, thus lessening the effectiveness of any participatory exercise within a non mission-critical e-government context. Alexander, Robertson and Maiden [9] have studied the main factors that influence the selection and adoption of requirements development processes in industry. Their findings corroborate the author's experience whereby practices in industry vary considerably from the "*elaborate techniques*" discussed in academic literature. Furthermore, the authors have also noticed that the government sector suffers from a "*combination of apparently rigid procedures, low education and limited use of available process knowledge*". Based on these observations, the author believes that unless guided by capable analysts who are also good communicators, government stakeholders should be handled with care – using a simple to follow yet rigorous and controlled requirements development process that adopts intuitive, familiar and possibly non-technical notations, jargon and schematics.

van Lamsweerde [88] argues that "*although goal-based reasoning is highly appropriate for requirements engineering, goals are sometimes hard to elicit. Stakeholders may have difficulties expressing them in abstracto*". Through the groundwork of Jacobson, *scenarios* were first utilised in human-computer interaction research in the late 1990's. This has put additional focus on the area. Ivar Jacobson used the term *use case* instead of *type scenario*. The derivation of the term *use case* started much before the 90's.

Through his work in Ericsson's AXE system (1967) Jacobson started using scenarios in an informal and sometimes *"coarse-grained"* manner. There was no formal structure for the usage of use cases or type scenarios; but they simply indicated what the system's workings were suppose to be.

From a historical perspective, Cockburn [32] outlines the evolution of the term *"Use Case"*:

1. *Anvendningsfall* (Swedish for situation of usage)
2. Usage Case (Did not sound good in publications)
3. Use Case (as used today)

There are a number of benefits or advantages pertaining to RE based on scenarios, and Glinz [63] identifies the following:

1. *Taking a user's viewpoint*: This can help validate requirements in terms of adequacy, as they give users a feel of what they will eventually get, in contrast to the classical and more formal techniques such as entity-relationship diagrams or class diagrams, including UML.
2. *Partial specifications*: Scenarios can provide a decomposition of the system into functions (user-system interaction representing a transaction) through the user's perspective, whereas such functions can be treated separately.
3. *Ease of understanding*: Avoid the problems pertinent to pure narrative specifications techniques and formal methods, and allows for requirements to be elicited as natural language specifications. Glinz denotes that the use of user-system interaction descriptions is a natural way for understanding and discussing (eventually refining) requirements [63].
4. *Short feedback cycles*: Treating each user function, which has been elicited through user consultancy, separately will allow for better understanding and faster feedback between the requirements engineers and the users.
5. *Basis for system test*: Having scenarios which define interaction sequences can serve as a good starting point for system test planning; verifiable test cases can be derived from scenarios.

Jarke and Kurki-Suonio [80] note that scenarios are informal and inexpensive tools that may also provide a "*middle-ground abstraction*" to encourage team-members from different disciplinary backgrounds to participate in the requirements development process (e.g., offering common grounds for HCI practitioners and software engineers) [85]. Furthermore, Maiden and Robertson [100] suggest that scenarios as well as use cases are effective tools for the elicitation of stakeholder requirements. The informal nature of scenarios tend to "*suspended commitment*" and thus can be used to encourage experimentation while fuelling innovation [85]. During system design, especially in the development stages, scenarios can serve as a handy reference point against which the development team can derive a sequence of events and exceptions which the system or sub-system or a particular function or sub-function has to abide with. From a methodological perspective, Alistair Sutcliffe [161] argues that scenario-based

design "*is the closest that HCI comes to a systematic method*", which unlike methods from software engineering and requirements engineering, is mainly used to support the thought process rather than to offer "*step-by-step guidance*" [161]. From personal experience it is felt that many stakeholders in the design of new e-government services are non-technical and can be considered as novices with respect to requirements engineering techniques and their deliverables, especially when specialised diagramming notation is used. This view is corroborated by Faily [52] and Hamilton, Pavan and McHale [68].

RESCUE [85] is a scenario based requirements development process that integrates various requirements elicitation, modelling and management techniques (and best practices) specifically built to specify requirements for complex socio-technical systems. RESCUE has been built specifically for the air traffic control domain, and has been successfully adopted in multiple projects (e.g., DMAN [86] and CORA-2 [85]). The authors argue that traditionally requirement specification techniques have evolved from distinct disciplines tackling domain-specific challenges (e.g., task-analysis from HCI and use cases from software engineering) however when it comes to safety-critical socio-technical systems (e.g., air traffic control) hybrid-processes are needed. RESCUE adopts this approach, combining different techniques to offset one's deficiencies with another's strengths [85]. This is obtained by integrating techniques and best-practices from different disciplines that handle different aspects of complex systems (e.g., realtime systems, ergonomics and human-computer interaction) while merging them into a cohesive process of discovery. RESCUE adopts techniques such as human activity modelling, i* models for system goal modelling, use cases, scenario walkthroughs and *Volere*-inspired activities (i.e., Quality Gateway) and templates (i.e., requirements snow card) for requirements specification, verification and management.

Considering the domain tackled by this thesis (i.e., non-critical public facing e-government services) it was decided to investigate James and Suzanne Robertson's work on *Volere* [138]. Here the authors present a requirements development and management framework that proposes simple constructs and techniques while maintaining and encouraging process rigour. On the other hand RESCUE adopts a significant level of formalism especially in the way requirements are elicited and modelled – in line with its application on safety-critical systems. Simplicity is important in the e-service development domain, however this should not be obtained at the expense of rigour. *Volere* provides this balance, and this has been shown through its extensive use in both the public and private sector [98]. RESCUE may be well suited for safety-critical e-government systems, however this discussion is beyond the scope of this thesis.

Volere adopts scenarios as well as actors (i.e., active and interested), goals and testable requirements through measurable fit-criteria [99]. It is designed in a way that allows analysts to have a full view of what led to a specific requirement, the rationale behind it, related events and business use cases. A change in a specific requirement can be assessed and validated against the related business events, business use cases and product use case(s) – and vice-versa. The *Volere* process does not enforce strict rules on the system analyst however it proposes a templated set of deliverables (and milestones) together with corresponding methods while allowing the use of different modelling techniques (e.g., UML). Diagramming techniques are left up to the team's discretion – who can adopt formal techniques such as UML models, BPMN and

flowcharts as well as informal techniques such as rich pictures - as long as the techniques adopted are deemed suitable for the task at hand and understood by all of the primary stakeholders. The process is adaptable to three levels of agility: (1) projects which require a high level of agility ("*Rabbit projects*"), (2) projects which need as much agility as possible but are constrained by project or organisational circumstances ("*Horse projects*") and finally (3) projects involving a large number of stakeholders and requiring formal documentation due to legal or contractual obligations ("*Elephant projects*") [138].

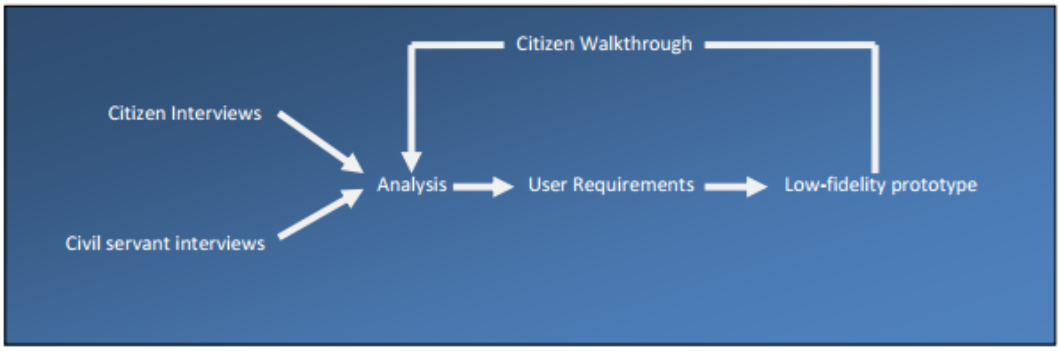
2.2.2 RE frameworks for usable e-government

In [43] Donzelli and Bresciani present *REF* (Requirement Engineering Framework) a goal oriented requirements development methodology that was adopted within the context of an e-government project. In *REF* organisations are modelled as actors, creating an organisational model through a network of actors in which collaboration and conflicts occur. On the other hand, goals define the relationship between such actors. The authors acknowledge that in e-government scenarios there are "*very diverse stakeholders, with very different skills and backgrounds*". REF is aimed at helping the designer discover, define, refine and reconcile requirements and goals, which are split into soft and hard goals. The introduction of H-Connections ('Hurting') in the modelling tool allows the designer to identify possible conflicting stakeholder points of view, and therefore allow for reconciliation efforts to be carried out at this stage. The main aim is to "*evolve the analyses only along the most promising alternatives*". An example of an H-Connection is given in [43], between the Employee's soft-goal 'Protect-My-Privacy' and the Head of Unit's goal 'Provides-employee's-Number-of-Documents'. In collaboration with the relevant stakeholders this can help the analyst focus on goal refinement through which potential solutions may be found (e.g., specify an average number of documents rather than an explicit value). Before taking a decision, all possible solutions are discussed with the stakeholders involved in this conflict (although further refinement iterations may still be required). Based on the *i* framework* and given the graphical tools provided, the issue of scalability still stands, and thus it makes it difficult to abstract and analyse a portion of the whole scenario at a time.

Seltsikas and Papas [149] propose a RE approach for trans-national government information systems (TN-GIS) that "*is particularly sensitive to political pressures and intent, and an approach that can endure extremely extended development timescales*" [149]. The authors pose a simple and yet a very important question; are traditional RE theories suitable within a TN-GIS context? They acknowledge that there might "*be hundreds and thousands of potential [unrelated] users from twenty-five or more countries*" [149]. Action research together with participant observation were used in their highly qualitative methodology. The *Volere* requirements development process was applied to gather and synthesise requirements. However the authors argue that *Volere* does not "*provide much detail how requirements can be elicited from stakeholders*" [149].

van Velsen et al. [166] introduce a citizen-centric RE process for e-government projects and at the stakeholder identification stage the model is divided into two major groups: 'Citizens' and 'Civil Servants' as show in Figure 2.1.

Figure 2.1: Citizen-centric RE for e-government projects [166]



Iterative prototyping is suggested in this approach encompassing citizen walkthroughs which will elicit feedback and scope for user requirement redesign. Scenarios are used in order to facilitate discussion and feedback between the design team and users. Personas are also introduced in this process in order to cater for the heterogeneous user groups which will be using e-services. Representatives for each persona are invited to participate in the walkthroughs. At the same time the model assumes that there are two major groups of stakeholders: citizens and civil servants. In reality, a large number of heterogeneous citizen groups exist, each with different needs and goals.

The authors reflect upon a number of issues with RE techniques, especially those applied in e-government scenarios:

1. Experts' interpretation may vary between one and another
2. Specialists with technical knowledge generally take the "*upper hand during the process of designing the systems*" [166]. This might result in systems which do not have a good fit with its highly heterogeneous users, citizens.
3. Since user groups are highly heterogeneous, the risk of neglecting sub-groups exists.
4. A statistically balanced set of representative users should be involved in the design process in order to have the widest view possible.
5. User requirement elicitation is a lengthy process and this needs to be accounted for in the planning process.
6. Technical development should not start before the requirements elicitation process is exhausted. Doing otherwise might result in re-design activities, if at all possible, or in expensive workarounds if the system has already been implemented up to a certain point.
7. It is hard to understand the return on investment of user-centric services. The only possible aim is to develop systems which are not 'disliked' or 'under-used'.

Rogers et al. [139] argue that most of the literature in RE emphasises the importance of identifying stakeholders, nonetheless "*none describes a model or a concrete approach for identifying stakeholders for a specific project*" [152]. Quality of knowledge acquisition for systems development highly depends

on the quality of the stakeholder identification process. *Volere*, a practitioner-centric requirements process [138] adopts Ian Alexander's Onion Model [7] for stakeholder discovery (see Section 5.5). This approach focuses on the project's stakeholder sociology which is in turn modelled using a spatial metaphor [7] – a series of concentric circles surrounding the system being developed. Each concentric circle acts as a placeholder for various stakeholder roles (e.g., the wider environment can contain regulators, developers, political beneficiaries, negative stakeholders and so forth). Stakeholders (i.e., humans, entities or other systems) are organised in terms of their distance from the system being developed, determined by the level of influence they have on the product and vice versa. The segregation of stakeholders into roles was also suggested by Sharp, Finklestein and Galal in [152].

The author then focused on system design processes and considered a set of collaborative workflows that would also make UX consideration an integral part of the primary design task-set. Janowski et al. [79] conclude that the lack of methodologies and models, as well as the lack of cohesion between related projects are two of the major causes of e-government project failure. Measuring success by the actual delivery of products may not necessarily mean that the overall e-government transformational strategy is being met (i.e., *project management* vs *programme management*). A cohesive methodology provides a standardised approach across projects affording cumulative knowledge across projects for use in future projects. Whitten and Bentley (cited in [6]) also suggest that methodologies should store knowledge across projects in order to avoid cold start issues in future projects, especially when development teams change. Based on this it can be argued that a requirements elaboration and system design methodology should not just help the team to plan, manage, execute and control the development of a system, but should also provide mechanisms to monitor and flag any decisions that may have short and long-term implications on UX.

Rahaman and Sasse [131], Beautemant et al. [12] as well as Porter et al. [127] have presented frameworks and techniques to assess the impact of design decisions on ULX, tackling the psychological impact, compliance to imposed security, perceived workload and task completion respectively. Software development methodologies are there to guide the development team to work efficiently, however products will finally be used by different categories of end users and it is only logical to conclude that such methodologies should include user-centric techniques, tools and interim deliverables. Personas as well as user stories are often adopted to inform the decision making process. For instance, Seffah et al. [82] presented *UX-P – User Experiences to Pattern* – a user-centred design framework based on the use of personas, associated users' experiences and HCI (mainly UI) patterns (referred to as building blocks in [81]) to create pattern-oriented designs at different levels of abstractions (i.e., site navigation to screen layout). The *Persona to Pattern (P2P) Mapper* tool is used to help developers pick suitable building blocks (e.g., UI components) based on a scoring system that considers the respective users' experiences (represented within the persona specification) as well as the context of use. Through this technique and associated tools, the authors have proposed a way by which design, development and usability considerations are streamlined. A case study in which *UX-P* was adopted is presented in [81]. *UX-P* proposes the adoption of user experiences to inform the selection of UI building blocks to generate an early con-

ceptual design, but implicitly assumes that such designs will always result in a positive user experience – during and beyond the interaction time and space (i.e., UX and ULX, as defined in this thesis).

Agile methodologies are incorporating user stories (i.e., high-level statements about specific aspects of a system) to tackle specific UX issues as part of the design process, and on this matter Dimmick [42] states that this can lead to short-sighted design decisions aiming to fix immediate UX issues in specific user stories within the bounds of a specific sprint[1] while risking the neglect of larger design questions.

In an e-government context, the bigger picture UX decisions are as important as a design decision at the micro-level. According to Dimmick, certain big design issues "*don't fit neatly into an existing user story or an individual sprint*" and so he proposes *time-boxed design spikes* as part of an agile framework (SCRUM) allowing designers to focus on complex UX issues at specific *time bubbles* in-between sprints [42]. Design spikes may be used by designers to tackle over-arching design questions that may invalidate, or back-up, the whole development effort. Nonetheless, design spikes depend on the application of user-centred techniques and tools – inheriting their shortcomings (see Section 2.2.4). Unfortunately although personas are considered to be the de facto tool for user-centric design they may not satisfy the bigger picture in terms of designing for interaction in long-lived systems. Grocki and Thomson [66] consider personas to be snapshots of target users' traits, intentions, needs and behaviours and as rich as they can be they do not follow changes in target users' real needs and traits over time. Storyboards, task flows and scenarios are traditional tools used to show time progression, however these tell short-term stories of single or multiple interactions over a short period of time (i.e., systems of transactions). The authors advocate for what they term as "*illustrating the full story of engagement*" (i.e., systems of engagement and personal fulfilment) spanning across longer time-frames and multiple channels of interaction – journey modelling. This thesis acknowledges this view and uses it to support the selection of an established yet appropriate requirements development process (see Section 5.3).

Van Velsen, et al. [166] provide a list of considerations for the elicitation of user requirements in e-government projects:

1. **Heterogeneous user groups**: in terms of cultures, skills, political opinions and disabilities. Different demographics are also a crucial element.

2. **Incidental use**: most e-services are rarely used or used sporadically, and thus proper guidance should be given throughout the process.

3. **Complicated content**: especially since users come from heterogeneous backgrounds and situations. This might make it difficult to accommodate all instances for e-service provision. A particular design decision may apply for one specific group but not for another due to inherent differences, such as technological skills amongst others.

4. **Interoperability**: an e-service should be provided as a "*coherent and logical whole for the user*", irrespective of where data is coming from and the number of departments sharing or providing

[1] "*A short (ideally two to four week) period in which the development team implements and delivers a discrete product increment, e.g. a working milestone version.*" – Source: https://confluence.atlassian.com/display/AGILE/Sprint, (accessed February 2015)

information for the successful provision of the e-service.

5. **No competition**: designers are generally tempted to give less attention to HCI issues, such as user-friendliness and service attractiveness. This may affect the acceptance of the e-service and lowers usability. It is important to add that the way a policy is implemented (even its attractiveness) will reflect on the policy itself as well as on the policy maker [75].

6. **Return on Investment**: is very difficult to assess in e-government projects, and the authors suggest that ROI "*should be assessed using subjective user-satisfaction criteria*". These ROI evaluation criteria might be defined earlier on during the requirements elicitation phase. These criteria are also used in the approach suggested in [166].

When designing large systems such as national or transnational governmental systems, one may find that certain traditional RE techniques may not suffice due to numerous additional factors. Seltsikas and Papas [149] outline a number of issues with respect to the identification of user requirements in 'Trans-National Government Information Systems' or TN-GIS, including (1) the involvement of hundreds or thousands of stakeholders (across various stakeholder types), (2) higher political influence and pressures, (3) higher difficulty for consensus building, (4) extended development time-scales and (5) the need for a mixture of traditional requirements elicitation methods and user centred design techniques.

National systems carry with them a higher level of complexity, and therefore traditional RE approaches may display a certain amount of inadequacy, such as incomplete requirement coverage, myopic design, context unaware systems and so on. Lapouchnian [93] argues that the more traditional RE approaches[2] focus too much on a system as solely a software component, modelling data and processes, giving minimal importance to the underlying system purpose and domain (or environmental) related concerns, views and issues. Cost overruns, delivery delays, incomplete systems and system failures traceable back to defective specifications are just a few of the problems which might be encountered [92]. Lapouchnian [93] also denotes that non-functional requirements are generally omitted from the original requirements documents.

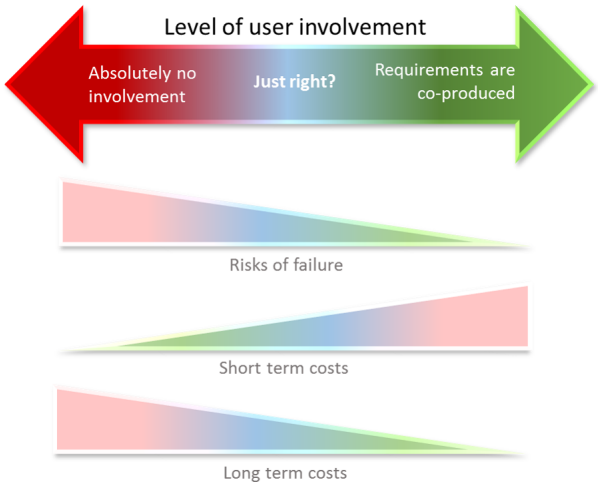
2.2.3 User involvement in RE

During the design and development of an e-service user feedback is unquestionably essential and several techniques such as brainstorming, interviews, workshops and scenarios may provide the project team with essential information – directly from the primary source [151]. Seyff and Maiden [151] argue that such feedback is especially relevant in the earlier activities, in particular during the requirements development process, increasing the chances of formulating better requirements. However obtaining end-user feedback can be slow and expensive. End users are generally involved at specific points throughout the project, at least at the requirements stage and at the user acceptance stage, however there is a large gap in-between in which design decisions can go critically wrong (e.g., decision to adopt the highest level of assurance in the enrolment process without considering the users' expectations, context of use and capabilities). A best case scenario would be to have a representative group of users present throughout

[2] Structured methods such as Structured Analysis and Design Technique (SADT) and Data Flow Diagrams (DFDs)

the design process, however for public facing e-government projects this may be overwhelming, unmanageable and expensive. Figure 2.2 represents this challenge through three contrasting effects. The cost of the requirements exercise would be relatively low if users are not involved, however this introduces a considerable risk of failure (e.g., users decide not to use the product, resulting in costly corrective measures). If these risks materialise the project may incur high correctional costs especially since late corrections may cost up to 200 times more than if these are carried out at the requirements stage [16].

Figure 2.2: Finding the right balance for user involvement during the requirements development process – building on Boehm and Papaccio's theory on the cost of late corrections [16]



Several techniques are proposed to improve user involvement. *iRequire* [150] adopts mobile technologies to allow end users to document requirements themselves in situ via an application installed on their mobile device. The authors assimilate the technique to an HCI method called *cultural probes* whereby users self-report their activities together with contextual information (gathered automatically – through sensors – or manually) without the direct involvement and influence of a requirements analyst. Nonetheless this technique targets the earliest stages of the process whereby information sent in by users is treated as an informal and incomplete set of requirements which sets the foundations for "*consecutive refinements activities*" [151]. Self-reporting may be especially effective for cognitive activities which are in themselves difficult to observe. ART-SCENE [177] is another user-centric technique which involves end users directly within the requirements development process. It adopts a systematic technique called *scenario walkthroughs* whereby normal-course and alternative scenarios (generated via the tool and based on ART-SCENE specific use case templates) are used as aides to help in the discovery of requirements. Supported by web-based tools this technique is grounded on the idea that people are "*better at identifying errors of commission rather than omission*" whereby participant ideas, thoughts and feedback are prompted through textual and visual cues [178]. This technique has been adopted for systems related to air-traffic management.

To mitigate risks arising from bad decision making due to limited access to end users during the requirements development process (e.g., due to logistical or financial reasons), alternative user-centric tools and techniques are adopted (e.g., personas). These tools however have their own shortcoming

especially if used improperly (see Section 2.2.4). User walkthroughs and focus groups can generate interesting insights during the design process, however care is required to balance out their cost with the benefits obtained (number of people to involve and timing/frequency of sessions).

2.2.4 User-centred design techniques

Several user centric software design techniques have been devised and used in e-government projects over the years, including personas, user walkthroughs, use cases, scenarios, wire-framing and low/hi-fidelity prototyping. Furthermore, Buie and Murray state that governments also use tools such as focus groups, surveys, server log analytics and regulatory checklists to align themselves towards usable designs [24]. van Velsen et al. [166] present a list of citizen-centric RE activities for the development of e-government scenarios, including studies on life-events, interviews with experts, surveys, focus groups and think aloud sessions.

Inviting participants to a focus group is extremely useful, however there's always the risk of getting outliers who can skew the focus onto trivial issues. Personas are used to encourage the design team to focus their attention on the end user throughout the design stage, mitigating the risk of building systems that appeal to the developers' own interests [55]. Cooper [36] defines personas as archetypal representations of specific user groups, each bringing onto the drawing board more scope for discussion, and a deeper understanding of what the persona might want, like and dislike. This representation is built on empirical evidence rather than on mere speculation and stereotyping (*archetypes* vs *stereotypes*). Cooper distinguishes personas from user profiles ("*a brief biographical sketch*") and market segments (based on demographic, distribution channels and purchasing behaviour). By contrast personas are based on ethnographic data and focus around behaviour and motivations (goals). Without understanding and encapsulating motivations and goals, personas may still serve as an effective communication tool within the design team however it will not contribute as an effective requirements development and design tool [38]. Faily and Fléchais argue that personas "*do not look like specifications*" and developers might not relate to aspects such as names, jobs, goals and feelings [55].

Personas are meant to facilitate group discussion while directing the designers' energy towards a unified direction. Various requirements elaboration and design techniques are proposed in literature and many of these methods suggest the use of personas as a technique to understand the eventual user and inform design decisions [138, 54, 53, 55, 123, 117, 166]. Auto-makers Ford use personas in the design of new car models and in an article in The New York Times[3] Phil Patten states that the company first designs the driver before designing the car. In fact Ford propose various driver profiles, including details about their social lives based on demographic research. Design teams at Ford believe that personas help get everyone on the same page, while providing a common focus "*from the clay modeller to the chief executive*". This statement was made by Murat Yalman, Ford's director of global advanced product strategy. Yalman encourages the representation or personalisation of the car driver so that everyone understands who they're working for, creating "*very memorable ideas that live on after the meeting or presentation*".

[3]Phil Patten, The New York Times, http://www.nytimes.com/2009/07/19/automobiles/19design.html?pagewanted=all, (accessed September 2013)

Personas were initially referred to as CustomerPrints at OgilvyOne and were used to help build market segmentation strategies. Alan Cooper [34] first introduced the term persona in the software field in his efforts to make software more "*human and forgiving*". According to Cooper personas are defined by goals and these are generally discovered as a by-product of a rigorous investigation process of the problem domain based on successive refinements. Dotan et al. [44] state that personas make assumptions about end users more explicit, however the authors argue that a persona cannot provide information on critical aspects such as system usability and usefulness. For this reason, other user-centred technique, such as user testing, are still required [44].

To build effective personas a rigorous process of discovery is required, collecting behavioural data across multiple sources and adopting various techniques, including in-context interviews with and observation of potential users, demographic research, historical information such as actual user help-desk requests or complaints and so forth. Adlin and Pruitt [2] take this further and propose the *persona lifecycle*, a metaphoric and cyclical framework that assimilates persona-related development processes with the human lifecycle. The authors suggest five phases: (1) *family planning* (why is this persona required and what data sources are available?), (2) *conception and gestation* (systematically turning data and assumptions into personas), (3) *birth and maturation* (personas are introduced in the organisation), (4) *adulthood* (adoption of personas at several stages of a project, including system design, development, evaluation and post-launch) and (5) *lifetime achievement and retirement* (the point where persona-related efforts are measured and plans are put in place for its re-use, re-incarnation or retirement). Adlin and Pruitt suggest a number of reason as to why persona efforts fail in a practical setting, and these include (1) low management acceptance and support, (2) poor communication of personas during projects and (3) lack of understanding of the technique by the product design and development team [2]. In their book [2], Adlin and Pruitt provide detailed guidelines on how to build and use personas – making this an important resource for both HCI researchers and (especially for) practitioners. Their experience-based, data-driven and methodical approach fills the gap in HCI literature on the proper use of this user centred design tool. The authors continue to argue that personas (and associated processes) should be integrated into existing development processes and tools. This may in turn encourage adoption and improve persona-based communication.

Personas also bring forth scope for qualitative discourse, and it is up to the designers to accept or reject statements and assumptions brought forward throughout the design process. Arguments based on personas are subjective as much as the persona itself and its validity can be threatened [52]. Developers may also refute aspects of a persona to argue against a specific product feature. This weakens the persona's legitimacy especially if other characteristics are called into question [55]. Other risks exist; what if the designers come up with a stereotypical persona which is not representative of actual users? Ethnographic research might be useful at this point however it might still not validate personas since "*specific examples of data cannot be provided to prove their accuracy*" [55]. Faily and Fléchais [55] present Persona Cases, a technique that attempt to legitimise the validity of personas by grounding their characteristics in, while making them traceable to the originating source of empirical [qualitative]

data. This technique adopts Grounded Theory as its underlying development method and frames persona characteristics as arguments (see Table 2.1) following Toulmin's Model of Argumentation [164] (cited in [55]). Persona Cases call for a rigorous analytical process whereby thematic concepts are explored, inter-related (after being analysed), and linked to their respective primary source of evidence (refer to Section 3.1.2). Faily and Fléchais believe that Grounded Theory artefacts (i.e., evidence) are helpful when questions about a persona's characteristics are raised during design meetings [55]. Persona Cases are based on argumentative techniques, adopting terms such as *claims* (persona characteristics representing any of the following behavioural variables; activities, attitudes, aptitudes, motivations and skills), *grounds* (evidence for these claims), *warrants* (describing how grounds contribute to claims) and their *backing* (origin of warrant's assumption). Assumption may also act as *rebuttals* (challenging a claim's validity) whereby a *modal qualifier* determines the degree of certainty for a given claim.

The process for creating Persona Cases involves three steps [55]: the first step (summarise propositions) involves the identification of salient themes (thematic concepts) based on how grounded (in empirical data) and networked (with other themes) these themes are. Propositions are based on quotations underlying each theme in the resulting GT model. The second step is to argue characteristics, describe claims, justify their existence with propositions (acting as the claims' grounds/warrant) and specify modal qualifiers (based on the analyst's confidence on the relationship between claims and propositions). Propositions that rebut a claim are also listed. These may then be used while debating a persona's characteristic. Finally persona narratives based on elicited characteristics are written for each behavioural variable type. This links such narratives back to the Grounded Theory model as well as supporting artefacts. This evidence provides further validation to the persona, however care must be taken not to build "*elastic*" personas as some characteristics may be irrelevant to the current context of analysis or are poorly grounded – possibly encouraging team members to argue in favour of, or against any design decision using vague or irrelevant attributes. Additional personas might be required to reflect newly discovered characteristics elicited in previous steps. The authors suggest the use of affinity diagrams to organise characteristics into "*natural groupings*" that form the basis for new personas [55].

Table 2.1: A persona characteristic argument example (adapted from [55])

Relationship	Information Security Indifference is a cause of Island Mentality
Characteristics	SCADA isolation make hacking unlikely
Behaviour Variable Type	Attitude
Grounds	II-5, IM-3 [*links to evidence*]
Warrant	IM-4 [*links to evidence*]
Backing	Island Mindset concept
Qualifier	Probably
Rebuttal	None

Persona Cases provide solid grounds for discussion and debate, nonetheless this design technique is subject to issues that may inhibit adoption by practitioners (including developers and policy makers). The first issue is related to the level of specialist knowledge required to build Persona Cases (i.e., knowledge of Grounded Theory and thematic coding). Having a team member with such knowledge on board is a rare occurrence in industry especially within small project teams. From personal experience, teams for

public facing e-government services are generally small in size (e.g., up to five members for medium sized projects, which may consist of an analyst, a lead developer, a domain expert(s), product owner and a representative from the government's IT agency). Furthermore, such teams would be operating on a tight budget and within short time-frames, which limitations may be tightened even further due to political pressures. Unless collaboration agreements with research entities are established, adoption of this technique, or indeed of any technique requiring specialist knowledge, may be threatened by issues of practicality. From the perspective of developers and product owners, spending time reasoning about personas may not *feel* as productive as sketching prototypes.

Productivity should be central to any requirements, design and development methodology, tool or technique, and any proposed user-centric technique for use in industry must itself be practitioner-centric, considering the practitioners' limitations and impact on their lived experience (ULX). Any proposed activity should contribute towards reaching the desired goals and anything that is perceived as breaking the flow (or budget) may be considered a hurdle, and potentially circumvented (refer to Section 2.4 for a review on current practices in e-service procurement and development). The creation of personas requires "*significant investment of time and resources*" [2] in order to benefit from their full value.

The Calibrated Persona technique was introduced by the author in [127] (see Chapter 4) as a way to support project teams make informed design decisions on critical aspects of an e-service (e.g., enrolment) at the earliest stages of a system's lifecycle, possibly before HCI specialists are introduced to the project. Re-usable behavioural models (for specific groups of users) are associated with traditional project personas making it possible for the team to predict user reactions towards different design decisions and for each groups of end users. This does not rely on the development of complete personas, whereby behavioural models (which may have been generated for use in previous projects) can be reused and associated with persona skeletons sharing similar characteristics with participants who were involved in the model generation process (i.e., user group calibration). Persona skeletons can then be further developed at a later stage in order to assist in low-level design decision making (internally or externally in case of contracted projects). Chapter 5 presents *Sentire*, a requirements framework that extends an industry-standard requirements process while adopting Calibrated Personas as part of its central workflow.

2.3 Enrolment Processes and User Behaviour

Enrolment refers to the set of actions a user needs to perform before being granted access to an e-service. This involves the verification of the user's identity as well as the generation of credentials (and shared secrets) to be used for authentication purposes in future interactions. This generally requires a user to provide the necessary proof of identity. The level of proof required by a service provider (identity assurance) depends on the risk levels involved (see Table 2.2). In low assurance level scenarios users may enrol by simply providing a valid email address, however when assurance level requirements increase more proof of identity may be required. In high assurance-level scenarios users may also be required to visit an enrolment centre in person while presenting other forms of identity (e.g., driver's license and biometric data). *Enrolment process* will be used as an umbrella term to represent other commonly used

terms, including *signing up* and *registration process*.

Establishing a proportional level of identity assurance for any given e-service can be a challenging task, wherein care is needed not to introduce overzealous levels of security that may have a negative impact on user adoption. Imposing identity assurance level requirements for an e-service without proper consideration for the users' experience may come at a cost and evidence for this is presented in Section 2.3.1. Nonetheless, risk levels pertaining to an e-service should also be respected and there are circumstances where a high level of identity assurance would be required. Finding the right identity mechanism to deliver such assurance, while keeping in line with user capabilities, perceptions, expectations as well as limitations and opportunities afforded by the context, is a non-trivial problem.

Beautement et al. [12] introduce the *compliance budget* as an organisational technique to understand, manage and potentially influence user behaviour when introducing security policies within an organisation. While studying employee security behaviour at two major commercial organisation the authors noticed that the employees' perception of costs and benefits of complying with organisational security requirements can have a direct influence on their behaviour – which can ultimately have a major impact on the effectiveness of the company's overall security goals. There is a limit on the effort individuals are ready to make (perceived costs) to comply with organisational security requirements and the higher the perceived benefits are, the better the chances are that a user would comply. Would the cost of installing a security patch – requiring employees to terminate all open sessions and rebooting their computer at noon – justify the organisation's security goals for a safer working environment? Would the requirement of a daily virus scan be justifiable if this has a direct impact on productivity (e.g., slows down machine)? If the amount of perceived effort required from the user is higher than the individual's level of perceived benefits (threshold) there is a higher risk that the user will not comply with the required security measure(s) – unless compulsion exists. This can also result in disruption, frustration and ultimately in failure to complete the task at hand. Beautement et al. [12] define friction as the imbalance between the business process (user goals) and security behaviour required, including any inherent cognitive and physical workload. Friction can cause resentment, and ultimately non-compliance, which can lead to service abandonment. The author believes that although the compliance budget has been introduced as an organisational technique this term can also be used at an individual-user level – describing a users' level of security workload she is ready to endure while performing a primary task. This level may differ depending on the context of use and the associated level of perceived benefits.

The four Levels of Identity Assurance (LoIA) presented in Table 2.2 are not prescriptive in terms of specific identity verification techniques to adopt (at each level), however the project team is expected to produce an enrolment process that (1) respects the LoIA recommended by policy makers (assuming it is in line with the e-service's risk levels) and (2) minimises the level of friction (perceived workload and benefits obtained). The author is not aware of any systematic and analytic process that could assist the project team design acceptable enrolment processes that respect these two basic goals. For instance, an enrolment process for a *LoIA 2* e-service may take several forms, however, while still complying with *LoIA 2* requirements, one process may be more disruptive than the other. Requiring an identifying

document that is not generally readily available (e.g., a birth certificate) would result in higher levels of perceived workload (i.e., user has to abandon the primary task to obtain a birth certificate before being able to resume). Would a driver's license provide the same level of identity assurance in this case? Are users ready to go through this hassle to gain access to this particular e-service? Disruption may also be caused by delays – are users ready to wait three days before receiving an activation code by post? Could alternative out-of-band techniques (e.g., SMS or voice) be used to reduce this delay? Can account verification be deferred to a later stage, following the completion of the primary task?

Table 2.2: Levels of Identity Assurance (LoIA) compiled from recommendations published by the UK [65], US [61] and Canadian [121] governments

LoIA	Description	Enrolment example	Risk levels
1	Little or no confidence in the asserted identity. Service providers require an identifier to link different online sessions to the same user but without the need to know who the user actually is	E.g., requiring a valid email address and password	Compromise results in no or minimal harm
2	Some confidence is required in the asserted identity. This is to make sure that the person is eligible to interact with the service provider	E.g., requiring a copy of the driver's license (scanned) before an activation PIN is sent by post	Compromise causes minimal to moderate harm
3	High confidence in asserted identity (beyond reasonable doubt)	E.g., out-of-band verification of one's identity (in person) before generating a digital certificate for authentication or signing purposes	Compromise could cause moderate to serious harm
4	Highest level of confidence in asserted identity (LoIA 3 + Biometric data)	E.g., out-of-band verification of one's identity (in person) including the recording of biometric information (such as fingerprints) before provisioning a hard cryptographic token for authentication and signing purposes	Compromise could cause serious to catastrophic harm

User enrolment can have a serious impact on the success of online government services. Different services require different levels of identity assurance (see Table 2.2) and enrolment processes are put in place to deliver them. From the citizens' perspective these processes often require an amount of effort that exceeds the level of perceived risk (and benefits) associated with the respective e-service. This disproportionality causes friction and is one of the root causes for low e-service take-up. Sasse and Fléchais argue that in reality security mechanisms are chosen without considering the production task and its performance requirements [144]. The latter must inform the selection of security mechanisms, within the acceptable levels of identity assurance (see Table 2.2).

2.3.1 Evidence from the field

France achieved a 32% increase in the number of tax forms submitted online when it changed from digital certificate-based enrolment and authentication to e-mail and password [120]. This was contrasted to the previous year's zero growth in electronic service usage. According to OECD, the French government intends to "*give up the electronic certificate totally*", in favour of this simpler identity mechanism. Business tax-payers were initially introduced to the email/password authentication model in October 2010 (in parallel with the digital certificate), resulting in a 44% increment in business users within six months, with over 90% choosing the email/password mechanism over the digital certificate. No security issues have been reported [120].

In the early 2000, the then HM Customs and Excise agency in the UK had initially required businesses to purchase a digital certificate in order to be able to file their VAT returns online. This resulted in extremely low levels of active users filing their sales tax online [136]. This problem added pressure on HMRC's (HM Revenue and Customs) management whose original goal was to improve the effectiveness of their own business processes. The department cited Gartner's recommendations which stated that user ID and passwords are adequate for most applications for which identification and authentication is required [136]. Digital certificates were not supported anymore and the HM Revenue & Customs department introduced a username/password mechanism which is easier to use and cheaper to administer. This resulted in higher adoption rates of the services on offer [51]. By the 31st of January 2009, the HM Revenue and Customs agency received over 5.8 million online self-assessment (SA) returns [135], a 52% increase over the previous year, amounting to 69% of all SA returns filed. This result was preceded by the creation of the Carter Programme, upon Lord Carter's recommendations to increase takeup of online services [134].

Other countries, such as Austria have attempted to tackle the takeup problem by dishing out millions of smart cards (with an embedded digital certificate) to their citizens. However this might not have provided the desired results and this can be tied down to the fact that such smart cards, although useful from a government perspective, did not meet any of the citizens' immediate goals. Rahaman and Sasse [131] denote that primarily the lack of opportunity to make use of the identity was conducive to a lack of perceived benefits. On average, Austrian citizens interacted with public entities 1.7 times a year back in 2010 [5]. Furthermore, the low level of understanding in the technology used (e.g., Public Key Infrastructure, roaming servers and multiple identity providers) made the system even more difficult to comprehend and use, thus limiting adoption rates even further. The authors argue that despite the fact that such technologies are portrayed as human-centred, the solution developed was based on what was "*technically feasible, and convenient from an administrative point of view*" [131].

The evaluation of a system's impact on the user experience is generally done as part of a post-mortem exercise, and is rarely considered as part of the design process. Despite the Austrian government's efforts to make the solution as human-centred as possible, between 2005 and 2009 only 74,000 citizens activated their digital identities and signatures, that is, 0.9% of the population [103]. Citizens were given multiple options, and could have obtained the digital credentials from a number of providers

and activated them over various media, including their Bank ATM card and mobile phones.

Mike Just who served as Director of Innovation with the Government of Canada stated that the push towards the use of digital certificates (*Epass*) in Canada was driven by two main beliefs: (1) PKI provides a more secure alternative than passwords, and (2) its support for digital signatures was in line with the policy and legislative activities undertaken at the time (personal communication, December 4, 2012). Just explained that citizens were given the option to either enrol for a single *Epass* or for multiple pseudo-anonymous ones, one for each federal government service. Initially the government's focus was predominantly on services offered by the tax department (Canada Revenue Agency) and the human resources department (Human Resources Development Canada). Just states that over six million *Epasses* were issued, however he argues that it is not clear whether these were actually used for annual tax submissions. PKI solutions present issues such as mobility, restricting users to a single machine on which their private keys would be stored (introducing challenges when it comes to shared computers or multiple devices). Just explained that the Canadian government attempted to solve this problem by introducing a "roaming server", a central server that stores citizens' *Epass* credentials and private keys. Citizens would then be able to access their private keys using a username and password. This in turn had an impact on the system's credibility and on the strengths that are usually associated with a public key infrastructure. Furthermore, in order to use their *Epass* for signing purposes a Java applet had to be downloaded, and this presented a number of usability challenges (such as download time, the need to download and install the Java runtime, browser support and local security configurations). The contract for *Epass* expired in 2012 and this was replaced by two "*robust and more cost-effective service options*"[4]: *Sign-In Partner*, a federated identity mechanism using existing online banking credentials (bank-card numbers or username/password) and *GCKey* (Government of Canada user ID/password). The move towards these two options was to encourage citizens to adopt online services while reusing existent and regularly used credentials (in the case of *Sign-In Partners*). Finally, Mike Just remarks that too much weight was given to the PKI-versus-password debate and that in practice the security of different mechanisms is largely affected by other factors, mainly usability.

2.3.2 Evidence from literature

In an OECD report issued in 2012 [120] the authors emphasise the importance of finding a balance between security safeguards (e.g., enrolment, identification and authentication) and usability, making sure that "*safeguards put in place do not themselves become a barrier to take-up of the service*" [120]. The report outlines a number of issues that EU member states are facing with regards to identity mechanisms in e-services currently being implemented or planned. Three key findings from this report are of great relevance to this discussion [120]:

1. Digital certificates provide the highest level of assurance however revenue bodies are looking at other identity mechanisms due to "*their negative impact on the uptake of electronic services*".

2. Shared secrets, such as passwords, PINs are considered as "*perfectly adequate*" by several national administrations to assure identities for "*most – if not all – secure electronic services*".

[4]Government of Canada, http://www.servicecanada.gc.ca/eng/ei/employers/roeweb/faq_login.shtml, (accessed June 2013)

3. Some administrations are shifting from digital certificates to shared secrets.

In a three year project funded by the European Commission and led by Oxford University's OII (Oxford Internet Institute), a set of barriers to the effective expansion of e-government services were identified. Of great interest to this thesis, enrolment and authentication processes were regarded as potential barriers "*if they are too cumbersome, costly or insecure*" [155]. Ease of use as well as price (if digital certificates need to be purchased) are important considerations which can help to avoid exclusion of certain user groups from engaging in e-government transactions [50].

Electronic signatures are generally viewed as "*big obstacles*" for the acceptance of e-government services and as a resolution to this issue the project team at OII suggested that low-trust options should be provided wherever possible [49]. This implies that authentication and identification requirements should be minimised as much as possible (or removed altogether) as long as the most basic legal and practical requirements are met. As a parallel example, many major online commercial service providers allow their users to complete a primary task without the need to enrol and authenticate (e.g., guest checkout in e-commerce sites).

Arguably services which are more sensitive or those which carry more risk to the service provider and the end user require higher levels of identity assurance, and thus more strength would be required in the identification (during enrolment) and authentication processes. Nonetheless, this strength should be commensurate to the actual risks involved. For instance, authentication should not be enforced whenever a simpler identification process (such as providing one's vehicle's registration number), would suffice (e.g., to pay a city centre congestion charge). It is reasonable to assume that no one would "*fraudulently pay a parking fine*" [51] although this might present some risks for non-enrolment based authentication processes (e.g., bank's help-desks asking for latest transaction details in order to provide sensitive information over the phone). On the other hand, applying for a passport calls for stronger identification processes.

The report also states that trust level requirements need to be flexible enough to adapt to the current level of risk and user behavioural patterns. Low levels of security could be used at the outset (e.g., registration), thus encouraging take-up, and then through the application of smart systems, realtime adjustments are carried out depending on the risk-level of specific situations and on customer behavioural patterns. Such systems may upkeep a high level of security while reducing the amount of "*trust hurdles*" imposed on citizens [51]. As an example, Symantec introduced systems capable of learning user behaviour, and carrying out automated interventions to build profiles of normal user behaviour, detect fraudulent transactions and in turn prompt users automatically to authenticate using a specific level of trust to "*complete a suspicious transaction*" [162]. Other solution providers (such as Entrust) are working on risk-based authentication and fraud detection platforms [48]. In e-government, risk-based authentication could be adopted to auto-adjust depending on the type of transaction, its sensitivity and trust requirements. Furthermore, service providers should strive to offload the burden generated by identification and authentication processes off the citizen's shoulder and onto e-government service backends.

A major challenge in selecting e-government identification and authentication mechanisms is the frequency of use. Some services are used several times in a citizen's lifetime however other services might be used on an infrequent or one-off basis. In the latter case, expensive or cumbersome identification routines together with strong authentication mechanisms might discourage the citizen to enrol for and use the service while increasing the chances of reverting to more traditional channels (e.g., in-person or by post).

OECD [119] proposes an operational principle relating to authentication mechanisms' fitness for purpose in e-government projects. Depending on the risk involved there "*should be enough security to address risk in an acceptable fashion, but not be unreasonably burdensome to accomplish the electronic communication*" [119]. Users can be under pressure to complete a production task, and any additional burdens imposed by security mechanisms might tempt users to cut corners [144].

In its *Action Plan 2011-2015*[5] for "*the provision of a new generation of e-government services*", the European Commission did not consider the problem of disruption in takeup caused by enrolment and authentication mechanisms in any one of its four political priorities. This could hinder the achievement of the 2015 targets which state that 50% of citizens and 80% of businesses should already be transacting with government online. In Action 29 of the same plan, the commission aims to encourage the implementation of the "*once-only*" enrolment principle by sharing implementation experience across member states.

2.3.3 Lessons from the private sector

In 2008, Forrester reported that 23% of checkout abandonment occurred when customers were asked to register for an account. For this reason many online stores adopted a federated identity policy. *OpenID* and *OAuth* are enabling protocols behind identity federation. These are used by major players including Google, Yahoo, *PayPal* and *Microsoft Live* to ease the workload associated with enrolment and authentication on third party sites. Service providers (SPs) and identity providers (IdP) interact within an identity federation IdF). After a user selects an IdP from the SP's enrolment page, the SP redirects the user to the respective IdP's authentication page. Following successful authentication the IdP informs the SP that the user is legitimate while providing any additional information as requested by the SP (disclosed by the IdP following explicit user consent). This eliminates the need for users to re-key personal information across different online services, avoids the need to manage multiple accounts and improves data accuracy across service providers (i.e., only one account needs to be maintained (IdP account) and all relying parties (SPs) will get the latest updates, if requested).

In an interview[6] Jared Spool, founder and CEO of User Interface Engineering (www.uie.com), discussed the "$300 million button" and his experience with a large online retailer. He observed that after users fill up their shopping basket and prepare for checkout they are generally presented with two option: (1) *Sign-in* or (2) *Register*. Spool suggested that users are only interested in purchasing products

[5] European Commission, http://ec.europa.eu/digital-agenda/en/european-egovernment-action-plan-2011-2015, (accessed May 2012)

[6] Vikki Morgan, https://econsultancy.com/blog/9614-profile-ogilvy-s-rory-sutherland-on-human-behaviour-and-innovation, (accessed May 2012)

(primary task) and resent the fact that they have to register (secondary task) before being able to complete their task. To test this hypothesis, the *Register* button's text was changed to *Continue*. According to Spool, the $25bn online retailer experienced $300 million increase in sales the following year. This was the result of a small and seemingly inconsequential fix. Discussing this slight change in the workflow, Rory Sutherland emphasised the fact that "*we're too preoccupied with technical innovation and too little with what people actually want*".

In March 2012 Experian reported that around £1.02 billion worth of online transactions were abandoned in 2011 by British consumers. 44% of UK shoppers abandoned at least one shopping transaction in 2011. This was the direct result of frustration caused by the "*length and complexity*" of "*old and inefficient identity [verification] measures*". Large retailers, such as House of Fraser have realised that this has a direct impact on revenues. For this reason users are given the option to sign-up, explaining the benefits they would receive in doing so, however they are also given the option to checkout without having to go through the whole enrolment process.

In support of this argument, *The Guardian*[7] reported that after *The Times* (thetimes.co.uk) introduced compulsory registration for its online readers, only 25.6% of unregistered users actually created an account, and it has also reported that online traffic to the site has fallen significantly, from 15% to 4.16% in approximately one month after making registration compulsory.

In 2012 Wetransfer.com (an online file transfer service) dropped its compulsory enrolment process, allowing users to transfer files without having to create an account first[8]. At the time of writing, Wetransfer.com had over 15 million monthly active users transferring over 1.5 million files per day. The service provider reported positive feedback from its users stating that the new *no-account-needed policy* was well received with the understanding that enrolment is only necessary if advanced features are required at a later stage.

2.3.4 HCI-Sec and identity mechanisms

Authors such as Rahaman [131], Malheiros [101, 102], Sasse [141], Fléchais [58], Adams and Sasse [1], Beautameant et al. [12], Inglesant [75], Brosthoff [21], Cooper et al. [38], Rogers et al. [139], Buie and Murray [24] and Earthy et al. [47] are working to identify potential user experience issues in identity processes covering aspects such as privacy and willingness to disclose personal information, trust, security-induced friction, economics of security, accessibility, policy making, usability, interaction design, cultural contexts, mobility, content management, national infrastructures, legalities and others. Sasse, Steves, Krol and Chisnell [145] introduce the issue of *authentication fatigue* related to the amount of cognitive workload posed on users from the introduction of authentication processes in online services. These processes interfere, or rather increase the effort required to complete the primary task (a mid-task hurdle that takes the user out of the flow – making it more difficult to backtrack to the original task). Frank Stajano from the University of Cambridge is working to "*get rid of all passwords*" through project *Pico* [156]. In Bonneau et al. [17] the authors review a series of web authentication schemes in an effort

[7] Josh Halliday, http://www.guardian.co.uk/media/2010/jul/20/times-paywall-readership, (accessed May 2012)

[8] Wetransfer.com, Introducing the new WeTransfer, http://wetransfer.pressdoc.com/33459-introducing-the-new-wetransfer, (accessed July 2013)

to replace traditional text-password based schemes. They suggest an evaluation framework based on 25 usability, deployability and security benefits in an effort to find the ideal scheme.

Although all of these efforts are driving towards better identity processes a gap still exists on how these various studies, heuristics and techniques can be adopted by non-HCI practitioners in the field. Adoption of user-centred design techniques in e-government service projects is hindered by threats originating from two fronts: *processes used* and *people involved*.

2.3.5 Human factors in security – workload

Hart describes workload as "*the cost of accomplishing mission requirements for the human operator*" [69]. In aviation, human costs to maintain performance could be fatigue, stress and accidents amongst others. What are the human costs in e-government services? The first cost for citizens is related to not being able to use a service, which could either result in sanctions (e.g., fine for not paying a congestion charge on time [75]) or loss of opportunities (e.g., having to use otherwise productive time to visit a government department in person). However the risk is also on the service owner. If human cost for a service is such that e-services are not used, the government will also have to absorb costs for handling that particular transaction via traditional channels, not to mention the low level of resource maximisation due to under-used e-services. This can also have political consequences. A negative experience will generally reflect and negatively impact the government's image of efficiency and competence.

Cain [25] argues that different workload measurement techniques actually assess different aspects of workload and this heterogeneity of focus stems from the "*lack of an accepted definition of workload*". According to the author different people have different perspectives on the meaning of workload, including (1) the task demands imposed on the user, (2) the effort the user needs to make to satisfy such demands and (3) the consequences of attempting to meet such demands.

Sasse, Steves, Krol and Chisnell [145] adopted GOMS-KLM (Goals, Operators, Methods and Selection rules – Keystroke-level Modelling) to assess the workload imposed by authentication events in terms of the time taken for a user to complete them [29]. GOMS-KLM, introduced by Card, Moran and Newell [29], evaluates workload by decomposing tasks into a set of basic actions or steps, on which time measurements are taken. Actions could be both physical (e.g., pressing keys, pointing the mouse) as well as cognitive (e.g., mentally preparing for the task) [158]. This technique then helps researchers predict user performance for a set of tasks while estimating the time required for a skilled user to accomplish the task [158]. Time is a reasonably simple, yet effective metric to evaluate the cost associated with security mechanisms, and this uni-dimensional workload estimation technique makes KLM especially easy to conduct and "*flexible enough to be applied in practical design and evaluation situations*" [29]. A number of prototyping tools are also available for GOMS-KLM to support user-centred design activities. *CogTool*, an open-source user interface prototyping tool developed at Carnegie-Mellon University, assists researchers and practitioners by abstracting the underlying GOMS-KLM theory to help them make better design decisions while taking user effort into consideration [158]. KLM Form Analyser (*KLM-FA*) [87] is another tool that aims to help designers build efficient (or improve existing) web-forms by predicting form-filling execution times. This tool offers a degree of flexibility by allowing designers to switch

user profiles for testing purposes (e.g., age and typing skills) as well as vary KLM rules and parameters for more advanced analysis (e.g., typing speed and age related time-adjustment multipliers). GOMS-KLM is an important benchmarking technique that can help practitioners and researchers determine the best and worst case scenarios (in terms of user performance and effort) for a given task (e.g., online form filling and authentication) however its simplicity might deter from its potential to provide measurable information on aspects such as frustration and self-confidence which, from a ULX perspective are also important considerations for the design of better security mechanisms. The author believes that the time taken to fill in a form does not necessarily imply a negative user experience, especially if the benefits obtained from using the e-service offset the cost associated with accessing it. For instance, a tax return e-service requiring users to authenticate by selecting a digital certificate, submitting a one time password and filling in several other fields might still be worth the while for a professional who would otherwise need to regularly fill in and post paper-based forms (refer to the *Type of Service* modifier discussed in Section 4.1.1). Workload can affect users in different ways (and for different reasons) and this impact may also vary across contexts of use.

For this reason, the author turned his attention to NASA-TLX – a multi-dimensional and subjective workload assessment technique. While developing NASA-TLX, Hart and Staveland [70] examined ten workload related factors, retrieved from sixteen experiments. Six of these factors were then proposed as a multi-dimensional rating scale combining magnitude and source information "*to derive a sensitive and reliable estimate of workload*" [25]. This was accomplished after a series of statistical tasks, mainly to determine the sensitivity of each factor on workload. In NASA-TLX, both physical and mental workload are measured rather than just cognitive workload. Rubio et al. [140] surveyed a number of studies which adopted subjective workload (cognitive) rating techniques. The authors ranked NASA-TLX at the forefront of sensitivity to experimental changes in workload conditions. This is also confirmed in Garteur's 'Handbook of Mental Workload Measurement' [3]. Hill et al. [73] rated NASA-TLX as the most sensitive to workload changes, followed by MCH (Modified Cooper-Harper) and finally SWAT (Subjective Workload Assessment Technique). Quoted in [3] Byers stated that MCH was found to be worse than SWAT as a measure of subjective workload. NASA-TLX allows subjects to record data post-task, and thus certain physiological and timespan-dependent effects may be in conflict to what is recalled by the subject. Techniques to counteract this issue include (1) screen-recording playback and (2) video-recording playback of the tasks performed. These techniques are designed to facilitate retrospective workload rating [3].

NASA-TLX uses six workload factors or dimensions and measures their relative contribution in influencing the user's perceived overall workload. Twenty years after presenting NASA-TLX, Hart [69] reviewed the current state of the technique. It was found that most recent studies using this technique handled investigations on interface design and evaluation, with 31% focusing on visual and auditory displays and 11% on input devices. 7% of the studies were carried out with users of personal computers. It is important to note that the technique was initially developed for use in aviation (e.g., flight-deck design) however nowadays it has been widely adopted for alternative uses and is also being used as a

benchmark against which other workload measuring techniques are evaluated.

Hart denotes that NASA-TLX can be used in various situations, from aircraft certification to website design. This thesis proposes the use of NASA-TLX to measure enrolment-specific workload, primarily because of its multi-dimensional nature and overall performance (i.e., sensitivity). Various other advantages of NASA-TLX exist, including ease of use, practicality of the method, reduction of between-rater variability (due to the adoption of weighted rankings) and the availability of clear instructions, supporting tools and case studies.

2.3.5.1 NASA-TLX

NASA-TLX requires the user to rate six workload factors ex post facto. This is generally a pen-and-paper exercise (evaluation sheet) containing six bipolar scales, one for each workload factor listed below:

Mental Demand (MD)	How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc)? Was the task easy or demanding, simple or complex, exacting or forgiving?
Physical Demand (PD)	How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
Temporal Demand (TD)	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
Own Performance (P)	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
Effort (E)	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Frustration (F)	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

Several tasks might be presented to the participant and after all the tasks have been completed, together with their corresponding evaluation sheets, the participant is required to complete a pairwise comparison of the six workload factors. A total 15 pairs representing all the combinations of these six factors are presented and for each pair the participant has to pick one that is perceived as contributing more towards workload. This process is required only once whenever the tasks performed are of a similar nature. This will provide a weighting on each factor for each participant, which will eventually be used to generate a mean weighted workload measurement (MWW) as a score between 0 and 100. This weighting indicates the emphasis each user places on the different workload dimensions (as contributors to workload) allowing for an overall weighted result. This helps the researcher present more realistic

insights into how users perceive a task's workload caused by security requirements. The pairwise comparison creates a tally count of the number of times each factor was selected, as shown in the example below:

Table 2.3: Example of a tally count resulting from a pairwise comparison exercise of the six workload scales

Subject	MD[1]	PD[2]	TD[3]	OP[4]	E[5]	F[6]
Tanya	0	5	2	1	3	4

[1] Mental Demand
[2] Physical Demand
[3] Temporal Demand
[4] Own Performance
[5] Effort
[6] Frustration

Once all the data has been collected, the evaluator works out the NASA-TLX mean weighted workload for each task. For each task one must (1) multiply each workload dimension by its respective tally value (i.e., $(R_i \times T_i)$), (2) summate the resulting values for the six dimensions (i.e., $\sum_{i=1}^{6}(R_i \times T_i)$) and (3) divide the total by 15 (total number of unique pair combinations). This produces the mean weighted workload (MWW) for each task. This can be formalised by the following equation:

$$MWW = \frac{\sum_{i=1}^{6}(R_i \times T_i)}{15} \tag{2.1}$$

Where R_i is the rating given by the user for the i^{th} workload factor (e.g., Mental Demand), and T_i is the total number of times the workload factor was selected in the pairwise comparison (tally count, contributing to the weight for the i^{th} factor).

Table 2.4: NASA-TLX data for one participant

Subject	Task	MD[1]	PD[2]	TD[3]	OP[4]	E[5]	F[6]	MWW[7]
Tanya	1	45	0	0	100	0	80	28
Tanya	2	20	90	100	80	100	100	95
Tanya	3	75	0	0	100	80	100	49
Tanya	4	0	0	0	0	0	90	24
Tanya	5	95	95	95	100	100	100	98

[1] Mental Demand
[2] Physical Demand
[3] Temporal Demand
[4] Own Performance
[5] Effort
[6] Frustration
[7] Mean Weighted Workload

Table 2.4 shows the rating for the six workload factors across all tasks which together with the tally measurements shown in Table 2.3 provides enough data to calculate the mean weighted workload (MWW – see equation 2.1).

2.4 Process-Related Threats to UX in E-Service Projects

This section tackles issues related to (1) e-service contract awards, (2) process knowledge reuse and (3) process consolidation.

2.4.1 E-service project contracts and processes

Jokela and Buie [84] list four types of e-government project contracts, namely; *in-house*, *scoped contract with multiple contractors*, *umbrella contract with one contractor* and *umbrella contract with multiple contractors*. Contracted projects are generally selected on the lowest bidder principle (known as the Lowest Prices Technically Acceptable – LPTA – in the US). Unless UX activities and requirements are clearly defined in the Request for Proposal (RFP) together with explicit mechanisms to verify their delivery, it is more likely that bidders would cut costs on UX activities in order to qualify as the lowest bidder. Lohfeld [97] argues that although LPTA is suitable for certain procurement exercises it is "*definitely*" not suitable for "*procurements with complex services or uncertain performance risk*". This has also been reported by Chvotkin as cited in Jokela and Buie [84]. In their recent work Pfleeger, Sasse and Caputo [126] state that even though usability and usable security guidelines are readily available from various research bodies, developers do not always take this knowledge into practice. Along with other quality attributes, user experience and usable security considerations "*suffer cuts*" because of constrained project budgets and schedules [126]. In the introductory chapter of the Human Centred Software Engineering book [4], titled "*Integrating Usability in the Development Process*", the authors claim that one of the major obstacles for effective integration of usability engineering (UE) within standard development practices is its own perceived *dispensability*. This perception may lead software managers to treat UE iterations as an unjustifiable expense without considering the long-term implications on software quality.

A mechanism to define measurable, verifiable, valid and comprehensive UX requirements does not exist and thus this risk persists. This need has been partially highlighted by Seffah, Gulliksen and Desmarais in [147] whereby the authors call for measurable usability objectives to be specified within the project plan. Viscusi et al. [167] list several authors who propose the use of SERVQUAL – a service quality scoring tool based on the mismatch between expected and perceived service levels – for *existing* information systems. No literature could be found on the adoption of SERVQUAL at the requirement stage for *new* systems. Although it may provide useful insights on existing e-services, it cannot be used directly to inform the requirements development process for new e-services.

Viscusi et al. [167] also propose a methodology for the specification of new administrative macro-processes whereby three main goals are specified:

1. *Process formalisation*: conceptual description of a process using modelling tools (e.g., Business Process Modelling Notation (BPMN) and Unified Modelling Notation (UML) diagrams) and the description of operations enacted by this process using suitable technological languages (e.g., Business Process Execution Language (BPEL) and XML).

2. *Interfacing systems*: description of different services involved in the process and of the data exchanged between them

3. *Implementation*

The proposed process is heavily oriented towards the selection of appropriate technologies to automate existing or re-engineered business processes, leaving the user experience out of the equation.

Building user-centred e-government services is not a want, but a need to execute a truly transformational strategy, whereby government gets closer to the citizen while reducing costs and enabling effective citizen interaction. Users have an expendable budget for security compliance, and in order to avoid 'spending' too much time on security they develop habits [12]. Such habits can lead to non-compliance or circumvention of security processes if the cost of complying with security requirements exceeds the perceived value derived from the interaction (e.g., externalising passwords). This cost can be expressed in many ways, including workload, time spent on security tasks and back-tracking to the primary task following a security step [145]. Nonetheless empirical testing with potential users on new public facing government systems is expensive and can easily push the launch date even further, possibly conflicting with the policy-maker goals (i.e., the product owner). Unforeseen problems are annoying, expensive and can lead to political blunders. Because of this, the author believes that certain problems can and should be caught earlier through systematic and analytical methods.

Empirical user-experience studies are still considered a must later on in the system's lifecycle (pre-launch) helping the project team catch the unexpected or the things that couldn't have necessarily been predicted based on existing knowledge, for which extrapolation is not possible. However, and wherever possible, analytical methods should be adopted as part of the process used to design and build public facing e-government services. Angela Sasse, head of the Information Security Research Group at UCL insisted that developers must informally but systematically list the exact steps that a user must perform in order to complete the primary task on a two columned sheet of paper (personal communication, October 16, 2013). In the second column designers and developers should list any potential errors that may occur at each step. This could also be augmented with workload measurements such as average step duration using keystroke-level modelling (KLM-GOMS) [142]. This systematic workload audit allows team members to explore the direct impact that processes will have on users (i.e., visually). A similar technique is shown in Figure 2.3, taken from a report published by the National Institute of Standards and Technology (NIST) [158].

This technique offers an opportunity for system designers to uncover potential friction points within security processes (at design-time). However this assessment is conducted independently from the end user and does not provide actionable indications on how to reduce friction for the various groups of potential end users, who may in turn respond differently to different workload levels in different contexts.

2.4.2 "Getting UX into the contract" — and the process

Government agencies do not specify usability at the requirements stage, or if they do, they do not ensure, or rather are not prepared to ensure usability throughout the project development lifecycle [84]. Jokela and Buie [84] agree that UX in government projects depends on defining clear requirements upfront, however by doing so one would not be imposing any guarantees of these being delivered unless measurable and testable fit-criteria are specified. Neil Maiden [99] argues that requirement quantification is

Figure 2.3: Keystroke Level Modelling sequence for a manual login process (taken from NISTIR7983 [158])

Step	Activity	GOMS-KLM Task Symbol	Time (in seconds)
1	Mentally prepare	M	1.35
2	Home hand on mouse	H	0.4
3	Position the cursor over the bookmark	P	1.1
4	Click mouse	P1	0.2
5	Position the cursor over the userid field	P	1.1
6	Click mouse	P1	0.2
7	Home hands on the keyboard	H	0.4
8	Recall userid	M	1.35
9	Enter userid (8 characters)	8 (K)	1.76
10	Home hand on the mouse	H	0.4
11	Position the cursor over the password field	P	1.1
12	Click mouse	P1	0.2
13	Home hands on the keyboard	H	0.4
14	Recall password	M	1.35

one of the "*biggest challenges*" that analysts face. It is difficult to specify measurable fit-criteria at the requirements stage, however not attempting to do so may result in vague requirements that leave ample room for interpretation (e.g., "the application must be usable"). By rejecting broad and non-measurable statements during the requirements development process (through healthy debate) the project team would be working towards a set of requirements that would implicitly encourage developer compliance who would be motivated to "*find ways to deliver on these requirements*" [99]. Maiden proposes *Volere* and *Planguage* as two techniques that can help requirements analysts quantify requirements. Nonetheless the author also argues that although both are effective at this, they may not be well suited to guide analysts in quantifying requirements that are applicable across a potentially heterogeneous range of users and contexts of use. For this reason Maiden proposes the use of use cases and scenarios.

Saucken et al. [168] argue that this problem is especially visible when specifying big picture UX rather than small or micro UX requirements. The main difference lies in the level of specification abstraction and the extent of impact such requirements may have on users. *Macro UX* deals with goals, purposes, perceptions and feelings whereas *micro UX* provides "*concrete hints for design*". The authors also argue that macro UX related requirements are "*often psychology-driven*" and consequently are "*hardly implemented in industrial practice*" [168]. This thesis extends this argument to includes ULX considerations which deal with the impact on users beyond the point of interaction with an e-service. Regular usability measurements (e.g., user walkthroughs) would be ideal, however that might turn out to be an expensive and time consuming exercise. Viscusi et al. [167] consider *appropriateness* to be an important consideration at both the macro and micro levels of e-government strategy allowing designers to evaluate capabilities of a system to achieve states that promote users' well-being, and turning these into utilities. However it is difficult to test for and verify *appropriateness* in a measurable way.

There is a disconnect between the various aspects of e-government service development, which include policy making, software engineering and user-centric practices. These are considered to be distinct and often sequential activities, with the latter merely treated as an afterthought in an effort to improve e-service appeal. Seffah and Metzker [148] refer to the *people gap* in which usability experts and software engineers "*do not share the same culture*". This gap is further accentuated when policy makers (who are generally not knowledgeable in either field) are involved. In [147] the authors also state that this gap is evident in each group's perspective and awareness of each other's operational constraints. Furthermore, and according to Hamilton et al. [148], wherever user-centred design expertise have been brought in by government decision makers a skills gap was evident between the two. This gap manifests itself as a push towards process-centric requirements, with decision makers framing the problem in a way to satisfy policy objectives rather than citizen needs and goals. User goals, requirements and usage contexts are not taken into consideration by default and if considered they may also be used incorrectly (e.g., user journey testing after launching a service) [68]. The authors also argue that even if applied at the earlier stages of the system's lifecycle, conclusions derived from applying user-centred techniques may be misinterpreted and "*wrongly operationalised by staff who lack UCD knowledge*". Nonetheless it is important to acknowledge the fact that over the years governments across the EU have recognised the importance of user-centricity as a core aspect of service design, corroborating with Hamilton's views who stated that there is a "*nascent appreciation of UCD [User-Centred Design]*" in the public sector, however the authors continue to state that UCD is "*by no means institutionalised in government departments. For many, the usability paradigm is unknown*" [68].

UX should start at the requirements stage embedding experience expectations within the initial specification document or request for proposals (RFP) in the case of contracted projects. Nonetheless mechanisms need to be in place (within the government agency) to ensure adherence to these initial requirements and this is where complexity is compounded. During a discussion with Maltese government officials (in the role of information system architects) it became evident that it is not always possible to ensure a positive user experience throughout and across government projects. Power relations, internal politics and resource constraints may contribute towards design decisions that may not always be favourable from a users' perspective.

Placing UX requirements within the initial specification document does not necessarily guarantee a good user experience out of the resulting product [84]. Discussing this issue with industry experts at the 2013 IAAC symposium (London) it became clear that this is a critical problem and there is no exact science to guide, measure and report experience-related activities throughout the requirements development process.

There is a risk of knowledge-loss following project completion especially when government bodies do not accumulate and share design and development knowledge in a structured way for use in future projects (internal and across entities). In the case of contracted projects, government bodies should request adherence to a standard process for system design and development that ensures consistency while contributing towards a government-wide knowledge base. Contractor experience would still play

a significant role in the overall selection criteria, however equipping contractors with a process, toolset as well as knowledge on known issues with different user groups (UX related data) would reduce the dependence on contractor experience while mitigating the risk of basing contractor-choice on subjective, generally bloated, claims. Responsibility for UX is generally a contentious issue since this also involves accountability for future fixes, potentially causing disputes between parties.

A unified yet flexible design process that mitigates against these risks deserves further investigation. Jokela and Buie [84] argue that it is not realistic to assign complete responsibility for usability and UX to the contractor since it is currently difficult to define verifiable, valid and comprehensive usability requirements and it is also currently very difficult to estimate the effort required when bidding for projects in a competitive environment. The purchaser must carry out interim usability tests to assess progress, however there is no systematic way to verify UX quality during the design and development process. Re-using past experience within a well-defined process might help.

2.5 People-Related Threats to Usability in E-Service Projects

Zurko and Simon [180] proposed guidelines for the design of user-centred security mechanisms. Here the authors suggest cognitive workload as an important factor to consider (and measure), nonetheless they do not provide metrics or techniques that can be used to inform the requirements development process in a pragmatic manner. Tools and techniques adopted by HCI people may not necessarily be equally useful for software developers, who must ultimately take (low-level) design decisions that will have a direct impact on users. Pfleeger and Caputo [125] argue that by understanding and leveraging knowledge on human behaviour (i.e., behavioural science) designers and developers would be in a better position to design and build more secure and usable systems that are both efficient and effective from a user's perspective. Understanding aspects such as cognitive load and the impact that proposed systems may have on the end user is a fundamental design activity, but one that is rarely taken into account [125].

2.5.1 The disconnect between software engineering and HCI

Sutcliffe [161] believes that user-experience (UX) considerations need to be embraced by the requirements engineering community. Seffah et al. [147] list several requirements for human-centred software engineering, in an attempt to bridge the gap between HCI and software engineering. These include an understanding that users are a valuable resource, who however may not always be able to express their needs effectively. For this purpose, methods to observe and analyse user behaviour are needed as evidence upon which to ground requirements, supplemented by active user involvement. The second requirement for effective integration of HCI within development practices is the need for better developer tools to turn requirements into usable designs given unavoidable technological and resource constraints. The lack of production tools that integrate the two domains effectively will results in hesitation from developers to "*spend too much of their time*" with actual users [147]. Seffah and Metzker [148] outline a set of obstacles for the effective integration of usability in the development processes. One of the major obstacles is "*the meaning of usability*" itself. Although standards exist to operationalise the term it is generally vaguely used as a standalone non-functional requirement amongst many others – "*as though*

this term encompasses all there is to know about the field" [147]. Human or user-centred design, usability engineering and user experience design are sometimes used interchangeably. The authors continue to state that user-centred activities within the development process are often seen as mere frills by software engineers who are in the end responsible to deliver the final product.

The "*responsibilities gap*" is another major obstacle. This refers to the idea that when the 'real system' is delivered (i.e., the software artefact) usability experts come in to "*make the interface layer user-friendly*" [147]. On the other hand, UX designers believe that the concept comes first (including the intended experience that needs to be delivered) upon which the product is then built. Seffah, Gulliksen and Desmarais re-iterate that usability specialists must operate following some core engineering values — in both thought and action [147]. This is further explained by the "*modularity fallacy*" which is listed as another major obstacle — interface design cannot be treated as a standalone task and usability engineers must work closely with software engineers to design a system that provides both usefulness and a positive experience from the ground up. Development patterns such as MVC (Model-View-Controller) [132] promote the separation of concerns in which the view can be developed independently from the controller portion of a system (i.e., core logic), however in real terms, the view depends on the functionality offered by the controller, and no interface-candy could ever make a cumbersome work-flow (dictated by the core application logic) friendly. This leads to the conclusion that usability is not a property of the user interface – but it depends on a higher level of abstraction at which UX and usability considerations are built-in the requirements, design and development process – guiding the selection and design of views, specification of models and the development of controllers (in an MVC architectural context). Furthermore, non-functional requirements should not be vaguely defined, but specified in a testable manner.

Dowell and Long [45] suggest that human factors (HF) practices are not well-addressed in software engineering processes – (1) poor integration between these two fields, as well as (2) the lack of a formal structure within which experience can be accumulated for re-use across projects, are presented as two of the main shortcomings in HF practices [45]. Dowell and Long also state that the lack of integration between HF considerations and requirements development processes lead to poor design decisions, generally taken by developers who are mostly not specialists in human factors and ergonomics [45]. The authors also argue that advice on human factors can be easily ignored at the early stages of development – and usability assessment is generally "*relegated to the closing stages of development programmes*" at which point it would be too expensive to make "*even modest re-implementations*". In a NIST Workshop Report [124] several authors argue that actions related to usability and security must be integrated within the various stages of the software development lifecycle. In this report, Jeremy Epstein also argues that security and usability requirements suffer from a lack of metrics, making it difficult to understand how usable (and secure) a system is. Epstein believes in the need for a "*reform*" in software procurement processes. In the same report [124] Carol Woody argues that a comprehensive software development process that encompasses security and usability is a necessity mainly because user-specific security practices do not "*effectively consider the capabilities and realities of human actions*". Security and usability are gen-

erally treated as add-on features, generally designed for completeness' sake and to claim that a system is usable and secure. Epstein contrasts this approach to one in which security and usability are "*emergent properties of software systems*". This report was the outcome of the first Software And Usable Security Aligned for Good Engineering (SAUSAGE) workshop. This edition of the workshop was sponsored by the National Institute for Standards and Technology (NIST) to promote usability and security actions as fundamental components in "*all stages of the software development lifecycle*" [124]. A set of recommendations was also presented, and two recommendations are of particular importance to this thesis: (1) the need to identify and develop methods and tools to support usable security across the various development activities, including the requirements and design stage and (2) the need to develop a framework for understanding usability and security preferences, and for enabling trade-off analysis among usability and security choices.

2.6 Glossary of Terms

This section provides a glossary of terms and their working definition based on existing literature.

Accessibility

Accessible systems are those designed to promote social inclusion, allowing people with disabilities to adopt mainstream technologies through alternative or enhanced interfaces. The objective of accessibility is to allow the majority of users, possibly all, to achieve their goals irrespective of any possible visual, auditory, motor or cognitive disability. Standards and techniques exist to help system designers specify measurable and verifiable accessibility requirements while tools are available to assure compliance. Tim Berners-Lee, director at W3C states that the power of the web is in its universality and access by everyone regardless of disability is an essential aspect. When a system is not designed for accessibility then it automatically excludes specific groups of users, or at best makes their lives more difficult. Furthermore, accessibility goes further in its social inclusivity goal and W3C include these non-disability related aspects as important considerations for accessible systems: old age, literacy, language barriers, use of older technologies and bandwidth, mobile and resource constrained devices, new and infrequent web users and also limited access to technology in rural areas. For all of these aspects W3C proposes a set of accessibility guidelines for web-based systems [71]. Legal requirements exist for public entities (e.g., federal agencies, government departments) to comply with accessibility guidelines, such as the Section 508 Laws in the US and the Web Accessibility Standards in Canada. Other countries provide guidelines, checklists as well as legal policies to encourage public agencies to ensure accessibility for their electronic services [20]. The International Standards Organisation (ISO) publishes guidance and pointers on software accessibility under ISO9241-171:2008 (Ergonomics of human-system interaction – Part 171: Guidance on software accessibility).

Attitudes vs *Behaviour*

The term *attitudes* refers to explicit feedback given by participants on the subject matter (i.e., based on past experiences) which would in turn be processed through qualitative methods. On the other hand *behaviour* refers to observable and measurable user activities which would in turn be processed through quantitative methods. Malheiros [101] observed that what users report (e.g., concerns on privacy) may vary significantly from their actual behaviour (e.g., while transacting online).

Critical design factors

These are elements within a system that may violate both *end-user goals* and *lived experience goals*. End-goals (i.e., what a user wants to do) represent user motivation for completing a task (e.g., "*save time by paying taxes online*") while lived experience goals (i.e., how a user should feel) are more universal and express feelings generated while interacting with the product (UX) but also feelings that go beyond that interaction (ULX), carrying a knock-on effect on the user's personal activities beyond that interaction (e.g., making a user feel '*stupid or uncomfortable*') – resulting in reduced effectiveness and self-esteem while increasing resentment towards the system and its provider [38].

Design

The stage in the requirements development process within which product use cases for the product-to-be are specified as a series of actions (executed by any active actor(s)) which would eventually lead to the completion of a primary task.

This should be an iterative process, ideally involving empirical exercises to establish the effectiveness of the process or processes being considered. To some extent this complies with the definition of interaction design provided by Rogers et al. [139] – "*designing interactive products to support the way people communicate and interact in their everyday working lives*" – however this thesis stops short from specifying visual design aspects such as colour schemes, typography and so forth. Non-functional requirements specified during the design process should be verifiable through the use of measurable fit-criteria – e.g., "*users shall be able to locate the search page in less than 20 seconds 80% of the time*".

Formative UX assessment

Crooks [40] defined formative assessment in education as an activity "*intended to promote further improvement in student attainment*" and assessment is defined as "*any process that provides information about the thinking, achievement or progress of students*". This principle is applied in this thesis and formative UX assessment is proposed as part of the design and development process to promote systematic and iterative UX improvements starting at the earliest stages of a system's lifecycle.

Framework

Charles Haley, cited by Faily in [52] defines a framework as "*a set of milestones indicating when artefacts should be produced*" without being too prescriptive on the intermediate steps required to produce them. A framework, as a supporting structure, should also recommend or provide optional tools and techniques to help its

adopters be more effective and efficient in the production of deliverables. Such recommendations must however comply with the framework's underlying philosophy (e.g., building user-centric systems).

Primary task vs Secondary task

Sasse and Fléchais [144] use the term *production tasks* to refer to those actions that a user performs to attain a goal or a desired state (e.g., obtain a birth certificate), while *supporting tasks* are actions that support production tasks but are not essential to their completion [144] (e.g., enrolment, identity verification and account activation for use in future interactions). This thesis will use the terms *primary tasks* and *secondary tasks* to refer to production and supporting tasks respectively. In the same paper [144], the authors state that human behaviour is goal driven, and any action that takes the user away from her goal, or that leaves the user to choose between complying with the security requirements imposed by a system or getting the job done, is essentially bad design. Systems must be designed in a way that encourage the completion of primary tasks in the most effective, efficient and secure way.

Requirements development process

This refers to the entire range of requirements-specific activities; from elicitation to quality assurance as well as requirements management and reuse. Wiegers [173] defines the requirements development process as a four stage process: starting from the *elicitation stage*. At this stage the team identifies project milestones and deliverables, specifies the project's scope and vision, identifies stakeholders, product champions, use cases as well as business events and associated responses. Among other techniques Wiegers suggests user observation (within the users' context) as well as requirements elicitation workshops (with stakeholders) as two important information harvesting techniques. The elicitation stage is followed by the *analysis stage* wherein the elicited information is analysed, requirements are prioritised, grouped, formalised, modelled (visually) and also ranked by the value each requirement contributes towards customer satisfaction. Prototypes may also be built during use case design activities. If clarifications are required the process would re-iterate from the elicitation stage. The *specification stage* is used to record requirements based on an officially accepted template, including each requirement's initiator and supporting artefacts as well as quality attributes (as guides towards customer satisfaction). Finally, the *validation stage* acts as the quality gateway wherein requirements are tested and validated for completeness, correctness, fitness, consistency (in terminology adopted) and traceability. During this stage, one should make sure that requirements are decorated with acceptance criteria [173] (at a minimum) and with fit-criteria for more rigorous processes [138]. The entire development process is iterative, and the validation stage may take the project team back to the specification stage (for re-writing), analysis stage (for re-evaluation)

or back to the elicitation stage (for corrections). The logical flow suggested by Wiegers [173] is largely congruent to processes proposed by Axel van Lamsweerde [91] and Suzanne and James Robertson [138].

Saturation – qualitative (codebook) and statistical

This thesis makes use of the concept of saturation to inform the researcher on when an investigation could be safely terminated without jeopardising the study's output quality. In particular, *two* types of saturation were used to guide the researcher across the various studies: *qualitative saturation* (see Section 4.1) and *statistical saturation* (see Section 4.3).

Qualitative saturation is the point at which additional qualitative studies would not yield significantly new and useful knowledge. Unlike quantitative studies, the qualitative researcher needs to assess her data as the study progresses in order to determine when to stop, or if necessary, when to extend a study. Guest et al. [67] provide a number of insights on this matter and the reader is urged to consider their recommendations when embarking on qualitative research. This thesis makes use of qualitative saturation to determine the point at which any additional interviews (analysed thematically) would not yield significantly new codes, or situations which might affect the ranking of existing codes in the codebook (i.e., the number of times a code is used).

Statistical saturation is the point at which any new input data would not yield any significant improvement on the statistical models being produced (e.g., regression models) [128]. If the change in prediction power (or model parameters) following the injection of new data is negligible, then the collection of further data might not be justifiable (depending on the model's purpose and domain of use). Saturation is hereby defined as the point at which statistical models do not exhibit significant improvements in prediction with the addition of more data (e.g., an additional 10 participants will not yield more than 2% improvement over predictions generated by the original model). Out-of-sample tests can be used for this purpose whereby updated models are tested against a set of known observations. A score based on residual values (predictions vs actual observations) will help determine by how much the model has improved with the addition of new calibration data – based on the original model's score (without the new data).

Security friction

Beautement et al. [12] define friction as the imbalance between the business process (user goals) and security behaviour required, including any inherent cognitive and physical workload. Friction can cause resentment, and ultimately non-compliance, which can lead to service abandonment.

2.7 Conclusions

This chapter presents evidence of a clear relationship between enrolment-specific design decisions and user adoption, supporting the hypothesis that over-protection has a negative impact on takeup (see Section 2.3). These findings shed more light on the first sub-question (SRQ1) presented in Section 1.2.

A review of current practices in e-government service procurement and development (see Section 2.4) reveals that the lowest bidder approach is generally adopted for e-service procurement (or Lowest Prices Technically Acceptable in the US). Bidders are likely to cut costs on UX related activities unless these are specified in a measurable and verifiable way. This has been flagged as a major issue, motivating the rest of the investigation. Transforming experience requirements into a measurable and verifiable form is not trivial, although believed to be necessary. A number of standalone techniques exist that can be used to measure the impact on the users' lived experience, and tools such as GOMS-KLM can be utilised to assess workload levels associated with specific tasks. Other techniques (e.g., NASA-TLX) have been reviewed in Section 2.3.5. The literature reviewed for this thesis has not provided pragmatic indications on how enrolment processes can be designed in such a way that support the service provider's required level of identity assurance while respecting end-users' limitations and expectations.

This chapter has also shown that a culture gap exists between usability people and engineers (see Section 2.5) and this is further accentuated by the skills gap that exists within e-government project teams in which policy makers are not aware of the importance of UX activities to the success of e-government projects. Usability, HCI and user experience are frequently treated as a non-engineering discipline, or rather, an add-on activity following the 'real' software development process – to make an existing system 'user friendly'. This thesis is mainly driven by the belief that user experience has its roots at the core of the development process, starting at the requirements stage. A badly designed use case will have a negative snowball effect on end users and service providers alike, and no matter how much work is put into the improvement of the user interface (UI), the lived experience (ULX) will not improve.

Several user-centred techniques and tools exist, however more emphasis is required on the development of efficient and pragmatic practitioner-oriented techniques. These should also abstract complex theory and be encapsulated within usable and production-ready tools to assist in the specification of testable experience requirements for e-government services.

Chapter 3

Methodology

This chapter starts by reviewing research methods generally adopted in HCI and requirements engineering (RE). This is then followed by an outline of the methodology used to develop and evaluate the main contributions presented in this thesis.

3.1 Research Methods in HCI, Requirements and Design

According to Blandford [14] there is a "*huge number of qualitative methods, often minor variants of each other*" and choosing an appropriate method for a study may be a challenging task. Woolrych et al (cited in [14]) compare qualitative techniques to ingredients in recipes whereby to obtain effectiveness one must "*construct a recipe [with the available ingredients] that is right for the occasion [addressing the study's purpose]*". Furthermore, the use of qualitative techniques in HCI is not mature [14] and HCI researchers may find it difficult to make informed decisions on which techniques to adopt in any given study. Qualitative methods cannot be too prescriptive and researchers should be able to mix and match techniques according to the study's objectives, the nature of the data and resources at hand.

Cooper [38] argues that qualitative research in interaction design helps to provide a deeper understanding of the domain or problem at hand through the discovery of behavioural patterns. This corroborates the idea that most usability issues are generally discovered within a few in-depth interviews with participants [115, 67]. Several qualitative methods exist, including stakeholder interviews, subject matter expert interviews, contextual inquiries or master-apprentice learning [13], literature review and user feedback via focus groups and low or high-fidelity prototyping.

According to Faily [52] formative and summative evaluation have been adopted by the HCI community – "*to provide feedback to Software Engineering activities*" on the usability aspects of a design, rather than on the design activity itself. During the design process formative evaluation is used to test a design – prototyping – while summative evaluation is adopted to obtain insights into the quality of a final product – even if this is through controlled usability tests. The former informs the design process while the latter measures the quality of deliverables. As opposed to summative evaluation, formative evaluation aids in test-running various aspects of a theory or proposal. As Robert Stakes puts it, "*when the cook [or cooks] tastes the soup, that's formative. When the food critic tastes the soup, that's summative*" [10]. Drawing on this metaphor, any negative report made at this point by the food critic would not just

highlight issues with the product (e.g., soup's taste or texture) but could also cause extensive damage to the restaurant's goodwill.

Rogers et al. [139] suggest triangulation as a strategy to provide "*different perspectives and corroboration of findings across techniques*" producing defensible findings through rigorous inquiry. Triangulation involves more than one technique to investigate a phenomena (e.g., interviews, questionnaires, and focus groups). The authors also suggest piloting as a technique to gather data which can in turn inform the design of the research method itself (e.g., questionnaire structure).

Qualitative research is quintessential as a methodology to understand a phenomenon, but the challenge lies with transforming such knowledge into practical and measurable guidelines for designers to adopt. These practical guidelines would help practitioners design and build products that meet users' expectations while taking their behavioural patterns into consideration. Sections 3.1.2 and 3.1.2.3 take this argument a step further.

3.1.1 Data collection techniques

Data collection techniques deriving from both qualitative and quantitative research domains can be adopted to generate the necessary data upon which claims are made and through which concepts and theories are developed. Semi-structured interviews, focus groups and participant observations are three techniques that provide the qualitative researcher with a high level of flexibility. These techniques also provide interviewees with a stronger voice which can in turn influence the formulation of an initial research idea [23]. While interviews and focus groups are effective in providing insights into people's perspectives and experiences, they are not adequate enough to reconstruct a detailed understanding of participant practices [14] (when compared to participant observation). On the other hand, the level of flexibility afforded by semi-structured interviews (and focus groups) allows the researcher to cover potentially new or unknown grounds, over and above the targeted discussion topics set within the initial interview guide. Participants are given the liberty to discuss a phenomenon as seen from their own perspective and as experienced within their social context – seeking out the participants' world view [23]. This can be contrasted with quantitative research whereby the researchers' concerns are materialised within the data collection mechanism itself (e.g., questionnaire) and the inquiry is driven around and limited to these concerns.

Focus groups introduce more participants to an interview whereby the interviewer becomes a moderator, facilitating the discussion based on social cues (e.g., arising from participant interaction) while following an interview guide with salient aspects that need to be covered during the session. Focus groups introduce further opportunities over semi-structured interviews since participants may (1) probe and challenge each other on particular views, (2) change or qualify their own views based on others', as well as (3) agree with or expand on views which might not have surfaced in a one-to-one setting [23, 14]. Group dynamics may provide additional insights to the inquiry, however this may also be disruptive especially when participants take control or pose a strong influence on opinions expressed by others. Bryman suggests that focus group participants are more likely to express "*culturally expected*" views [23] and this may influence the reliability of the emerging data.

Semi-structured interviews as well as focus groups are important tools however they share some challenges. Setting up meetings, conducting and transcribing sessions is time consuming [14, 23]. This problem is compounded even further when organising a series of focus groups with different participants. In both cases, there is also the risk of participants not showing up, which can be mitigated by over-recruiting or by offering small payments or store tokens [23].

Both types of data collection methods require the researcher to record and transcribe the discussions before analysis can commence. This is a time consuming exercise however it allows for repeated and in-depth examinations of participant conversations, preserving minute details which may turn out to be important insights providing useful nuances on the phenomenon being investigated.

Quantitative research is on the other hand concerned with explaining reality through objective measurements about, as well as causal relationships between events [101, 23] (e.g., A causes B). This may not be an appropriate approach to study phenomena within a social world (or socio-technical contexts) where it is understood that people are driven by their own interpretation and perception of the world around them [23]. Nonetheless, quantitative techniques (e.g., surveys) can act as important instruments through which the researcher can investigate aspects of an apparent causal relationships or test new hypotheses arising through qualitative research methods.

3.1.2 Qualitative data analysis and interpretation for design

3.1.2.1 Grounded theory and thematic analysis

Grounded theory, originally introduced by Glaser and Strauss [62], refers to a method by which a theory is derived directly from data following a recursive process of data collection and analysis [23]. Following methodical disagreements, the original authors parted directions. Strauss [159] recommended different philosophical perspectives on procedural aspects of the method and these were then re-established together with Juliet Corbin in 1990 [160]. The goal of grounded theory is to generate a rich understanding of a specific aspect of social-life [169]. Strauss states that grounded theory is based on a concept-indicator model [159], whereby empirical indicators, or rather empirical data, informs and directs the construction of a concept. Through coding, the researcher would then mark aspects of the data as indicators of a class of events or behavioural actions which may eventually result in coded categories represented through indicators carrying a similar or consistent meaning. Dey [41] observed that there are "*probably as many versions of grounded theory as there were grounded theorists*". Thomas and James [163] criticise grounded theory and suggest that the idea of approaching the process with an open mind and without any preconceptions is virtually impossible.

> "*How are grounded theorists to quarantine themselves, as social selves, from the data they are analysing and re-analysing to enable 'theory' to emerge?*" – [163]

The search for themes is a central premise of most qualitative analysis methods, and thematic analysis is often used to refer to the coding process within such methods rather than a method in its own right. Braun and Clarke [18] propose a six-phase process for thematic analysis, guiding the researcher to (1) familiarise herself with the data, (2) label interesting aspects within the data (initial codes), (3) gather

groups of related labels to construct themes, (4) iteratively review and (5) refine themes and finally (6) produce a report containing further reflections on each themes, supporting examples and corresponding narratives [14]. Thematic analysis helps the researcher review the data and sort it into categories in order to garner a "*deeper appreciation of the content*" through the discovery of patterns and developing themes [122]. Malheiros argues that this technique is similar in nature to the open coding phase in grounded theory [101]. Bryman [23] also states that the word 'theme' (i.e., patterns of interest within the data related to the research question [18]) is often misused by qualitative researchers (e.g., used to refer to individual codes) and suggests that it could be the direct result of a lack of detailed procedural guidelines on the method itself. Braun and Clarke [18] also suggest that there are no hard rules to distinguish stronger themes from weaker ones (e.g., using prevalence thresholds) therefore placing the researcher's judgment at the centre of the thematic analysis process.

3.1.2.2 Coding

Coding is a common practice adopted across most qualitative data analysis techniques [23]. This practice assists data analysts to develop an understanding of data corpora (e.g., interview transcripts) generated via semi-structured and unstructured data collection techniques. Coding presents an opportunity to highlight important aspects within the text that may contribute towards the research question(s) being answered. Codes explain what a data item within the text refers to (i.e., line-by-line or incident-by-incident), in terms of topics, actions, sentiments and so forth. Several coding techniques have been suggested by various qualitative researchers [31, 18, 23, 160, 62] and their adoption by the research community has been further supported through the development of various computer aided qualitative data analysis (CAQDA) tools. *Atlas.ti*[1] and *NVivo*[2] are two of the main commercial CAQDA tools offering numerous features and functions to support coding and analysis on datasets containing text, images, video and audio (e.g., code-by-list, object crawling for patterns, word occurrence reports). Visualisation tools (e.g., cloud tags, code network views and frequency bars) provide fresh views on the data, potentially uncovering interesting, new and unexpected insights which may go unseen using manual coding methods. Also, an active community provides most of the answers to questions that arise while working with such toolsets, especially as a newcomer to qualitative research. Alternative CAQDA tools exist, both as freeware and open-sourced[3].

Initial coding gives the researcher an opportunity to familiarise herself with the data at hand. Following an initial read-through of the entire dataset the researcher would then take an exploratory stance by making marginal notes and observations about "*significant remarks and observations*" arising directly from the data [23]. This process of labelling and highlighting elements within the data is the starting point for the generation of a codebook – an index of terms that helps the researcher formulate theoretical interpretations of the data [23]. Kathy Charmaz [31] suggests that codes should be "*short, simple, active and analytic*", staying close to the data while preserving action, meaning and possibly

[1] Atlas.ti Scientific Software Development GmbH, Atlas.ti Qualitative Data Analysis, http://www.atlasti.com/, (accessed September 2014)

[2] QSR International, NVivo, http://www.qsrinternational.com/products_nvivo.aspx, (accessed September 2014)

[3] Huang Ronggui. RQDA: R-based Qualitative Data Analysis. R package version 0.2-3. http://rqda.r-forge.r-project.org/, (accessed September 2014)

context (or social setting). As the coding process progresses the researcher will start to observe common elements across the dataset, afforded through a coding-by-list approach (re-using previously developed codes) rather than through free-style or open-coding. Following the first few iterations, the researcher will then start consolidating codes (e.g., merging similar or removing redundant codes) while noting possible criteria through which codes could be classified. Code families, or categories allow the researchers to group similar (or closely-related) codes in order to generate a high-level structure representing the nature and essence of a potentially vast dataset. This structure could in turn assist the researcher to interpret, theorise and reflect upon findings inline with the initial research questions posed [23]. This coding style is reflected in the initial coding stages of grounded theory (open coding) as well as in the thematic analysis process recommended by Braun and Clarke [18] – whereby the data is systematically broken down and re-structured for further analysis. Bryman observes that codes "*can do more than simply manage the data you have gathered*" [23] and refers to Strauss and Corbin's practice of establishing the interconnection between codes and code categories in an effort to discover dimensions that could explain a ''*broader phenomenon*" [23].

Strauss and Corbin's practice of axial coding [160] attempts to relate identified code categories and sub-categories in an effort to synthesise and organise the data and to understand the main relationships, properties and dimensions of each category. During the initial coding stages the corpus of data is disassembled into separate and distinct codes whereby axial coding reconstructs the data as a coherent whole [31]. This process materialises, or rather visualises the inter-categorical links through the adoption of a set of scientific terms, including;

1. **Conditions**: *A is associated with B* – mainly answering the why, where, how come and when questions.

2. **Actions/interactions**: *B leads to C* – mainly answering the by whom and how questions.

3. **Consequences**: *C is cause of A* – mainly answering the what happens questions following an action or interaction

Axial coding adds discipline to the theme searching phase suggested by Braun and Clarke [18], reducing ambiguity by encouraging researchers to apply an analytical frame to their data [31]. This can also serve as a means to reduce sole dependence on the researcher's intuition in this sense making exercise – keeping as close to the data as possible without imposing, as much as possible, personal leads or ideas. Building codes and categories from the data helps the researcher stay close to the facts. At the same time, relying solely on axial coding to extract themes may become a myopic analytical exercise whereby the larger picture is hidden from view due to a low-level and finer-grained sense making effort. The author believes that a level of personal judgement is still required (i.e., along with the risks this introduces). Braun and Clarke [18] suggest the use of visual techniques to complete this phase, including mind-maps, tables and "*theme piles*". The latter is a technique similar to card-sorting (generally applied in interaction design) collating relevant and inter-related codes into themes and sub-themes through the process of organising sticky notes (representing individual codes) in separate piles.

3.1.2.3 Practical issues with qualitative techniques

One of the major challenges in moving from qualitative research to design is to determine the right number of participants required to generate significant insights upon which to base design decisions. There's no objective answer to this. Nonetheless there are signals that need to be considered. One of the main signals is data saturation. There exists a natural point at which additional studies would not yield significantly new and useful knowledge. Unlike quantitative studies, the qualitative researcher needs to assess her data as the study progresses in order to determine when to stop, or if necessary, when to extend a study. Guest et al. [67] provide a number of insights on this matter. Their main premise is that there are virtually no guidelines to determine non-probabilistic sample sizes and through a number of studies they explored saturation with the intent to offer evidence-based recommendations. In principle, the authors suggest that 12 interviews are generally conducive to data saturation, however overarching themes would already be present by the sixth transcript. The studies presented in this thesis exhibited a similar pattern and in all cases over 90% of codes were uncovered by the sixth interview. Nonetheless this was mainly due to a narrow focus in the exercise. Charmaz [31] states that a "*small study with modest claims*" can help you justify saturation earlier while not affecting your credibility, as opposed to studies that span a number of possible phenomena. Researcher experience might also shape the definition of saturation [104] whereby a seasoned researcher might find interesting and new insights in data corpora that are otherwise seemingly saturated for the newcomer. Because of this Mason [104] believes that the concept of saturation can be inappropriate. The argument brought forth is that with further assessment there will always be potential for new concepts to emerge and the author in turn proposes saturation to be the point at which data analysis becomes counter-productive and new discoveries add no value to the overall study. This is especially the case when there is an excess of data, rather than the lack of it.

This thesis considers thematic analysis to be the technique of choice when analysing qualitative data however there are no strict rules imposed on the researcher, which might raise some concerns. By adopting a high-level coding strategy (coarse grained) one may dilute or lose important details, however low-level coding (fine grained) may consume too much time and the data may become too fragmented and thus very difficult to synthesise at a later stage. Furthermore, coding as an abstraction discipline is a subjective exercise and it highly depends on the study's objectives, the phenomena being investigated and the researcher's experience. There is also the risk of unconsciously guiding the coding effort to corroborate one's perceptions, goals and intuition. One way to mitigate this risk is to involve third parties in the coding process to sample the strategy and confirm that one's abstractions are related to those of other researchers, and thus desensitised from personal preconditions. Furthermore, techniques inspired from grounded theory are also used along with thematic analysis. Axial coding is used to synthesise a story from existing codes by establishing relationships and code families. Nonetheless one may choose to stop short from generating a theory since this might not be a primary goal. Adopting parts of a larger framework or process can at times be more suitable than the whole set of recommendations, as long as claims on their use are substantiated with proper justification.

Building codes and categories from the data helps researchers stay as close to the facts as possible

and any code or theory developed would be to some degree grounded in empirical evidence (mitigating influence from preconceived ideas). During informal discussions with other qualitative researchers it transpired that the best strategy to analyse qualitative data is to select the tools and techniques that best fit your (1) data (e.g., *visual* vs *textual*), (2) experience and (3) goals (e.g., *granular* vs *coarse coding*) and to present emerging results in a transparent and honest way while outlining possible shortcomings and acknowledging the challenges encountered.

Another challenge in qualitative research is generalisability. This highly depends on the researcher's goals, however it would be simpler to investigate manageable and focussed aspects of a phenomenon rather than aiming for widely generalisable results. Generalisability may however be an interesting by-product and triangulation might validate such claims and offer insights into the results' adaptability within other contexts. With this respect, the HCI researcher may opt for techniques such as user evaluation studies, field observations, questionnaires, follow-up interviews and focus-groups.

Embedding qualitative findings within the requirements and design workflow is another major challenge. Finding a way to embed design insights in a practical and systematic way could encourage practitioners to consider the users' experience earlier on in the software development process. Research in HCI has flourished over the years and is readily available online and in print, however it is only recently that UX experts have been involved in major e-government initiatives (refer to the UK's *Government Digital Service Design Principles*[4] and the US's *DigitalGov UX Program*[5]). This thesis aims to make user experience considerations more ubiquitous within the requirements development process, as a core engineering discipline. The goal is to introduce a streamlined, immediate and reasonably indicative UX feedback cycle (on critical aspects of design) within the requirements development process – encouraging early, low cost and frequent design iterations based on simulated user feedback. A usable CASE tool is also essential to help designers, developers and policy makers compile, specify and manage system requirements in a collaborative manner.

Rogers et al. [139] argue that turning qualitative data into numbers may be a dangerous practice. Misrepresentation and manipulation to promote a theory can damage the research community's reputation, but also practitioners' deliverables and ultimately the end users' experience. For this reason, triangulation and evidence of corroboration is extremely important. Results should be presented in a genuine and transparent manner (e.g., providing the necessary statistical tests, medians and modes as well as means and raw figures when sample numbers are small).

3.1.3 Case study as a research method

Case studies allow the researcher to experience an event first hand while observing contextual nuances or as Flyvbjerg puts it, "*nuanced view[s] of reality*" [59]. Immersion does not necessarily need to result in lengthy and cumbersome field studies (e.g., using ethnographic or participant-observation data collection methods) and Yin argues that, depending on the topic, high-quality case studies may still be produced without spending too much time in the field. However, immersion gives a deeper meaning to the interpre-

[4]Government Digital Services, https://www.gov.uk/design-principles, (accessed January 2014)

[5]Government Services Administration, DigitalGov User Experience Program, http://www.usability.gov/how-to-and-tools/guidance/gsa-first-fridays-program.html, (accessed February 2015)

tative process and improvements or changes may also be studied across subsequent interventions. Case study research affords a learning experience to the researcher, acting as a pedagogical technique as well as a research method. Flyvbjerg [59] states that situating the learning process within the context of the problem domain can enhance the researcher's experience, as opposed to context-independent training. The latter is based on textbooks and information from secondary sources, through which "*virtuosity and true expertise*" can never be obtained. Context-dependent knowledge and expertise is at the "*centre of the case study as a research method*" [59]. Experiencing a phenomenon first-hand, revisiting the theory and re-applying it within a subsequent study can foster a deeper appreciation of the general research problem and to the specific issues faced by the stakeholders involved. Thus adopting case studies can also act as mitigation against the risk of reaching "*academic blind alleys, where the effect and usefulness of research becomes unclear and untested*" [59]. Using the case study method requires rigour in its undertaking, as in all other methods. Yin [176] argues that researchers must follow a systematic process in order to ensure validity and repeatability of the investigation.

According to Yin [176] case study research is one of the "*most challenging of all social science endeavours*" but one that can answer the 'how' and 'why' questions for a specific phenomenon and its treatment – potentially complementing other methods of inquiry based on statistical evidence. A main critique of the method is that case studies are mostly applicable in the preliminary, or rather exploratory phase of a research exercise, rather than in the explanatory phase [176]. The author argues that this critique, or rather prejudice, stems mainly from the idea that case studies are not conducted in a systematic and rigorous manner. Yin agrees that several case study-based research exercises were "sloppy" in their inquiry and he proposes that this was the direct result of a lack of methodological texts on the subject. Small-*N* studies are generally criticised for the lack of potential generalisation, especially when one case study is used ($N = 1$). In his book on the limitations of social research, Shipman [153] differentiates between cases and samples whereby cases are not designed for generalisation and abstraction but for reflective inquiry with a level of perceived relationship between researcher and researched – continuing that simply abstracting evidence and reporting it in "*academic jargon, ignoring its potential for helping those who provided the access, is exploitation*". This encourages the researcher to consider case studies as being more than simple studies, but as interventions to learn and potentially be of help to the stakeholders within the context — even if by reporting back the results "*as a courtesy*" [153]. Multiple-case studies can be used as a method to strengthen the theory, however Yin argues that, like other scientific methods of experimentation, case studies "*are generalisable to theoretical propositions and not to populations or universes*". By adopting case studies researchers can work towards analytic or theoretical generalisation, rather than statistical generalisation. Flybjerg [59] believes that the case study as a research method adds immense value to the researcher especially because it provides immersive first-hand experiences within the context being investigated. This benefit is compounded when research is tackled iteratively over multiple-cases with incremental adjustments to the theory being developed.

Considering HCI as a socio-technical domain, case study research should be adopted as an opportunity to learn more about a specific phenomenon, its reaction to applied theories and also about its

impact on stakeholders. The researcher must be open to unexpected knowledge and willing to pursue a thread that could lead to the discovery of a new phenomenon altogether. Choosing the case study as a methodology should follow a desire to learn through immersion while being willing to bend or drop pre-conceived ideas or assumptions as more experience is gained. This method can also serve as a channel for hypothesis testing and theoretical evolution through multiple interventions. This is synthesised by Flyvbjerg in the following statement: "*the advantage of the case study is that it can 'close in' on real-life situations and test views directly in relation to phenomena as they unfold in practice*" [59].

Case study research is therefore quintessential when the researcher needs to learn more about the socio-technical context and how it affects and is affected by the theory being developed. Quantitative methods can then be adopted to expand on, confirm or reject emerging aspects of a theory.

3.2 Methodology Adopted for this Thesis

Action research, coined by Kurt Lewin in 1944 [96] was the first methodology considered for this thesis, and according to Cohen et al. [33] it is a "*powerful tool for change and improvement* ***at the local level***" [emphasis added]. The intention was however to carry out interventions across different e-government contexts and scenarios to assess the techniques being developed by this thesis (involving different user groups and stakeholders). Action research was thus abandoned in favour of a multiple-case study methodology. This still ensures that the theory being developed is iteratively revised and potentially improved, while validating its applicability, practicality and adaptability across different contexts of use. Iterative interventions can also inform the development and evolution of supporting tools.

Action research and case study based research are similar in nature, both helping the researcher gain "*an in-depth understanding of particular phenomena in real-world settings*" [15]. The two however have some distinctive differences. Action research mostly starts off by identifying specific issues or situations within a social context. These are then studied and tackled iteratively within the same context. On the other hand, case study research begins with an interest in a set of phenomena [15] and the researcher examines how they affect or are affected by participants in a case study intervention, following a data collection protocol. The case study researcher can use the outcome of the studies for independent dissemination within the academic community, whereas action researchers are obliged to feed the results back into the social scenario in which the collaborative intervention has been conducted following an iterative process of amelioration [15, 33]. Collaborators in action research have an equal footing whereby everyone involved (i.e., researcher and researched) can influence the direction of the study. Action research suggests single-context interventions out of which context-specific observations may be extracted. These observations can be used to inform the academic community but more importantly to assist in the stabilisation or improvement of the social scenario in which the collaborative investigation is taking place. On the other hand, by adopting a multiple-case study method researchers seek to arrive to some form of analytical generalisation through cross-case observations.

Figure 3.1 outlines the overarching methodology adopted for this thesis. A multiple-case study approach was selected wherein each case study has its own strategy of inquiry – all contributing towards the development of *Sentire* – referred to as *the theory* for the rest of this chapter. This is not to be confused

with the output of a grounded theory based study whereby a theory emerges without any pre-established theoretical foundation or assumptions. Aspects of grounded theory were adopted in the initial stages of this thesis to learn about user attitudes towards enrolment processes.

For the first phase of this thesis (see Section 4.1) a series of semi-structured interviews were organised to construct a better understanding of user perceptions towards enrolment processes, including reported attitudes, behaviours and experiences across several types of online services. Batches of interviews were transcribed and coded iteratively using open and axial coding. This iterative data collection and coding process was monitored to determine the point at which no new substantial knowledge was being discovered and no major changes to the codebook were made (i.e., stabilised). Code categories were then adopted and operationalised during the user behaviour modelling phase (see Section 4.2), upon which the analytical and simulation elements of the theory were then based (see Chapter 5). The overall strategy adopted was to establish an initial theory and evolve it through a number of interventions across different contexts (i.e., multiple case studies).

Each of the four case studies presented in this thesis is unique in its own right, carrying distinct sub-goals (in line with overall objectives), involving different stakeholders, across various contexts of use and groups of end-users. Combining multiple methods allows the researcher to triangulate findings, construct a richer understanding of the domain while providing intermediary checks to highlight issues such as misinterpretation or miscalculation. Quantitative methods were adopted within each individual case study to evaluate the evolving user behaviour modelling technique by comparing predicted user behaviour against actual user feedback (i.e., online surveys). Statistical tests were also used throughout the thesis to assess the strength of (1) individual models generated for the different groups of users as well as (2) the associated data collection and model generation protocol. In Chapter 7 focus groups were used to uncover attitudes of younger user groups towards enrolment processes (i.e., undergraduate students) while in Chapter 8 questionnaires were used in an attempt to identify the existence (and extent) of a causal relationship between the introduction of a compulsory and high-effort enrolment process and its impact on the users' lived experience (ULX) (i.e., compulsory exam registration e-service for 16-18 year old students). Each intervention informed the development of the theory itself as well as the design of subsequent interventions. Intermediate results were also validated (and criticised) through peer-reviews and expert feedback.

The method adopted contains a significant path-finding element since no directly related research exists on the core constructs proposed within this thesis, including user experience modelling (see Chapter 4), model calibration (see Section 4.2), knowledge reuse (see Section 4.4) and ULX simulations as part of the requirements development process (see Section 5.4). Observations from the different studies inform a single underlying theory which is then reviewed in the final chapter through cross-case conclusions (see Chapter 10).

The multiple-case study method adopted follows Yin's guidelines [176] whereby first a theory is developed and cases are then selected together with corresponding data collection protocols. Case studies are conducted iteratively and not in parallel due to the possibility that one study may produce discoveries

that could lead to a re-design of the data collection protocol or to the selection of alternative subsequent cases. This is an important aspect of the multiple-case study method which suggests that insights and discoveries should not be ignored. Researchers are encouraged not to be rigid in their approach with the intention to accommodate their original theory or design, but should be ready and open for change based on the understanding that theories evolve as more knowledge is gained.

Each case study is reported individually upon which cross-case conclusions were derived. Following each study the theory was modified to reflect any lessons learnt. Case studies were initially selected based on prior knowledge and assumptions of the various contexts in which e-government services are conceived, designed and built, as well as the environments in which they operate. Four cases were ultimately selected through which the author was able to learn more about the implications of the theory being developed as well as the modelling and requirements elaboration techniques being proposed. Mixed methods were adopted in some cases, nesting other methods within the inquiry to drill down into specific observations or results (e.g., using quantitative methods to verify observations). Details about the use of mixed methods are given separately for each case study in Chapters 6, 7, 8 and 9 (whenever applicable).

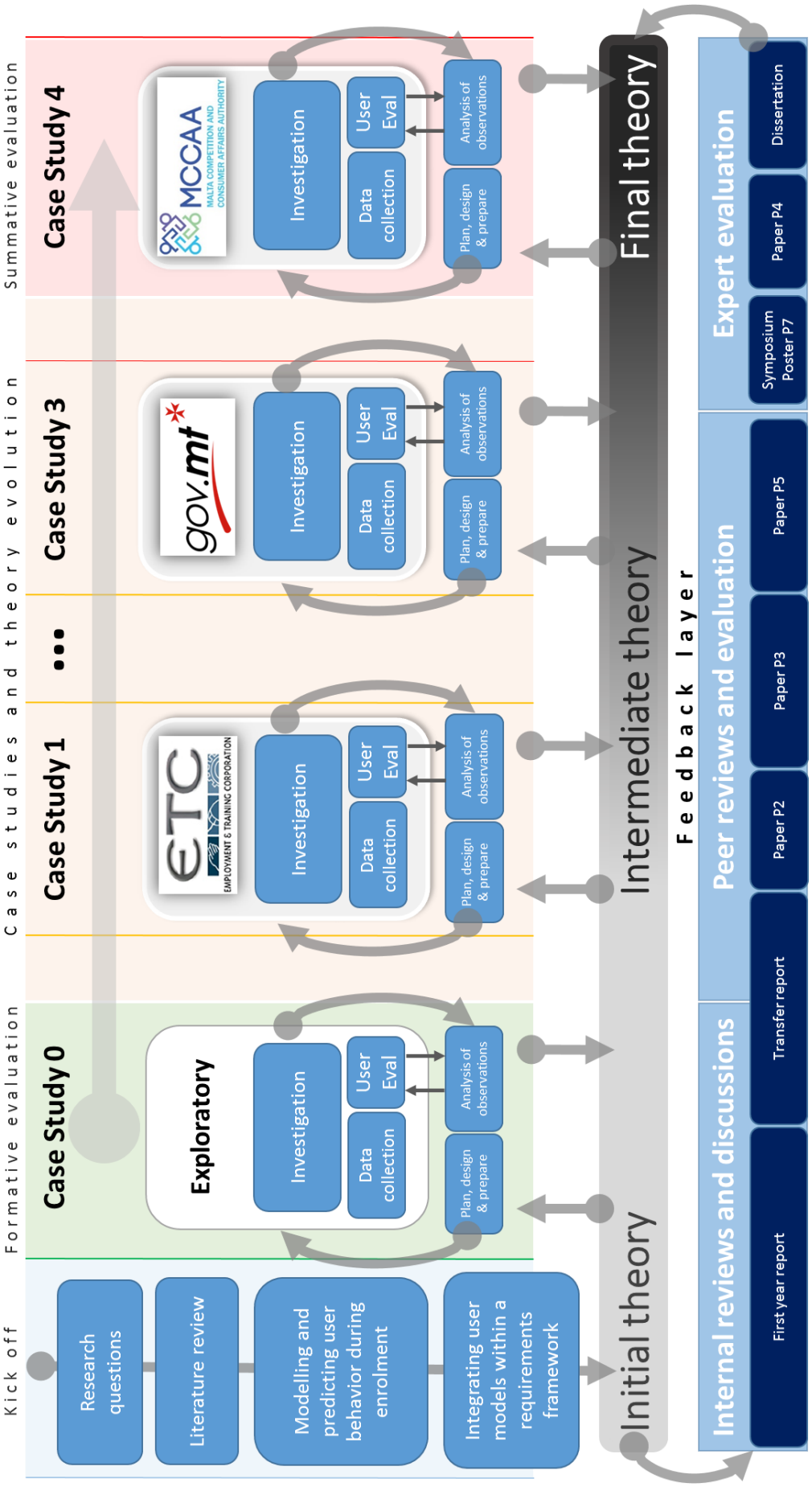


Figure 3.1: The overarching multiple-case methodology used for this thesis

The methodology adopted allowed for the evolution of the theory presented in this thesis (see Chapters 4 and 5) together with the associated deliverables as contributions for both researchers and practitioners. Figure 3.1 outlines the main stages that characterised a five year period. The research direction and initial research questions evolved primarily in the first 18 months throughout which relevant literature was reviewed, modelling techniques were investigated together with literature on requirements frameworks. The formative study (i.e., case study 0) helped to understand the mechanics of the evolving modelling techniques as used within a requirements development process. Feedback was received at different points in time, particularly through internal reviewing mechanisms (i.e., first year report and transfer viva report), supervisory meetings as well as informally from fellow researchers.

It was decided to move the investigation forward by establishing whether these techniques may have any impact in the real world. This led the author to plan a series of studies (i.e., cases 1 to 4) to better understand the nature of various real-world contexts, the interaction between stakeholders as well as the effect of the proposed techniques on stakeholders' practices. Each study was planned, designed and executed following a pre-established workflow (this is presented in each respective chapter). Following the completion of each study the data collected was then analysed in a systematic manner. This generally required the researcher to revise the developing theory (and supporting tools) to accommodate new insights. This in turn informed the planning stage for the subsequent case study within which the developing theory would be re-applied (more details in Chapters 6, 7, 8 and 9).

Quality assurance throughout the process was maintained through a feedback layer (shown in Figure 3.1). Following or during each case study the author published emerging findings, experiences and observations. This was an important mechanism by which valuable feedback was obtained from both the HCI and Requirements Engineering communities (both internal and external to UCL). This acted as an intermediary layer of assessment which allowed the author to learn more about the implications this thesis might have for both research and practice. Salient aspects of these papers and reports are reproduced in the respective chapters. In the final two iterations feedback was also requested from industry experts, including the authors of *Volere* (James and Suzanne Robertson), members of the Information Assurance Advisory Council (at the IAAC 2013 Symposium) as well as industry track reviewers at the 22nd IEEE International Requirements Engineering conference. The various stages outlined in Figure 3.1 also contributed to the evolution of the supporting CASE tools, presented in Section 5.7.

3.2.1 Exemplar study to test proposed design framework

The model was initially evaluated using an exemplar case study, namely the *employment license permit service* used in Malta. In his work Faily [52] suggests that exemplar studies provide an "*initial validation of the theoretical research contributions, and to put these contributions in context*".

At this stage the author had an early understanding of the research context as well as an incomplete view on how to model user behaviour and simulate user feedback for enrolment-specific design decisions. The initial inquiry involved a series of interviews with stakeholders from various government entities as well as end-user representatives, including regular and irregular immigrants seeking work in Malta. The *Volere* process was used to guide the investigation for a new (fictitious) e-service – *applying for an*

employment license online. This helped the author establish a better understanding of both the *Volere* process as well as on the potential impact that enrolment can have on the various user groups within this specific e-government scenario.

3.3 Summary

This chapter reviewed research approaches adopted in HCI as well as in the requirements and design disciplines. This informed the selection of research methods adopted for this thesis, including both qualitative and quantitative data collection and analysis techniques.

An outline of the research strategy is provided in Figure 3.1 and discussed in Section 3.2. This is mainly inspired by case study research (see Section 3.1.3) and includes a feedback loop with academic communities and industry experts to aid in the evolution and validation of both theoretical and practical contributions. The theory presented in this thesis (see Chapters 4 and 5) as well as the corresponding CASE tools (see Section 5.7.1) were developed iteratively following a series of interventions across different contexts (i.e., case studies).

Chapter 4

Understanding User Attitudes Towards Enrolment to Build Behaviour Based Models

This chapter presents a set of design factors that cause friction during enrolment. The first section of this chapter introduces the study that uncovered such factors whereas the second part of the discussion presents how such factors can be operationalised for modeling purposes. The discussion then considers a number of techniques that can be adopted to build behavioural models (around these design factors) to explain and predict users' perception of and reaction to existing and new enrolment processes across different contexts. A data collection and analysis protocol is also presented (i.e., user group calibration process) together with a discussion on its inherent risks and potential mitigation strategies. Finally, this chapter concludes by discussing the possibility of building a national or regional behavioural model knowledge base for reuse across e-government projects.

4.1 Identifying Design Factors within Enrolment Processes

4.1.1 Aims

The idea of empowering citizens and providing them with better access to government services has been around for a number of years, however the same mistakes are seen repeating themselves in new e-service deployments around the world. Security is treated as a secondary task and its impact on user experience follows as a mere afterthought, if at all. This thesis suggests that enrolment is a major hurdle which can influence the user's decision making process leading to e-service adoption (e.g., "*is it worth going through this enrolment process?*"). Excessive enrolment-related workload can overshadow any perceived benefits associated with an e-service, encouraging potential users to revert to traditional service delivery channels (e.g., snail-mail or in-person). This conflict between user goals (goal achievement) and organisational goals (higher levels of identity assurance) causes friction. When such friction is unavoidable (i.e., compulsory use) frustration and resentment towards the service provider may arise. Therefore, it is proposed that government entities cannot consider the *number of successful enrolments* as a sole metric for success. Sections 2.3.1 and 2.3.2 present evidence on the impact of enrolment processes on e-service take-up.

Before considering ways to improve the requirements development process for new e-services it was

decided to investigate and uncover design elements that may cause security friction during enrolment. For this reason a study was carried out to define and operationalise a set of enrolment specific design factors that could have a negative impact on e-service adoption. Maximising successful enrolments while striking a balance between acceptable levels of identity assurance (product owner requirements) and workload levels (on end users) is not trivial. The following question was thus posed: *What design decisions discourage users from completing the enrolment process for an online service*?

4.1.2 Method

4.1.2.1 Participants

A total of 20 participants were recruited for this qualitative study which consisted of a series of semi-structured interviews. Participants were in the 18 to 35 age range (median age was 33 years old) and had at least a secondary level of education. All participants were regular internet users (daily), 12 of whom were male and eight female. Convenience sampling was used since the study was highly exploratory in nature. However participants were also filtered through judgment sampling based on the perceived level of IT proficiency (i.e., basic, intermediate and advanced). Six participants were considered to have a basic level of IT proficiency (i.e., basic internet surfing), another six were considered to have an intermediate level (i.e., daily use of productivity tools), while the remaining eight participants were expert computer users (i.e., IT practitioners and researchers). This range of experiences provided a wider scope for discussion from which raw data was generated for further analysis.

Interviews were driven by several guiding questions (see Appendix A.1) in order to touch upon various types of services and their respective enrolment processes (e.g., e-banking and e-government). The interviews were designed to elicit general negative experiences and issues encountered during online enrolment. The main stages of this exercise are outlined in the following section (see Section 4.1.2.2). The main outcomes (a set of themes) are outlined in Table 4.1.

4.1.2.2 Process

Phase 1: Semi structured interviews Unlike thematic analysis (TA), a readily available and compiled corpus of data was not assumed and an incremental approach to data gathering and analysis was adopted – mainly inspired from the interleaved approach used in grounded theory [14, 31]. Rather than selecting one participant for each iteration, interviews were conducted in batches of participants across multiple rounds. Each round of semi-structured interviews produced a set of transcripts which were then analysed (phases two through five). The outcome of the analysis informed the planning of the next round (iterating back to phase one for another round of interviews).

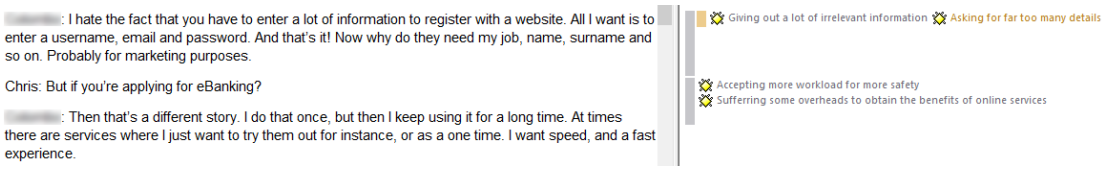
For this thesis two rounds of interviews were conducted. Following the first round of interviews the guiding questions were refined to reflect newly acquired knowledge while keeping in line with the initial goals. The study's topic was quite specific "*experiences with online enrolment processes*", and wide variances in people's responses was not envisaged.

Phase 2: Initial Coding All interviews were recorded, transcribed (in full) and coded. At first an exploratory stance was taken without adopting any specific coding discipline. This was taken as an

opportunity to get to know and understand the data. As coding progressed the author started to understand common elements across participants' views on the topic at hand, and this also afforded a coding-by-list approach (re-using previously used codes from the codebook). In general, incident-by-incident coding was adopted however in some cases participants went into some important detail during their interview (on some aspect or another), in which case line-by-line coding was adopted.

Atlas.ti, a qualitative data analysis tool, was used to store, analyse and annotate datasets (e.g., using memos).

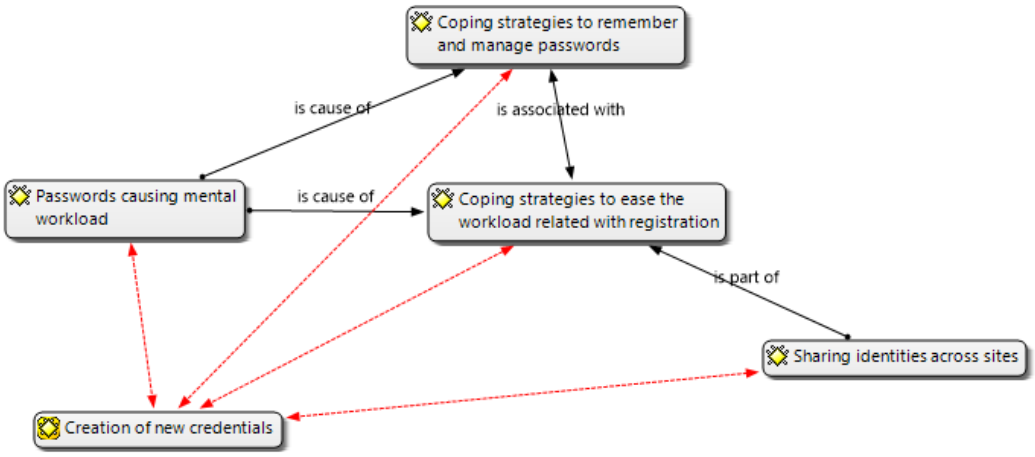
Figure 4.1: Initial coding using Atlas.ti



Phase 3: Axial Coding The author then proceeded with Strauss and Corbin's practice of axial coding [160] by relating identified code categories and sub-categories in an effort to synthesise and organise the data and to understand the main relationships, properties and dimensions of each category. During the initial coding stages the corpus of data is disassembled into separate and distinct codes whereby axial coding reconstructs the data as a coherent whole [31].

Dominant codes were immediately visible, and themes started to emerge, especially when relationships between codes and code families were constructed.

Figure 4.2: Emergence of related codes through axial coding using Atlas.ti

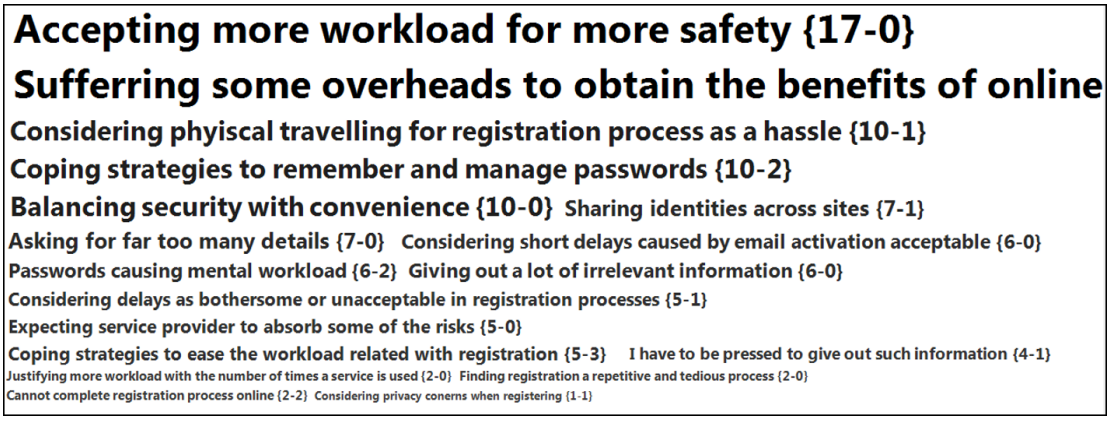


The built-in set of code-to-code link types (available in *Atlas.ti*) was used to create a networked view of the codebook (see Figure 4.2).

Phase 4: Reviewing Themes Given a set of candidate themes the author then started to investigate their level of representation for the identified concepts, using the associated codes' groundedness within

the underlying data as a primary indicator. The metrics used were the number of quotations associated with each code (for each theme) and the level of coupling between nodes (codes) within each theme. This examination was possible through *Atlas.ti*'s visual tools, including code tag-clouds in which the size of a code's tag reflects its frequency of use (see Figure 4.3).

Figure 4.3: Thematic analysis – codebook for the first round of interviews visualised as a code cloud (size of text indicates frequency of occurrence)



Through this visual assessment the author could assess the need for new nodes (themes and codes), merge closely related ones as well as modify and remove nodes altogether. Although not an exact science, it is still in line with Braun and Clarke's recommendation of determining a candidate theme's necessity or correctness by examining the supporting data (Is the data too diverse? Is there enough data to justify a theme's existence?) [18]. Braun and Clarke refer to the concepts of internal homogeneity and external heterogeneity to judge the value of a theme (Are the theme's codes coherent enough? Is the theme distinctive enough from other themes?). The researcher should stop when a satisfactory set of themes is obtained with a thematic map representative of the underlying data. One may risk wasting too much time in this process especially since recoding may be required if new themes are introduced.

Phase 5: Refining Themes Once satisfied with the thematic map Braun and Clarke [18] suggest that the researcher should go back to the codes and underlying data to identify the essence of each theme (unit testing) and that of the overall thematic map (system testing) – making sure that themes are healthy, in a way that they do not try to "*to do too much, or to be too diverse and complex*".

At this point the author decided to test for codebook saturation to determine the need for another round of interviews. Saturation is operationalised as "*the point in data collection and analysis when new information produces little or no change to the codebook*" [67]. Guest, Bunce and Johnson state that "*if the goal is to describe a shared perception, belief, or behaviour among a relatively homogeneous group, then a sample of twelve will likely be sufficient*" [67]. Based on this the author tried to determine whether the number of participants in the dataset was sufficient enough to get a sense of thematic exhaustion from the data. Although code exhaustion was evident following the first round of interviews (i.e., and by the fifth interview) the author believed that with

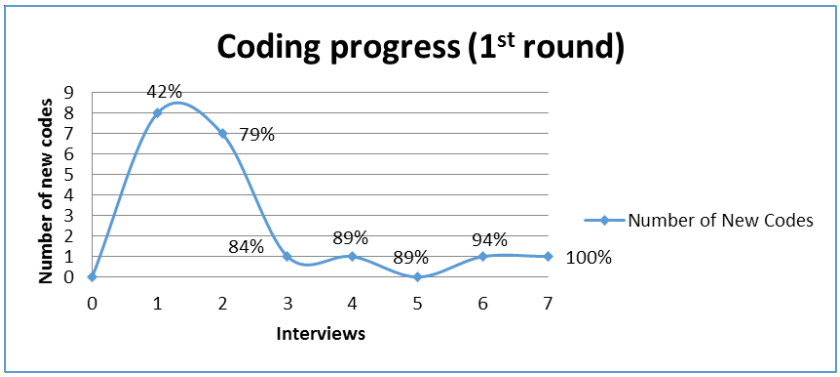
the resources at hand this thesis could benefit highly from a second round of interviews. This was also based on the fact that by now more experience was gained and a second confirmatory exercise would improve the author's understanding of the domain, confirm (or reject) existing themes and possibly provide new insights which may have been missed in the previous round.

The second set of interviews was based on an updated set of guiding questions (interview guide), informed from the experience and knowledge gained so far (see Appendix A.1). This enables the researcher to focus on pressing issues as well as on aspects that may shed more light on weaker themes (and miscellaneous codes) found on a thematic map.

An additional 11 participants were recruited for the second round of interviews, generating an additional three hours of transcribed audio. Phases one to five were repeated. As in the first round, saturation was evident by the fifth interview, and no new codes were uncovered after that (reverted to code-by-list since existing codes were mostly applicable to the new data). In this round, 25 codes (83%) were identified by the third transcript. All codes have been identified by the sixth interview, with no changes to the codebook thereafter. At this point the author concluded that more interviews would not yield any new knowledge. Some codes became predominant after only a few interviews, and only a few shifts in code groundedness were observed throughout the exercise.

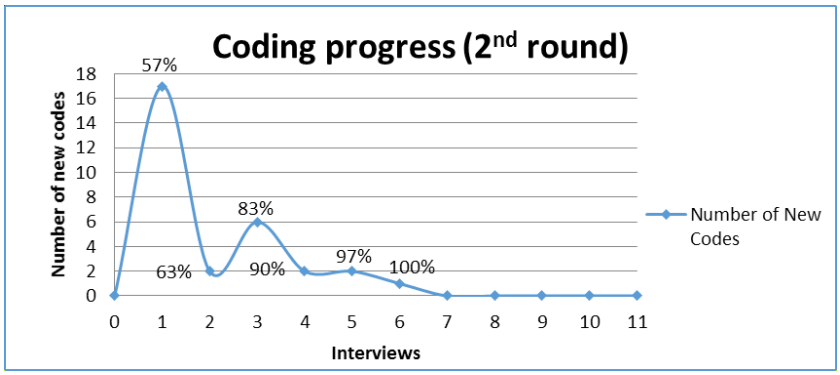
For the initial dataset (nine participants) a total of 19 codes have emerged. All codes have been applied to at least one transcript. Of these codes 16 (84%) were identified by the third transcript with an additional 3 (16%) codes identified within the next four interviews. Figures 4.4 and 4.5 give a % cumulative indication of codes found after each round of interviews.

Figure 4.4: Coding progress shows that most of the codes emerged by the third interview for the first set of interviews



Several authors have earmarked six as the number of participants needed to uncover the majority of themes within a research exercise of a qualitative nature [67]. For instance, Nielsen and Landauer [115] have stated that in usability testing the most severe issues are detected after the first few interviews and that most problems are then detected after interviewing three to five subjects [115]. In their findings it is clearly shown that by the third subject the majority of problems (codes) are uncovered, and after the fifth there is little variation on the proportion of usability problems being

Figure 4.5: Coding progress shows that most of the codes also emerged by the third interview for the second set of interviews



found. In this thesis, 84% and 83% of the codes in the respective datasets were uncovered by the third transcript, after which little to no changes in code rankings have occurred (i.e., considering the number of times specific codes were linked to different quotations from the transcripts).

Following the generation of the thematic map on this second set of data the author was sufficiently satisfied that in essence it was identical to the first map (even though there were some semantic differences between the two), and together with the saturation tests on both datasets the author was now confident that a representative blueprint of the entire corpus of data (in line with this study's objectives) was produced.

It was now time to move on with the process and start developing on these principles to assist designers and developers in their decision making process: moving from research to design.

Phase 6: Implementing design principles At this stage, a set of themes was established which in themselves can be considered to be a set of design principles (see Table 4.1). At this point the author could have opted to stop and report these principles using an analytical narrative to "*convince the reader of the merit and validity of [the] analysis*" in a concise, interesting, logical, evidence-based and coherent way [18].

However the author opted to go beyond an analytical narrative and decided to use the resulting themes, or design principles, to build a practitioner-oriented decision support system (and a computer-aided software engineering (CASE) tool) to help project teams make informed decisions when designing enrolment-based e-government services (see Chapter 5).

In principle these themes were operationalised in order to be able to quantify them within known constraints, measure users' reaction towards them (through an empirical exercise in situ or in a lab environment) and use this data to build prediction models to assess new scenarios (i.e., new e-services) in terms of the users' willingness to complete the primary task online and the associated level of perceived workload.

4.1.3 Study outcomes

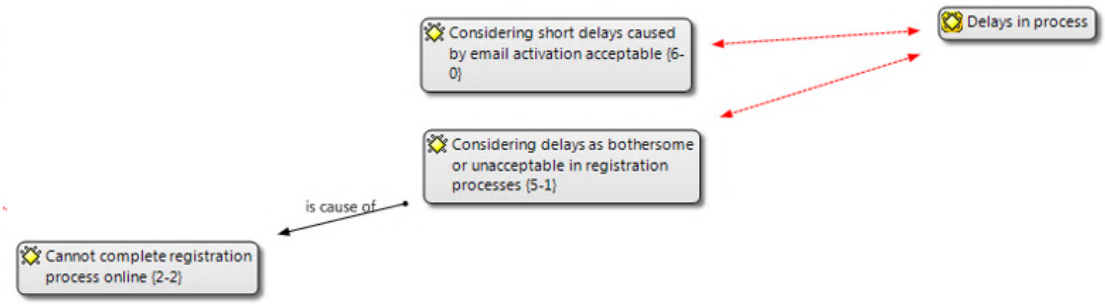
After coding and analysing these two data-sets, six common meta-themes emerged. These were then used to elicit a set of high-level design factors that can explain friction caused by enrolment steps within the primary task at first-interaction with an e-service. A *primary task* incorporates a set of steps which a user needs to execute before attaining a goal (e.g., apply for a birth certificate), excluding steps that need to be taken by the service provider to honour user requests (e.g., verify request and issue certificate). On the other hand *secondary tasks* are supporting activities which are however not essential for the attainment of the users' immediate objectives (e.g., creating an account for future interaction). Sasse and Fléchais [144] state that human behaviour is goal driven, and any action that takes the user away from her goal, or that leaves the user to choose between complying with the security requirements imposed by a system or getting the job done, is essentially bad design. Systems must be designed in a way that encourage the completion of primary tasks in the most effective, efficient and secure way. First interaction represents a user's initial attempt to access and use an e-service in order to accomplish a primary task. Secondary tasks at first interaction (e.g., enrolment) may introduce perceptions of workload whereby an internal cost-benefit assessment may lead users to (1) complete the primary task online (i.e., adopt service) [12], (2) adopt an alternative service provision channel (e.g., in person, by phone) or (3) abandon the task altogether (i.e., non-compliance).

The following is an outline of the six meta-themes identified, together with supporting evidence and intermediate outcomes.

1. Frustration caused by design elements that introduce *delays to the primary task* (T1

   Delays are defined as disruptions that temporarily suspend a primary task for a given period of time, but without disrupting daily routines and other unrelated tasks. Although waiting is required, users do not have to go out of their way to complete the primary task (e.g., *slight delay* – activation email sent after a couple of minutes *or major delay* – three day period for account activation following manual verification). Once the waiting period is over (e.g., activation email received), users may then resume with the original task. This is related to and contrasted with theme T2.

Figure 4.6: Theme discovery through code maps (first coding round) – code family for delays



   Initially a parent theme *Interruptions* was considered, however the distinction between minor and major interruptions started to emerge, especially through the idea of "*show-stoppers*" ver-

Figure 4.7: Theme discovery through code maps (second coding round) – code family for delays

sus "*slight delays*". These two types of interruptions have therefore been made distinct via two separate themes: (1) frustration caused by delays and (2) frustration caused by interruptions to daily routines. These two interruption types have a different impact on the users' lives, meriting distinct treatment during the modelling phase.

Participants seemed to be quite forgiving when short delays to the primary task were introduced (e.g., account activation via email verification), however less so when major delays are enforced (e.g., account activation carried out by service provider).

- "*It doesn't bother me. Generally after you click you're granted access on the next step.*" [Participant 16]
- "*If it's 1 minute I don't mind, but 3 days yes!*" [Participant 10]

2. Frustration caused by design elements that *interrupt the primary task and disrupt daily routines* (T2)

   This theme refers to disruptions that temporarily suspend a primary task but which also disrupt daily routines and other unrelated tasks. When interruptions are introduced, users are required to go out of their way to complete the primary task (e.g., user is asked to visit a government office in person to complete the enrolment process). Interruptions may also be combined with delays (e.g., in-person service enrolment (*interruption*) does not grant immediate access to the e-service since a three day activation period (*major delay*) is required)

Figure 4.8: Theme discovery through code maps (first coding round) – code family for interruptions

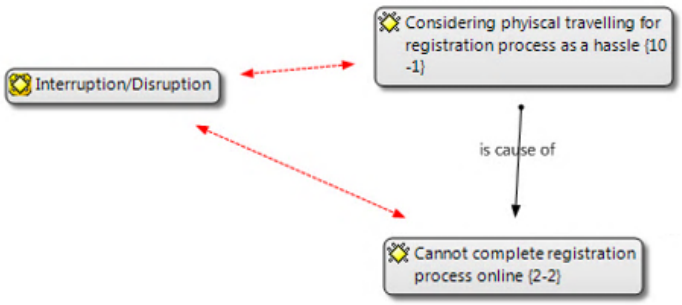
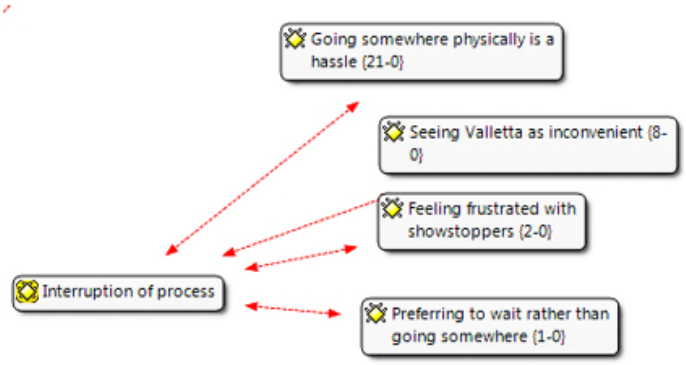


Figure 4.9: Theme discovery through code maps (second coding round) – code family for interruptions



The following are excerpts from the data illustrating codes related to this particular theme.

- "*But while registering I'd find that I can't complete the process I would go crazy!*" [Participant 10]
- "*No. To tell you the truth I clicked on it, thinking that I could do it online, but then I realised that I had to go myself to Evans building [Valletta]. That is a hassle.*" [Participant 18]
- "*It does make sense since it is not a show-stopper, if then you cannot do anything, then it's a different story.*" [Participant 1]

3. Frustration caused by the *number of form fields in online forms to recall* (T3)

   The amount of perceived work associated with enrolment appears to be affected by the tediousness associated with online forms, including the number and type of data items requested.

Figure 4.10: Theme discovery through code maps (first coding round) – code family for items to recall in enrolment form

Figure 4.11: Theme discovery through code maps (second coding round) – code family for items to recall in enrolment form

The following are excerpts from the data illustrating codes related to this particular theme.

- "*I hate the fact that you have to enter a lot of information to register with a website.*" [Participant 2]
- "*It's not just the basics, but additional data. They seem like they want to know about you. I registered for Shelfari... they asked for just some details, like email address and passwords, nothing else. There was no need to submit more information. At times you want to register quickly and not fill up forms.*" [Participant 15]

4. Frustration caused by the *number of new credentials to create*, irrespective of strictness imposed by the security policy (T4)

   Participants expressed concern with the amount of secrets they have established with different online service providers. Coping strategies are generally adopted to counteract this issue, including password externalisation, adoption of credential managers, data re-use across different sites and so forth.

Figure 4.12: Theme discovery through code maps (first coding round) – code family for items to generate

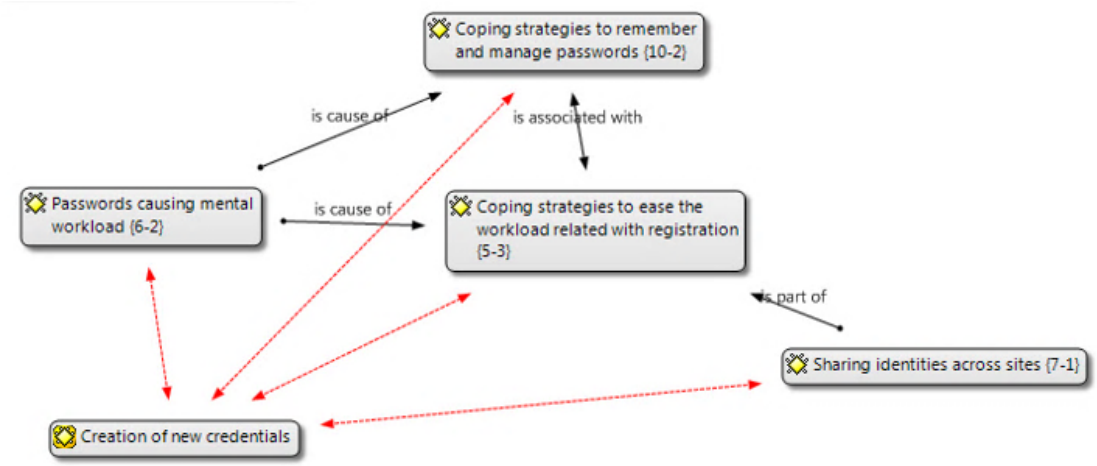
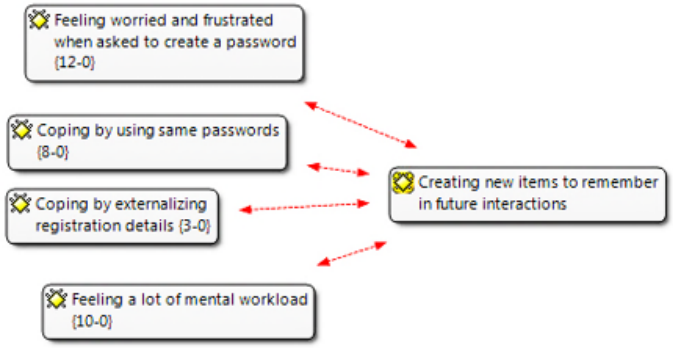


Figure 4.13: Theme discovery through code maps (second coding round) – code family for items to generate



The following are excerpts from the data illustrating codes related to this particular theme.

- "*The fact that you have to create a username bothers me. It's simpler to simply use the email address.*" [Participant 11]

- "*...you end up using the same pool of passwords. I stratified them, I did variations on 3 passwords.*" [Participant 1]
- "*What I find very convenient is authentication with open ID. I use SUSEGallery, where I can manage various appliances and machines, and I only need to sign-in one time. It could be Google ['s identity], and usually, I sign in there and that helps me sign in to other services.*" [Participant 7]

5. *Service compulsion and frequency of use (type of service)* (T5)

   Participants seemed to accept more workload and be more forgiving when services are (1) used more frequently and (2) carry a higher level of compulsion (i.e., users are legally required to comply). The higher the frequency of use and service compulsion, the more elastic participants seem to be (i.e., accepting higher levels of perceived workload).

Figure 4.14: Theme discovery through code maps (first coding round) – code family for type of service

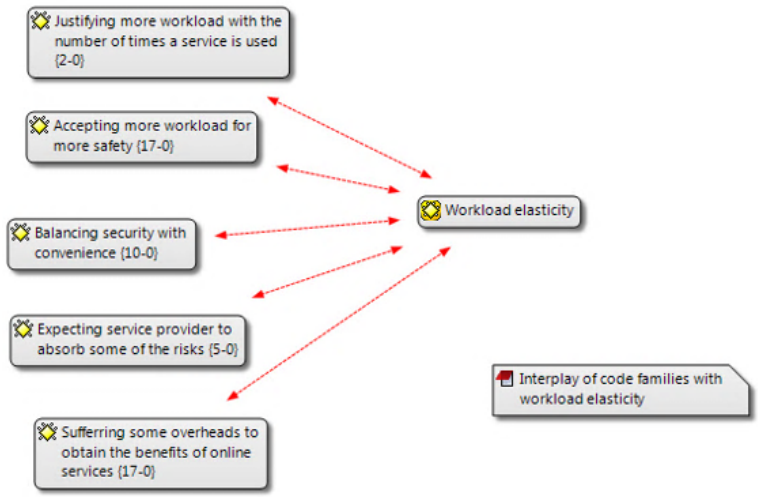
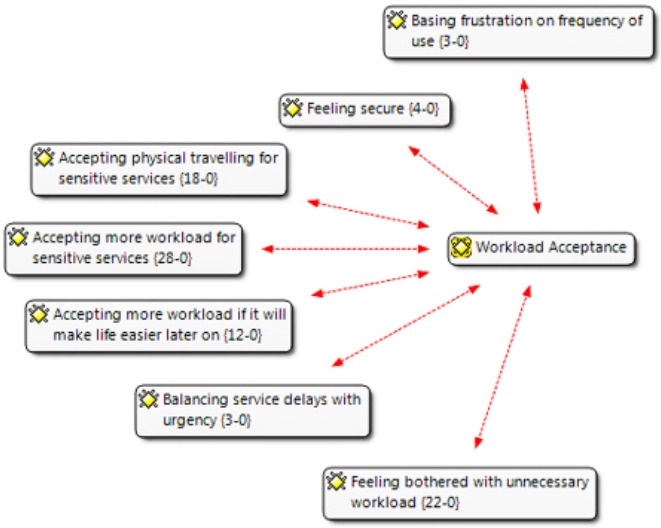


Figure 4.15: Theme discovery through code maps (second coding round) – code family for type of service



The following are excerpts from the data illustrating codes related to this particular theme.

- "*Ok how many times would you need a birth certificate? It depends on the amount of usage. If it saves me from queueing then that's worth the while.*" [Participant 2]
- "*Primarily I understand the need for sacrificing some things in order to benefit from online services. Otherwise you have to go in person, or write a cheque or correspond manually.*" [Participant 5]
- "*If you do it once, and then you've got a life-time paying income tax online from home..*" [Participant 4]

6. The number of *alternative providers* (T6)

   Participants expressed their frustration with having only one provider for e-government services (i.e., at a local, regional and national level). This is contrasted with commercial services where several providers may exist, offering comparable services with less adoption hassle. Although this theme was not strong, it offers interesting insights into the impact that e-government services may have on users' lived experience. Switching service providers may not be an option, and this may also result in frustration and possibly resentment towards the service provider.

   On the other hand, if public sector service providers do not build effective online services that encourage adoption and user retention, they would have to absorb the cost of handling transactions offline via traditional channels (e.g., walk-in, phone and snail mail), as well the cost associated with supporting the inefficient online services (e.g., help-desks and damage to political goodwill).

   - "*It depends how urgent the matter is. If it is urgent then I would simply find another service provider (online service).*" [Participant 15]
   - "*Personally the cost of registration vs the benefit obtained wasn't that good compared to other services.*" [Participant 5]

A number of low-level design aspects that affect the users' experience were also suggested by some participants (e.g., usability issues with password strength, interface design, lack of clear instructions, help-desk support and technological compatibility issues).

- "*Things things things ... a lot of words, small fonts... it is not clear at all. It's not clear how to apply for services.*" [Participant 16]
- "*Online it's more [finger snap] immediate. [If] you need some support, and you call and they tell you that no one's there to help...*" [Participant 10]
- "*I hate it when sites give you passwords which you cannot change, especially when it is alphanumerical*" [Participant 20]
- "*The thing that annoys me most is the fact that you have to create passwords on sites you rarely use. For example, those you use 3 times a year. Those are the most difficult*" [Participant 10]

- "*No. They were using some form of browser which I don't have, and since I use IE, I tried to register, but I couldn't.*" [Participant 18]

Although these insights were derived through a systematic process and are also highly intuitive, they are not pragmatic enough to guide designers to find the right balance between the required level of identity assurance and a positive user experience. This makes it even more complex when considering that attitudes and reactions towards the same design decisions may vary across user groups and across e-service contexts. This complexity is further compounded by themes T5 and T6 which can be considered to be behavioural modifiers – exerting influence on the users' decision making process. A modelling technique was devised (see Section 4.2) in order to build objective and quantitative representations explaining user groups' reactions to and perceptions of different enrolment scenarios within different e-service contexts. This technique produces statistical user models that can explain users' reactions towards the different design factors outlined in Table 4.1, which models could then be used to predict the impact of enrolment-specific design decisions for new e-services across the various targeted user groups. The design factors outlined in Table 4.1 contribute towards sub-research question SRQ2 (see Section 1.2) – *which enrolment-specific design factors contribute to friction?*

Table 4.1: Enrolment-specific design factors, operationalised for modelling purposes

Theme	Design Factor	Description	Type
T1	Delays (*D*)	Disruptions that delay goal achievement, but without disrupting daily routines and other unrelated tasks. Although waiting is required, users do not have to go out of their way to complete the task (e.g., *slight delay* – activation email sent after a couple of minutes *or major delay* – three day period for account activation following manual verification).	Ordinal[1]
T2	Interruption (*I*)	Disruptions that delay goal achievement while also disrupting daily routines and other unrelated tasks. When interruptions are introduced (I = true), users are required to go out of their way to complete the primary task (e.g., user is asked to visit a government office in person to complete the enrolment process). Interruptions are combined with specific delay intensities (e.g., D = major, I = true).	Nominal[2]
T3	Items to Recall (*ItR*)	The total number of identity-related attributes requested from the user (e.g., enrolment form fields such as addresses, contact numbers and dates)	Interval
T4	Items to Generate (*ItG*)	The total number of secrets that need to be generated by the user and shared with the service provider (e.g. usernames, passwords, secret questions/answers and PINs)	Interval
T5	Type of Service (*ToS*)	Users seem to accept or are ready to endure more workload when a service is compulsory (e.g., penalties apply for non or untimely compliance) and also when its frequency of use increases (e.g., user is legally bound to submit a form on a quarterly basis, as opposed to one-off services)	Ordinal[3]
T6	Replaceability (*R*)	Users will more likely abandon a task if an alternative and equally reputed service provider provides an equal or comparable service with lower barriers to entry (i.e., perceived workload)	N/A[4]

[1] Possible values: none, slight, major
[2] Possible values: true, false
[3] Possible values: Type 1, Type 2, Type 3 and Type 4
[4] Choosing an alternative (competing) service provider is difficult or impossible within an e-government context, therefore this factor has not been considered for modelling purposes

The **Type of Service (ToS)** design factor is an ordinal variable prescribing four types of generic and channel-agnostic government services (i.e., could be provided offline or online). Each type represents a service with a different level of compulsion and usage frequency.

Type 1: Citizens can do without this service, with no legal implications (e.g., government gazette/newsletter).

Type 2: Service is required a few times in a lifetime (e.g., applying for a good conduct certificate).

Type 3: Service is required once or more per year within a legal time frame (e.g., individual tax returns).

Type 4: Service is required regularly within legal time-frames (e.g., city centre congestion charge).

Services bound by a legal time frame are generally enforced through penalties (e.g., interest or late payment fee) however service providers may also encourage compliance through rewards (positive reinforcement). An interaction was noticed between the level of workload users are willing to endure and the benefits obtained after enrolling for a service. In theory, the higher the ToS is, the larger the benefits are if one enrolls for the e-service channel. In doing so, one would be lowering the chances of negative consequences (such as penalties for untimely compliance) while minimising the hassle associated with the need to visit a government office in person at each interaction. Workload is the combined effect of the other design factors identified during the coding exercise (see Chapter 4.1) however its potential impact on users' willingness to use an e-service can only be determined when the type of service (ToS) is established. People might accept more workload to enrol for an e-service when the service provider requires (1) frequent interaction and (2) strict compliance, irrespective of the delivery channel(s) used (e.g., in person, by post or online). The higher the frequency of compulsory use is, the higher the perceived convenience becomes for the adoption of an e-service (as opposed to traditional channels). This may result in people accepting more workload in the enrolment process, potentially with reluctance and contempt.

On the other hand, **Replaceability (R)** comes into effect when alternative service providers exist. For instance, if a delay exists during the enrolment process for a specific e-service (e.g., 3 day verification period) one may simply decide to move on to the next best reputable e-service provider. This also holds true for most of the other design factors (at different levels). For this reason replaceability can also modify the effect that the other design factors have on the users' decision making process. In e-government there is no real competition and thus replaceability has minimal to no effect on user behaviour (i.e., switching between service providers would be difficult or impossible). In the e-government domain users can generally choose between using an e-service or its manual counterpart, both offered by the same service provider.

Based on these findings, two metrics have been developed – *given an enrolment process for any given e-service*:

1. What is the user's level of *perceived enrolment workload* (PEW)?
2. What is the user's *willingness to complete the primary task* (WCT)?

These metrics can be represented through the following functions:

$$PEW = f(D, I, ItR, ItG) \tag{4.1}$$

$$WCT = f(PEW, ToS, R) \tag{4.2}$$

At first one may believe that WCT and PEW are linearly related, whereby an increase in perceived workload would result in a decrease in the user's willingness to complete the primary task online. This is however not always true especially for services for which users are willing to endure more workload to obtain additional convenience (e.g., *ToS* 3 and *ToS* 4 services). Even if WCT for a specific user group

turns out to be encouraging (i.e., high), the *PEW* value should still be considered since it can shed more light on the potential impact of a proposed enrolment process on the users' experience. High levels of *PEW* when *WCT* is high (e.g., in a *ToS* 4 service) may cause security-induced friction, which could in turn result in feelings of frustration and resentment towards the policy maker. High *WCT* does not necessarily equate to a positive user experience, and multi-dimensional workload rating mechanisms can help determine the nature and extent of the impact an e-service enrolment process may have on different groups of users.

4.1.4 Abstracting low level details for modelling purposes

Without detracting from their importance, it was decided to abstract away from low level design details (derived via the coding process) and synthesise outcomes using six high level themes (T1–T6). This in turn resulted in a set of broad design factors (see Table 4.1) underlying the modelling process discussed in Section 4.2. This decision is based on the following arguments:

1. Modelling the entire spectrum of design considerations (uncovered from this exercise along with generic HCI and HCISec heuristics) would result in a large and unmanageable problem space. User modelling is a complex activity, and this thesis does not aim to predict the entire spectrum of user experience issues surrounding enrolment.

2. Introducing more modelling variables (predictors) would generally result in the need for additional treatments to the data in order to generate enough information to explain the impact of each variable on the expected outcome. The more treatments are required, the longer the data collection process becomes. Since data collection for user modelling revolves around participant involvement one needs to be careful not to end up with an expensive and time consuming process. This requires a balance between the amount of data treatments required (through which behavioural models are generated) and the quality of resulting models.

3. User models are not meant to replace heuristic evaluation and discount usability methods suggested by the HCI community (i.e., practice and research). The latter techniques are meant to capture generic usability issues and are introduced once low or hi-fidelity prototypes are available. User models on the other hand are meant to highlight egregious design issues, in this case about perceived workload associated with enrolment processes, and predict their impact on the users' lived experience – at the requirements stage when they are still relatively cheaper to rectify, yet difficult to capture.

4. Enrolment-process related perceived workload may affect user behaviour before user interaction takes place. The length and complexity of enrolment forms and the type of enrolment data required may lead users to abandon the task before any information is actually submitted. Even if security policies are lenient (e.g., allowing weak passwords) users will still be discouraged from enroling if the process is perceived as unnecessarily laborious.

5. Modelling techniques should not tackle everything at once – the problem space grows exponen-

tially with the introduction of new predictor variables. Therefore this thesis proposes a protocol through which other behavioural models may be developed. Such models could tackle other critical aspects of e-service design such as privacy concerns and password policies. This thesis suggests a divide and conquer approach to user modelling, whereby further models can be generated, validated and added to the Calibrated Persona construct.

4.2 Building User Group Behavioural Models for Reuse

Participants have mainly expressed their concerns on explicit factors surrounding enrolment processes and by observing these as generic guidelines one would automatically be working towards producing workable and acceptable e-services, in which security goals do not overshadow, but rather enable users' primary goals. However there is no systematic way by which one could assess the impact of the various design factors (at different intensities) on the user's experience and on the e-service itself (e.g., *activation by email* vs *activation by post*). This is especially important when a minimum level of identity assurance is required and omitting enrolment is not an option. The author believes that thematic maps or analytical narratives do not provide practical and objective design guidelines and for this reason the themes outlined in Table 4.1 were used to build a quantitative model with which designers could objectively measure and predict the impact of specific design decisions on users.

The author hypothesised that through these design factors a set of re-usable statistical models could be built to explain the different user groups' reactions towards enrolment process variants associated with different types of e-services. This hypothesis was based on the understanding that different user groups do not exhibit the same behaviour when presented with an enrolment task. For instance some people might not be too bothered with a specific level of security-related workload while others would immediately shy away from the same demands. This is based on a further hypothesis whereby it is believed that different people place different weighting on the various workload dimensions. There needs to be a practical approach by which this information is conveyed to practitioners rather than basing design decisions on assumptions and subjective generalisations. For this reason a systematic calibration exercise was devised to shed more light on the various user groups and their behaviour in different enrolment situations. Different enrolment processes may contain varying intensities of the design factors identified earlier (listed in Table 4.1) and different groups of users may react differently to such factors across e-service contexts.

The User Group Calibration (UGC) exercise is carried out on groups of people who are potential users for a given system. This measures their attitudes (based on behaviour) towards different enrolment processes, their willingness to complete the task (at the four type of service (ToS) levels), while capturing NASA-TLX specific workload measurements for a series of fictitious tasks. Given specific enrolment process configurations (adopting various combinations of design factors and intensities thereof), the UGC exercise provides a set of user-specific measurements which includes (1) the impact on the users' willingness to complete the primary task and (2) workload related information based on Hart and Staveland's NASA-TLX [70]. This information is collected from a number of UGC participants falling under a particular user group under investigation which is then processed in a statistical package to fit

two regression models (i.e., a multiple linear regression model for perceived workload data and a binary logistic regression model for the users' willingness to adopt the e-service, explained by the probability that a user will complete the enrolment process). These models provide regression coefficients which explain the user group's attitudes towards the different enrolment-specific design factors. These coefficients could then be used to calculate the potential impact of different enrolment process configurations on users.

4.2.1 Modelling for prediction

Several machine learning techniques and algorithms exist to predict outcomes from input parameters, including Artificial Neural Networks (ANN), k-Nearest Neighbour (k-NN) and Support Vector Machines (SVMs). Machine learning techniques are generally adopted as black-boxes as opposed to the more traditional statistical techniques, nonetheless they are still used in both classification and regression problems (i.e., for categorical and continuous variable outcomes respectively). Breiman [19] contrasts two modelling cultures: (1) *data modelling* (i.e., traditional statistical techniques) and (2) *algorithmic modelling* (i.e., machine learning techniques). and lists the following major differentiating factors:

- *Data modelling culture*
    1. **Black-box:** no
    2. **Model validation:** goodness of fit tests and residual examination
    3. **Estimated population:** majority of statisticians
- *Algorithmic modelling culture*
    1. **Black-box:** yes
    2. **Model validation:** predictive accuracy
    3. **Estimated population:** minority of statisticians

Breiman [19] continues to argue in favour of algorithmic modelling, stating that data modelling has kept statisticians from using more suitable algorithmic models while preventing statisticians from working on exciting new problems [19]. A level of path-finding was required in order to understand which of the identified predictors (i.e., design factors listed in Table 4.1) are actually meaningful for different groups of users and across contexts of use. The goal was not to merely predict outcomes but to understand the nature of the data and the interaction between the various predictors through a complete and transparent view of the underlying modelling process. The willingness to complete the task (WCT) and perceived enrolment workload (PEW) are both dependent variables whereby the design factors identified in Table 4.1 are on the other hand independent variables (predictors). Based on this it was decided to adopt multiple-linear and binary-logistic regression modelling (as white-box techniques). White-box techniques also provide more control, or rather, more insights throughout the entire model-fitting process (e.g., variable significance testing informing which predictors to include or exclude for specific groups of users). Furthermore, regression models are quick and easy to generate, test, use,

troubleshoot and deploy as part of a wider design technique in which immediate and repeatable results are desirable for effective decision making. In a separate study [108] the authors observed that ANN (as a machine learning technique) does not produce better results over multiple linear regression, especially when the underlying data is linear in nature. Dreiseitl and Ohno-Machado [46] also observed that there exists a 5:2 ratio for which the adoption of artificial neural networks does not provide any statistically significant benefit over the use of logistic regression analysis. The authors argue that the popularity of regression is attributable to its ease of use. By mid-2014 there were approximately 500,000 publications adopting regression modelling in PubMed's medical publication repositories[1]. Neural networks were used in approximately 40,000 publications, followed by decision trees (approx. 10,000), support vector machines (approx. 7,000) and finally k-nearest neighbours (approx. 5,000).

The following sections will introduce the necessary regression techniques that will be used to model users' reactions to and attitude towards enrolment process design factors across different types of e-government services. To maximise process transparency this chapter will also outline different statistical tests that can be applied in an effort to disclose the strength and validity for each model produced.

4.2.2 Regression models

4.2.2.1 Linear regression

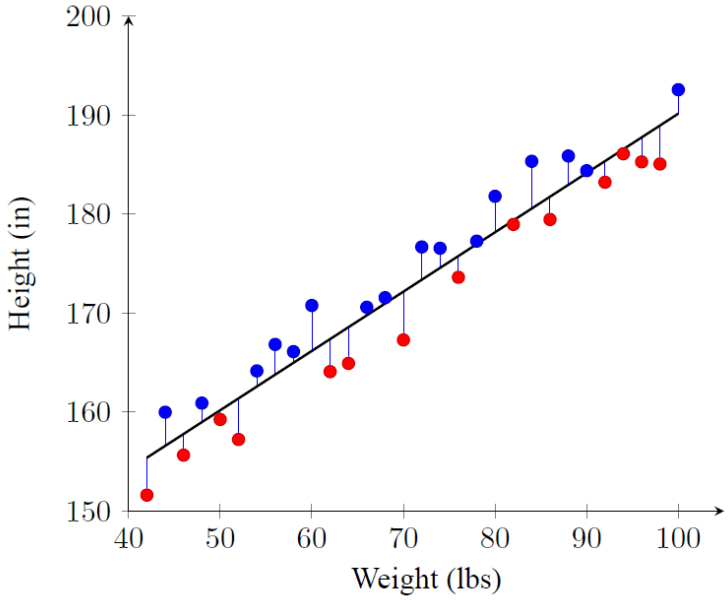
Understanding correlation between two variables is a useful practice however predicting one variable from another presents the researcher with the possibility to learn more about unknown and potential scenarios [57]. By fitting a model to the data the researcher can then attempt to predict values of one variable (dependent variable) from other variables (independent variables or predictors). Simple regression is adopted when predicting an outcome from one predictor, whereas multiple regression is used when there are various predictor variables. Thus by fitting a model to the data one can then (1) explain the nature of the data and (2) predict outcomes for specific scenarios.

The data itself dictates the modelling strategy to adopt. A linear model is used to describe the relationship between variables using a straight line (over the observed data), and thus this is represented with a simple linear equation. This could be used to explain, for instance, the relationship between the chances of rain and the time of year. Models do not always fit perfectly to the observed data and it is up to the researcher to decide which model best describes such data. For this purpose there are multiple tests one can conduct to determine the line of best fit. A case in point is the sum of least squares, wherein the line (model) that goes through (or is closest to) most of the data points in the observed data is selected. This test helps to minimise the distance between actual data points and the line representing values predicted by the model. These differences are referred to as residuals [57] (see Figure 4.16). The less residual values a model leaves the more representative it is to the underlying data, minimising over and under estimations. The residuals are then squared to avoid cancelling of negative with positive residuals and the results are added to produce the residual sum of squares. The smaller this value is, the better the model is in representing the underlying data. This method is referred to as the method of least squares. Goodness of fit tests would then give a *grade* to the best fitting line indicating the variance in outcome

[1] PubMed.gov, http://www.ncbi.nlm.nih.gov/pubmed/, (accessed April 2014)

that this model is able to explain (R^2) while the *F-test* explains the model's improvement relative to how much error remains [57]. In principle, R^2 is the squared value of the Pearson Correlation Coefficient between the observed data and the model's predictions. These tests are used to highlight the best fitting model which can then be used to predict possible outcomes based on past observations.

Figure 4.16: A linear regression model for medical measurements on a group of students (*weight* vs *height*). Residuals are marked as vertical lines between actual data points (dots) and the model (line) (Image source: http://goo.gl/Cb0CJ3)



Simple regression is represented by the following model for a random sample of the population.

$$\hat{y}_i = b_0 + b_1 X_i \tag{4.3}$$

$\hat{y}_i$ is the outcome variable (e.g., *expected* height of participant i) which is predicted by multiplying b_1 (the gradient of the model) by X_i (the weight of participant i) and adding the resultant value to b_0 (the point at which the model intercepts the y axis – the intercept). Table 4.2 outlines the values associated with a prediction model, followed by an explanation of how this can be used to predict new outcomes.

Table 4.2: Parameter estimates for weight (outcome variable) and height (predictor) data

Parameter Estimates		
	Coefficients	Sig
Intercept (b_0)	-126.715	.000
Height (b_1)	3.708	.000

The expected *weight* when *height* is 0 is -126.715, which clearly is not a possible outcome. This is because only measurements from students who stood between 50 to 72 inches tall were taken during this particular study. For every additional inch in *height*, *weight* is expected to increase by almost 4lbs (3.708). So when the *height* (X_i) is 60 (inches) then *weight* ($\hat{y}_i$) is expected to be 95.77 (lbs) as shown in

equation 4.4.

$$Weight(expected) = -126.715 + (3.708 \times 60) = 95.77 \tag{4.4}$$

Simple regression helps the researcher understand the nature of the data and through the generated models one could then extrapolate and gain insights into new scenarios by modifying the predictor variables to approximate possible outcomes. Nonetheless prediction quality highly depends on the quality of the models and the underlying observations.

4.2.2.2 Multiple regression

Multiple regression extends the principles applied for simple regression by introducing more than one predictor (independent variables) into the equation. These predictors will be considered when fitting a model to explain the observed data across multiple dimensions (e.g., *weight* vs *height* vs *age group*). The multiple regression model for a random sample of the population is shown below, whereby "*we seek to find the linear combination of predictors that correlate maximally with the outcome variable*" [57]. Instead of a line, the model is represented by a regression plane (as shown in Figure 4.17). Model fitting adopts the same strategy of minimising residual error (between actual data point and the regression plane).

$$\hat{y}_i = b_0 + b_1 X_{1_i} + b_2 X_{2_i} + b_3 X_{3_i} + ... + b_n X_{n_i} \tag{4.5}$$

$\hat{y}_i$ is the outcome variable (e.g., *expected* weight of participant *i*) which is predicted from the summation of all the predictors multiplied by their respective coefficients. X_{n_i} is the value of the n^{th} predictor for the i^{th} observation. As a case in point the height of participant *i* can be predicted by first calculating the summation of $b_1 X_{1_i}$ (i.e., height of participant *i* (X_{1_i}) multiplied by the height coefficient b_1) and $b_2 X_{2_i}$ (i.e., age of participant *i* multiplied by the age coefficient b_2) and then adding this value to the model's *y intercept* b_0 (value of y when all other variables are 0). If the regression model is strong and representative of the data, then the predicted value should be close to the actual reading for that participant. This precision depends on the model's fitness and on the goodness of such fit (multiple R^2). Following this technique one can confidently predict weight values for new participants that fall anywhere within the range of values used to build this model (this might not be the case for predictors that fall outside the original range of observations). Table 4.3 outlines the parameter estimates for a multiple regression model involving two predictors (*height* and *age*) for the outcome variable *weight*.

Table 4.3: Parameter estimates for *weight* (outcome variable) and *height* and *age* (predictors)

Parameter Estimates		
	Coefficients	Sig
Intercept (b_0)	-114.781	.000
Height (b_1)	2.714	.000
Age (b_2)	0.303	.000

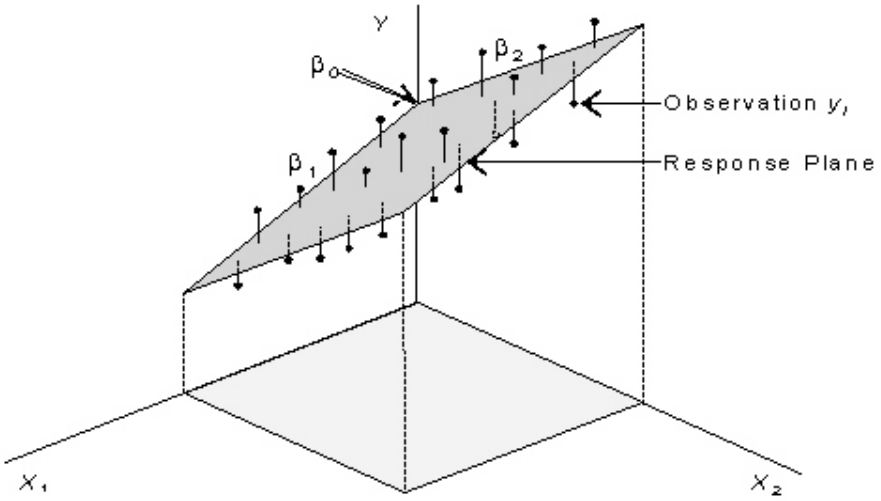
For every additional inch in *height*, *weight* is expected to increase by 2.71lbs and for every additional month in *age*, *weight* is expected to increase by 0.3lbs. So when *height* (X_{1_i}) is 60 (inches) and

age (X_{2_i}) is 189 (months) then *weight* ($\hat{y}_i$) is expected to be 105.3 (lbs) as shown in equation 4.6.

$$Weight(expected) = -114.781 + (2.714 \times 60) + (0.303 \times 189) = 105.326 \tag{4.6}$$

Multiple regression models provide a blueprint of the observed data which can then be used to predict outcomes in different scenarios by varying predictor values. With two predictors and an outcome variable the model could be represented as a three dimensional plane in which the *y axis* is the outcome variable and the *x* and *z axes* are the predictors.

Figure 4.17: A multiple regression model with two predictors (Image source: http://goo.gl/t0UVfD)



The regression plane for two predictors can be easily plotted on a three dimensional scatter-plot allowing the researcher to visually determine the predicted outcome, however with additional predictors (dimensions) visualisation becomes extremely difficult.

4.2.2.3 Linear regression and Generalised Linear Models

Linear regression is based on a linear response model whereby a constant change in predictor values results in a constant change in outcome – thus expressing a linear relationship [157]. A person's weight could be predicted from height and age via the respective regression coefficients generated from a set of observations (recordings of height, age and weight of a representative group of people). Linear regression is suitable for this purpose, however this technique assumes that the outcome variable follows a normal distribution. Observations do not always follow this assumption and there are instances in which a linear response model does not give meaningful or realistic results.

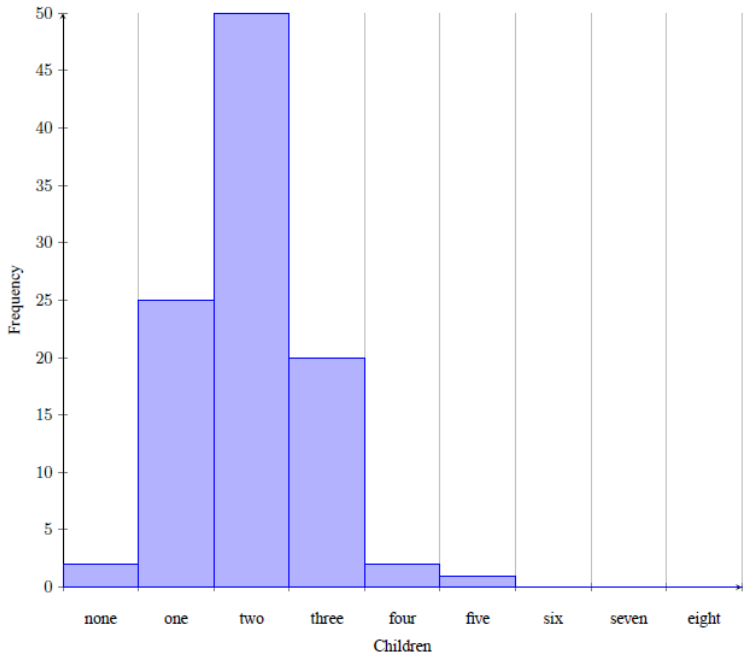
Nelder and Wedderburn introduced Generalised Linear Models in 1972, and it is considered to be one of the most important contributions to statistics in the last 60 years [74]. Linear regression models are special cases of Generalised Linear Models, however GLMs are able to handle observations that depart from the assumptions of traditional regression models. Generalised Linear Models were introduced to tackle the following situations [157]:

1. **Dependent variable is non-continuous**. A case in point would be a study on family planning

choices based on socio-economic indicators (see Figure 4.18). The outcome variable (number of children) would be discrete (unlike continuous figures such as 1.2 children) and most likely the choice of number of children would be right skewed (following a Poisson distribution) since most families would opt to have up to two children, whereby less will plan to have three, four or five – with fewer still planning for six or more children.

2. **Dependent variable is non-linearly related to predictors**. The effect of the predictor variables on the outcome variable is not linear (e.g., *age* vs *health*) whereby a change in age in early adulthood may have a lower impact on health as opposed to a change in age at a more advanced stage in life. A *link function* between age and health would be required to reflect the change in their relationship at different points.

Figure 4.18: Outcome variable following a Poisson distribution (fictitious example)



GLMs can thus be used to predict outcomes of a non-continuous nature (discrete and possibly following a Poisson distribution) which are also not necessarily linearly related to their predictors (e.g., predicting health status on age). Before generating regression models for *perceived workload* and *willingness to complete a task* the data needs to be tested for any of the above conditions, in which case different modelling parameters would be applied in line with the GLM framework.

4.2.2.4 Logistic regression

Logistic regression extends regression by allowing the researcher to build models to predict categorical outcomes based on past data (e.g., *will a specific user finish the task at hand?*). Multiple predictors (independent variables) can also be used to predict this outcome, and these can be both continuous or categorical. Logistic regression can be used across a large number of domains including healthcare [57] (e.g., *given these variables, is this patient's tumour benign or cancerous*?). These are life-saving

techniques, however they can also be adopted in other domains to answer critical questions such as *will users abandon the task (give up) when they see this enrolment page?* This kind of assessment requires a rigorous empirical effort to understand the different user groups, their behaviour and attitudes while determining and selecting techniques to build representative models.

Logistic regression can predict a binary output (binary logistic regression – e.g., *yes* or *no*) or a categorical output with more than two categories (multinomial logistic regression – e.g., *morning*, *afternoon*, *evening* or *night*). In binary logistic regression, models are built to help the researcher understand the probability of an outcome occurring given the independent variable's values. Binary logistic regression is also adopted when the underlying data moves away from the inter-variable linearity assumption of simple or multiple regression. When the outcome variable is binary (e.g., *true* or *false*) then a linear relationship with predictor variables is not possible. Berry and Feldman (cited in [57]) suggest the use of non-linear transformations on variables and provide logarithmic transformation as an example by which one could express non-linear relationships in a linear way [57]. The following equation outlines this transformation.

$$f(x) = \frac{1}{1+e^{-x}} \tag{4.7}$$

For this reason, and assuming multiple predictors, the logistic regression equation takes the following form.

$$P(\hat{y}_i) = \frac{1}{1+e^{-(b_0+b_1X_{1_i}+b_2X_{2_i}+b_3X_{3_i}+\ldots+b_nX_{n_i})}} \tag{4.8}$$

$P(\hat{y}_i)$ is the probability of $\hat{y}_i$ occurring, e is the base of natural logarithms (approx. 2.718), b_0 is the y intercept for the model and b_n is the coefficient (weight) for the n^{th} predictor (X_{n_i}). X_{n_i} is the value of the n^{th} predictor for the i^{th} observation. b_0 determines the outcome ($P(\hat{y}_i)$) when all predictors are set to 0 (X_{n_i} = 0), whereas b_n adjusts the rate of change in probability of $\hat{y}_i$ occurring when X_{n_i} is incremented. An outcome close to 0 means that the event is unlikely to occur. Model fitting is based on maximum-likelihood estimation (MLE) whereby a good fit is obtained when the values predicted by the model given specific predictors are closest to the actual observations. This is an iterative process and terminates when convergence has been reached, or when minor or no improvements on the previous model has been found. If convergence is not found it might indicate that the given predictors are not good enough to predict an outcome (e.g., high levels of collinearity making it difficult to assess the effects of individual predictors). The log-likelihood statistic can be used to determine the model's goodness of fit. The log-likelihood test "*is analogous to the residual sum of squares*" used in multiple regression – indicating the extent of unexplained information in a given model [57]. MLE is obtained by maximizing the log-likelihood statistic for a model [111]. Log-likelihood compares the actual outcomes (data points in data set) with values generated through the model ($P(\hat{y}_i)$). A large log-likelihood statistic indicates a weak model (poor fit) and a value closer to 0 indicates less to no unexplained observations (perfect fit). Log-likelihood values are not meaningful on their own, but are used to compare models during model fitting iterations – until convergence is achieved. Convergence based on the log-likelihood statistic can be

specified as the point at which the percentage difference in log-likelihood between iterations falls below a specific value, unless a maximum number of iterations is specified [39] in which case the iterative process stops (assuming the use of a statistical package such as SPSS).

Model fitting in binary logistic regression requires particular attention. Since the outcome variable is binary (i.e., *true* or *false*) the base model (against which the quality of new models is assessed) cannot be the mean of outcome values (unlike in linear regression). In binary logistic regression the base model or best guess (against which the quality of subsequently generated models is measured) would be the outcome category that has occurred more often in the observed data. The base model assumes no predictors and is made up of the y intercept only (b_0) – the best guess when no predictors are available. Predictors are then introduced to the base model while monitoring for improvements using the following equation.

$$x^2 = 2[LL(newmodel) - LL(basemodel)] \tag{4.9}$$

Various techniques exist to introduce predictors: (1) forced entry, introducing blocks of predictors at one go, or (2) stepwise entry, starting off with the constant and introducing predictors gradually, determined by a score (forward stepwise). Alternatively the process can start off with all the predictors in the model and then removing predictors that have the least impact on the model's fit to the data (backward stepwise). Some predictors might not be significant and may be excluded to improve the overall model, leaving only those that have a significant contribution to the model's predictive power (using the likelihood ratio test). The Wald statistic can be used to measure the utility of each predictor in improving the model and outcome predictions although this may be misleading especially when regression coefficients for predictors are large, increasing the standard error which is in turn used to compute the Wald statistic ($\frac{b}{SE_b}$) [57]. The likelihood ratio is the most expensive statistic (computationally) but it is more reliable than the Wald statistic. The odds ratio (*Exp(B)*) can also be used to determine the importance of a predictor in its contribution towards predicting outcomes. This is a measure that indicates a change in odds (of an event outcome occurring) given a unit change in a predictor (X_n). If *Exp(B)* for *age* is 1.5 then a one year increase in *age* increases the odds of the outcome by 50% ($(OR-1)\times 100$). However if the odds ratio for height is 0.99, a unit increase in height would have no significant impact on the change of the outcome occurring (1% decrease in odds i.e., $(0.99-1)\times 100$). When OR is greater than 1 it transpires that an increase in the predictor results in an increases in the odds of the outcome occurring. According to Field [57] backward stepwise methods are well suited when carrying out exploratory work on new sets of data, as opposed to data for which previous research exists which can be used as a basis for hypothesis testing. Furthermore, backward stepwise is favourable to the forward method [107, 57] mainly because by using this method all independent variables are initially included in the model and if any variable is highly significant only with the inclusion of another variable this will not be excluded from the final model. Forward stepwise might potentially exclude important variables (i.e., suppressor effect).

Tests such as the Hosmer and Lameshow's R_L^2, Cox and Snell's R_{CS}^2 and Nagelkerke's R_N^2 provide a

"*gauge of the substantive significance of the model*" [57]. These tests produce different measurements, however their interpretation is conceptually consistent with the goodness of fit test (R^2) used in linear regression. The Nagelkerke statistic can be interpreted as the extent by which a model explains variability in the data. For instance a value of .469 indicates that the model can explain 46.9% of the variability. There are various statistics to measure goodness of fit, however these models, including both Cox and Snell and Nagelkerke's, are pseudo R^2 statistics since they are only conceptually analogous to the goodness of fit R^2 measure used in linear regression [22].

Consider a model that explains the user's willingness to enrol for and use an e-service given a particular enrolment process designed according to the factors outlined in Table 4.1. Table 4.4 outlines the parameter estimates for a logistic regression model involving a subset of (significant) predictors (Items to Generate (*ItR*), Items to Recall (*ItR*) and Type of Service (*ToS*)) for a categorical outcome variable (willingness to complete task (*WCT*)).

Table 4.4: Example parameter estimates for the *willingness to complete task* outcome variable

Parameter Estimates		
	Coefficients	Sig
Intercept (b_0)	-3.201	.000
ItG (b_1)	0.878	.000
ItR (b_2)	-0.224	.000
ToS 1 (b_3)	2.635	.000
ToS 2 (b_4)	1.646	.000
ToS 3 (b_5)	0.119	.808

R^2 for this model is of 0.23 (Cox & Snell) and .33 (Nagelkerke). This means that according to the Nagelkerke statistic, the binary logistic regression model with these predictors explains 33% of the variability in the data.

Logistic regression was carried out using the backward step-wise (likelihood ratio) method. The Delays (*D*) and Interruptions (*I*) predictors were excluded from this model (at the final iteration) for this particular set of observations (retrieved from a study involving a group of undergraduate students). Table 4.6 shows how the model performs in comparison to actual observations, contrasting predicted outcomes with actual readings. Predicted outcome follows internal encoding shown in Table 4.5 (SPSS uses these codes internally).

Table 4.5: Dependent variable encoding (*willingness to complete task*)

Original Value	**Internal Value**
Yes	0
No	1

Table 4.6: A small sample of actual observations together with their respective modelled outcomes (*expected*). Workings for values in bold are shown in equations 4.10 and 4.11

Model testing							
Observations						*Model Outcomes*	
ItG	ItR	D	I	ToS	Complete Task?	P(Outcome)	Complete Task?
3	8	2	2	0	No	.56819	No
3	8	2	2	1	No	.32856	Yes
3	8	2	2	2	Yes	.09607	Yes
3	8	2	2	3	Yes	.08625	Yes
3	0	1	1	0	No	**.88783**	No
3	0	1	1	1	No	.74643	No
3	0	1	1	2	Yes	.39000	Yes
3	0	1	1	3	Yes	**.36217**	Yes
...	...	...	...	...	...	...	...

Modelled outcomes are generated using equation 4.8 and worked examples are given below.

$$P(WCT) = 0.887 = \frac{1}{1+e^{-(-3.201+(0.878\times 3)+(-0.224\times 0)+2.635)}} \tag{4.10}$$

$$P(WCT) = 0.362 = \frac{1}{1+e^{-(-3.201+(0.878\times 3)+(-0.224\times 0)+0)}} \tag{4.11}$$

Logistic regression allows the researcher to build models for data with categorical outcomes. This however requires careful assessment of the nature of both the dependent and independent variables in order to adopt appropriate modelling techniques for optimum model fitting.

4.2.3 Modelling willingness to complete a task

The willingness of a user to complete a primary task given a specific enrolment process can be abstracted as a binary variable; either 0 (*abandon task*) or 1 (*complete task*). To explain this binary value a binary logistic regression model was adopted. The predictors, or independent variables, are the design factors identified in Table 4.1. The model to be used is explained by the following equation.

$$P(\hat{y}_i) = \frac{1}{1+e^{-(b_0+b_1X_{1_i}+b_2X_{2_i}+b_3X_{3_i}+\ldots+b_nX_{n_i})}} \tag{4.12}$$

$P(\hat{y}_i)$ is the *expected* probability of a user deciding to complete the task at hand while X_{n_i} is the value of the n^{th} predictor for the i^{th} observation. The following example displays the value of two possible predictors:

X_{2_i} Number of items to recall from memory for the i^{th} observation (e.g., date of birth, passport number and mother's maiden name: X_{2_i} = 3)

X_{4_i} Delayed caused by some security check for the i^{th} observation (e.g., verification code sent by email: : X_{4_i} = Minor)

b_0 is the y intercept for the model while b_n is the regression coefficient for the corresponding variable X_n. The regression coefficients are required to calculate the probability of $\hat{y}_i$ occurring, that is the probability of the user deciding to complete the task. These coefficients will vary across different user

groups because different groups of people may react differently to the various predictors (enrolment design factors). This depends on how users behave during the data collection stage and on the quality of the data collected (see Chapter 4.3 for user group calibration). Statistical tests help mitigate the risk of producing ill-fitting models and this may also result in the elimination of statistically insignificant predictors from the model itself.

4.2.4 Modelling perceived workload

Perceived workload for a given task can be abstracted as a continuous variable, being any number between 0 (*no perceived workload*) and 100 (*extreme workload*). To explain this a linear regression model was adopted.

The predictors, or independent variables, are the design factors listed in Table 4.1. The model to predict workload is a multiple linear regression model where multiple predictors are used, as shown in the following equation.

$$\hat{y}_i = b_0 + b_1 X_{1_i} + b_2 X_{2_i} + b_3 X_{3_i} + ... + b_n X_{n_i} \tag{4.13}$$

$\hat{y}_i$ is the outcome variable (i.e., *expected* perceived workload) which is predicted from the summation of all the predictors multiplied by their respective coefficients [57]. X_{n_i} is the value of the n^{th} predictor for the i^{th} observation. The following example displays the value of two possible predictors

X_{2_i} Number of items to recall from memory for the i^{th} observation (e.g., date of birth, passport number and mother's maiden name: X_{2_i} = 3)

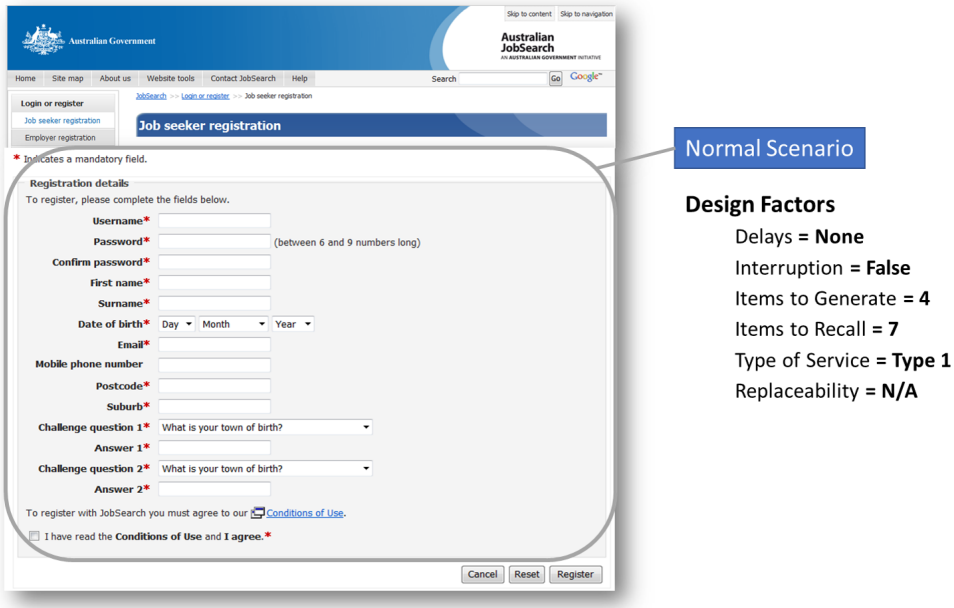
X_{4_i} Interruption caused by some security check for the i^{th} observation (e.g., visit a registration authority in person to verify identity: X_{4_i} = Major)

Specific tests and plots can be used to determine whether any of the assumptions for regression analysis are violated. These assumptions are [57]: (1) errors (residuals) should be normally distributed, (2) predictors must be quantitative or categorical with a quantitative outcome variable, (3) predictors must have varying values (i.e., non-zero variance), (4) there is no "perfect" multi-collinearity between predictors, (5) there is no correlation with variables that are not included in the model, (6) residuals have a constant variance at each predictor level (homoscedasticity), (7) there is no correlation in adjacent residuals (independent errors), (8) outcome values originate from different entities, and finally (9) the mean values of the outcome variable for subsequent predictor increments lie along a straight line (linearity). The goal of this thesis is not to generate highly accurate prediction models – nonetheless models are introduced to guide and inform project teams throughout the requirements development process on potentially severe user experience issues arising from specific design decisions (see Chapter 5). For this reason some assumptions may be violated (e.g., this thesis does not consider aspects such as interface design in the modelling process and thus the fifth assumption may be violated), however tests will be conducted to ensure that any generated models are sensibly fit-for-purpose and not unjustifiably biased.

4.3 User Group Calibration (UGC)

In order to measure user groups' behavioural patterns when facing e-service enrolment processes it was first important to survey the different enrolment process configurations adopted across different types of e-services around the globe. For this reason a number of e-government services were studied and for each service the enrolment-specific design factors were measured as shown in Figures 4.19.

Figure 4.19: Surveying enrolment pages (*Jobsearch.gov.au*)



From this exercise the most common enrolment processes used in e-services were elicited and generalised into nine fictitious enrolment tasks, as shown in Table 4.7. These scenarios would allow the researcher to record and analyse user behaviour across a wide range of enrolment processes, each of which carries distinct intensities of the different design factors. A standard set of tasks would also allow the researcher to compare results across different groups of users. This number (nine) was found to provide a good balance between task variety (in terms of workload) and the time required for a regular internet user to complete them.

Table 4.7: Set of nine enrolment tasks generalised from a survey of commonly found design configurations across various e-services (from low to high workload and assurance levels)

Design Factors	Fictitious enrolment tasks								
	Low			Medium			High		
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Items to Generate	0	1	1	2	2	2	3	3	3
Items to Recall	1	2	3	4	5	6	7	8	NA
Delays	0	0	1	0	0	1	0	0	1
Interruptions	0	0	0	1	0	1	1	0	1
Type of Service	*Tasks are repeated for each ToS*								
Replaceability	*N/A*								

Referring to table 4.7, columns A to I represent the nine fictitious enrolment pages (secondary tasks) that users may find on different e-government service portals, while the four rows represent the different design factors constructed earlier. Each fictitious enrolment task is also associated with all four service types (i.e., *ToS* 1 to 4). These tasks are categorised under three levels – *low*, *medium* and *high* – representing increasing levels of identity assurance requirements and workload. These nine tasks also cover the four levels of identity assurance listed in Table 2.2 (also suggested by Williamson et al. in [174]). For instance, *task I* requires the user to generate three new credentials for future use such as a password, security question and a username. The process also takes a couple of days to complete due to a manual verification procedure, thus introducing a major delay. Furthermore, it also requires the user to visit a registration office to complete the enrolment process thus interrupting daily routines. On the other hand, *task B* takes only a few seconds to complete, requires the user to generate just one new artefact (e.g., personal identification number) and there are no delays or interruptions in the process. Eventually these tasks were revised following the second intervention (see Chapter 7) and multiple levels of delay were also introduced, as shown in Table 4.8. These tasks are originally based on real-world e-services. Table 4.9 exemplifies the relationship between these fictitious tasks and equivalent real-world e-services.

Table 4.8: The set of nine enrolment tasks were modified to include multiple delay intensities

Task	ItR	ItG	I	D
A	1	0	No	No
B	2	1	No	No
C	5	1	No [1]	Minor [2]
D	4	2	No	Major [3]
E	5	2	Yes [4]	Major [4]
F	6	3	No	Minor [5]
G	6	4	No	No
H	9	3	No	Minor [6]
I	NA	3	Yes [7]	Major [8]

[1] Credit card details are required
[2] Wait a few minutes for activation email
[3] Wait three days before account is activated
[4] Visit closest outlet to confirm identity
[5] Upload recent photo
[6] Call free-phone to activate account
[7] Visit registration office during specific opening hours
[8] Three day waiting period till PIN is received

Based on these pre-defined tasks a mechanism was devised to help the researcher capture user behaviour across different enrolment process configurations generally used within different types of e-government services. This data would then be used to compute the required prediction models for perceived workload and the willingness to enrol and complete the primary task online. An online portal was created offering nine fictitious e-services with different enrolment processes. User Group Calibration (UGC) participants are asked to go through each enrolment process. After each task participants are asked to rate six workload scales assessing the different dimensions as specified in NASA-TLX (*i.e.*, *Mental Demand*, *Physical Demand*, *Temporal Demand*, *Performance*, *Effort* and *Frustration*). Following

Table 4.9: Examples of real-world e-services adopting enrolment processes similar to the ones presented in Table 4.8

Task	*Based on...*
A	Directorate of labour (Iceland)
B	Estonian e-government portal (Estonia)
C	Birth certificates (Ontario)
D	Comune di Milano (Italy)
E	Student finance (England)
F	Study permits (Canada)
G	Inland revenue (Italy)
H	Access key registration (Canada)
I	e-ID registration (Malta)

the nine tasks participants are then asked to give a weighting for each of the six scales by completing a short comparison exercise (see Section 4.3.2 for more details). This provides enough information to generate a participant-specific weighting for each dimension representing the level of contribution towards perceived workload. These weightings are meant to reduce in-between rater variance [70]. Following each task participants have to rate their willingness to complete the enrolment process by answering the following question: "*Given this enrolment process, would you consider using this service*?". Users are provided with alternative means that can also be used to achieve their primary goal (e.g., free-phone or traditional post). Four 10-point Likert scales ranging from 0 to 1 (with increments of 0.1) are presented, one for each of the four types of service (*ToS*).

After completing the nine tasks, the participant's data is transferred to a spreadsheet for further processing. Each sheet contains nine rows, (one for each task) whereby each row holds the task ID, rating for each NASA-TLX workload scale, a computed overall mean weighted workload for each task, and the willingness to complete the task for the four types of service. Data from all UGC participants (grouped by user group) is merged and prepared for further processing. To be able to explain user behaviour, two regression models need to be generated based on the data collected: (1) a linear regression model for perceived workload and (2) a binary logistic regression model for task completion. After fitting these two models on the data (using a statistical package), a y intercept (b_0) together with a set of regression coefficients (for each design factor) are generated. Calibrating participants from a specific user group would provide the researcher with data related to that group's reactions towards various enrolment-related design factors (across various types of e-services).

The UGC exercise implements a mechanical procedure within which users' appraisal of the situation they're facing is modelled. The generated models and coefficients are not related to visual design aspects, but are abstracting the different factors that may increase perceived workload for end users while attempting to complete their primary task. This means that the test is agnostic to the small or micro UX aspects of the e-service (e.g., colour schemes, visual appeal, typography as well as text, image and form field placement). It measures and models the level of perceived workload as well as the users' willingness to enrol and use the e-service.

4.3.1 Calibration process

The workflow in Figure 4.20 outlines the steps required to calibrate a user group followed by a description of each step.

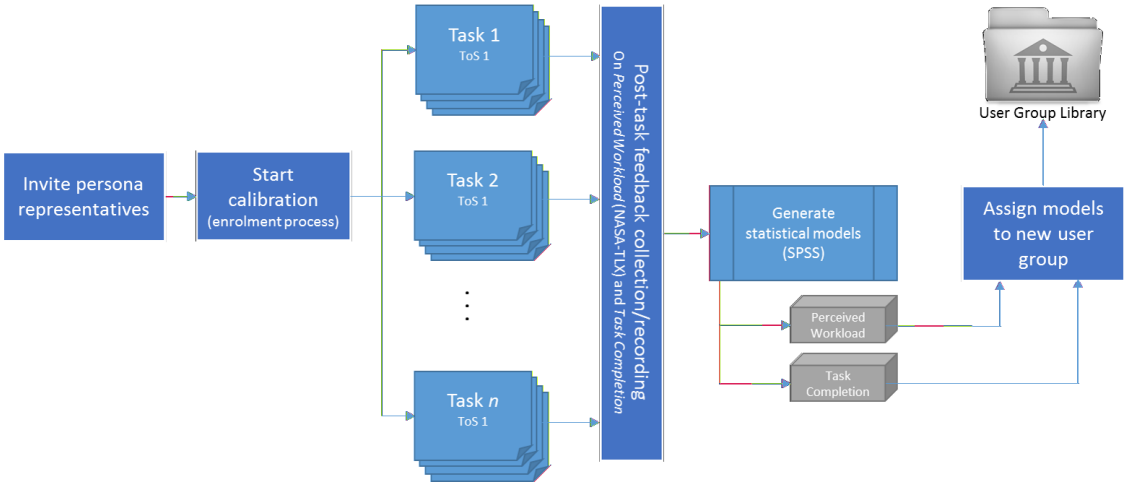


Figure 4.20: Calibration process overview

1. Invite a number of representatives from each of the various user groups associated with project specific personas. At the time of writing the calibration process was hosted at http://calibrate.devbell.com (see Figures 4.21, 4.22 and 4.23). Tasks are presented in a random order so as to avoid unnecessary bias based on apparent patterns (e.g., expecting subsequent tasks to be more demanding or intensive).

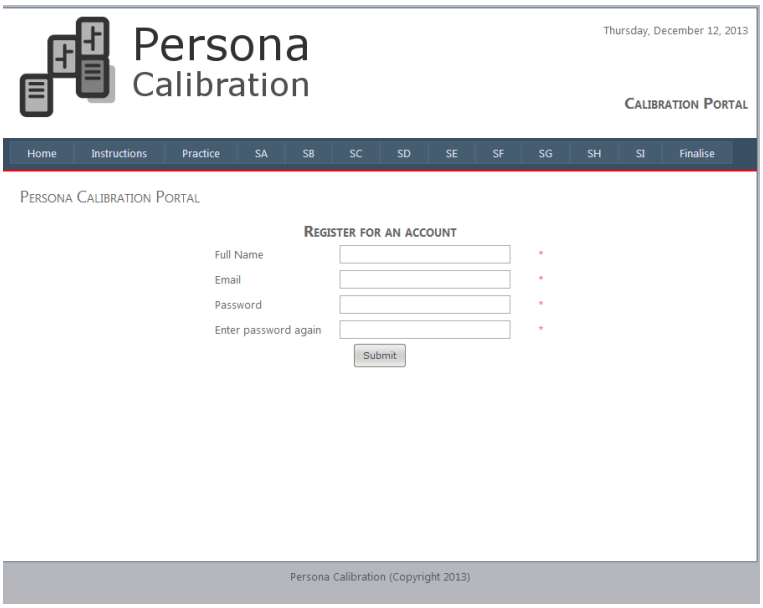


Figure 4.21: One of the nine tasks presented during calibration

2. After each task, participants are asked to complete an online feedback form (see Figure 4.24) which captures NASA-TLX specific workload ratings as suggested by Hart and Staveland [70].

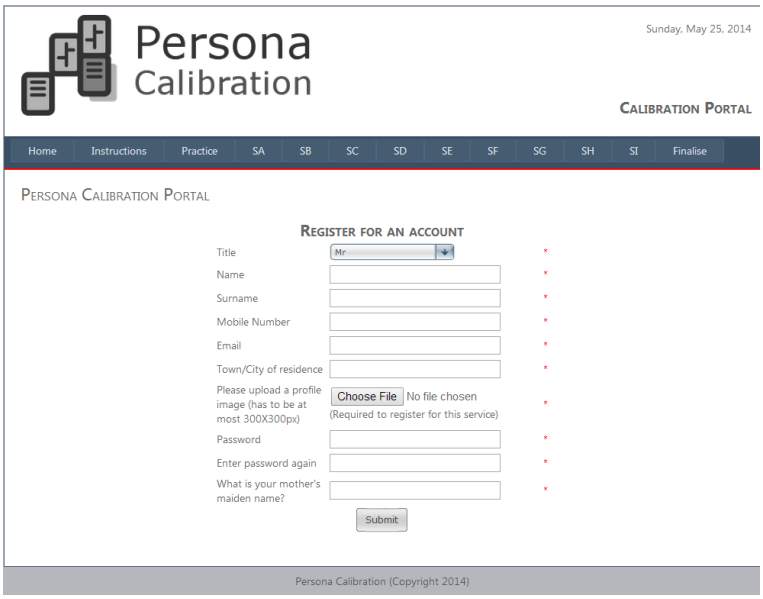


Figure 4.22: Another task presented during calibration

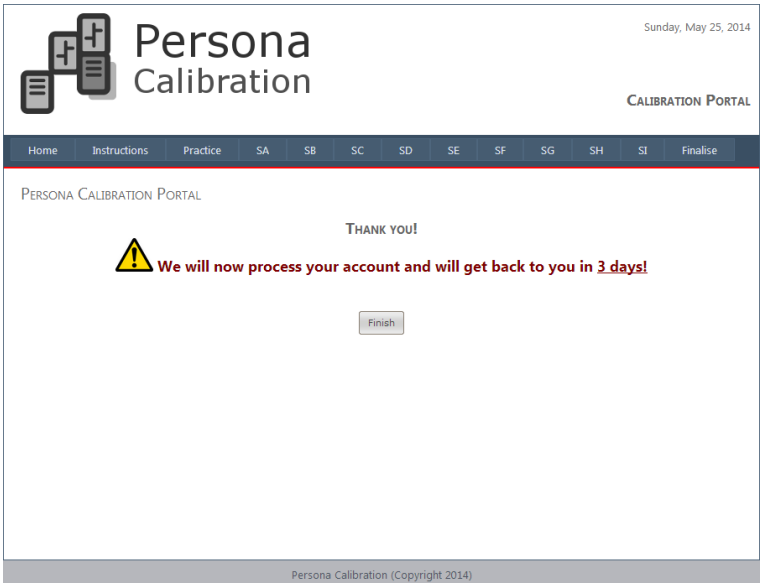


Figure 4.23: Participants are presented with notifications whenever interruptions or delays are present

This form also captures the participants' willingness to complete the task in four different scenarios, reflecting the four Types of Service (*ToS*) identified in Section 4.1.1. The first scenario involves no legal obligation for use while the other three pose incrementing frequencies of use and levels of compulsion (i.e., requiring compliance within specific legal time frames). Scenario four represents an e-service that is used frequently and citizens are required to comply in a timely manner otherwise penalties apply. In all cases the alternative is to visit a government office or to send forms by post. These four scenarios are customised according to the user group under investigation (e.g., *filing tax returns* as an example of a *ToS* 3 service may not be relevant to students, in which case an alternative example is provided – *e.g., study-unit add/drop form which needs to be submitted 2 weeks before the start-of-term*).

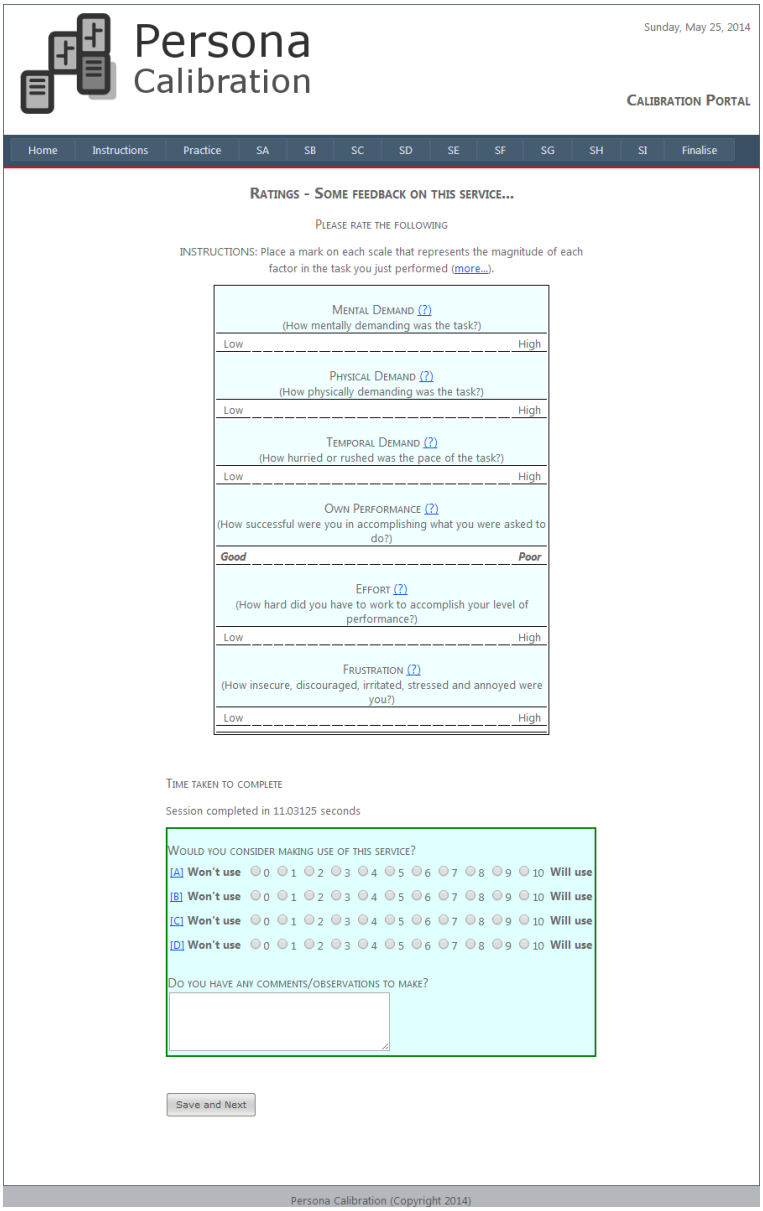


Figure 4.24: Calibration task evaluation form

3. The collected data is preprocessed, grouped (by user group) and prepared for model fitting (see Table 4.10).

4. Using a statistical package (e.g., SPSS) two regression models are then generated: one for perceived workload (multiple linear regression model) and one for the willingness to complete the task (binary logistic regression model). The modelling step will provide the researcher with a number of regression coefficients that explain the user groups' reactions towards the various design factors (e.g., for each additional *item to recall* introduced in the enrolment process, the likelihood that the task is abandoned increases by 35%). These user group regression coefficients are then associated with the respective project persona(s) and eventually used to predict the expected perceived workload as well as the number of people that might complete the task for a given e-service being developed (i.e., *ToS*), per persona (i.e., Calibrated Personas), and for each alternative

Table 4.10: Sample output from a calibration exercise

	NASA-TLX measurements													*Task completion measurements*			
		Weighting (tally count)						Rating							Willingness to complete task			
	Task	MD	PD	TD	P	E	F	MD	PD	TD	P	E	F	MWW	Type 1	Type 2	Type 3	Type 4
Participant A	A	1	2	2	1	4	5	0	0	0	0	0	0	0	0	1	0.9	0.9
Participant A	H	1	2	2	1	4	5	35	25	40	0	25	10	21	0.3	0.9	0.9	0.9
Participant A	B	1	2	2	1	4	5	0	0	15	10	5	10	7.33	0.5	0.7	0.8	1
Participant A	G	1	2	2	1	4	5	5	10	15	5	10	5	8.33	0.9	0.9	0.9	0.9
Participant A	C	1	2	2	1	4	5	15	15	25	10	10	20	16.33	0.2	0.7	0.8	1
Participant A	F	1	2	2	1	4	5	5	10	35	10	15	25	19.33	0.4	0.8	0.9	1
Participant A	E	1	2	2	1	4	5	15	65	60	20	45	30	41	0	0.7	0.8	0.9
Participant A	D	1	2	2	1	4	5	5	5	10	5	5	20	10.67	0.4	0.7	0.7	0.7
Participant A	I	1	2	2	1	4	5	25	65	70	30	50	40	48.33	0	0.7	0.8	1
Participant B	A	2	5	2	1	1	4	15	0	0	10	15	0	3.67	0.7	0.5	0.7	1
Participant B	H	2	5	2	1	1	4	75	15	60	30	50	35	37.67	0.5	0.6	0.7	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

enrolment process design.

It is important to note that at this stage the analyst should monitor the responses and identify any participants that provided data which is noticeably different from that of other participants within the same user group. This might indicate the existence of a new or different user group altogether. Thus, at the analysis stage it is important to consider outliers as evidence for the existence of potentially new user groups, leading to fresh insights into the e-service's target users.

SPSS (or any alternative thereof) can be used to support model fitting and generate the required coefficients. Manual model fitting is extremely difficult due to multiple dimensions introduced by the various predictors (*ItR*, *ItG*, *D*, *I* and *ToS*). Once the regression coefficients for the two models have been generated, they are assigned to the respective user group, which can in turn be associated with one or more project specific persona, turning them into Calibrated Personas. Recurring UGC exercises will result in more realistic and fine-tuned coefficients. It may be the case that certain predictors are statistically not significant in any of the models for a particular user group. In that case, the respective portion of the model is removed (i.e.$b_n X_{n_i}$ for which X_{n_i} is the value of the insignificant predictor for the i^{th} observation while b_n is the regression coefficient for the same predictor).

Table 4.11: Regression coefficients generated for the *young urban professionals (30–40)* user group. These coefficients explain the user group's reactions to the various enrolment-related design factors

	Regression coefficients	
	Task completion (see Figure 4.25)	Perceived workload (see Figure 4.26)
B-Coefficient	5.866	3.888
Items to Generate	-0.78	NA
Items to Recall	NA	2.183
Delays	-1.434	34.332
Interruption	-1.925	24.127
Type of Service 1	-2.339	NA
Type of Service 2	-1.448	NA
Type of Service 3	-0.718	NA
Type of Service 4	NA	NA

Figure 4.25: Task completion parameter estimates for the *young urban professionals (30–40)* user group

Parameter Estimates

Decision[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
Yes	Intercept	5.866	.922	40.438	1	.000			
	[NatureOfService=0]	-2.339	.607	14.837	1	.000	.096	.029	.317
	[NatureOfService=1]	-1.448	.596	5.893	1	.015	.235	.073	.757
	[NatureOfService=2]	-.718	.608	1.396	1	.237	.488	.148	1.605
	[NatureOfService=3]	0[b]	.	.	0	.	.	.	.
	[Delays=1]	-1.434	.413	12.021	1	.001	.238	.106	.536
	[Delays=2]	0[b]	.	.	0	.	.	.	.
	[Interrupts=1]	-1.925	.507	14.446	1	.000	.146	.054	.394
	[Interrupts=2]	0[b]	.	.	0	.	.	.	.
	NewFacts	-.780	.305	6.545	1	.011	.458	.252	.833

a. The reference category is: No.

b. This parameter is set to zero because it is redundant.

Figure 4.26: Perceived workload parameter estimates for the *young urban professionals (30–40)* user group

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	3.888	1.4646	1.018	6.759	7.049	1	.008
[Delays=1]	34.332	6.5205	21.552	47.112	27.722	1	.000
[Delays=2]	0[a]	.	.	.	.	.	.
[Interrupts=1]	24.127	4.8212	14.678	33.577	25.044	1	.000
[Interrupts=2]	0[a]	.	.	.	.	.	.
Recall	2.183	.4188	1.363	3.004	27.183	1	.000
(Scale)	.542[b]	.0527	.447	.655			

Dependent Variable: Workload

Model: (Intercept), Delays, Interrupts, Recall

a. Set to zero because this parameter is redundant.

b. Maximum likelihood estimate.

More participants to a calibration exercise will yield stronger predictive models, however a statistical saturation point exists [128]. Saturation is hereby defined as the point at which statistical models do not exhibit significant improvements in prediction with the addition of more calibration data (e.g., an additional 10 participants will not yield more than 2% improvement over predictions generated by the original model). Out-of-sample tests can be used for this purpose whereby updated models are tested against a set of known observations. A score based on residual values (predictions vs actual observations) will help determine by how much the model has improved with the addition of new calibration data – based on the original model's score (without the new data).

Further to this the calibration process may uncover unexpected clusters of behavioural patterns from within the same set of participants (who might have initially been assumed to share common

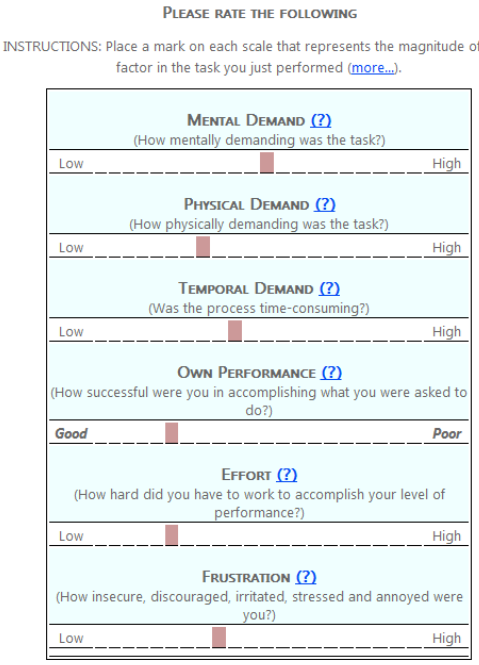
behavioural patterns). Clustered behavioural patterns may indicate the existence of different user groups and further investigation might be required, which may then lead to the discovery of new (and unexpected) user groups altogether. A case in point can be observed in figures 4.29a, 4.29b, 4.30a and 4.30b in which potential patterns in workload weighting is evident.

4.3.2 A note on perceived workload calibration

The perception of workload can vary across different user groups. As an example, asking for the Social Security Number during the enrolment process may have a different impact on an *accountant* persona as opposed to a *student* persona. An accountant will most probably have this information readily available however a university student would probably need to do some additional work to obtain it, if available at all. However these assumptions do not provide objective guidance to support design decisions. The UGC exercise helps to generate quantitative insights on users' reactions and perceptions towards enrolment factors across different scenarios.

The NASA-TLX pen and paper process was automated using a web scripting language providing the user with a point and click evaluation sheet following each of the nine tasks, as shown in Figure 4.27.

Figure 4.27: The NASA-TLX evaluation section presented in the UGC evaluation sheet following each task

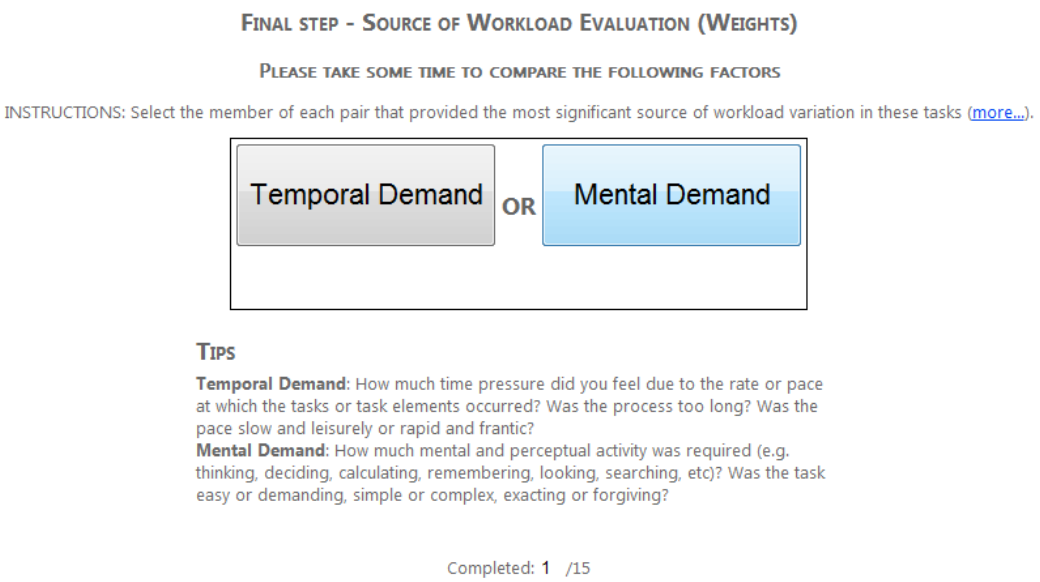


After all the tasks have been completed, together with their corresponding evaluation sheets, the participant is presented with 15 pairs containing all the possible combinations of the six workload dimensions. For each pair the participant selects the scale that is perceived as the larger contributor towards workload. This produces a weighting for each scale which would then be used to generate a weighted overall workload score. Theoretically this should reduce in-between-rater variability through the generation of a weighted workload score that factors in each participant's understanding of what contributes most towards workload. This process is only required once since the nine tasks are of a similar nature

(i.e., enrolment pages).

The resultant weighting for each workload scale (as given by each participant) will eventually be used to generate a weighted workload measurement as a score between 0 to 100.

Figure 4.28: Each participant needs to complete a pairwise comparison of all the workload scales in order to generate a weighted workload measure, thus reducing between-rater variability

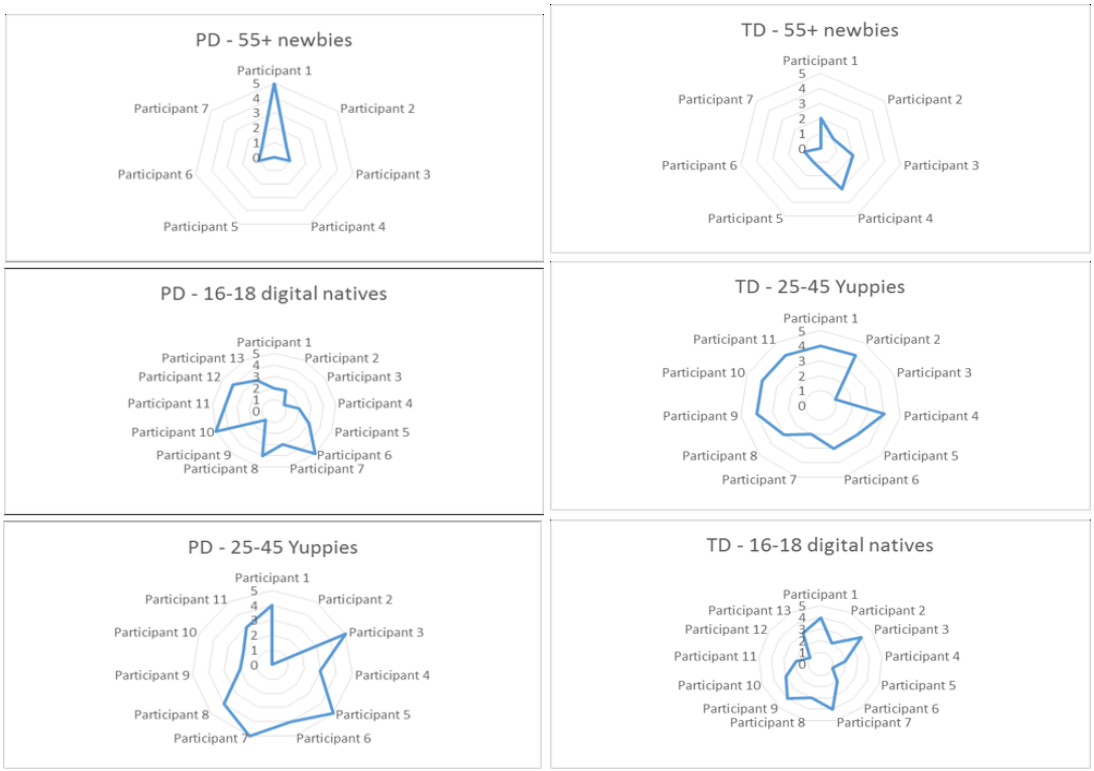


The pairwise comparison creates a tally for the number of times each factor was selected, as shown in Table 4.10 (weighting section). Once all the data has been collected, the evaluation tool works out the NASA-TLX mean weighted workload (MWW – see equation 4.14), and this together with all the other UGC data is automatically exported into a spreadsheet for further processing. Mean weighted workload results are shown under the column marked MWW in Table 4.10.

$$MWW = \frac{\sum_{i=1}^{6} (R_i \times T_i)}{15} \tag{4.14}$$

A distinctive difference exists in the way different groups of users assign weight to the various workload dimensions and this can be demonstrated visually using radar-charts. The charts shown in Figures 4.29 and 4.30 were generated from various calibration exercises across three different groups of users. These charts map out the weighting given by UGC participants for all TLX workload dimensions as a number from one to five following the pairwise comparison of the six sources of workload. This number is used as a weighting factor when measuring overall workload based on the hypothesis that different people have different perspectives of what constitutes workload.

Through this type of representation it may also be possible to notice divergent patterns arising from within the same user group. This may be an indication that a distinctive sub-group(s) exists, carrying a different perception of what constitutes workload. These sub-groups may become evident when their data consistently differs from the data of other participants within the same group. This is a clear sign that further investigation might be required to determine whether these patterns are simply caused by



(a) Workload weighting for *Physical Demand* across three user groups. *55+ newbies* are generally less bothered with physical demand as opposed to the *30–40 young urban professionals* group

(b) Workload weighting for *Temporal Demand* across three user groups. *55+ newbies* consistently gave less importance to temporal demand as opposed to the *30–40 young urban professionals* group

Figure 4.29: Weighting for *Physical Demand* and *Temporal Demand* across three user groups

outliers or whether they actually represent a case were the user group was initially overgeneralised. This informs the calibration process while fine-tuning the understanding of the various user groups represented by project personas. In turn this can also inform the selection and development of these personas (e.g., introducing new (and unexpected) personas based on fresh knowledge arising from the calibration exercise).

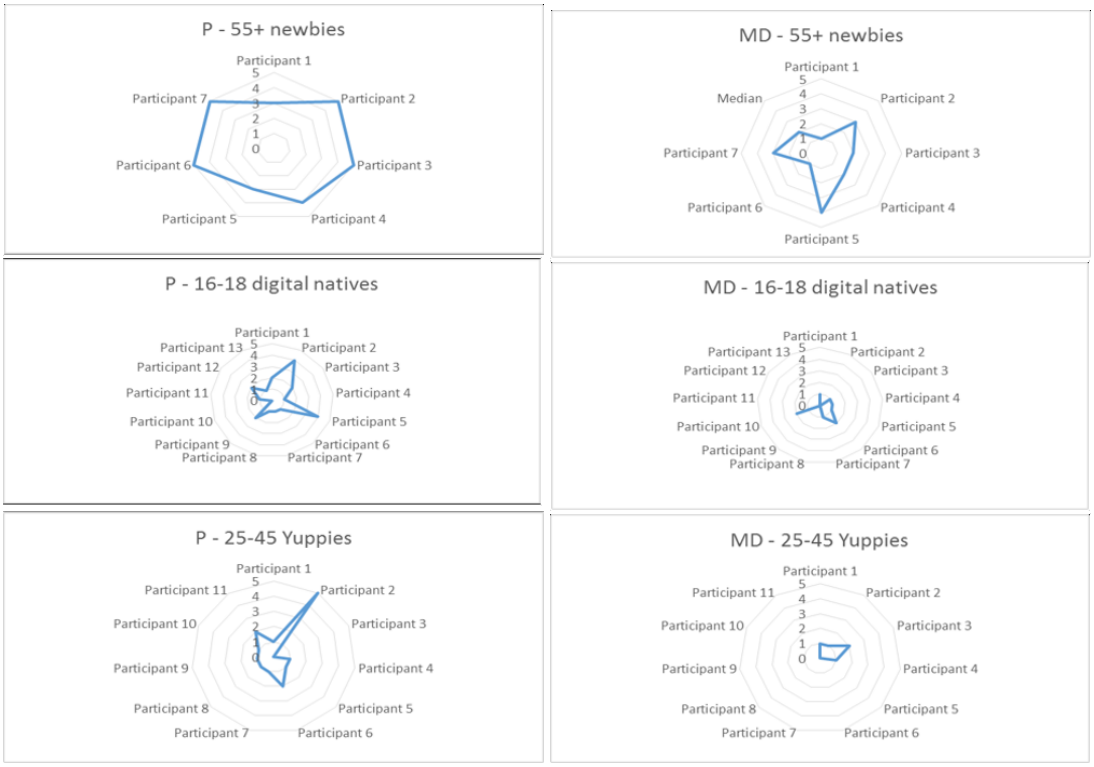
Designers need to be aware of the risks that arise when introducing unnecessary workload, including non-compliance and non-adoption however this needs to be balanced with the required level of assurance as well as with the user groups' behavioural patterns.

4.3.3 User group calibration and the context of use

User Group Calibration (UGC) exercises can be conducted in three ways: *lab-based* (supervised), *in-context* (supervised) and *online* (unsupervised). In all three cases the same online calibration portal is used to collect and store data.

Lab-based calibration allows for simultaneous calibration of an entire group of participants. Although cheaper and faster, this approach has two issues:

1. It is generally very difficult to get a group of people in one room at the same time, especially when dealing with professionals.



(a) Workload weighting for *Performance* across three user groups. Conversely to the other cases, *55+ newbies* gave more importance to performance (e.g., lack of confidence resulting in perceived workload)

(b) Workload weighting for *Mental Demand* across three user groups. In this case *55+ newbies* were more concerned with mental demand as opposed to the other groups

Figure 4.30: Weighting for *Performance* and *Mental Demand* across three user groups

2. Calibration is conducted outside the normal context of use and this may present validity issues. Participants can be asked to bring their own devices for the session (e.g., laptop or tablet) however this might not always be possible. Also, the lack of automatic form-filling features available in a personal browser as well as differences in keyboard layout may add further pressure on UGC participants.

On the other hand *facilitated in-context* calibration inherently considers the users' working/living context. This requires the researcher to schedule visits to the participants' own environment (e.g., place of work) to conduct one or more calibration sessions, either one-on-one or in small groups. In turn this makes this method more expensive to conduct (and more time consuming). Nonetheless the presence of a facilitator, who may offer immediate guidance and clarifications, may further improve the validity of the data collected. The resulting models should be decorated with a *calibration context* attribute, clearly indicating where calibration took place (e.g., workplace, home) for future reference and reuse.

Facilitated in-context calibration is sometimes necessary when dealing with user groups who are less flexible or who have strict time constraints. In one of the case studies, a user group called *young urban professionals (30–40)* was calibrated by visiting all participants, individually, at their place of work. In this case it was noted that decisions related to the different calibration tasks were highly influenced by contextual nuances whereby participants could refer to their physical surroundings and

work-conditions before submitting their feedback on perceived workload and task completion (e.g., "*how hard would it be to scan a utility bill over here?*"). Behaviours and attitudes might change in different contexts, however in-context calibration mitigates this risk.

Finally *online in-context calibration* is the most flexible and cost effective method, however care must be taken when opting for this method. Online calibration may yield incorrect data due to misunderstandings, boredom or disinterest from the participants. This can be mitigated by providing remote assistance via VoIP or a realtime chat facility. Conferencing systems such as Google's *Hangouts*, *Skype* and *AnyMeeting* can be used to facilitate multiple sessions simultaneously.

4.3.4 Sampling for calibration

Sampling for user group calibration participants follows three guiding principles: representation, agility and cost. Stratified[2] and convenience sampling is highly effective to kick off the sampling exercise. Depending on the user group, various other techniques can be adopted to expand the sample size (within the set constraints of time and budget). For instance, it was noticed that for most user groups snowball sampling turned out to be the most effective recruitment method, whereby participants recommended and invited other potential participants (generally friends or business partners). Other user groups (e.g., managerial roles) were more responsive to a personal invitation over the phone followed by additional information sent via email. In this particular case, facilitated in-context calibration was the only feasible technique from a participant's perspective and the researcher had to visit participants' places of work to conduct the exercise. This however provides additional benefits: in-context calibration results in the recording of highly nuanced behavioural patterns affected by the participants' environment (increased calibration loyalty). Also, facilitated in-context calibration affords the opportunity for the researcher to gain additional insights on participants, their behaviour within the context of use, the participants' motivations as well as constraints.

In some instances recruitment was encouraged with the use of rewards, however given the costs involved there were no real benefits over snowball sampling without rewards. Random (passive) sampling for participants was also used however response rates were generally low, with or without rewards (e.g., flyers and mail-shots to thousands of participants). Furthermore, active random sampling was found to be an expensive exercise and the cost (mainly time) to recruit participants outweighed the benefits (e.g., campus visits and in-class calls for *undergraduate students* user group participants).

Finding a good initial set of participants who fall under the required parameters (e.g., *self-employed with basic computer knowledge*) is not always trivial, nonetheless it is a crucial phase in the process of establishing a representative sample. Personal connections and 'asking-around' are generally two of the most effective techniques to kick-off the sampling process, setting the pace for snowball-driven recruitment.

By being empathetic and understanding towards participants it was felt that they in turn adopted a highly responsive and collaborative stance, referring the researcher to other people who may be valid user-group representatives. A level of personal judgement, discussion with peers and an optional short

[2]The population is stratified according to the user groups being investigated. A number of participants will then be selected from each stratum, ideally in proportion to the size of the other strata in the population.

set of screening questions can help to determine whether a particular participant falls within the required parameters specific to the target user group being investigated (determined by the persona's characteristics).

Sample sizes vary, and anywhere between 10 to 20 participants were involved during the various user group calibration exercises conducted throughout this thesis. Intermediary statistical tests must be conducted in order to determine whether the underlying data is useful enough to generate statistically significant models, and this should inform the recruitment process and the resulting sample size. In theory one should obtain stronger predictive models when increasing the number of UGC participants, however there is a natural point at which new observations will not yield any significant improvements to the model's predictive power (i.e., statistical saturation). In this case, the cost of conducting additional UGC exercises would not be justified. Nonetheless user groups should be re-evaluated from time to time to avoid the risk of generating misleading simulations due to generational shifts in behaviours and attitudes (see Section 4.4).

Refer to Section 10.3.3 for a discussion on issues related to user group clustering and sampling.

4.4 User Group Knowledge Base

For each user group in a given population, behaviour is modelled following a calibration exercise involving several user group representatives (participants). User groups are generally established based on citizen demographics, however the calibration process may shape the development of such groups through emerging behavioural traits, informing the creation, modification, merging or retirement of groups. This may in turn affect the requirements development process through the introduction of new (and possibly unexpected) project personas, which would in turn inform the development of use cases and scenarios. Table 4.12 provides an example of several project personas that share three user groups' behavioural traits (i.e., formalised through statistical behavioural models).

Table 4.12: Multiple project personas linked with different user groups, distinguished by some common factor(s)

User Group	Project personas
Undergraduate students (18–22)	Chris Christine Miguel
Young urban professionals (30–40)	Carol Maryanne Ingolf
55+ year old newbies	Doris May

Care is required to avoid over generalisation when constructing user groups. For instance, a *Secondary school teachers* user group (using the plural, not to be confused with personas) can be considered to be an over-generalised group. Although secondary school teachers generally share a common educational background, teachers across different age groups may exhibit significant behavioural differences when interacting with computer systems (e.g., due to different levels of experience and confidence). A 50 year old teacher working in a secondary school might have had less exposure to electronic means

of learning and teaching, and thus system designers may want to distinguish him from other teachers in the 25–35 year-old bracket. For this reason, calibration should focus on multiple user groups: *secondary school teachers (22–40)*, *secondary school teachers (40–55)*, and *secondary school teachers (55+)*. During calibration one might notice that different user groups share similar behavioural patterns and workload perceptions, and thus these could be merged (e.g., *secondary school teachers (40+)*). User groups should be re-usable across projects and thus generic group names are recommended. For instance, instead of referring to *secondary school teachers (22-40)* one should abstract this group even further and consider *graduate professional (22–40)*. This makes this user group persona-agnostic and thus re-usable across project personas (i.e., not necessarily related to e-learning projects). A project may require several personas however these may share one or more calibrated user groups. This means that a user group called *university graduates (under 25)* could be linked with a number of unrelated project personas (e.g., nurse persona for an e-health system and teacher persona for an e-learning platform) as long as contextual differences are taken into account for the e-services being developed. Both personas can be university graduates and under 25 years of age, and although they are used in separate projects they may still share the same behavioural patterns and attitudes when dealing with security tasks. This does not apply when contextual calibration is adopted for a specific user group since the resulting models could be influenced by contextual nuances that may not be applicable outside that specific environment (consider calibration conducted in a hospital emergency room as opposed to a clerical environment). In these specific cases, user groups should be flagged with the context in which calibration took place (e.g., the underlying models for the *35-45 business-person (office)* user group may differ significantly from the *35-45 business-person (travelling)* user group). In situ and contextual calibration is discussed in Sections 4.3.3 and 9.4.1.4.

A number of user group differentiating factors exist, which could include both demographic and biographic user properties (e.g., age group, education level, geographic context and computer literacy). Algorithmic techniques (e.g., clustering) may also be used to uncover subtle differences in behaviour between participants (from a pool of participants with the same set of properties). This may help fine tune the user group creation process by exposing potential clusters of users from less obvious behavioural patterns in data (see Section 10.3.3). User groups are not user archetypes, but are mere containers for behavioural models created for (and through representatives of) individual groups of users. User groups are in turn associated with project personas (turning them into Calibrated Personas) which can then be used to simulate user feedback to assist in early-stage decision making. The existence and quality of this feedback depends on (1) the models available for each group of users and (2) the strength of such models.

The upkeep of the user group library is strongly suggested, by:

1. Conducting routine re-calibration exercises on existing groups (e.g., yearly). The model's predictive power may increase with more participants although a saturation point exists.

2. User groups may need to be updated (or retired altogether) especially when participants who were used to build the original behavioural models (through calibration) are replaced with a new gen-

eration of users who may exhibit different behavioural patterns and attitudes towards enrolment processes. Tests such as the Wilcoxon Signed Rank test (no assumptions of normal distribution) or the paired-samples t-test (assuming a normal distribution) can help the researcher determine whether two sets of data from the same population (assumed) are significantly different from each other. This may happen when a user group has been in existence for a number of years (e.g., five or more years).

3. Creating new user groups when specific patterns emerge from the calibration data. Data can hold interesting insights on behavioural patterns and through clustering techniques one may identify recurring deviations in data collected from a single user group.

Given a shared user base across government agencies, the author argues that government entities would benefit from the creation (and upkeeping) of a user group knowledge base. For instance, over 200 e-services were rolled out (or updated) in the last two years in Malta, all serving a specific subset of the same population.

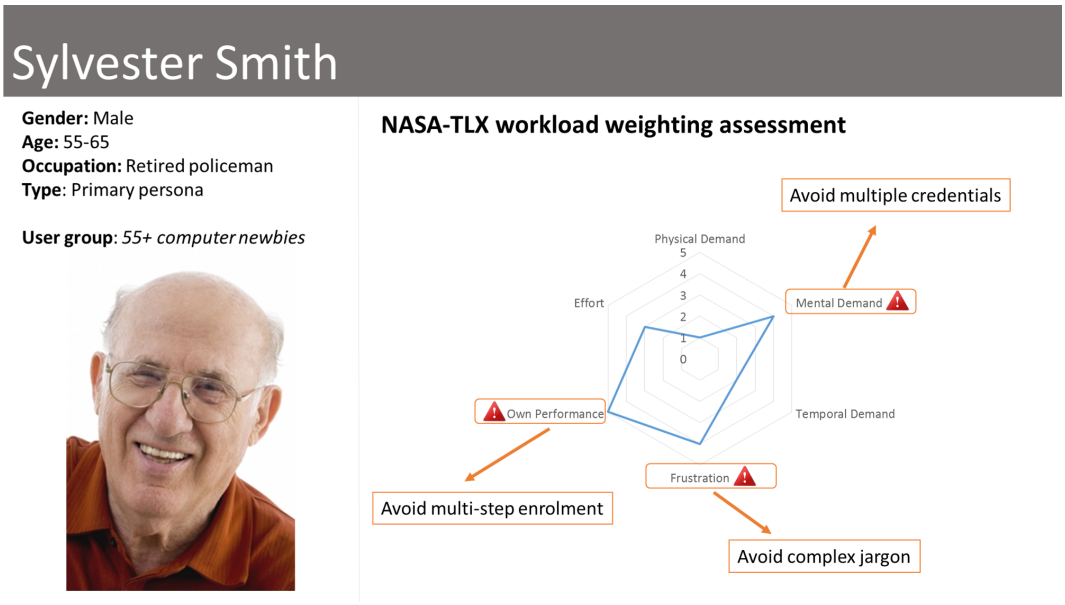


Figure 4.31: Weighting for the various NASA-TLX workload dimensions generated from the user group calibration exercise conducted with *55+ year old computer newbies*. These are then used to generate warnings on design aspects that can have a stronger negative impact on this particular group of users

In practical terms the user group library consists of a database of user groups together with their statistical models explaining their reaction towards the various design factors – in this case related to enrolment. When personas are constructed for a new e-service, the project team should determine whether an applicable user group exists within the library. User groups give a 'voice' to the persona construct (turning it into a Calibrated Persona), adding analytical power which could be used to simulate user feedback on critical design decisions during the requirements development process. Furthermore, Calibrated Personas can also be used to glean practical design recommendations for the different user groups involved. These recommendations would be automatically generated based on the underlying data produced during the calibration process. Figure 4.31 and 4.32 depict two Calibrated Personas together with

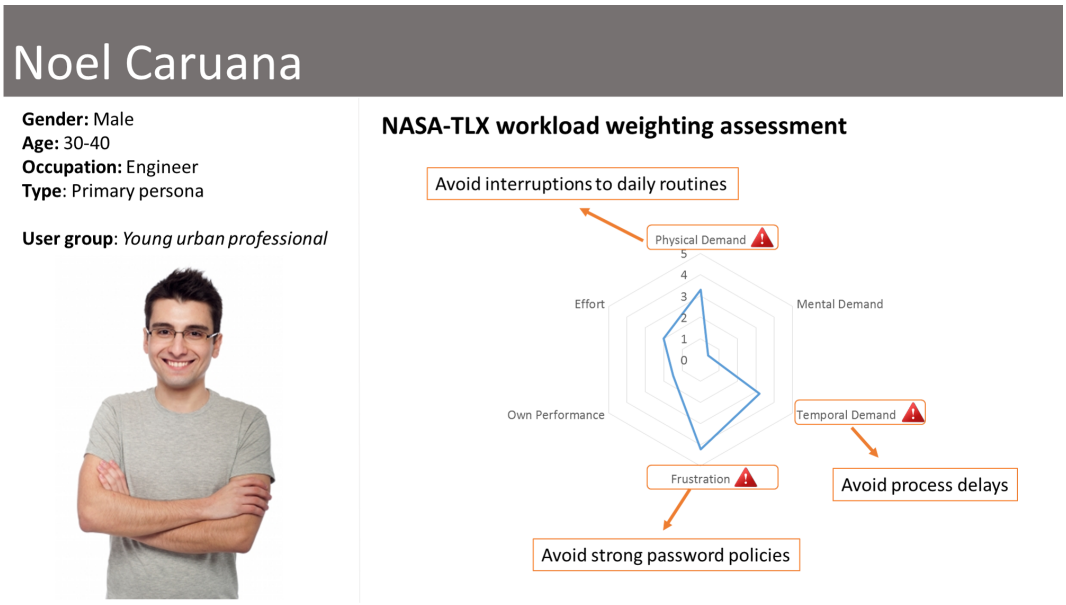


Figure 4.32: Weighting for the various NASA-TLX workload dimensions generated from the user group calibration exercise conducted with *young urban professionals (30–40 years old)*. These are then used to generate warnings on design aspects that can have a stronger negative impact on this particular group of users

practical design recommendations based on the respective groups' workload weightings. These warnings are generated for the project team's consideration, highlighting design aspects that have a negative impact for each of the two project personas. Weightings shown are based on median values derived from NASA-TLX data (final pairwise comparison) generated during the user group calibration exercise with each respective group of participants.

4.5 Summary

The enrolment process is considered to be a critical aspect for e-service design and development, potentially affecting adoption rates and overall success (see Section 2.3). A number of enrolment-specific design factors have been presented (see Table 4.1). These factors have been uncovered through an empirical exercise designed to explore issues and negative experiences encountered by regular internet users during enrolment (see Section 4.1). Based on these design factors as well as a survey of common e-service enrolment processes, this chapter then presents a technique to build behavioural models for different groups of users that help explain and predict the level of perceived workload and their willingness to enrol for and use an e-service (i.e., given a specific enrolment process and type of service). Data modelling (i.e., regression), as opposed to black-box algorithmic modelling (e.g., Artificial Neural Networks) was found to be the more practical and sensible approach for this thesis. This was mainly due to the exploratory nature of the study whereby full control of the underlying processes as well as the availability of well documented tests (to monitor output quality) were highly desirable. Two regression techniques were used, one for perceived workload (multiple linear regression) and one for the users' willingness to complete the task (binary logistic regression).

Individual enrolment-related design factors may have a different impact (i.e., different intensity) on the users' lived experience (ULX), and in turn this impact may also vary across the different groups

of users. For this purpose the User Group Calibration exercise (UGC) was developed – a systematic data collection protocol to measure and aggregate behavioural patterns exhibited by different groups of users when facing different enrolment processes. The calibration exercise consists of a set of fictitious enrolment tasks based on a survey of commonly found e-service enrolment processes across the globe. Following each tasks participants provide their feedback on the perceived workload involved, across six workload dimensions, as well as on their willingness to complete the task, across four types of e-services. This data is then pre-processed and used to fit the two regression models out of which two sets of regression coefficients are produced. These coefficients could then be associated with project specific personas (turning these into Calibrated Personas) which could in turn be used to predict (or rather, indicate) possible user reactions towards enrolment-centric e-services (see Chapter 5). Knowledge on different user groups is stored in a user group knowledge base for future reuse across different projects within the same governmental context (e.g., region or nation). As with other knowledge bases, maintenance is an important activity – and behavioural models need to be maintained by (1) re-calibrating them (to improve model strength) and (2) replacing them with new models to respect generational nuances (e.g., 25 year old undergraduates of today may not exhibit the same behaviour as their counterparts five years down the line). Risks to validity were also discussed and techniques such as UGC participant sampling as well as contextual calibration were presented as measures to mitigate these risks. The user group calibration exercise provides important insights on user attitudes (e.g., towards workload) which can be used to inform design decisions (see Figures 4.31 and 4.32).

Chapter 5 will outline *Sentire*, a requirements framework that adopts Calibrated Personas to generate simulated user feedback on critical design decisions (i.e., enrolment-based use cases). By embedding user feedback simulations at the requirements stage practitioners can test the impact of their design decisions before building the actual product (or prototype). This also introduces the idea of testable experience-related requirements which are decorated with measurable fit-criteria. Any proposed use case design must be run through the simulator to determine whether such requirements are observed (see Section 5.1).

Chapter 5

A Requirements and Design Framework for Enrolment Based Public Facing E-Services

Janowski, Estevez and Ojo's [79] concluded that the lack of methodologies and models, as well as the lack of cohesion between related projects are two of the major causes of e-government project failure. Furthermore, Seffah et al. [147] insist that an integrated framework, incorporating techniques from both the software engineering and usability engineering fields will be the catalyst for more effective use of user-centred techniques within software design and development processes. This chapter presents *Sentire* – a requirements and design framework for public facing and enrolment based e-government services. *Sentire* adopts user behavioural models, Calibrated Personas and simulated user feedback within a usable, flexible yet rigorous requirements and design framework based on the *Volere* requirements templates and process.

5.1 From Deferred to Realtime Feedback on Design Decisions

User feedback is quintessential and the frequency by which this is obtained can have a severe impact on the quality and cost of the final product. Design improvements can only be verified if they are measurable, and thus quantitative techniques are desirable. Finally all of this needs to respect the project constraints whereby additional pressure on resources may break the whole effort. Reasonably indicative but measurable user feedback needs to be obtained as often and as cost-effectively as possible, however this is a challenging proposition.

Sentire contributes to this end by helping designers simulate user feedback based on statistically sound behavioural models embedded in project personas (Calibrated Personas). This can shed light, with reasonable accuracy, on the impact of critical design issues on users by generating feedback in quantitative, measurable and thus comparable terms. This feedback can be generated at every design iteration or with each decision without the need to involve actual users each and every time, incurring additional overheads (time and money). The goal of *Sentire* is to capture sub-optimal design decisions on critical aspects as early in the process as possible.

Figure 5.1 presents a simplified representation of the Software Development Life Cycle (SDLC) for e-government services.

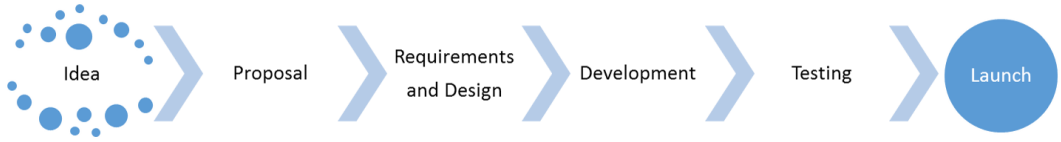


Figure 5.1: Generic SDLC for e-government services

The type of enrolment, informed by the level of identity assurance required, is considered to be a critical design decision (CDD). For this purpose the type of enrolment and other critical design decisions need to be tested earlier on in the software development workflow (see Figure 5.2).

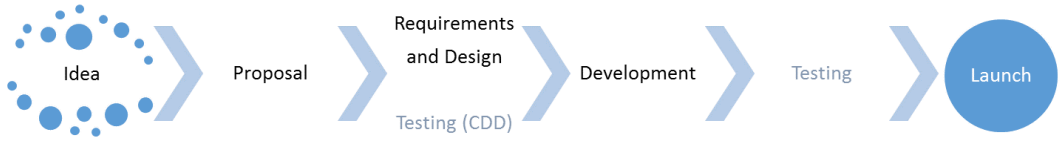


Figure 5.2: Amended SDLC workflow – introducing requirements and design-time testing for issues related to Critical Design Decisions (CDD)

Sentire shifts user testing on critical design decisions (in this case for issues with user enrolment processes) to an earlier stage within the SDLC. Calibrated Personas are used to simulate immediate user feedback on specific enrolment-related design decisions. Quantitative user insights generated through statistical behavioural models can help designers align their assumptions with actual (albeit simulated) user perceptions and attitudes. User behaviour modelling is adopted to give a voice to personas (with users in-absentia), while aiming to achieve a good balance of feedback speed, accuracy and cost. This would allow project teams to ask questions such as "*what impact would this process have on the various target user groups?*". Answers to such questions would be available at the requirements and design stage while being presented in an objective and actionable manner.

Unlike usability which can be measured following the ISO standard 9241, user experience (UX) is difficult to quantify. For this reason it makes it even more difficult to gauge potential impact on users at the design stage when high-fidelity prototypes are not yet available. Deferring feedback until an early release is available might result in expensive rework. This thesis attempts to address this problem by:

1. Building quantitative user models that explain user behaviour during enrolment, and

2. Integrate these user models as part of a widely used and industry strength requirements development process.

For this reason *Volere* was selected as the requirements development process of choice and a number of modifications were conducted in order to accommodate the Calibrated Persona technique. This resulted in a modified *Volere* process, which was named *Sentire*. This name is inspired from the Italian verb 'to listen', and reflects the idea of listening to target end users at design-time (represented by Calibrated Personas) to inform decision making and help avoid expensive changes at a later stage. This technique

does not replace traditional UX evaluation techniques (e.g., user walkthroughs and focus groups), but precedes them within an overarching framework wherein several tools may also be adopted to solve other usability issues. The choice of tools depends on the nature of the project, including the level of agility required. A comprehensive list of UX evaluation techniques is given at UXMastery.com[1] and by Morville [110]. These also intersect with a set of tools specifically built to ensure usability, suggested at UsabilityNet.org[2].

This thesis aims to help designers and policy makers understand the impact that critical design decisions can have on end users at the earliest possible stages of a system's lifecycle. This:

1. Provides a low-cost and early feedback mechanism (shortening the feedback loop on critical aspects at design-time),

2. Reduces the risk of major rework later on in the process, and

3. Instils user-centricity as a systematic discipline within the design process by highlighting the fact that every design decision can have an impact on the user experience, which impact could also vary in magnitude and direction (*negative* vs *positive*).

In line with Nielsen's task success rate measurements [113], *Sentire* provides indications of the probability that a given user group would be willing to complete the primary task online. At the same time *Sentire* also provides NASA-TLX measurements as well as other metrics as meta-information to explain the given indications. *Sentire* also allows product owners to specify task success rates as measurable fit-criteria for specific requirements (e.g., 95% confidence that 80% of new users shall be able, and willing, to enrol on our service on their first interaction). In collaboration with developers, the project team could then work iteratively towards reaching the desired experience goals, from day one.

5.2 Government Wide E-Service Requirements and Design Strategy

The author is presenting a systematic and analytical requirements framework for e-government services based on an industry strength requirements development process. This framework encourages knowledge accumulation, reuse and design-time user feedback simulations (via Calibrated Personas). Efficiency and process improvements can only be achieved through the adoption of a systematic, government-wide and consolidated requirements, design and development framework.

The author's objective is to ingrain elements of user-centred design, including active awareness of the user experience and human factors into the requirements development process, not as an optional tool or component reserved for UX specialists, but as a hard-step within an analytical process of discovery that informs decision makers on the impact that their actions could have on users. The *Volere* requirements development process and templates were extended to include user experience analytics through

[1] UXMastery.com, http://uxmastery.com/resources/techniques, (accessed March 2014)
[2] UsabilityNet.org, http://www.usabilitynet.org/tools/methods.htm, (accessed March 2014)

Calibrated Personas, associated meta-information (e.g., NASA-TLX data) and regression analysis. E-government project teams need to consider both the user experience as well as the users' lived experience (ULX) (see Chapter 2.6 for terminology and definitions).

5.3 Adopting and Extending *Volere*

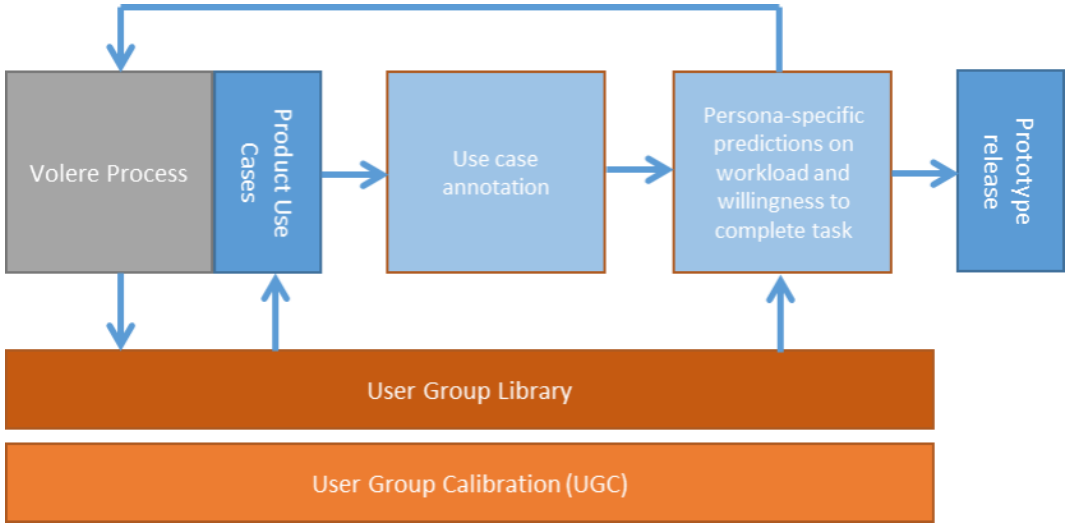


Figure 5.3: *Sentire*'s extensions to the *Volere* process

As shown in Figure 5.3, *Sentire* extends the *Volere* process in two directions, which eventually merge again following each design iteration. The first extension is the creation of a user group library, making re-usable user group behavioural models available for use in future projects. The idea behind the user group library is to have a knowledge base of behavioural models represented as self-contained statistical entities. Project specific personas are created during the *Volere* process and these may then be associated with existing user groups from the library. The resulting Calibrated Persona is then stored within the persona library for future use and adaptation. The user group library gives system designers a continuously evolving arsenal of user behavioural models representing the various user groups in a given country or region. The more one learns about these user groups, the more representative such models become. Through iterative and ongoing User Group Calibration (UGC) exercises the various user groups within the library are further enhanced, giving them a stronger 'voice'. This voice can then be assigned to the respective persona(s) which is in turn associated with product use cases as an active actor.

The second extension applied to the *Volere* process is the annotation of product use cases (PUCs) with enrolment-related design-factor measurements (as outlined in Table 4.1). These annotations, together with the respective behavioural models (associated with Calibrated Personas) can then be used to generate simulated user feedback. Each PUC contains normal case scenarios which are in turn defined as a series of steps (or tasks). The annotation exercise lets the project team flag enrolment related steps, each of which is then annotated accordingly by specifying the occurrence and intensity of each design factor. The team is also required to specify the Type of Service (*ToS*) for each PUC, indicating the use case's frequency of use and underlying legal requirements. Once the annotation process is complete and

actors (Calibrated Personas) are assigned to these use cases, one can then produce experience-related insights for each Calibrated Persona across the various use cases, initiating an iterative process of amelioration. Once design goals are reached and a balance between identity assurance, perceived workload and acceptability is struck, designers can then move on to (low or high-fidelity) prototyping which can help capture other non-critical usability issues (i.e., small UX).

Sentire is both iterative and evolutionary. At each design iteration analysts obtain measurable and comparable insights on how different users will react to different design alternatives. This will also help them expose areas of contention that could deter users from enrolling and completing the primary task online. At this point designers can go back to the design board to revisit the contentious product use cases and associated requirements while keeping an eye on possible repercussions caused by subsequent changes. Changes to existing product use cases or the addition of new ones (together with any associated requirements) need to be (re)assessed for user acceptance through simulated user feedback.

A central extension to *Volere* is the introduction of testable experience requirements which are especially important in contracted projects. Project teams can specify measurable fit-criteria for usability and humanity requirements and developers need to ensure that any proposed product use case for the new e-service respects the given parameters (e.g., *At least 80% of undergraduate students shall be willing to complete the task online with a perceived workload level of less than 35%*). In this case developers (e.g., contractor) must collaborate closely with product owners (e.g., government entity). To ensure transparency and process visibility between the two parties this thesis presents an online computer-aided software engineering (CASE) tool to facilitate *Sentire* specific workflows within a collaborative environment (see Section 5.7).

5.4 *Sentire*

This section outlines *Sentire*'s main milestones and deliverables with reference to the various sections marked in Figure 5.4.

Sentire was used in four case studies out of which several insights emerged. These will be discussed in Chapters 6, 7, 8 and 9. The mechanics were confirmed in the earlier studies and results were validated in subsequent interventions. The degree of agility is dependent on the project's number of stakeholders involved, the need for documentation, the geographical dispersion and other factors. *Volere* does not prescribe a mandatory set of tools, techniques or deliverables, however a discussion on their merit is given in Chapter 5.5. The following will outline the main stages shown in Figure 5.4, wherein the darker elements are *Volere* specific milestones and the lighter elements are *Sentire* specific extensions.

Volere related descriptions and process-related details are adapted from Robertson and Robertson's primary publication on this technique [138].

1 – Project blastoff This is also referred to as project initiation or kick-off during which the project is given its own life within a well established context. Specific deliverables include the project purpose and goals, scope of the project, list of stakeholders, constraints, glossary of terms, domain properties and assumptions, cost estimates, risks and a

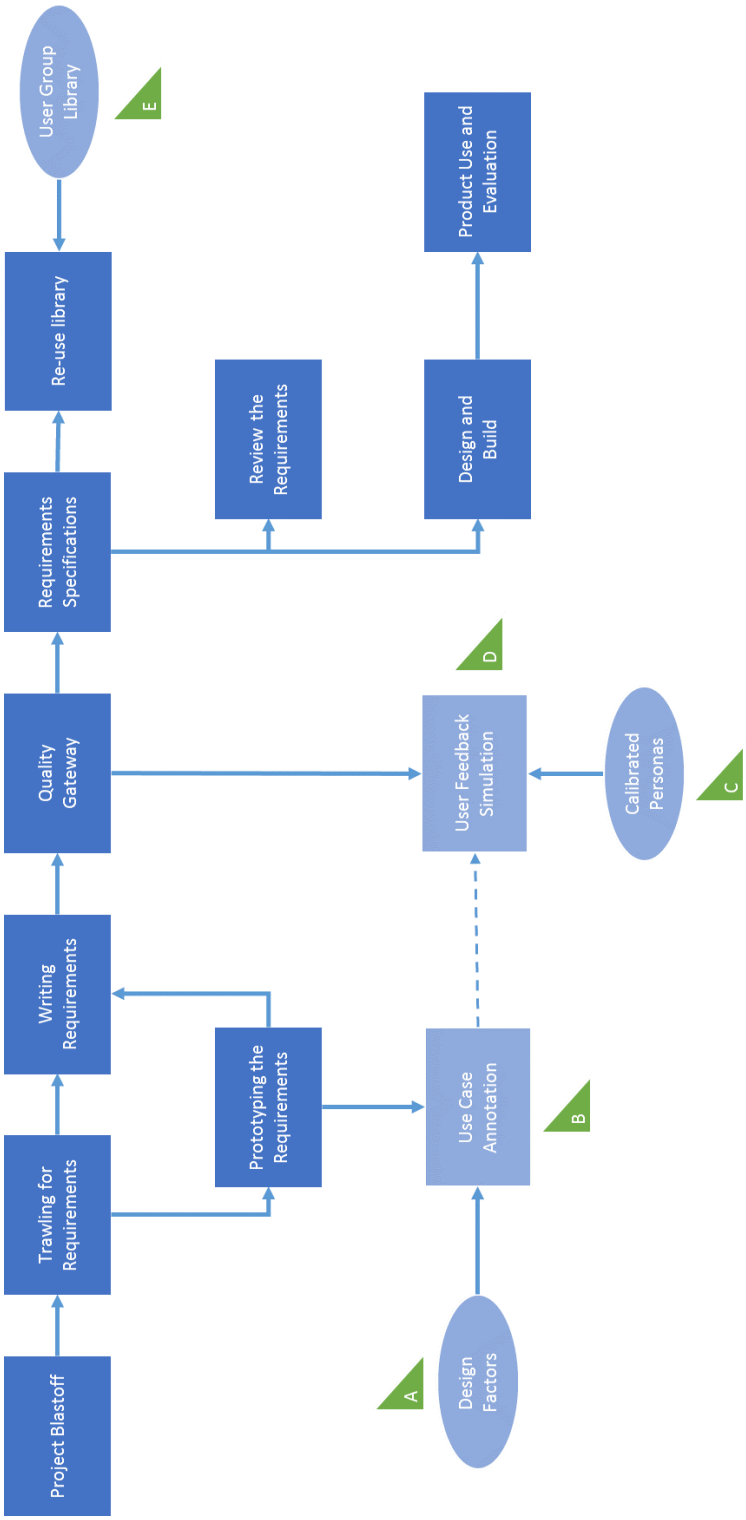


Figure 5.4: *Sentire*, a requirements framework that listens to personas – extending the *Volere* process. *Sentire*-specific steps are labelled from A to E and are drawn in a lighter shade.

low-fidelity prototype of the system-to-be (see Table 5.1). This provides a good foundation to take a go or no-go decision on the requirements process itself based on the project's perceived feasibility. A requirements analyst acts as a coordinator helping the principal stakeholders identify the business problem that the project will solve and draw a clear line around the scope of the solution under investigation. Principal stakeholders include the product client, product users, technical experts, business experts and requirements analyst. Any other stakeholder who might be critical to the project's success should be involved at this stage.

Table 5.1: Project blastoff (deliverables)

Deliverable	*Description*
Context model	Consensus on the scope of the business area under investigation and its interaction with external entities
Stakeholder analysis	The more primary stakeholder are identified at this stage the better the analyst will be able to maximise requirements discovery within the scope defined. The onion model for stakeholder discovery is generally adopted at this stage
Project purpose and goals	Consensus as to why this project is needed and its advantages. A set of PAM triplets are generally used (Purpose, Advantage, Measurement)
Facts and assumptions	This includes business rules, domain related facts and assumptions
Glossary of terms	Terminology and acronyms relevant to the specific business domain
Constraints	A list of solution constraints, external constraints, constraints from the use of off-the-shelf software, scheduling constraints and budget constraints
Risks	Potential project specific risks
Estimates	Cost estimates of the requirements development process
First-cut prototype	Generally a low-fidelity prototype indicating how the solution might materialise
Decision	Based on the above, should the team proceed with the requirements development effort and the project itself?

2 – Trawling for requirements

The project scope together with the list of stakeholders are indispensable inputs for this phase of the project, used as the primary starting point for the requirements development process. This phase is divided into four steps; identifying the business events, specifying business responses to these events (business use cases), specifying how the new product will assist the business response (product use cases) and finally the requirements for the product to ensure adherence to the original objectives. Several techniques can be adopted throughout this stage, including apprenticeship, workshops and interviews.

Events can be externally initiated (e.g., client sends request for quotation) or time triggered (e.g., end-of-month processing of salaries).

Business use cases are triggered by one or more events, affect or are affected by one or more stakeholders and is operated by one or more active stakeholders. The use of scenarios to specify business use cases is encouraged, including *normal*, *alternative*, *exception* and *misuse/negative* scenarios. The analyst should also specify

the expected outcome as well as the overall business rules guiding the business response.

Product use cases specify what the new product will do to assist in the business response (e.g., automating parts of the process or eliminating aspects of it). In formulating the product use cases, the team should not simply automate the existing business processes, but find ways to improve, innovate or obliterate them. The authors behind *Volere* state that the "*hardest part of requirements gathering is discovering the essence of the system .... the underlying business reason for having the product*", and using this to deliver "*a truly useful product*". This is contrasted with the more common practice of specifying a perceived solution to a perceived problem, by simply automating an inefficient process [138]. Scenarios are used in both business and product use cases. This helps both the analyst and the stakeholders agree on the steps required for the user to complete a task (use case). Scenarios in product use cases are used to determine how a particular feature of the system-to-be, as well as its users, should behave to achieve the desired objectives. These scenarios are then used to elicit the actual requirements. Deliverables for this stage are shown in Table 5.2.

Table 5.2: Trawling for requirements (deliverables)

Deliverable	*Description*
Events	A list of externally initiated and time triggered events, including expected inputs (what is required for the event to execute) and outputs (what is produced following the event)
Business use cases	A list of technology agnostic use cases describing how the business reacts to any of the identified events. Use cases also include normal, alternative, exception and misuse scenarios. The outcome is also defined, as well as business rules governing the use case. Artefacts related to this use case are also gathered for use in the subsequent stages
Product use cases	A list of technology-specific use cases describing the role of the system-to-be in assisting, improving or replacing the business response. Step by step scenarios are defined for each product use case

3 – Writing and prototyping requirements

Once the product-to-be is well defined through a set of product use cases the analyst moves on to specify the requirements that will guide the development of the new system. Requirements could be tackling aspects such as functionality, look and feel, usability, performance, maintainability, support, security as well as cultural, political and legal considerations. Stakeholders are generally capable to contribute to one or more of these requirement categories and it is up to the analyst to get the right feedback from the respective stakeholders. Requirements are specified using a concise, yet rigorous, template referred to as the requirements *shell* or *snow card* – see Figure 5.5. This template helps the team to agree upon and truly understand the rational behind the requirement and a simple technology-agnostic language is used. Conflicts with other requirements can be specified as well as tracking infor-

mation such as requirement originator, association with product use cases, supporting material, change history and prioritisation. Furthermore, requirements must be testable and for this purpose fit-criteria are specified. This helps developers understand what needs to be achieved while helping testers ensure that any deliverable meets the requirements.

Prototypes for product use cases or for associated requirements may be built throughout the process. These could help explain complex aspects of a requirement or to generate a better understanding of what can be achieved with the existing knowledge and tools. Low or high-fidelity prototyping techniques can be adopted which can in turn inform the requirements development process by providing input signals to generate further insights and elicit new ideas/feedback from the project team or from potential users (if available). New requirements may be developed as well as new product use cases in an iterative process of discovery.

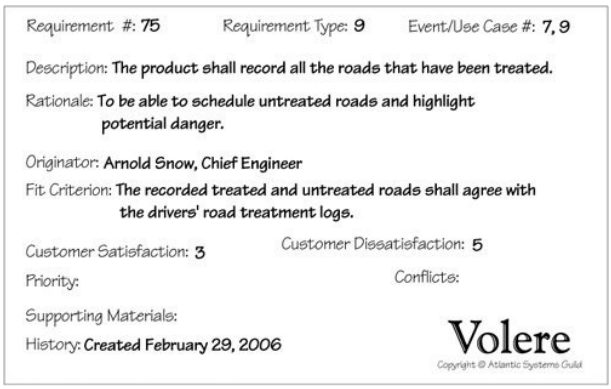


Figure 5.5: The *Volere* requirements shell or snow card

Table 5.3: Writing and prototyping requirements (deliverables)

Deliverable	*Description*
Requirement snow cards	A set of requirements following the requirements template and formatted using the requirements shell. These are traceable, testable and rigorously compiled in an iterative manner and based on scenarios defined in the system-to-be's product use cases.
Prototypes	Low or high-fidelity prototypes on product use cases or on specific requirements. These prototypes may inform the requirements development process by exposing unforeseen scenarios

3.1 – *Use case annotation*

Each product use case may contain multiple scenarios (e.g., normal, alternative, exception and misuse scenarios), each of which is defined as a series of steps. Use case annotation is carried out at step level, whereby steps containing enrolment-specific design factors are annotated accordingly using the units of measurement specified in Table 4.1. These annotations, together with active actors associated with use cases (i.e., Calibrated Personas) are then used to compute user feedback simulations.

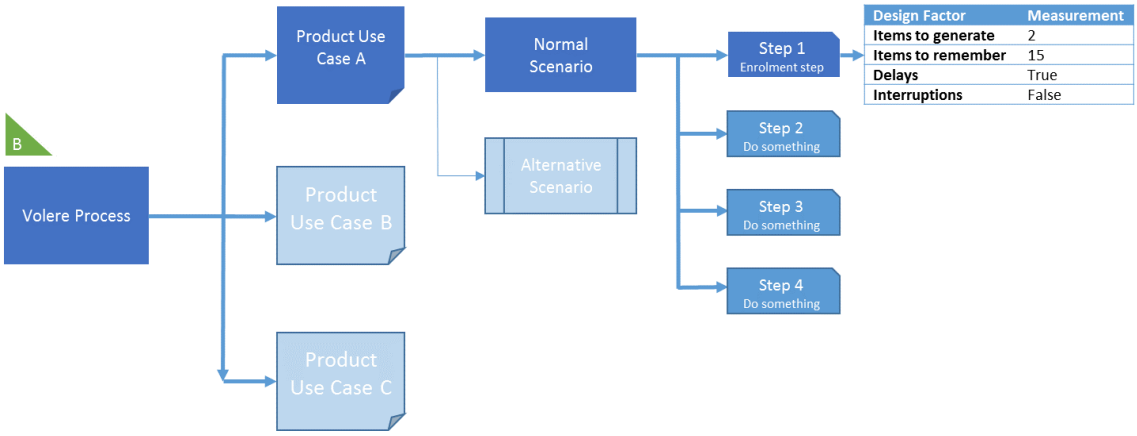


Figure 5.6: Enrolment-specific use case annotation

This thesis tackles enrolment as a critical design decision, however there are other aspects that may be explored and adopted as part of the framework (e.g., privacy issues and disclosure). To study and uncover other design factors for critical design decisions, the researcher can adopt the process outlined in Figure 5.7. More information on each individual step is given in Chapter 4.1.

Table 5.4 outlines the main deliverable for the use case annotation process.

Figure 5.7: Process to uncover design factors for critical design aspects

Table 5.4: Use case annotation (deliverables)

Deliverable	*Description*
Annotated product use cases	Product use case scenarios are studied and annotated (where applicable) with measurements related to the design factors under consideration. In this case, all enrolment steps within the various scenarios were annotated as shown in Figure 5.6

4 – Quality gateway

The quality gateway is a quality assurance step through which requirements are studied for correctness, thus encouraging an iterative process of refinement. *Volere* suggest a number of quick tests, which may just be yes/no answers to questions such as: *Is the fit-criterion specified?, Is the data needed for the requirement available as per the context model?, Is there a rationale for the requirement?, Is the requirement technology-agnostic?, Is the stakeholder value defined for each requirement?.*

James and Suzanne Robertson argue that making "*requirements visible, and testable, means that you discover any problems as early as possible, and correct*

them before they become major issues". Based on this philosophy the quality gateway is extended with user feedback simulations (section D in Figure 5.4). This sub-process depends on Calibrated Personas which are in turn based on existing statistical user behavioural models stored within the user-group library (see Chapter 4.4).

Table 5.5: Quality gateway (deliverables)

Deliverable	*Description*
Correct requirements	Following a quality assurance process, requirements are revisited for correctness, testability, fitness and validity. This process may also uncover new requirements or processes that might result in minor or major changes impacting the requirements development process. Scoping is a crucial exercise and the stakeholders must decide what to tackle as part of the current project and what to specify as future capabilities

4.1 – *User feedback simulation*

This step adds a second dimension to the quality assurance process whereby active stakeholders (actors which are represented by Calibrated Personas) 'voice their opinion' on scenarios containing enrolment steps. Figure 5.8 shows the steps required for this process.

Inputs required to simulate user feedback include the annotated product use cases and the Calibrate Personas' underlying regression coefficients. With this information one can then execute regression analysis to generate predictions for perceived workload and willingness to complete the task (for each Calibrated Persona and for each scenario). Figures 5.9 and 5.10 outline the algorithms used to generate both predictions.

Once (1) personas are constructed for the current system, (2) linked with the respective user group models (regression coefficients), (3) associated with use cases and (4) use case scenario steps are annotated, one can then simulate user feedback in order to assess how the current design decisions may perform. Design measurements together with regression coefficients associated with the respective actors (Calibrated Personas) are parameterised into the respective regression functions (linear and logistic) in order to generate predictions for the actors' perceived workload and on the willingness to complete the task for each scenario. This information is then presented in a visual and actionable layout (i.e., dashboard).

Table 5.6: User feedback simulation (deliverables)

Deliverable	*Description*
Simulated results for the willingness to complete the task	Willingness to complete the task online is calculated using the algorithm outlined in Figure 5.9 and results are presented by user group
Simulated results for perceived workload	Perceived workload is calculated using the algorithm outlined in Figure 5.10 and results are presented by user group

As part of the quality gateway, these results should inform the requirements de-

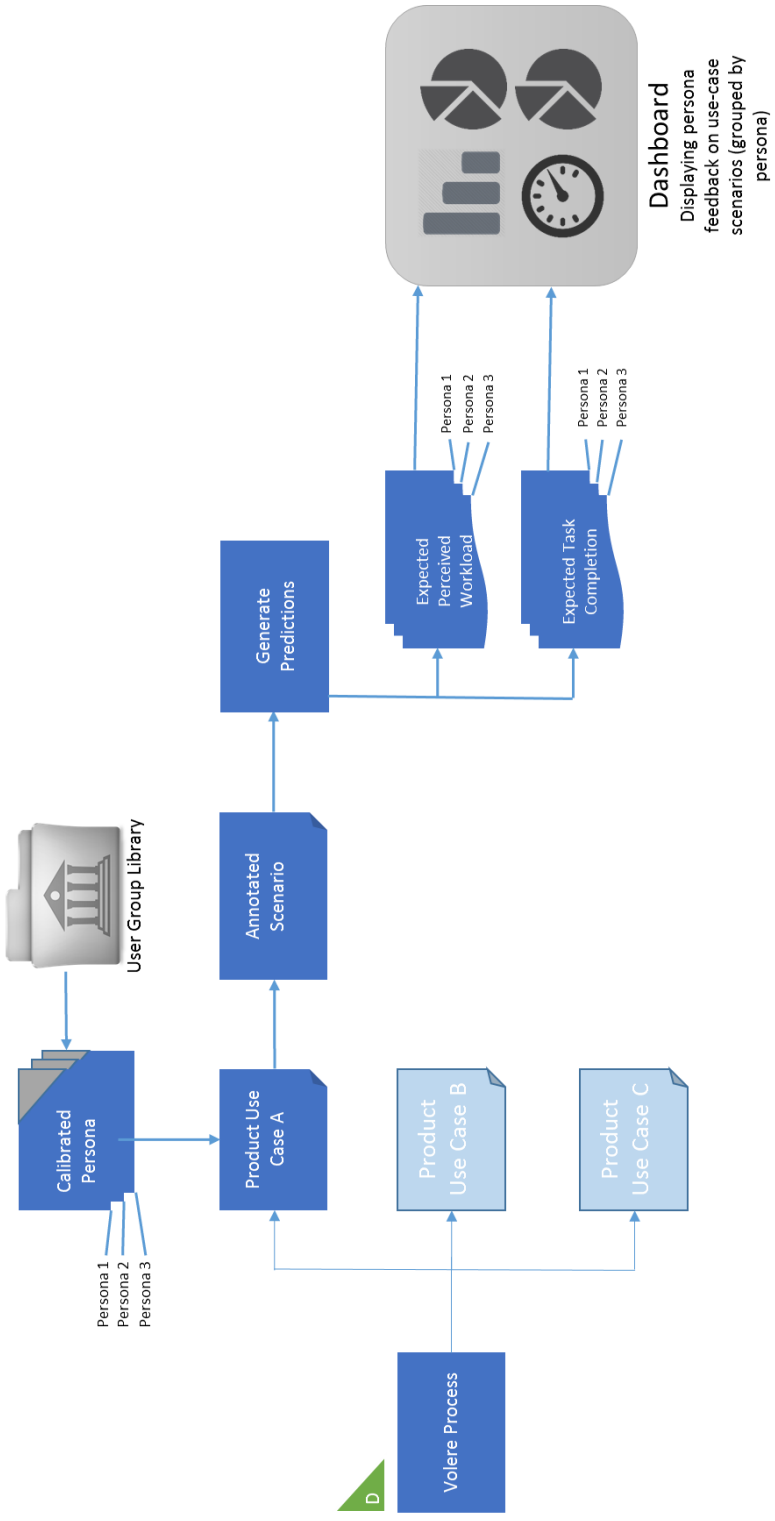


Figure 5.8: *Sentire* – generating simulated user feedback based on Calibrated Personas and annotated use case scenarios

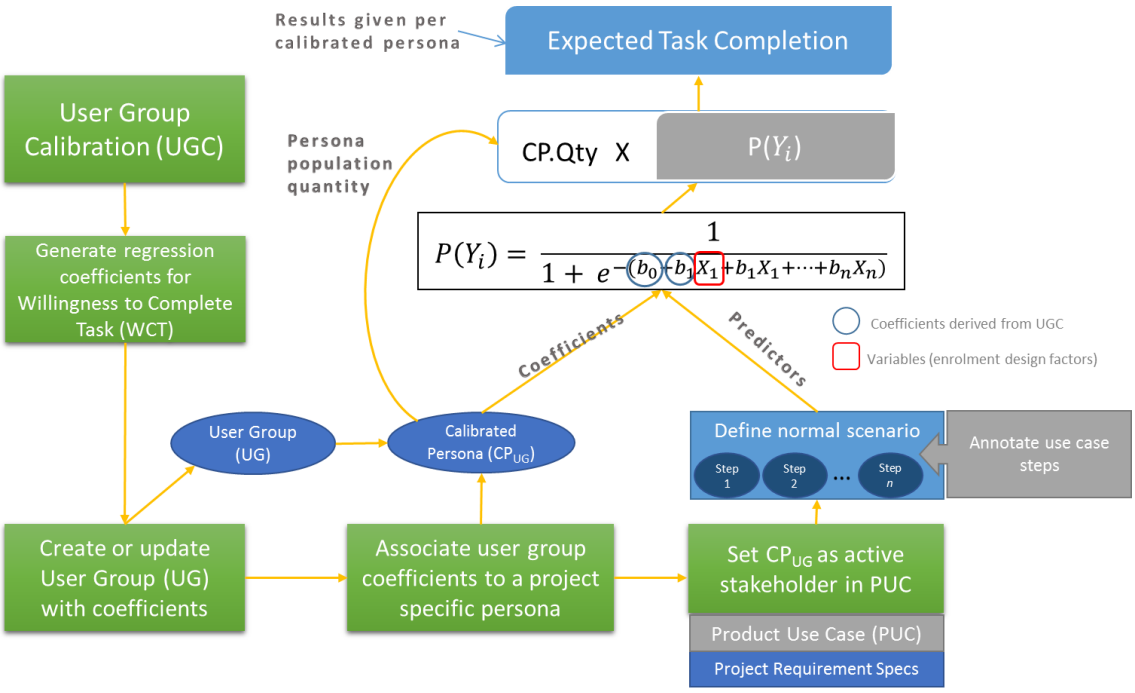


Figure 5.9: *Sentire* – simulating user feedback on the willingness to complete the task (for each user group)

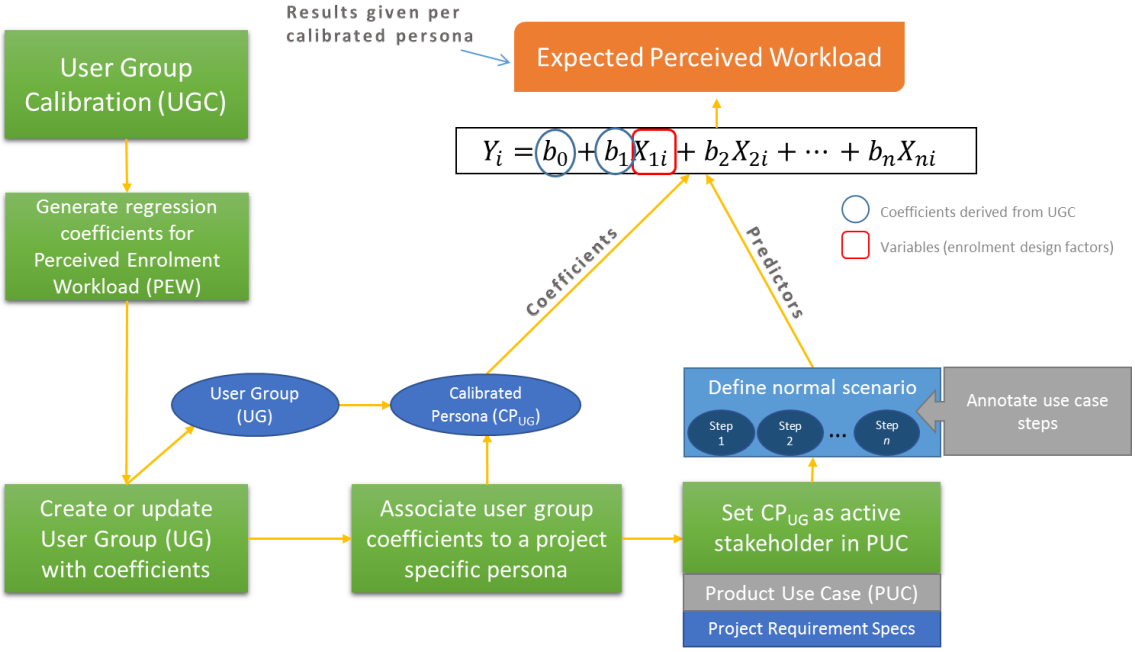


Figure 5.10: *Sentire* – simulating user feedback on perceived workload (for each user group)

velopment process and if a particular scenario results in unacceptable levels of perceived workload or low completion rates then the project team must revisit the respective use case and associated requirements to understand and tackle any issues that might inhibit take-up. Tweaking individual enrolment steps might help find the right balance between the required identity assurance and the users' acceptance thresholds for the different user groups involved in each scenario.

A CASE tool has been built in order to automate and facilitate user group calibration, the requirements development process as well as user feedback simulation. This is discussed in further detail in Chapter 5.7.

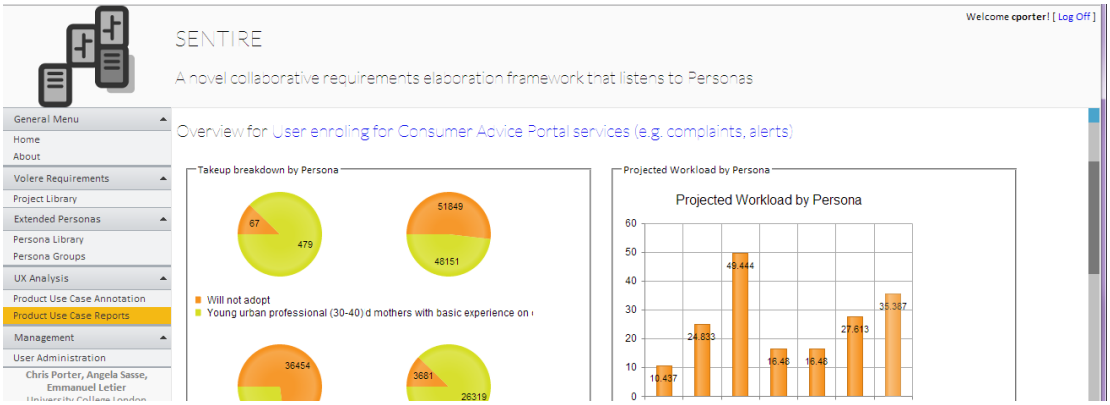


Figure 5.11: *Sentire* – dashboard for simulated user feedback generated via a custom built CASE tool. The pie charts on the left denote each user groups' predicted willingness to complete the current product use case while the histogram (right) represents perceived enrolment workload (the various user groups are represented by Calibrated Personas)

5 – Requirements Specification

Following a rigorous quality assurance process, including simulated user feedback, the quality gateway will determine which requirements are to be accepted and written formally within the specification and which requirements will be rejected or revisited in an iterative process of improvement. This process is entirely based on the checklist of quality attributes, which also includes simulated user feedback as an additional check on top of the original *Volere* recommendations.

These simulated figures can be adopted both as fit-criteria for specific requirements (thus ensuring that the requirement is testable) but also as a metric to inform the design of product use cases (providing intermediary figures indicating whether a specific enrolment-centric use case is acceptable by the various user groups involved).

Requirements must be testable, and thus measurable. If one cannot test a requirement then it is very difficult to determine whether the product meets the specifications. Any requirement must go through the quality gateway to determine its validity, testability and origin.

Especially when development is outsourced, the project team could specify the required levels of identity assurance while denoting an acceptable level of task-abandonment for different user groups, as shown in Figure 5.12. Without being too prescriptive on implementation details, requirements can be specified using measurable levels of impact that the solution can have on the various user groups. Developers must in turn produce a solution that adheres to these criteria while respecting users' behavioural patterns and workload acceptance thresholds (i.e.,

for specific types of services). In a collaborative and iterative effort, developers can eventually test out their proposed designs (i.e., product use cases) through the CASE tool's user feedback simulation module.

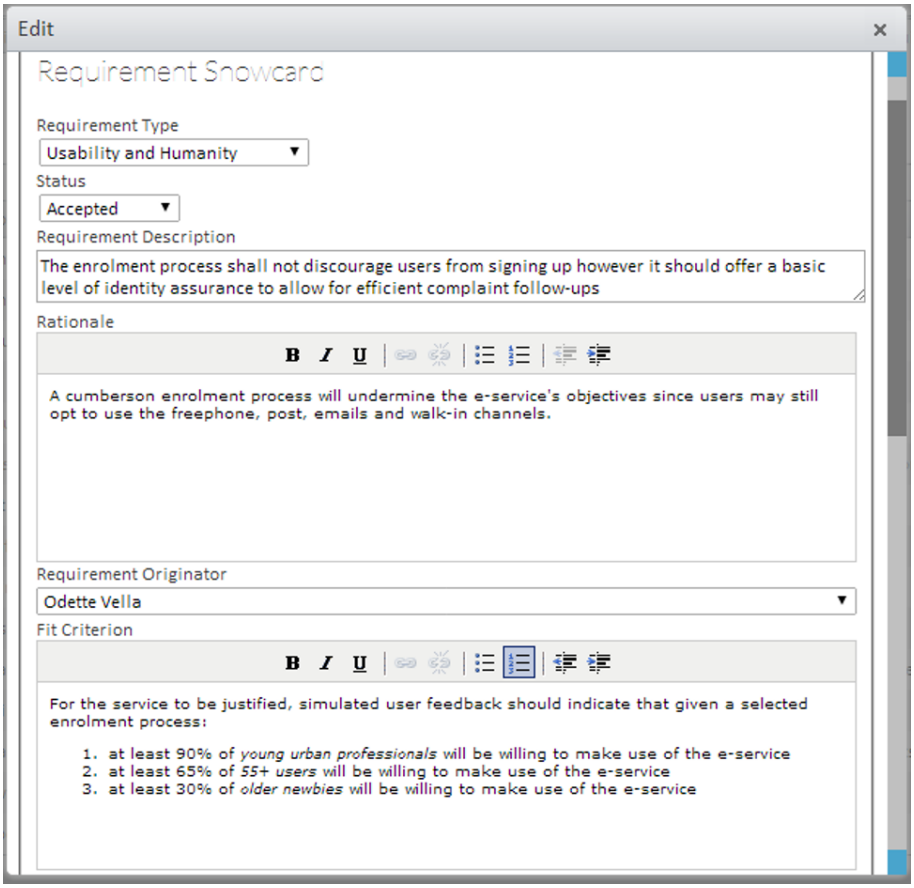


Figure 5.12: A requirements snow card indicating precise fit-criteria for a usability and humanity requirement based on the business owner's experience, insights and previous cost-benefit assessments. These will in turn inform an iterative product use case design process guided by verifiable base-conditions indicating when an acceptable design has been reached (measured through simulated user feedback)

Table 5.7: Requirements specification (deliverables)

Deliverable	*Description*
Requirements specification (*Volere* Template)	A set of accepted requirements are compiled following the *Volere* requirements specification template. The CASE tool built for *Sentire* produces output compliant with the *Volere* requirements specification template. This will be discussed in Chapter 5.7.

6 – Reuse library Reuse is central for a government-wide requirements and design framework. Following a number of case studies it was found that most Calibrated Personas in the library were directly re-usable for subsequent studies. This was also the case for most of the non-functional requirements. This promotes inter-departmental and inter-entity knowledge accumulation across several projects, minimising the risk of missing out on previous knowledge while streamlining the requirements trawling process. Given the way requirements are specified, including categorisation, links

to the originator and rationale, one can easily find and adapt existing requirements for use in new projects.

Sentire adds behavioural models to the *Volere* reuse library. Statistical models explaining user attitudes towards critical design decisions (i.e., enrolment) are stored within the reuse library and associated with relevant project personas whenever necessary (across different projects).

Table 5.8: Reuse library (deliverables)

Deliverable	*Description*
Re-usable requirements	Some requirements may be valid across several projects and domains. For this reason requirements are to be made available for reuse in future e-government projects. These requirements can then be reused without modifications or as a basis for some other requirements. The CASE tool for *Sentire* provides a library of requirements which can be imported into new projects.
Behavioural models for different user groups	A set of statistical models explaining user behaviour are produced for different groups of users (i.e., user group calibration). These models are stored in a knowledge base and used whenever relevant project personas (representing a specific user group) are constructed for a new e-service (i.e., models are associated with the project persona, turning it into a Calibrated Persona).

The user group library is discussed in some length in Chapter 4.4. Figures 5.13 and 5.14 depict the user group library as implemented within the *Sentire* CASE tool.

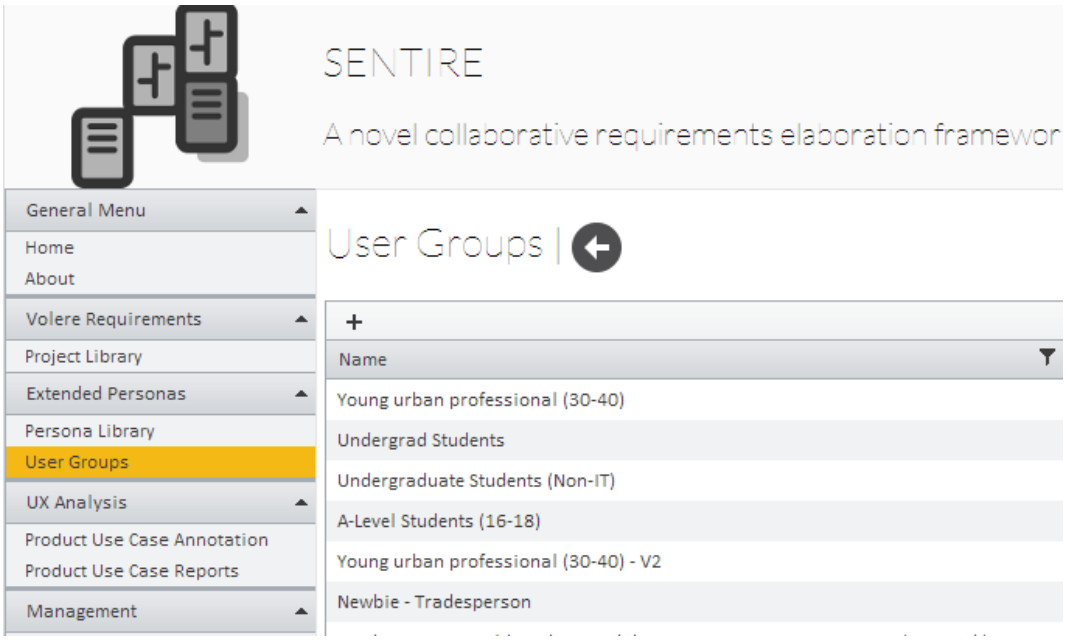


Figure 5.13: *Sentire* – user group library implemented within the CASE tool

7 – Review, Design, Build and Evaluation

So far, the process was iterative and incremental with requirements specified throughout the process and validated within the quality gateway. Re-design of certain aspects might require the team to go back to the requirements trawling stage to

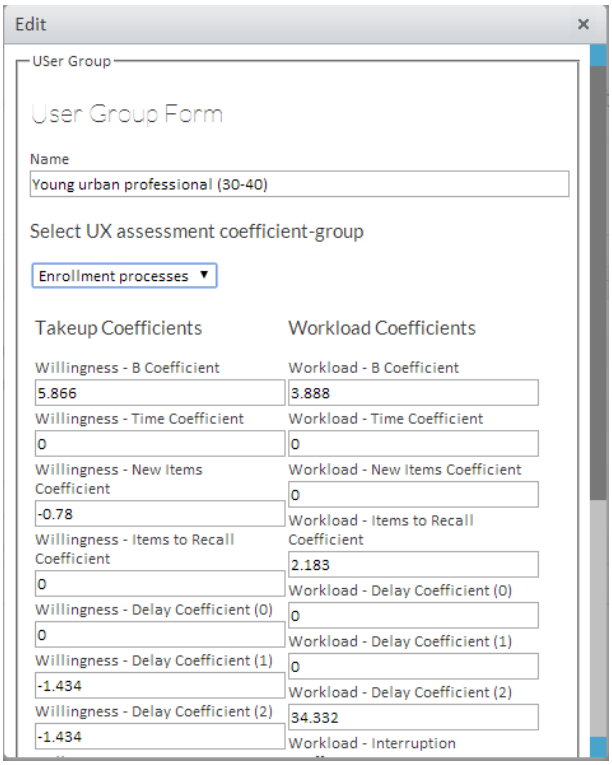


Figure 5.14: *Sentire* – assigning regression coefficients (for the *WCT* and *PEW* models) to a user group

dig for more information and produce or refine use cases. This could occur if the team finds missing requirements or unexamined business events (and in turn use cases). An iterative review process is suggested in the *Volere* process which also includes prioritisation of requirements and estimation of effort required to build the product.

Once the team is satisfied that the specification is robust (i.e., with no missing and conflicting requirements) it may now proceed to the development, tendering or contracting stage. It is important at this point that a complete picture is available whereby risks and costs can be determined confidently. If the final specification has outgrown the original estimates the team may decide to split the project into different phases, starting off with the core requirements.

James and Suzanne Robertson [138] argue that the *Volere* process is iterative and incremental, and development can start before the requirements development process is exhausted. The team might decide to investigate a few use cases on which developers can then start working. In the meantime, another set of use cases could be studied. Agile projects may use incremental prototypes to guide the requirements development process itself, which prototypes might eventually find themselves in the public domain. In this case, user feedback simulations are quintessential during the prototyping stage, informing the project team on critical design issues before these are actually built.

On the other hand, in projects for which external contractors will be involved (e.g., tendering process), requirements must be specified and guaranteed through testable fit-criteria as shown in Figure 5.12.

5.5 Supporting Tools and Techniques

A number of techniques can be adopted and adapted for use within the *Sentire* framework, and their selection is highly dependent on the level of agility required. If agility is a priority, user-centred prototyping tools such as eye-tracking and card-sorting can (and should) be adopted as part of the requirements, design and development iterations. Their use on low or high-fidelity prototypes would help the primary project stakeholders gain additional insights on current assumptions (e.g., *users shall be able to locate an answer within 30 seconds*). Eye tracking technology helps in measuring and assessing users' interaction with existing prototypes while card-sorting can be used to improve them mainly by building a better understanding of the users' mental-models with respect to the product being developed. Most of these techniques are already discussed elsewhere within this thesis, however a recommended set of tools will be provided together with their context of use as well as the main rationale for adoption.

Onion Model *Used in: Stakeholder discovery*

Ian Alexander's onion-ring model for stakeholder identification provides a structured way to maximise the identification of possible people, entities and systems that may have an impact on or are impacted by the system being developed. The *Volere* specification template provides several pre-defined categories that can be used to comb the business context for stakeholders.

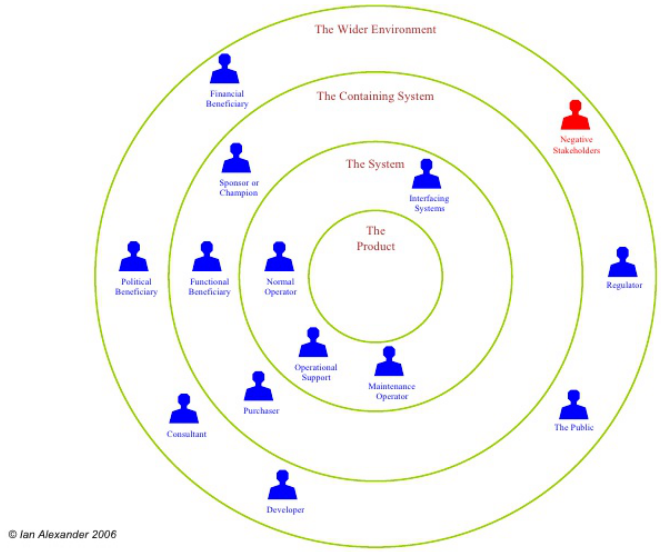


Figure 5.15: Ian Alexander's stakeholder taxonomy and the onion model

Personas *Used in: Trawling for requirements*

Personas are used to encourage the design team to make conscious decisions respecting human cognition levels, capabilities and emotions. In [38] Cooper stated

that personas help to unlock "*the power of visceral, behavioural and reflective design*". Secondary users, whose needs should also be met however not as explicitly as the primary personas', are also considered throughout the persona creation process. Persona creation should follow a participatory method whereby project stakeholders are encouraged to provide their own views on, and experiences with current clients (i.e., potential end users, initially considered as persona hypotheses). This information would enrich the corpus of data upon which primary and secondary personas are crafted. As Dotan et al. [44] suggest, direct stakeholder involvement in this creative process increases the chances of buy-in, persona believability due to familiarity and ultimately methodological success [44].

Persona Cases *Used in: Trawling for requirements*

Faily and Fléchais propose the use of Persona Cases which attempt to legitimise the validity of personas by grounding their characteristics in, while making them traceable to the originating source of empirical [qualitative] data [55]. Any persona attribute should be grounded in empirical evidence, making it harder to rebut persona-based design arguments while steering away from over-generalisation and stereotyping (see Section 2.2.4 for a complete discussion on Persona Cases).

Calibrated Personas *Used in: User Group Calibration*

Calibrated Personas can act as a design aide providing immediate and quantitative, thus measurable and comparable user feedback on the impact that critical design decisions (e.g., enrolment) can have on users. Through a calibration process (see Section 4.3.1), statistical models are built for different groups of users to explain and predict behaviour. These models are then associated with project personas to produce simulated user feedback on project-specific design decisions. This promotes objectivity in the decision making process by introducing testability within experience related requirements through the use of measurable fit-criteria (e.g., *85% of [user group] shall be able to complete the task in less than 5 minutes with a perceived workload level of less than 30%*).

Statistical packages *Used in: User Group Calibration*

Data collected during user group calibration sessions is consolidated and prepared for processing using a statistical package (such as SPSS). Two regression models are created: a linear regression model to explain perceived workload and a binary logistic regression model to predict the users' willingness to complete the enrolment task and use the e-service to complete the primary task online. Statistical packages simplify and speed-up model fitting and testing.

CASE tool *Used in: Entire process*

CASE tools assist the project team to manage the requirements development process while making it transparent to all of the primary project stakeholders. A CASE tool for *Sentire* was developed to support and manage the whole process through an online collaborative workspace. See Section 5.7.

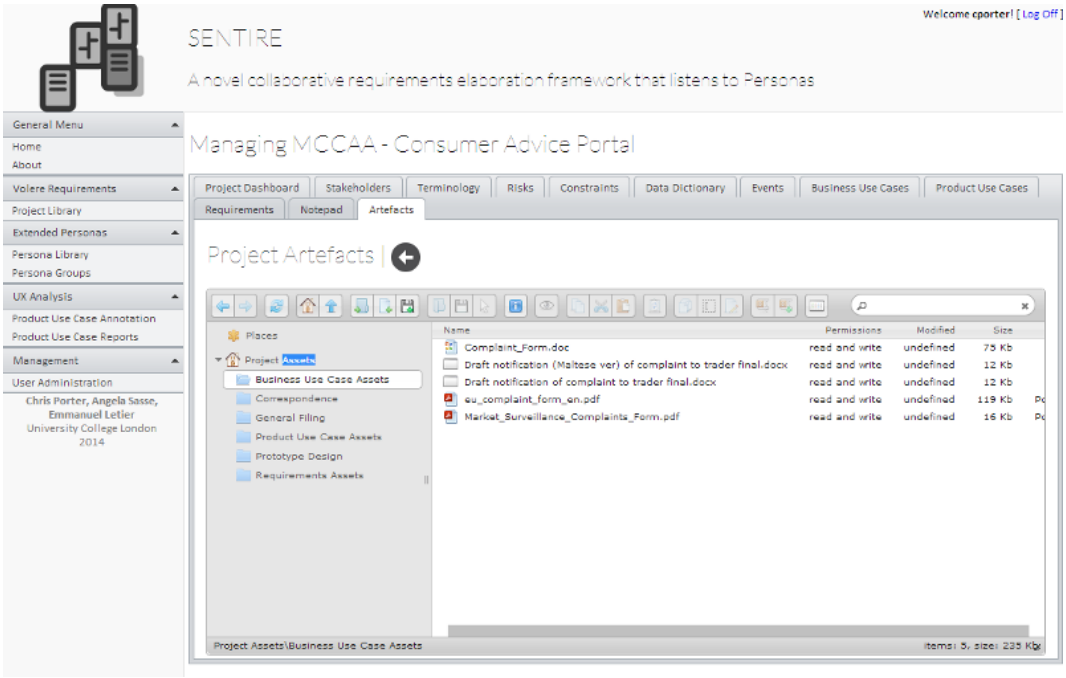


Figure 5.16: Artefacts library in *Sentire*'s CASE tool

Participatory design

Used in: Entire process

This is a design approach that promotes active involvement from users, which also extends to decision making. Further to this, participatory design considers systems to be "*networks of people, practices, and technology embedded in particular organisational contexts*"[3]. Design workshops should include as many stakeholders as possible. *Sentire* adheres to this philosophy since it also compensates for missing user group representatives through Calibrated Personas.

Card sorting

Used in: Prototyping (agile projects)

This technique helps designers discover the "*users' model of the information space*" [112]. The outcome of a card-sorting exercise is generally a reorganisation of the product's interface and the way information is organised hierarchically. Users are asked to organise cards representing various concepts (identified during the requirements development process) into piles reflecting their perception of similarity. These groups of concepts are given a name (using a separate card – generally colour coded). This process of organisation, categorisation, linking and

[3]CPSR (2000), http://cpsr.org/prevsite/program/workplace/PD.html/, (accessed April 2014)

naming provides the design team with insights into how users perceive product-related information areas or concepts. This process can be initially carried out within the primary stakeholders team to generate an initial architectural outline of the information and functions involved. Card sorting participants have to agree on naming conventions for groups of concepts, and this can inform the creation of menus, menu items and layouts for a low or high-fidelity prototype. Pre-determined categories can be used throughout the session, however new categories can also be introduced (*closed* versus *open/hybrid* card sorting).

Figure 5.17: Card sorting can be used to inform the e-service's information architecture

Eye-tracking and RTA

Used in: Prototyping (agile projects)

High-fidelity prototypes can be assessed for usability issues using eye-tracking technology. Eye-tracking sessions are generally conducted using a set of pre-determined tasks (goals) during which eye-gaze data is captured for a deeper assessment on findability, navigability and explicit pain-points (e.g., uncover failed attempts to find the search button through heat-maps). Retrospective Think Aloud (RTA) sessions provide deeper and invaluable insights and knowledge on what users expect, what they look for and the rationale behind their decisions. This data supplements the eye-tracking information. Results from these sessions can be used to inform the creation and evolution of both functional and non-functional requirements.

5.6 Observations and Criticisms

Analytical experience assessment (UX-analytics) on critical design decisions at the requirement stage brings forth a systematic process of understanding and knowledge refinement on the e-service being de-

veloped and the targeted user groups. This does not necessarily mean that *Sentire* will capture all issues in enrolment processes, however it offers a means to capture the more egregious and potentially disruptive problems based on current knowledge. User studies are still valuable to capture any unexpected design issues related to aspects such as screen layout, navigation and so forth. *Sentire* adds due diligence to the design process to mitigate late-changes on core design aspects of a system while at the same time placing the user and user experience at the core of the design workflow. This is mainly because *Sentire* computes and returns simulated user feedback (in this case on willingness to complete a task and perceived workload) at each design step (i.e., scenario building). User feedback is optimised through a process of calibration which produces statistical behavioural parameters for different groups of users. These parameters could then be reused across projects that target the same group or groups of users, thus improving the return on investment associated with the calibration process.

Sentire works on the null-hypothesis that the current process being designed for a new e-service is conducive to a positive user experience. For this reason Type I and Type II errors may occur. Type I errors occur when a flag is raised on a non-issue, thus incurring unnecessary expenses due to time and effort spent on attempting to identify the issue and resolve it. On the other hand Type II errors occur when critical issues in design are not flagged and the design issues go through unnoticed (hypothesis is accepted even if false). These may then be captured during user testing at a later stage or post-launch. Changes to the original design may turn out to be expensive, cause unnecessary delays, and possibly, political embarrassment. Traditionally, comprehensive and empirical user testing is not scalable unless enough time and money are injected into the process. The UK has recently run tests on its government portal with hundreds of user walkthroughs, an unprecedented effort in the public sector, however repeating it over a number of iterations to test fresh changes in design may become prohibitively expensive.

5.7 Tool Support

This chapter outlines the design principles adopted for the CASE tool built specifically to support the *Sentire* process, which ultimately produces *Volere*-compliant requirements specifications.

5.7.1 Policy makers and CASE tools

In a report on industrial practices, Alexander, Robertson and Maiden [9] note that only a third of respondents (out of 152) reported the use of tools to manage the requirements development process within their organisation. Some of the tools listed by practitioners include *DOORS*, *Requisite Pro*, *Rose*, *CORE*, *Cradle*, *Microsoft Word*, *Microsoft Visio* as well as home-grown software. The majority of respondents originated from industries such as aerospace (18), defence (43) and finance (21). Respondents from the public sector (8) reported the lowest level of tool adoption. Also, this group of respondents reported the lowest levels of formal education as well as professional training on requirements development processes [9].

Drawing on this information as well as personal observation, design decisions were based on the assumption that the majority of policy makers involved in the design of e-government services are (1)

generally not well versed with requirements engineering and user centric design processes, (2) have never or rarely used specialised CASE tools and (3) may also not be proficient in technology and associated practices, thus trusting other people with important decisions. Above all this, time is also a limited resource and big decisions have to be taken quickly. Policy makers should be actively involved in e-service development processes and any CASE tools adopted should offer full process transparency while being easy to set up and use.

5.7.2 Iterative prototyping

The CASE tool for *Sentire* was built iteratively following the assessment of learning outcomes across the four case studies presented in Chapters 6, 7, 8 and 9. Primarily the need for the tool was felt following the first case study (see Chapter 6) during which *Google Docs* was used as the main collaborative environment within which the various *Volere*-specific deliverables were maintained. The effort required to maintain and follow links between the different documents for events, use cases, risks, constraints, personas, stakeholders and requirements was compounded in time when more information was unearthed and formalised. Calibration and simulated user feedback generation was even more laborious since a complete disconnect existed between the *Google Docs* repository and the theory behind Calibrated Personas. User feedback prediction requires complex computations, prone to human-error and beyond the understanding of most primary stakeholders who were not involved in the construction of the theory itself.

Various tools exist which are to some degree compliant to the *Volere* requirements specification template however at the time of writing most available products required a commercial license, were closed-sourced and none of which offered extensibility capabilities (e.g., through application programming interfaces). The author's goal was to modularise, integrate and as much as possible automate the persona calibration and user feedback generation aspects of *Sentire*, while guiding the primary project stakeholders throughout the entire requirements development process following a simple workflow and within a collaborative environment.

5.7.3 First generation

Initially *Google Docs* was adopted, using a structured set of folders within which documents were stored. Spreadsheets were used for requirement snow-cards and business events, while regular text documents were used for business and product use cases. A separate tool was created to assist in product use case annotation while automating user feedback simulation. This was built as a throw-away proof of concept using the *.Net* framework (*C#* and *XAML*). It was completely detached from the *Google Docs* interface and couldn't be shared with the rest of the project team, however it resulted to be an important exercise that informed the evolution of the toolset. Due to this detachment all of the *Volere* process deliverables had to be re-keyed within this tool, including product use cases (re-typed from *Google Docs*), outcomes, actors, scenarios and so forth. Scenario step annotation could be carried out from within the tool itself, which values would then be parametrised into the regression module together with the regression coefficients for the active actor (i.e., Calibrated Persona). This prototype did not offer collaborative facilities, was too complex for stakeholders to comprehend and was not integrated within

the main requirements development workflow.

5.7.4 Learning from and evolving the CASE tool

Each case study presented a new set of challenges and lessons which informed the evolutionary path of *Sentire*'s CASE tool. Both *Google Docs* and the annotation/simulation proof-of-concept were discarded in favour of an online collaborative workspace built using the *ASP.Net MVC* framework and *SQL Server*. Several open-sourced JavaScript libraries were used to enhance the user experience, including *Cytoscape* (for complex network visualisation), *Elfinder* (a web-based file manager for project workspace assets), *Paintweb* (a web-based drawing tool) and *TinyMce* (a rich text editor).

The CASE tool was designed to be as usable as its underlying requirements development process, integrating the various sub-processes (including annotation and calibration) while abstracting complexities from the main workflows. Logically the tool provides four modules: *user group calibration*, *user group library*, *Sentire workflow* and a *Volere* compliant *reporting and simulation* module.

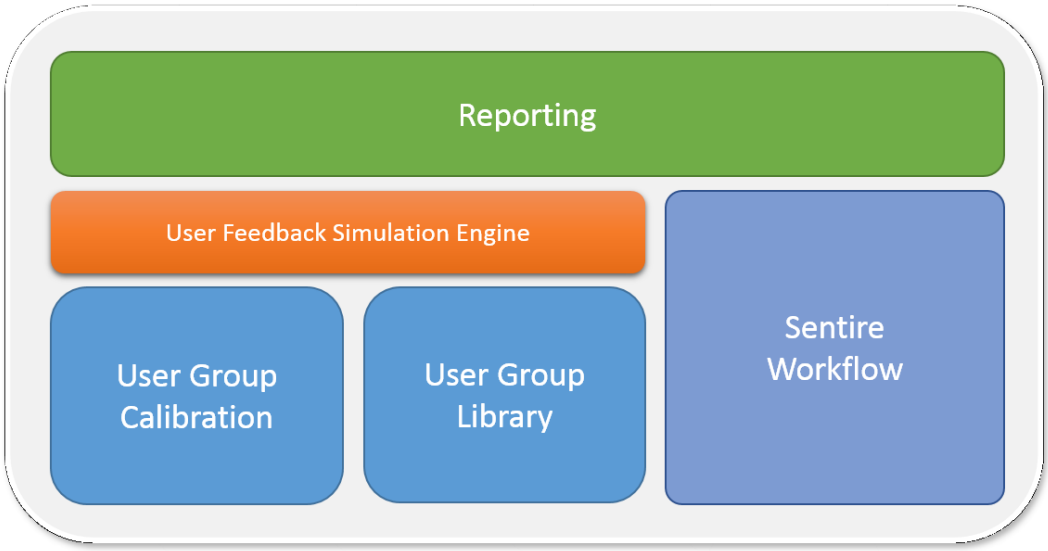


Figure 5.18: Conceptual view of *Sentire*'s architecture

The following sections will outline the various modules depicted in Figure 5.18.

5.7.4.1 User group calibration

User group calibration is offered as part of the main CASE tool, accessible via a specific sub-domain (at the time of writing this address was *http://calibrate.devbell.com*). Offering the calibration process as an online activity offers various advantages, including remote-assisted calibration, facilitated calibration in groups or mass calibration via a shared link. The user is guided through the process using a wizard-styled workflow, presenting calibration tasks in a random order while clearly explaining the terminology used.

5.7.4.2 User group library

User group models (i.e., two sets of regression coefficients) generated following a calibration exercise are stored in a central repository which are then available for (re)use across different projects (see Figures 5.13 and 5.14). Project specific personas can then be associated with any of the available user group

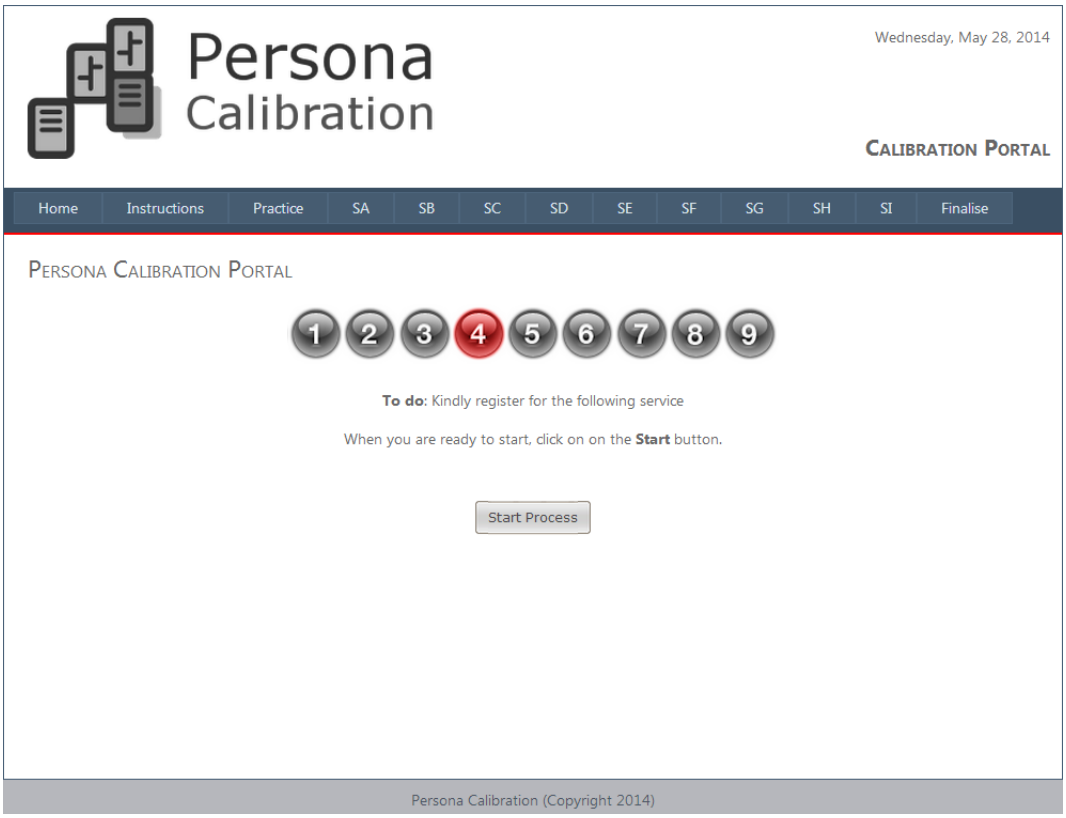


Figure 5.19: *Sentire* – presenting calibration tasks randomly to avoid the perception of incrementing workload

models, thereafter referred to as Calibrated Personas.

5.7.4.3 *Sentire* workflow

The guiding principle behind the design of the main screens were simplicity, findability and abstraction.

The layout presented in Figure 5.22 was developed over several iterations and a tile layout was finally adopted. This decision was based on several principles, including findability and familiarity [52]. Primary actions are simplified and made explicitly clear through the use of information tiles, aligning the tool to the "more with less" design principle[4]. Information tiles[5] provide a glimpse of the project's status (e.g., number of business events) while also doubling as buttons to open up the desired function point (e.g., business event management screen). Each tile represents a separate (yet sequential) section of the workflow.

5.7.4.4 User feedback simulation engine

This section is at the heart of *Sentire* and it abstracts the statistical complexities from the project team by providing meaningful, yet simple metrics on users' workload perceptions and willingness to complete the task at hand.

This section takes two sets of inputs – *user group models* (see Figure 5.13) and *use case annotations* (see Figure 5.23). These inputs are then used to generate user feedback simulations which are then passed on to the reporting module which is in turn responsible for visualisation.

[4]Microsoft (2014), http://www.microsoft.com/en-us/news/stories/design, (accessed March 2014)
[5]Microsoft (2014), https://community.dynamics.com/nav/w/designpatterns/110.instructions-in-the-ui.aspx, (accessed March

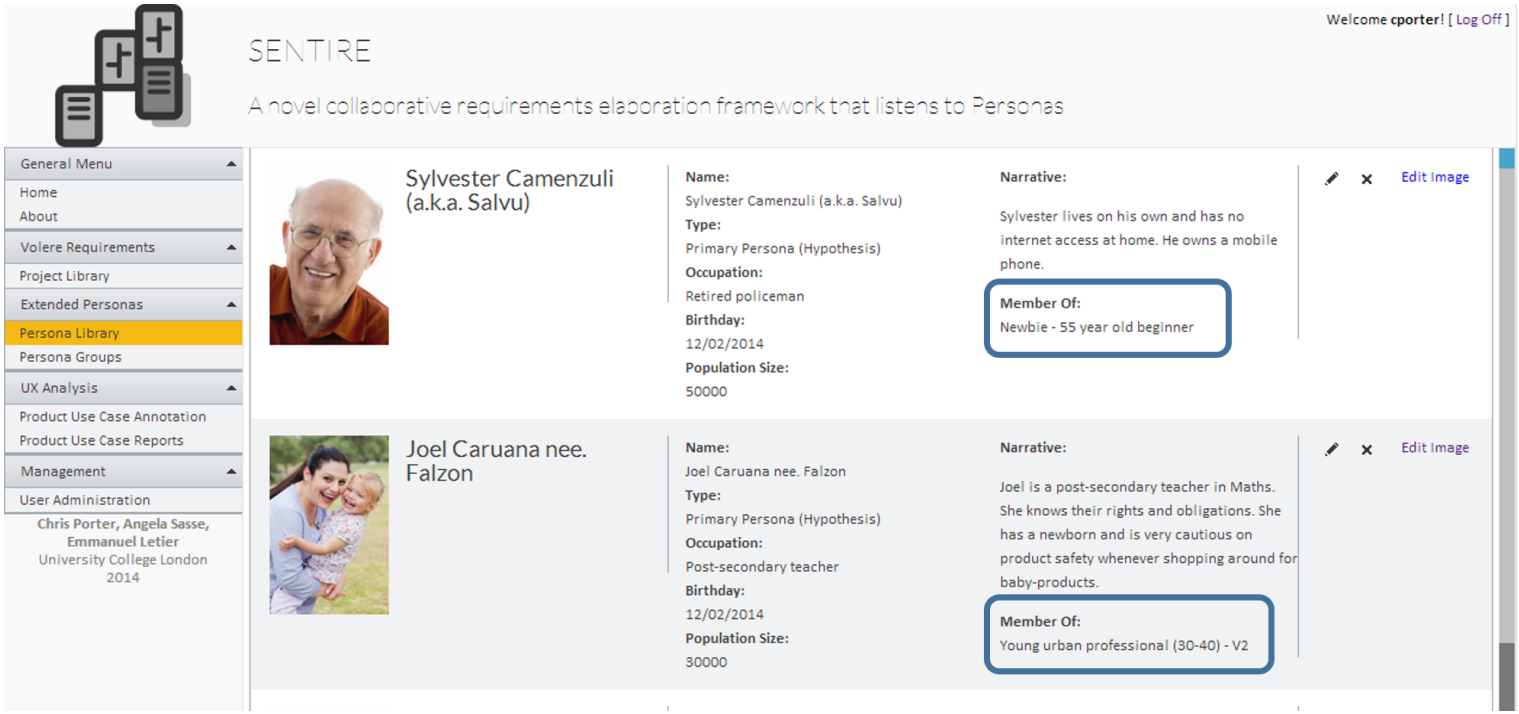


Figure 5.20: *Sentire* – Calibrated Personas (highlighted sections indicate the calibrated user group with which each project persona is associated – see Figure 5.21 for a user group example)

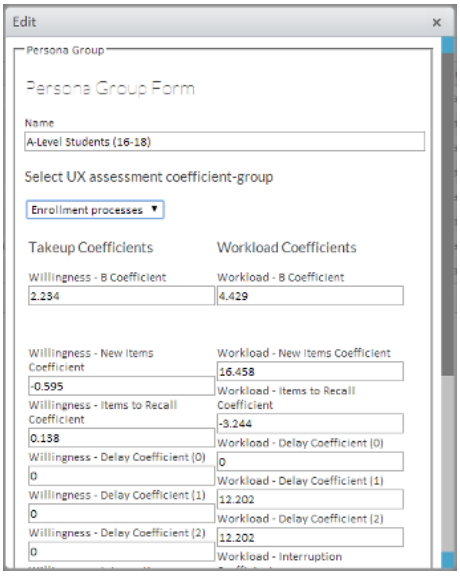


Figure 5.21: *Sentire* – user group dialog

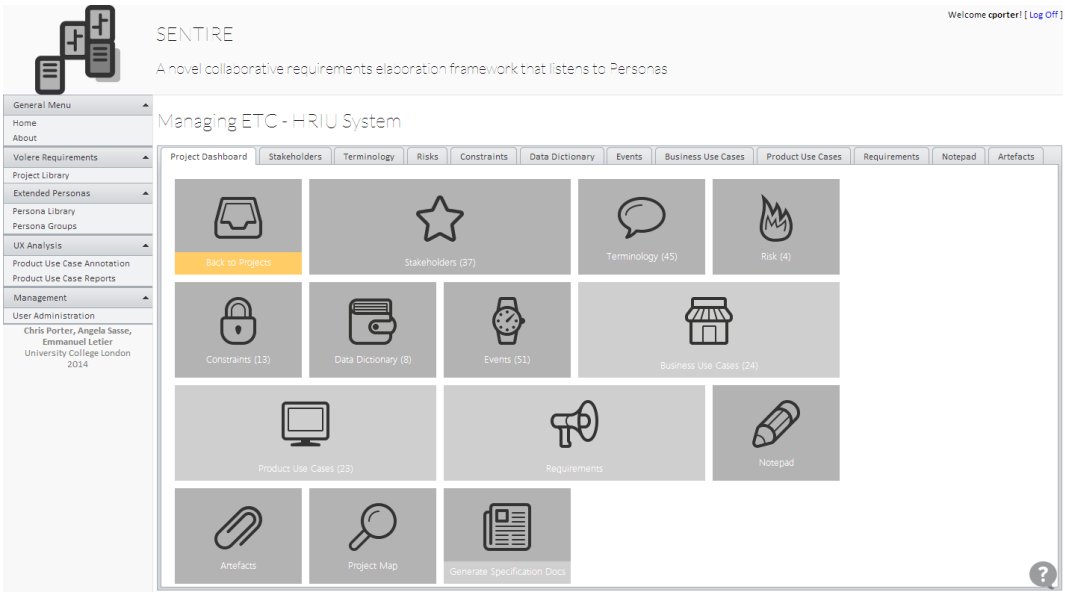


Figure 5.22: *Sentire* – project landing page

5.7.4.5 Reporting

The reporting module serves two purposes: *visualisation of simulated user feedback* and *generation of Volere-compliant requirements specifications*. Simulated user feedback is provided via a dashboard interface (see Figure 5.11). For each targeted group of users, represented by a Calibrated Persona, *Sentire* reports the amount of perceived workload (using a histogram) as well as the willingness to complete the enrolment task (using pie-charts). The dashboard interface allows the project team to assess the impact that a specific product use case might have on users without the need to carry out cumbersome calculations and without leaving the CASE tool (interrupting the workflow).

2014)

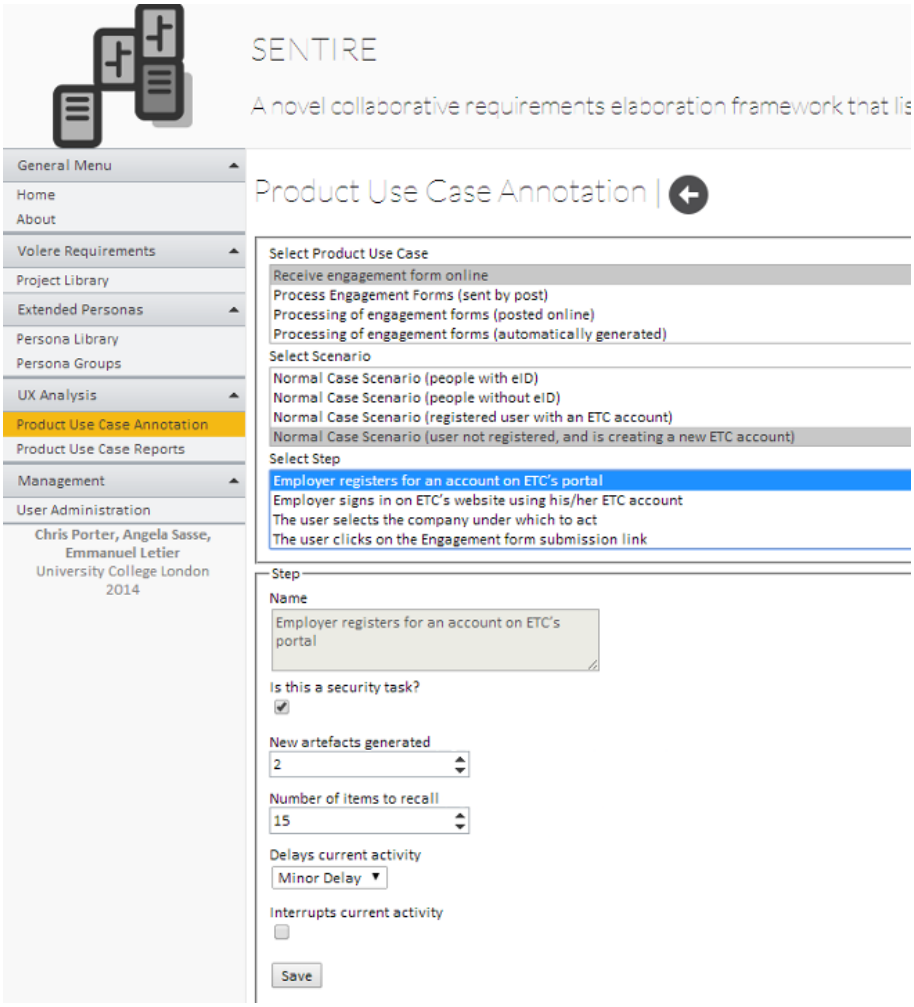


Figure 5.23: *Sentire* – use case annotation (average time is not being used in the calculations). Type of Service (*ToS*) is specified at product use case level (rather than at scenario level)

5.7.4.6 Domain visualisation

The underlying relational data model used for *Sentire* allows for the representation of the problem domain from various angles and perspectives. Traceability is a core *Volere* principle and this is promoted even further with dependency visualisation offered by the CASE tool. Figure 5.24 shows a network view of a project, allowing stakeholders to explore and view project associations and dependencies. This network view can be used to carry out impact assessments in case a requirement is modified, highlighting other requirements that might be affected through their associative links with product and business use cases. Similarly if a product use case is modified, team members would know which other entities should be reviewed. One can navigate and explore a project by selecting any of the various entities, which would in turn highlight the dependent, related or affected elements associated with the selected entity.

5.7.4.7 Collaboration

The CASE tool's workspace offers an online collaborative environment allowing stakeholders to co-produce requirements or simply monitor the requirements development process. A simple access control mechanism has been implemented allowing for teams to be easily set up allowing for stakeholders both

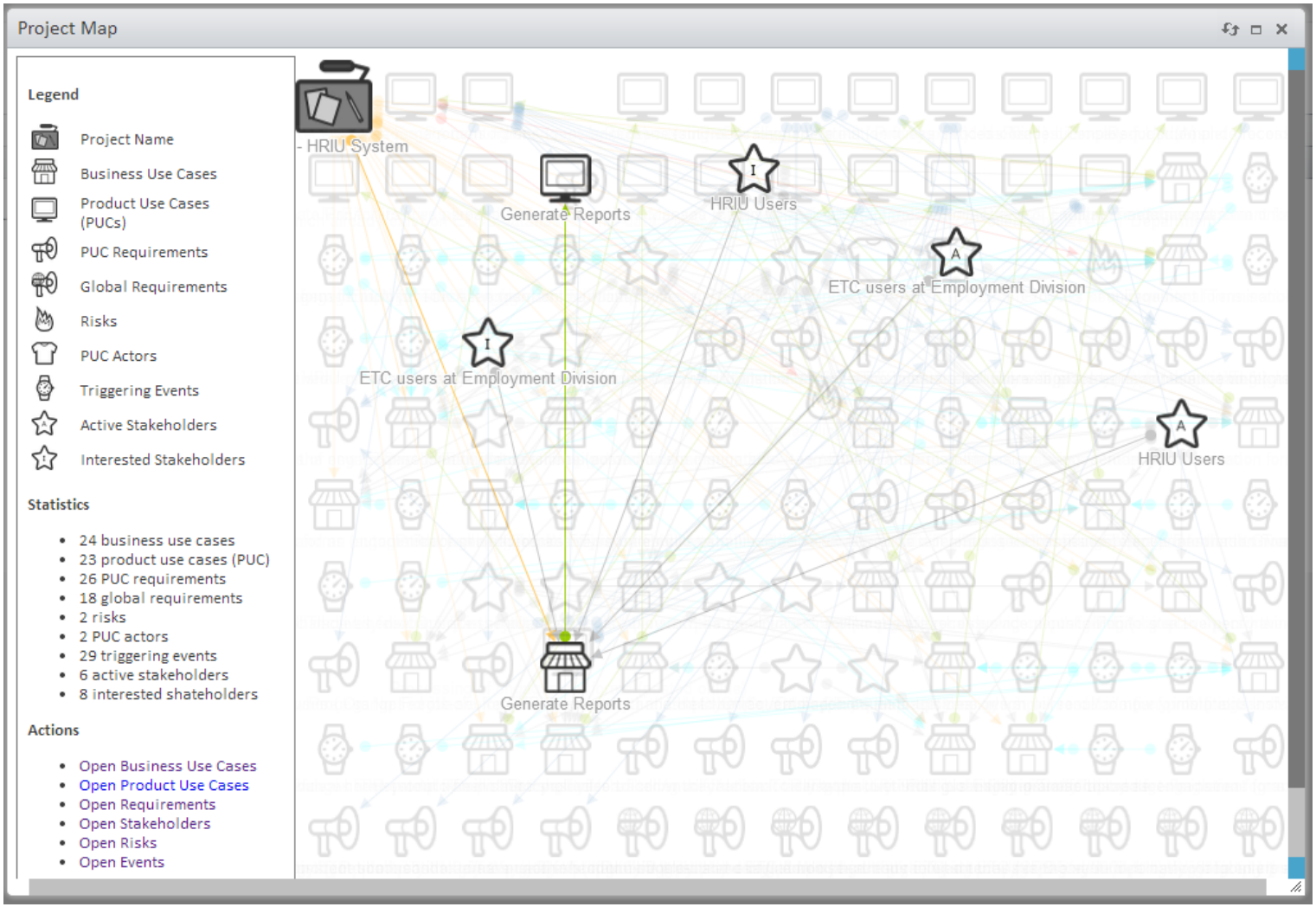


Figure 5.24: *Sentire* – project maps allowing for the visualisation of entities and their dependencies (from different perspectives). Various de-cluttering techniques are provided, especially for larger projects with hundreds of interlinked entities (e.g., highlighting of smaller project branches and dragging of overlapping items)

within and outside the institutional context to follow and possibly contribute to a project. Collaborators can be added *by invitation* using an email address as a unique identifier. Following an email check each collaborator will gain access to the respective project(s) for which an invitation was previously sent. Project access levels vary depending on the user's role (e.g., read-only, contributor or administrator).

5.8 Summary

This chapter presented *Sentire* – a framework devised to assist policy makers and system developers in the requirements development process for public facing and enrolment-centric e-government services. Extending the *Volere* requirements process, *Sentire* adopts Calibrated Personas (presented in Chapter 4) as part of the requirements development process, specifically within the *quality gateway* (see Section 5.4 – step 4.1). This allows the project team to generate user feedback simulations on enrolment-related use cases to test whether their design decisions are acceptable (from a users' perspective) or to ensure that a user experience requirement has been honoured in case of contracted projects. A practitioner-centric and collaborative online CASE tool was also built to support the processes and techniques presented in this chapter. Section 5.7 provides an outline of how this CASE tool evolved across the various stages of this thesis.

Sentire was validated and refined through a series of real-world case studies, presented in Chapters 6, 7, 8 and 9. Each study provides a number of insights that contributed towards the evolution of both the theoretical aspects of *Sentire* as well as its corresponding toolset.

Chapter 6

Case Study 1: Publishing a Tender for a National Employment Agency

This chapter discusses the experiences and lessons-learnt following the adoption of *Sentire* in a requirements development process for a new e-service commissioned by the national employment agency in Malta (ETC). This chapter begins by presenting the context and aims of the study. A discussion of the method adopted is then provided followed by an evaluation of results. The outcomes of this study will in turn inform the evolution of *Sentire* as well as the design of subsequent studies.

6.1 Defining the Context

A collaborative agreement was set up with Malta's Employment and Training Corporation (ETC), a semi-autonomous government entity that acts as a national hub and regulator for training and employment activities. ETC's management wanted to develop an e-service for one of its core activities handled by the Human Resource Information Unit (HRIU). This service enables the unit to monitor and manage all employee engagements and terminations across the islands (Malta and Gozo), requiring human resource managers to provide this information. This requirement has been enacted for a number of years (via legal notice L.N. 110 of 1993), however workflows for both end users and HRIU staff were, at the time of writing, still inefficient due to a significant amount of manual intervention. HRIU also provides statistical data to the government, and thus up-to-date and efficient reporting is a main priority. One of the few public interfaces of the new e-service is the *Engagement Forms Submission service* which is aimed at all local employers. Staff turnover as well as promotions and demotions must be reported to ETC's HRIU. This new project served as an ideal testing ground for *Sentire*. The core design team consisted of the IT manager, a system analyst, the author in the role of a requirements analyst, the HRIU manager and a principal user.

6.2 Aims

6.2.1 Research objectives

A primary goal of this study was to adopt the framework in a real-world environment to get more insights on its strengths, weaknesses, fitness for purpose and adaptability in a real-world context. The e-service

requires a high level of identity assurance since other government entities rely on the underlying transaction data for their own operations (e.g., social services use employment engagements and terminations as proof for social-benefit eligibility).

This provides an interesting case for *Sentire*, for which a balance between identity assurance and acceptable levels of workload needs to be obtained.

6.2.2 Practical objectives

ETC's objective was to draft a requirements document for the new HRIU e-service. This document would then be used to issue a call for tender and eventually guide the winning bidder.

6.3 Method

The main activities planned for and executed throughout this case study are listed in Figure 6.1. This section will focus on steps five, eight and nine (highlighted), which are central to the evaluation of *Sentire*.

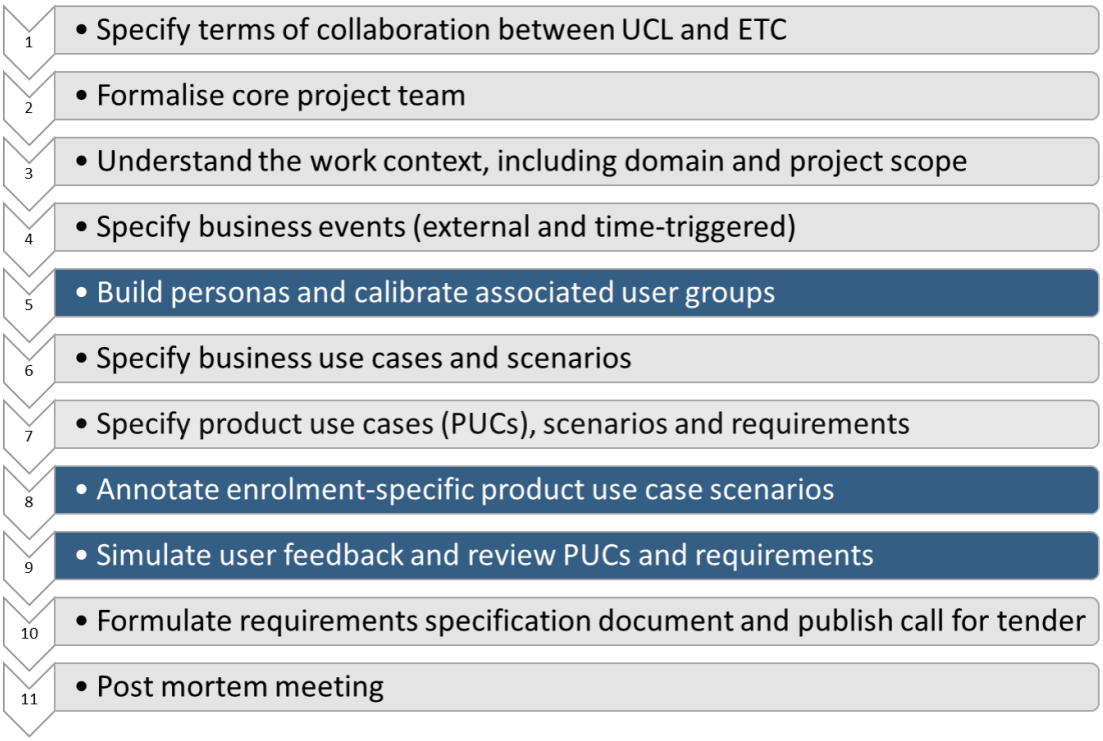


Figure 6.1: Plan of action for the HRIU case study – highlighted steps are discussed in this chapter

The first phase of the project largely followed the *Volere* process – *project blastoff* (steps 1, 2 and 3) and *requirements trawling* (steps 4, 5, 6 and 7). Facilitated workshops were conducted on a weekly basis at ETC's premises (see Figure 6.2) and *Google Docs* was used as the main collaborative document management tool. The setup included the use of a projector to allow for immersive and collaborative input. The requirements elicitation phase was completed in approximately two months and in the process 50 business events were identified and catalogued with 24 business responses to such events (Business Use Cases). Eventually 23 product use cases were proposed. Each business and

product use case had normal cases defined as well as alternative scenarios, exception scenarios and misuse cases. For each of the 23 product use cases a set of requirements was defined using *Volere* snow cards. Six personas were also identified including internal users (ETC staff), users from other government entities, and external users. The main user group (represented by Maryanne Jones – a Human Resource Manager) was calibrated after which simulated user feedback was generated to assist in the selection of an appropriate enrolment process(es) for the central (core) use cases. At the end of this process, the main *Volere* requirements document was delivered, from which a call for tender was derived and published.

Figure 6.2: Facilitated workshop at ETC

6.3.1 User group calibration (UGC)

One of the primary personas identified was *Maryanne Jones* (see Figure 6.4), a human resource manager. This persona was constructed using various primary and secondary sources of information, including interviews with human resource managers across a number of organisations who interact regularly with HRIU. These included the transport authority (Transport Malta) and a commercial recruitment agency (Pentasia), Their feedback was important to construct the persona, but also to fine-tune the proposed product use cases. It was decided to evaluate *Sentire*'s user feedback mechanisms with this persona since it represents the majority of public end users. For this purpose the persona's underlying user group was determined to be the *young urban professionals (30–40 years old*) group, determined from a demographic evaluation, consultation with the project team and interviews with a number of actual HRIU clients (which in turn informed the evolution of the persona itself). To conduct the user group calibration exercise several participants represented by this persona (and who could potentially make use of this e-service) were required. Eight HR managers working with major software companies accepted to collaborate in this first study, including Microsoft, GFI, CCBill EU, CrimsonWing Plc, Shireburn, ISL Ltd and LeCroupier.

The user group was named *young urban professionals (30–40)* since most of the participants fell within this demographic bracket (i.e., education, age and career). Once this group was calibrated a

Maryanne Jones

Name:
Maryanne Jones
Type:
Primary Persona (Actual)
Occupation:
Human Resources
Birthday:
01/12/1978
Population Size:
546

Narrative:

HR Manager at a relatively large firm employing 50+ people who are both technical and skilled labourers. Maryanne is computer literate but dislikes complex interfaces and disruptive processes. She is looking for ways to automate her tasks

Member Of:
Young urban professional (30-40)

Figure 6.3: Human resource manager persona. This persona is a member of the *young urban professionals (30–40)* user group

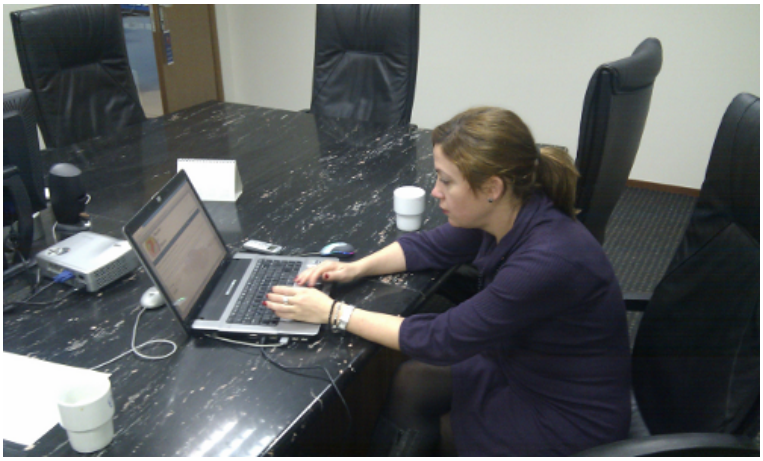


Figure 6.4: An HR manager participating in a user group calibration exercise. One-to-one in-context calibration was selected and calibration was conducted at the participants' premises

clearer picture of the users' reaction towards enrolment-specific design factors (and different intensities thereof) started to emerge.

Table 6.1: Regression coefficients generated for the *young urban professionals (30–40)* user group. These coefficients explain the user group's reactions to specific enrolment-related design factors

	Regression coefficients	
	Task completion (see Figure 6.5)	Perceived workload (see Figure 6.6)
B-Coefficient	5.866	3.888
Items to Generate	-0.78	NA
Items to Recall	NA	2.183
Delays	-1.434	34.332
Interruption	-1.925	24.127
Type of Service 1	-2.339	NA
Type of Service 2	-1.448	NA
Type of Service 3	-0.718	NA
Type of Service 4	NA	NA

Figure 6.5: Complete results for the task completion (*WCT*) regression coefficients generated for the *young urban professionals (30–40)* user group

Parameter Estimates

Decision[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
Yes	Intercept	5.866	.922	40.438	1	.000			
	[NatureOfService=0]	-2.339	.607	14.837	1	.000	.096	.029	.317
	[NatureOfService=1]	-1.448	.596	5.893	1	.015	.235	.073	.757
	[NatureOfService=2]	-.718	.608	1.396	1	.237	.488	.148	1.605
	[NatureOfService=3]	0[b]	.	.	0	.	.	.	.
	[Delays=1]	-1.434	.413	12.021	1	.001	.238	.106	.536
	[Delays=2]	0[b]	.	.	0	.	.	.	.
	[Interrupts=1]	-1.925	.507	14.446	1	.000	.146	.054	.394
	[Interrupts=2]	0[b]	.	.	0	.	.	.	.
	NewFacts	-.780	.305	6.545	1	.011	.458	.252	.833

a. The reference category is: No.

b. This parameter is set to zero because it is redundant.

Figure 6.6: Complete results for the perceived workload (*PEW*) regression coefficients generated for the *young urban professionals (30–40)* user group

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	3.888	1.4646	1.018	6.759	7.049	1	.008
[Delays=1]	34.332	6.5205	21.552	47.112	27.722	1	.000
[Delays=2]	0[a]	.	.	.	.	.	.
[Interrupts=1]	24.127	4.8212	14.678	33.577	25.044	1	.000
[Interrupts=2]	0[a]	.	.	.	.	.	.
Recall	2.183	.4188	1.363	3.004	27.183	1	.000
(Scale)	.542[b]	.0527	.447	.655			

Dependent Variable: Workload

Model: (Intercept), Delays, Interrupts, Recall

a. Set to zero because this parameter is redundant.

b. Maximum likelihood estimate.

These coefficients were eventually associated with *Maryanne Jones* (primary persona), thus turning it into a Calibrated Persona.

6.3.2 Product use case annotation

The second phase of the process was to annotate selected product use cases (in this case, the *Receive Engagement Form Online* product use case, which can also be considered as a Type of Service 3 for larger companies, within which *Maryanne Jones* operates). A number of scenarios were identified for this use case, each providing a different enrolment process for first-time users. This section will focus on two scenarios: the first scenario requires users to enrol on HRIU via the national e-ID ecosystem (federated login), assuming that users do not own an e-ID yet. On the other hand the second scenario requires users to enrol on HRIU by creating an account with ETC. The production task is common for

both scenarios (see Table 6.2). Tables 6.3 and 6.4 provide enrolment specific annotations for each of the two scenarios.

Table 6.2: Scenarios identified for the *Receive Engagement Form Online* Product Use Case. Scenarios vary from each other since they have different security tasks

ID	Scenario	Steps
1	*Normal Case Scenario (via national e-ID – users do not have an e-ID)*	• Register for an e-ID by visiting a Registration Authority – users have to visit a registration authority in person, after which an email is sent containing account activation instructions (including a password). This however requires a PIN which is sent by snail-mail (delivery takes around three days) • Employer signs in on ETC's website using the newly activated e-ID
2	*Normal Case Scenario (via ETC account)*	• Employer registers for an account on ETC's portal – user fills in and submits an online form with personal and company details (15 fields). The enrolment process is then completed in a day or two since ETC staff would require to validate the information submitted. Once the information is verified, the account is activated and the user is informed. • Employer signs in on ETC's web site using the newly activated ETC account
1&2	(Common to all scenarios)	• ... • The user selects the company under which to act (for group of companies or recruitment agencies) • The user clicks on the engagement form submission link • An online form is populated by the user (similar to Artefact A3 and A8 [referring to scanned forms]) • User selects a permit if form is from a semi-autonomous government entity • User submits form(s) and HRIU receive the data (coded as EE) for further processing

The enrolment steps for all scenarios shown in Table 6.2 were then annotated as shown below:

Table 6.3: Step 1 (first scenario): Employer enrols for an e-ID by visiting a Registration Authority (PIN is received by post)

Scenario 1	
Design element	**Measurement**
Items to Generate	2
Items to Recall	2
Delays	True
Interruptions to daily routines	True
Type of Service	3

Table 6.4: Step 1 (second scenario): Employer enrols for an account on ETC's portal (by submitting additional details for manual verification)

Scenario 2	
Design element	**Measurement**
Items to Generate	2
Items to Recall	15
Delays	True
Interruptions to daily routines	False
Type of Service	3

6.3.3 Simulating user feedback and revisiting PUCs and requirements

A throwaway prototype was developed to generate user feedback based on the techniques described in Chapters 4 and 5. Once use cases are annotated and regression coefficients for the Calibrated Persona are provided, initial results can then be generated. The following are the predictions generated for the above scenarios:

Table 6.5: Predictions for perceived workload and willingness to use the e-service based on Maryanne Jones (associated with the *young urban professionals (30–40)* user group)

Scenario	**Perceived Workload**	**Willingness to complete task**
Scenario 1	*71%*	*55.7%*
Scenario 2	*75%*	*89.6%*

Although perceived workload is high in both cases (see Table 6.5), the willingness to adopt the e-service varies considerably. Delay (*D*) is a major contributor to perceived workload, as well as the number of fields present in the enrolment form (*ItR*). However for this specific group of users, interruption (*I*) to daily routines has a high impact on the acceptability of a system, hence the lower percentage of users willing to adopt the e-service in the first scenario.

Given these results, the project team eventually decided to keep both options available to encourage takeup by those who do not have an e-ID, while keeping the federated login option open for those who do. Furthermore, these two enrolment options have been made available on the main ETC web site, and not just for the HRIU e-service (see Figure 6.7). Both enrolment processes provide the same levels of access and assurance.

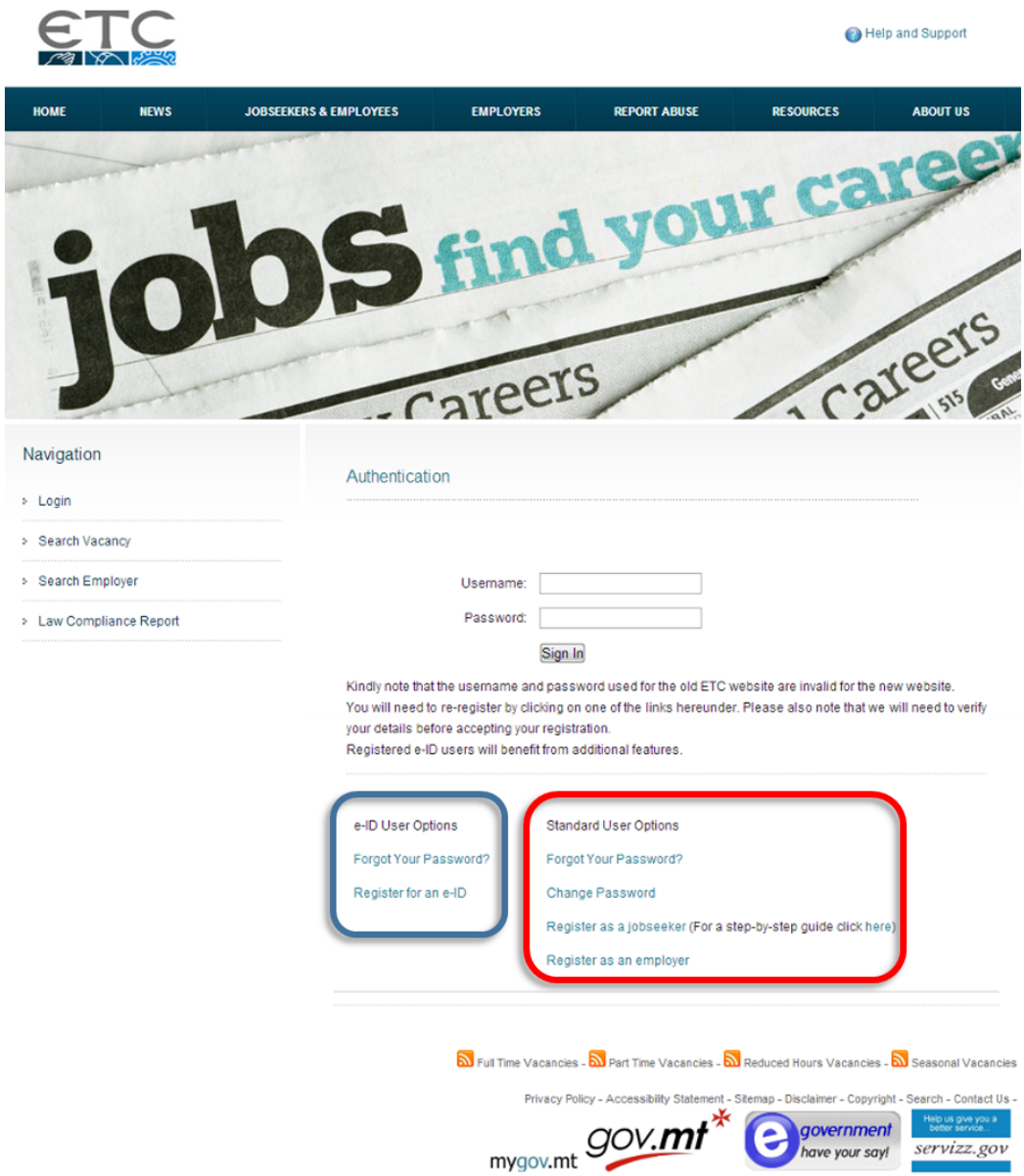


Figure 6.7: Login page at www.etc.gov.mt – both e-ID and ETC-specific enrolment processes are provided both of which provide the same level of access and identity assurance

6.4 Evaluation and Findings

6.4.1 Theoretical evaluation of calibration models

6.4.1.1 Willingness to complete task (WCT) model

Willingness to complete the task can be explained as a binary decision or outcome (*yes* or *no*). For this reason a binary logistic regression model was applied. Using the backward stepwise entry method, four significant predictors were found explaining almost 50% of the total variation in the yes/no response (Negelkerke = 0.484).

Table 6.6: *Young urban professionals' (30–40) WCT* model – testing fitness to the data

Pseudo R-Square	
Cox and Snell	.324
Nagelkerke	.484
McFadden	.354

6.4.1.2 Perceived workload (PEW) model

Before deciding on the regression model to adopt the outcome variable was tested to determine whether it follows a normal distribution. As shown in Table 6.7 the Shapiro-Wilk value for *workload* (outcome) is .811, rejecting the null-hypothesis that the population is normally distributed ($p\text{-value} < 0.05$). This was also confirmed through a visual check by plotting the data on a histogram (see Figure 6.8).

Table 6.7: *Young urban professionals' (30–40) PEW* model – tests of normality

Kolmogorov-Smirnov			*Shapiro-Wilk*		
Statistic	df	Sig	Statistic	df	Sig
.196	236	.000	.811	236	.000

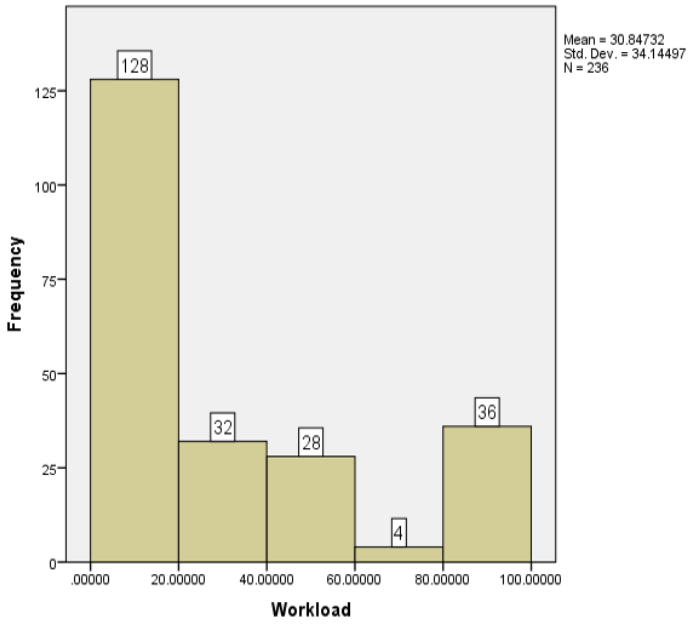


Figure 6.8: *Young urban professionals' (30–40) PEW* model – normality test for workload (visually right skewed – denoting a non-normal distribution)

A Generalised Linear Model with Gamma distribution was adopted (providing a better fit to right-skewed distributions). The Gamma Regression model identifies three significant predictors whereby Delays (*D*) is the best predictor followed by Recall (*ItR*) and Interruptions to daily routines (*I*). All have a large Wald Chi-square value (i.e., coefficients are statistically significant to the model). The expected perceived workload when there are no delays is 34.33% less than when there are delays (see Table 6.8). Similarly when there are no interruptions, perceived workload is expected to be 24.13% less than when there are interruptions. For every additional unit increase in recall (*ItR*), perceived workload is expected to increase by 2.18%.

Table 6.8: *Young urban professionals' (30–40) PEW* model – test of model effects

	Type III		
Source	*Wald Chi-Square*	*df*	*Sig*
(Intercept)	135.977	1	.000
D	27.722	1	.000
I	25.044	1	.000
ItR	27.183	1	.000

Tables 6.9 and 6.10 provide more information on the model's goodness of fit, including the Pearson Chi-Square measure as well as the Likelihood Ratio Chi-Square.

Table 6.9: *Young urban professionals' (30–40) PEW* model – testing goodness of fit

	Value	*df*	*Value/df*
Deviance	106.047	176	.603
Pearson Chi-Square	85.788	176	.487

Table 6.10: *Young urban professionals' (30–40) PEW* model – omnibus test

Likelihood Ratio Chi-Square	*df*	*Sig*
112.885	3	.000

6.4.2 Evaluation of task completion predictions

A short online questionnaire was used to present these alternative scenarios to a number of human resource managers, and for each scenario they were asked to state whether they would opt to enrol and use the e-service or keep using the traditional alternative (i.e., post). ETC's contact lists as well as personal contacts were used to distribute the link to the questionnaire and 17 respondents gave their feedback.

Scenario 1 The *WCT* model predicted that 55.7% would be willing to enrol and use the e-service. The user evaluation exercise has shown that 52.9% of the respondents would not be put off by the enrolment process and would be willing to use the e-service while the others would prefer to stick to the manual process (i.e., traditional post) – see Figure 6.9

Scenario 2 The *WCT* model predicted that almost 90% would enrol for the e-service. The user evaluation exercise has shown that all of the respondents who completed this question would not be put off by the enrolment process and would be willing to use the e-service (two respondents did not provide feedback for Scenario 2) – see Figure 6.10

These results are encouraging, however the rigour by which the requirements were specified was even more interesting. The use of Calibrated Personas gave the author the opportunity to learn more about the end user, specifically through the calibration process itself. This in turn informed the development of the primary persona. Simulated user feedback provided objective and quantitative insights allowing for design decisions to be tested throughout the evolution of product use cases and their underlying requirements.

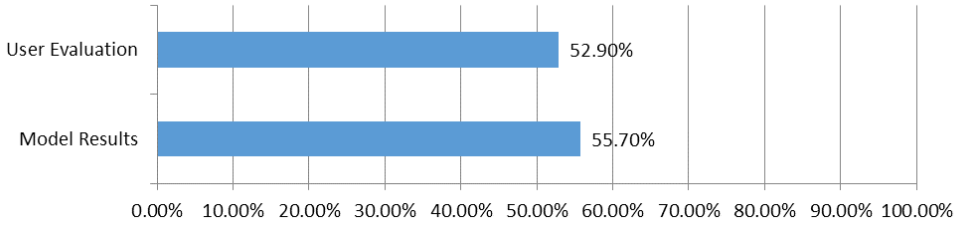


Figure 6.9: Feedback from actual users and the predictions generated via *Sentire* for Scenario 1

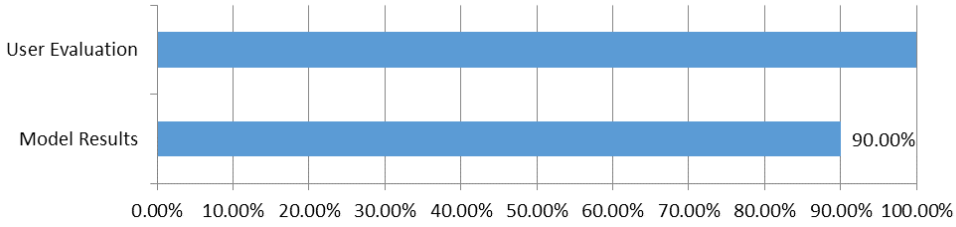


Figure 6.10: Feedback from actual users (excluding missing values) and the predictions generated via *Sentire* for Scenario 2. If missing values are considered as 'non-adopters' the user evaluation figure would go down to 88%.

.

6.4.3 Framework contributions and modifications

Further insights on aspects covering both the theoretical framework as well as the supporting CASE tools were obtained through this case study.

6.4.3.1 Redefining delay

It was noticed that UGC participants reacted in a considerably different way when confronted with minor delays (e.g., activation email received after a few minutes) as opposed to major delays (e.g., a three day period for account activation). This motivated the researcher to modify the calibration tasks to include three levels of delay: 0 – no delays, 1 – minor delays and 2 – major delays. Table 6.11 outlines the calibration tasks following the modification (reproduced from Chapter 4).

6.4.3.2 Collaborative workspace

Google Docs provided a convenient environment to create and co-author *Volere* compliant documents (on-site and remotely) however it was found to be insufficient when it came to maintaining links between entities. Hyper-linking was considered but found to be too cumbersome from a practitioner's point of view. Based on this experience a CASE tool was developed (see Chapter 5.7). The new CASE tool provides a collaborative environment for the co-production of *Volere* deliverables, leading to automated report generation, project visualisation and entity linking (e.g., relationships between actors, events, business responses, risks, stakeholders, product use cases and requirements, amongst others). Simulated user feedback was also integrated within the CASE tool as part of the main requirements workflow.

6.4.4 Plans for next study

The following high-level goals were specified for the subsequent case study:

1. Test the modelling technique on a different user group (e.g., undergraduates)

Table 6.11: Set of nine enrolment pages were modified to include multiple delay intensities

Task	ItR	ItG	I	D
A	1	0	No	No
B	2	1	No	No
C	5	1	No [1]	Minor [2]
D	4	2	No	Major [3]
E	5	2	Yes [4]	Major [4]
F	6	3	No	Minor [5]
G	6	4	No	No
H	9	3	No	Minor [6]
I	NA	3	Yes [7]	Major [8]

[1] Credit card details are required
[2] Wait a few minutes for activation email
[3] Wait three days before account is activated
[4] Visit closest outlet to confirm identity
[5] Upload recent photo
[6] Call free-phone to activate account
[7] Visit registration office during specific opening hours
[8] Three day waiting period till PIN is received

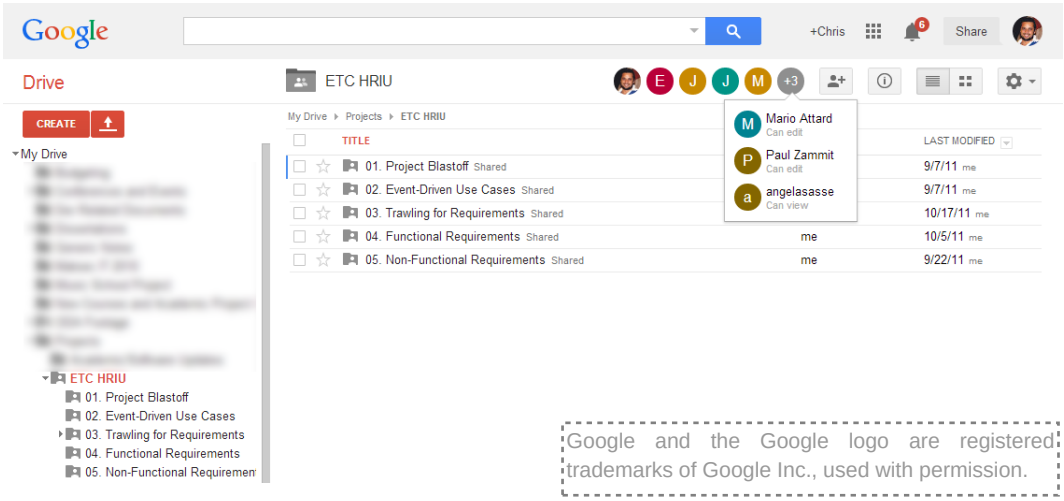


Figure 6.11: *Google Docs* was used for the first case study. It was convenient for document co-editing however the lack of resource linking capabilities provided a challenge especially when the number of use cases and associated requirements grew.

2. Test the sensitivity of behavioural models within a non e-government scenario

6.5 Summary

This chapter presented the first field-study in which *Sentire* was used to author a requirements specification document (and call for tender) for a new national e-service commissioned by the Employment and Training Corporation in Malta. At this point *Sentire* was evaluated with one calibrated user group – the *young urban professionals (30–40 years old)* user group, represented by Maryanne Jones as the primary persona. Several human resource managers from various organisations accepted to collaborate in this study and one-to-one in-context calibration was conducted within their work environment. The behavioural models generated were then associated with the primary persona, turning Maryanne Jones

into a Calibrated Persona.

The e-service being developed required a high level of identity assurance and *Sentire* was used to find a balance between this assurance level, the design of enrolment processes being considered for first-time users and the level of workload such users are willing to accept to gain access to this particular e-service. Simulated user feedback was based on the Calibrated Persona's underlying models. The results were evaluated using statistical tests as well as through a separate user-evaluation exercise to triangulate and assess *Sentire*'s predictive capabilities. Lessons learnt from the first case study, including calibration issues and lack of process management facilities, were used to inform the development of the theoretical framework and supporting toolset.

Chapter 7

Case Study 2: Non E-Government Service Evaluation with Undergraduate Students

This chapter outlines a number of insights that were generated following an intervention with a group of undergraduate students. These were invited to participate in a calibration exercise as well as a focus group session, out of which several observations were drawn. The behavioural models were used to simulate user feedback on the enrolment processes of a number of commercial services (i.e., *WordPress.com*, *Blog.com* and *LiveJournal.com*). This chapter starts by defining the main aims of the study followed by an outline of the method adopted to reach them. An evaluation of results together with findings are then presented.

7.1 Aims

7.1.1 Research objectives

The main aim of this case study is to test the sensitivity of the calibration process with a different set of participants. This also entails an assessment of the resultant models' quality in representing the user group's reactions towards enrolment processes and the various design factors involved. Another goal of this study is to assess the models' applicability outside an e-government context wherein services are *not compulsory* and *competing service providers exist*. This would shed more light on the sensitivity of the calibration process when service *Replaceability* is introduced (see Table 4.1).

It was decided to consider the following user group: *regular internet users aged between 18 and 22 who are currently following a bachelor's degree in a non-IT related area*. This set of demographics represents university undergraduate students who may not necessarily be advanced computer users, but who should at least have an intermediate level of IT proficiency, necessary at this level of education (e.g., equivalent to BCS Level 2 *ECDL*[1]). This also presents an opportunity to learn more about this user group's attitudes towards enrolment and enrolment-based services.

[1]BCS European Computer Driving Licence, http://www.bcs.org/upload/pdf/it-application-skills.pdf, (accessed January 2015)

7.2 Method

7.2.1 Participants

Following a call for participation at the University of Malta campus (using adverts in public spaces and announcements in lecture halls), 26 students accepted to participate in this study. 16 participants were male and 10 were female, all of whom were within the 18 to 22 age range. All participants were undergraduate students majoring in any non-IT related field, including chemistry, banking, statistics, history and accounting.

7.2.2 Process

The main activities planned and executed throughout this case study are listed in Figure 7.1.

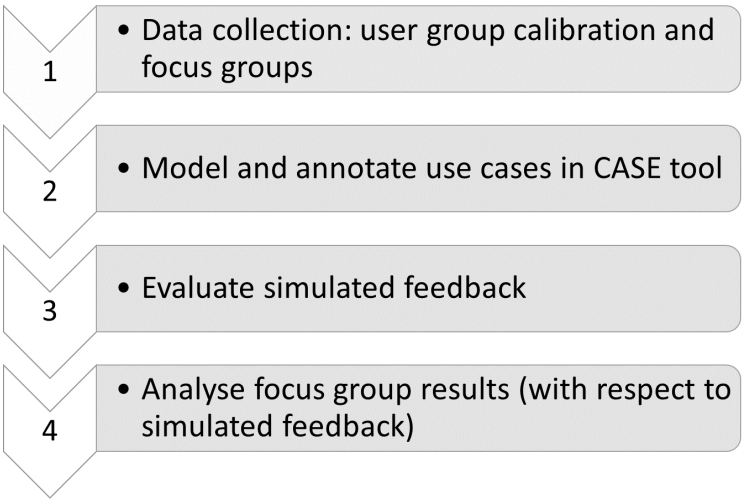


Figure 7.1: Plan of action for the *undergraduate students (non-IT)* user group case study

Two one-hour semi-structured focus group sessions were organised. The first 40 minutes of each session were dedicated to a discussion on the students' experiences (both positive and negative) during enrolment across different online services (including emailing, social networking, e-banking and e-government). The remaining 20 minutes were used to conduct a User Group Calibration (UGC) exercise in a lab environment. Six participants had to leave early (due to other commitments) leaving a total of 20 participants (12 male and eight female) who managed to complete the UGC process across the two sessions (nine and 11 participants respectively). All sessions were recorded (audio only) with the students' consent, which were then transcribed for further analysis.

The calibration data was then processed to generate behavioural user models for this user group. These models were then associated with the persona hypothesis[2] (Jane Smith). The CASE tool was then used to replicate and annotate three enrolment use cases from three major blogging engines, namely *Blog.com*, *WordPress.com* and *LiveJournal.com* in which Jane Smith was specified as an active stakeholder (primary user). The three use cases were categorised as *Type 2* services (ToS 2). *Sentire*'s CASE tool was then used to simulate user behaviour in all of the three use cases for Jane Smith. This was then followed up with a short survey (targeting undergraduate students in general) to evaluate their attitudes

[2] An early prototype of a primary persona (i.e., work in progress)

towards these enrolment processes and the results were compared to *Sentire*'s predictions on expected behaviour. This sheds more light on the sensitivity of the models and their representativeness of the respective user groups. Finally the focus group transcripts were analysed thematically to generate further insights and reflections on this user group's attitudes towards, and behaviour during e-service enrolment.

7.2.2.1 User group calibration (UGC)

Following the previous case study (see Chapter 6), the UGC exercise was modified to accommodate minor (1) and major (2) delays as well as no-delays (0) (see Table 6.11).

A persona hypothesis for the student user group was created (*Jane Smith* – see Figure 7.2) – a university undergraduate student majoring in Chemistry. She was associated with a new user group called – *undergraduate students (non-IT)*. Two calibration exercises, with nine and 11 participants respectively, were carried out in a lab environment wherein students had to go through a number of fictitious enrolment tasks designed to capture and model their behaviour.

Jane Smith

Name:
Jane Smith
Type:
Primary Persona (Hypothesis)
Occupation:
University Student
Birthday:
06/06/2014
Population Size:
10000

Narrative:
Undergrad student majoring in Chemistry.
Member Of:
Undergraduate Students (Non-IT)

Figure 7.2: University student persona (retrieved from *Sentire*'s persona library). This assumption (hypothesis) persona is a member of the *undergraduate students (non-IT)* user group

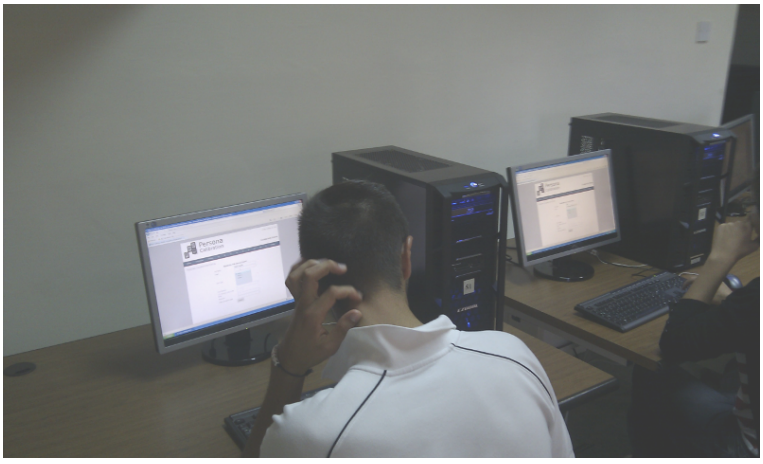


Figure 7.3: An undergraduate student participating in a user group calibration exercise. This was carried out in group within a lab environment.

Once calibration was completed, a clearer image of this group's reactions to the various enrolment process design factors started to emerge. Table 7.1 shows the regression coefficients generated for this user group.

Table 7.1: Regression coefficients for the *undergraduate students (non-IT)* user group. These coefficients explain the user group's reactions to specific enrolment-related design factors

	Regression coefficients	
	Task completion (see Figure 7.4)	Perceived workload (see Figure 7.5)
B-Coefficient	2.751	20.929
Items to Generate	-0.311	NA
Items to Recall	NA	2.628
No Delays	0.437	-10.706
Minor Delays	-0.348	-1.002
Major Delays	0	0
Interruption	-1.204	35.202
Type of Service 1	-2.459	NA
Type of Service 2	-1.709	NA
Type of Service 3	-0.696	NA
Type of Service 4	0	NA

Figure 7.4: Complete results for the task completion (*WCT*) regression coefficients generated for the *undergraduate students (non-IT)* user group

Parameter Estimates

Decision[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
Yes	Intercept	2.751	.395	48.572	1	.000			
	[NatureOfService=0]	-2.459	.290	72.024	1	.000	.086	.048	.151
	[NatureOfService=1]	-1.709	.282	36.641	1	.000	.181	.104	.315
	[NatureOfService=2]	-.696	.291	5.722	1	.017	.498	.282	.882
	[NatureOfService=3]	0[b]	.	.	0	.	.	.	.
	[Delays=0]	.437	.327	1.789	1	.181	1.548	.816	2.936
	[Delays=1]	-.348	.317	1.207	1	.272	.706	.379	1.314
	[Delays=2]	0[b]	.	.	0	.	.	.	.
	[Interrupts=1]	-1.204	.338	12.677	1	.000	.300	.155	.582
	[Interrupts=2]	0[b]	.	.	0	.	.	.	.
	NewFacts	-.311	.084	13.732	1	.000	.732	.621	.864

a. The reference category is: No.

b. This parameter is set to zero because it is redundant.

Figure 7.5: Complete results for the perceived workload (*PEW*) regression coefficients generated for the *undergraduate students (non-IT)* user group

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	20.929	3.1423	14.770	27.087	44.360	1	.000
[Delays=0]	-10.706	2.8364	-16.265	-5.147	14.247	1	.000
[Delays=1]	-1.002	3.2746	-7.420	5.417	.094	1	.760
[Delays=2]	0[a]	.	.	.	.	.	.
[Interrupts=1]	35.202	4.5979	26.191	44.214	58.618	1	.000
[Interrupts=2]	0[a]	.	.	.	.	.	.
Recall	2.628	.4525	1.742	3.515	33.741	1	.000
(Scale)	.507[b]	.0269	.457	.563			

Dependent Variable: Workload

Model: (Intercept), Delays, Interrupts, Recall

a. Set to zero because this parameter is redundant.

b. Maximum likelihood estimate.

7.2.2.2 Product use case annotation

Three different use cases were reproduced in the CASE tool, reflecting the design of three different blogging engines' enrolment processes. These were then annotated as shown in Tables 7.2, 7.3 and 7.4.

Table 7.2: *Blog.com*'s enrolment process

Use Case 1 – Blog.com	
Design element	**Measurement**
Items to Generate	1
Items to Recall	2
Delays	None
Interruptions to daily routines	False
Type of Service	2

Table 7.3: *WordPress.com*'s enrolment process

Use Case 2 – WordPress.com	
Design element	**Measurement**
Items to Generate	3
Items to Recall	1
Delays	None
Interruptions to daily routines	False
Type of Service	2

Table 7.4: *LiveJournal.com*'s enrolment process

Use Case 3 – LiveJournal.com	
Design element	**Measurement**
Items to Generate	2
Items to Recall	4
Delays	None
Interruptions to daily routines	False
Type of Service	2

7.2.2.3 Simulating user feedback

Once all of this information was specified within the CASE tool, user feedback simulations were then generated as shown in Table 7.5. Perceived workload is considerably low mainly due to the fact that there are no delays (*D*) or interruptions (*I*) in any of the three use cases. On the other hand, the willingness to complete the task is considerably high in all cases except for the second use case (*WordPress.com*). These results are evaluated in the following section (see Section 7.3.1).

Table 7.5: Predictions for perceived workload and the willingness to complete the task, generated for the above use cases and based on Jane Smith's behavioural models (derived from the *undergraduate students (non-IT)* user group)

Scenario	**Perceived Workload**	**Willingness to complete task**
Use Case 1 - *Blog.com*	*15.5%*	*76.3%*
Use Case 2 - *WordPress.com*	*12.9%*	*63.3%*
Use Case 3 - *LiveJournal.com*	*20.7%*	*70.2%*

7.3 Evaluation and Findings

7.3.1 Theoretical evaluation of calibration models

7.3.1.1 Willingness to complete task (WCT) model

The backward stepwise entry method was adopted for the binary logistic regression analysis (i.e., outcome is binary – yes/no). Four significant predictors were found explaining over 30% of the total variation in the yes/no response (Negelkerke = 0.305 – see Table 7.6).

Table 7.6: *Undergraduate students' (non-IT) WCT* model – testing fitness to the data

Pseudo R-Square	
Cox and Snell	.224
Nagelkerke	.305
McFadden	.191

The values shown in Table 7.6 show that this model is not as powerful as the one generated for the *young urban professionals (30–40)* user group. The author suggests that other significant factors influencing the willingness to complete a task exist for this group of users, especially when competing service providers exist (unlike in the e-government domain). A discussion on this is presented in Section 7.3.3.

Table 7.7: *Undergraduate students' (non-IT) WCT* model – testing goodness of fit

	Chi-Square	*df*	*Sig*
Pearson	21.224	24	.625
Deviance	21.849	24	.588

Table 7.8: *Undergraduate students' (non-IT) WCT* model – likelihood ratio tests

Effect	*Chi-Square*	*df*	*Sig*
ToS	104.150	3	.000
D	11.292	2	.004
ItG	14.020	1	.000
I	13.334	1	.000

7.3.1.2 Perceived workload (PEW) model

Before deciding on a regression model to use, the outcome variable was tested to determine whether it follows a normal distribution. As shown in Table 7.9 the Shapiro-Wilk value for *workload* (outcome) is .911, strongly rejecting the null-hypothesis that the population is normally distributed ($p\text{-value} < 0.05$). This was also confirmed through a visual check by plotting the data using a histogram (see Figure 7.6).

Table 7.9: *Undergraduate students' (non-IT) PEW* model – tests of normality

Kolmogorov-Smirnov			*Shapiro-Wilk*		
Statistic	df	Sig	Statistic	df	Sig
.128	684	.000	.911	684	.000

A Generalised Linear Model with Gamma distribution was therefore adopted since it fits better to right skewed distributions. The Gamma Regression model identifies three significant predictors whereby

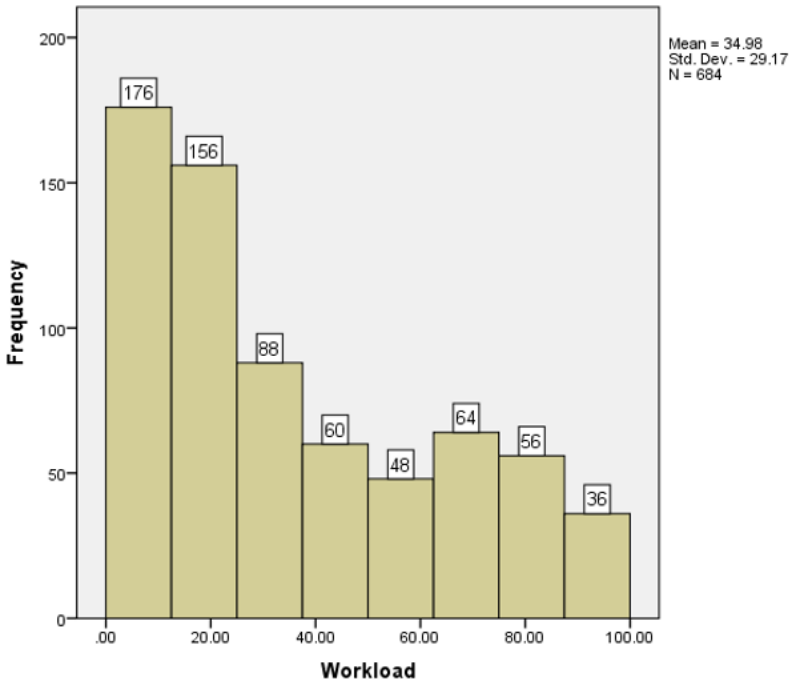


Figure 7.6: Undergraduate students' (non-IT) *PEW* model – normality test for workload (visually right skewed, denoting a non-normal distribution)

Interruptions to daily routines (*I*) is the best predictor followed by Recall (*ItR*) and Delays (*D*). All have a large Chi-square value (i.e., coefficients are statistically significant to the model). The expected perceived workload when there are no interruptions (*I*) is 35.20% less than when interruptions are introduced (see Table 7.1) . Similarly when there are no delays (*D = 0*), perceived workload is expected to be 10.71% less than when there are major delays (*D = 2*). For every additional unit increase in recall (*ItR*), perceived workload is expected to increase by 2.63%.

Table 7.10: *Undergraduate students' (non-IT) PEW* model – test of model effects

	Type III		
Source	*Wald Chi-Square*	*df*	*Sig*
(Intercept)	160.595	1	.000
D	21.393	2	.000
I	58.618	1	.000
ItR	33.741	1	.000

Tables 7.11 and 7.12 provide more information on the model's goodness of fit, including the Pearson Chi-Square measure as well as the Likelihood Ratio Chi-Square (see Section 4.2.2).

Table 7.11: *Undergraduate students' (non-IT) PEW* model – testing goodness of fit

	Value	*df*	*Value/df*
Deviance	336.065	607	.554
Pearson Chi-Square	277.408	607	.457

Table 7.12: *Undergraduate students' (non-IT) PEW* model – omnibus test

Likelihood Ratio Chi-Square	*df*	*Sig*
217.339	4	.000

7.3.2 Evaluation of task completion predictions

It was decided to follow up this study with a simple user evaluation exercise. An online questionnaire was distributed among undergraduate students through social media and email. This questionnaire contained screenshots from the three blogging engines' sign-up pages and for each screenshot respondents had to indicate whether they would consider signing-up (or otherwise) while indicating the reasons behind their decisions. 15 respondents completed this questionnaire, and comparisons were then made between reported behaviour and model predictions:

Enrolment process at *Blog.com* The *WCT* model predicted that 76.3% would be willing to enrol and use the service. The user evaluation exercise has shown that 78.6% of the respondents would not be put off by the enrolment process and would be willing to use this blogging engine – see Figure 7.7

Enrolment process at *WordPress.com* The *WCT* model predicted that 63.3% would be willing to enrol and use the service. The user evaluation exercise has shown that 71.4% of the respondents would not be put off by the enrolment process and would be willing to use this blogging engine – see Figure 7.8

Enrolment process at *LiveJournal.com* The *WCT* model predicted that 70.2% would be willing to enrol and use the service. The user evaluation exercise has shown that only 42.9% of the respondents would not be put off by the enrolment process and would be willing to use this blogging engine – see Figure 7.9.

Comments on the third case were varied, and the discrepancy between the predicted value and users' indications could be explained by a number of factors, including interface attractiveness (when compared to the alternative options) and privacy concerns. These are some of the students' responses:

- "*It seems too serious & the layout for filling in the information is a bit dull.*" [Questionnaire respondent]
- "*My answer is not really related to the information asked for in the registration process, it is more due to the fact that I did not like the user interface... it's not welcoming.. So I would only register if I wouldn't have managed to find better sources.*" [Questionnaire respondent] – see the **Replaceability (R)** behavioural modifier discussed in Section 4.1.1.
- "*But the combined information is quite a bit, enough to put you in a demographic. Probably results in targeted advertising. Also, whole feel is too formal.*" [Questionnaire respondent]
- "*The same old boring registration process and layout (one field for every row). The 'Captcha'*

at the end of the form can be really frustrating sometimes; at points I would not be able to get it right!" [Questionnaire respondent]

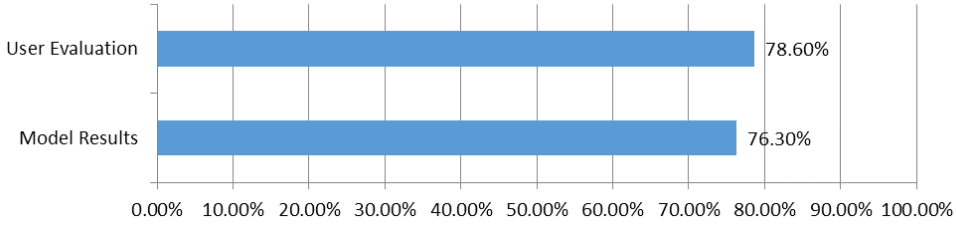


Figure 7.7: Feedback from actual users and the predictions generated via *Sentire*. This chart shows the percentage of undergraduate students (non-IT) who would be willing to enrol on *Blog.com*

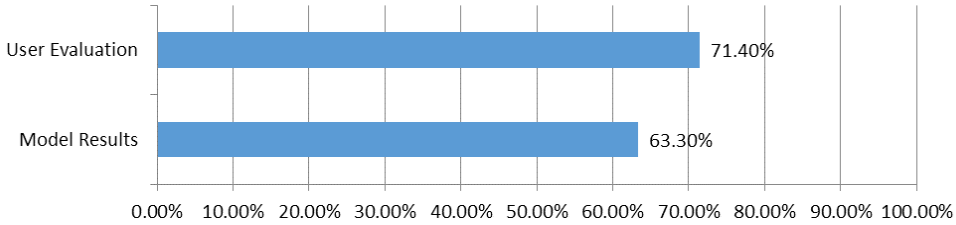


Figure 7.8: Feedback from actual users and the predictions generated via *Sentire*. This chart shows the percentage of undergraduate students (non-IT) who would be willing to enrol on *WordPress.com*

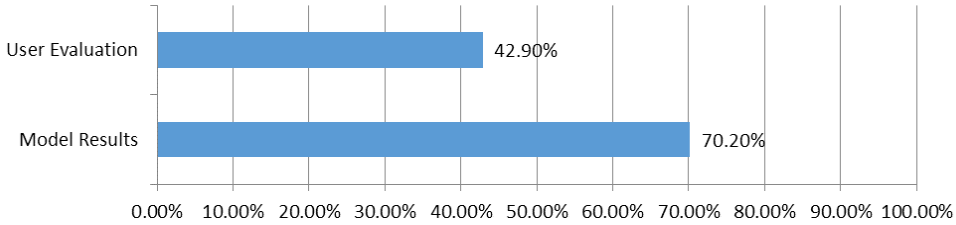


Figure 7.9: Feedback from actual users and the predictions generated via *Sentire*. This chart shows the percentage of undergraduate students (non-IT) who would be willing to enrol on *LiveJournal.com*

User interface (UI) design plays a crucial role in the way an e-service is perceived by this group of users and could potentially affect their decision making process. This is especially true with services that have a high *replaceability* (↑R) factor (i.e., competing service providers) and do not carry any legal usage requirement (i.e., *ToS* 1 and 2). This goes beyond the impact that identity processes alone can have on users. Portraying the right image can help to improve user perception, while softening the negative impact created by demanding enrolment tasks. In this case study, the undergraduate students (non-IT) user group took interface attractiveness into consideration while evaluating their options. The following are some comments from different respondents focusing on the look and feel of enrolment pages as well as on the underlying workload.

- "*Very simple and short process with a nice UI*" [Questionnaire respondent]

- "*It makes registering very easy and it has a clean and simple interface.*" [Questionnaire respondent]
- "[It has] *an attractive sign-up form*" [Questionnaire respondent]
- "*Quite straight-forward to use and very user friendly mostly because of the big boxes used.*" [Questionnaire respondent]
- "*It seems too serious and the layout for filling in the information is a bit dull.*" [Questionnaire respondent]

7.3.3 Focus group findings

The two focus-group sessions (preceding the user group calibration process) were recorded and transcribed. This provided the author with an opportunity to learn more about this user group's attitudes towards e-service adoption (including enrolment) as well as factors that influence their decision making process. Thematic analysis was used to analyse the data, out of which six themes emerged.

7.3.3.1 Theme 1: Competition, trust and service features

In a competitive domain, wherein alternative service providers exist, participants feel more confident to enrol on a service which has received good reviews by peers (or via online channels). Furthermore, participants noted that in order to encourage enrolment, benefits should be made explicitly clear.

It is more challenging to encourage a member of this user group to enrol if the service (or service provider) (1) is not popular, (2) service features and benefits are not explicitly outlined and (3) alternative and comparable service providers exist.

- "*If that site has good [reviews] then I would register*" [Focus group participant]
- "*If it's a famous site, then you trust it a little bit more because other people are using it*" [Focus group participant]
- "*I would use to Gmail because its linked to Youtube also. And then its simple to use also.*" [Focus group participant]
- "*I would go to Gmail, because Gmail groups emails together, forming a thread and that's more neat.*" [Focus group participant]
- "*...if there were more sites [offering the same service without enrolment], then I would go to the next one.*" [Focus group participant]

7.3.3.2 Theme 2: Social influence

Participants are also influenced by what their friends do (rather than just what they would recommend). This theme relates to a sense of social-belonging and interaction. If the majority of their friends have adopted a particular service provider (e.g., social network) then the decision to enrol on that social network would be much easier since it offers more opportunities for social engagement. This is also influenced by and influences social trends.

- "*More people had Facebook ... and you do it without recognizing that deep down you don't really need it*" [Focus group participant]

- "*I would use Hotmail, because first MSN was in trend and I used to chat, and then you use Hotmail as well*" [Focus group participant]

In addition, participants highlighted the fact that they also consider how their personal image is affected and how other people perceive them if they are associated with a specific service provider.

- "*It's more professional to use Gmail than Hotmail.*" [Focus group participant]

7.3.3.3 Theme 3: Privacy in security tasks

Participants get frustrated when asked for seemingly irrelevant information. This ties directly with Raha-man's *subject coupling* property outlined in [131]. This defines the "*representativeness between the captured identity [information] and the relevant partial identity of the subject in relation to the purpose and context*".

- "*It depends! It depends what the site is. For instance I hate it when they ask to register just to see the prices. In this case it's not a need, but it's just giving out the address to see prices.*" [Focus group participant]

- "*When the elections came I was using a political party's online system to give my opinion, but to do so they asked for my ID number. It was disguised as 'give your opinion' but in reality it wasn't just that. That's asking to give very important information.*" [Focus group participant]

Privacy issues were re-iterated several times by different participants, however they understand that a trade-off is necessary whenever a personalised service is desired.

- "*The data Google collects is to make your life better. They say they keep your search history, and then they use it to personalize your experience, e.g. auto complete while typing the search. They also provide you with relevant page result. But facebook, whatever data they take they either sell it.*" [Focus group participants]

Tying with privacy is the provision of a clear exit strategy. This is based on the idea that ahead of enrolment users are informed on possible ways by which their account could be closed, subscription terminated and associated personal data permanently removed (i.e., including data retention policies).

- "*... they give you the option to unsubscribe for instance after you register.*" [Focus group participant]

7.3.3.4 Theme 4: Workload

Major delays (e.g., 3 days waiting period) during enrolment were regarded negatively by participants and this was flagged as a main cause for frustration. Participants generally accept minor delays if this results in better security (e.g., email verification), however it was made clear that security measures must not stop them from completing the primary task in 'one sitting'.

- "*[Discussing activation email] It's just a link, it looks like a bit more professional, more secure? It gives that impression.*" [Focus group participant]

- "*[Discussing internet banking] I don't have it. Just for that reason [visiting local branch] – you can't apply for it online!*" [Focus group participant]

Lengthy forms, especially when more fields are presented in subsequent pages, were considered as annoying.

- "*You think you're ready from the questions, you click on proceed and then you find another set of questions. Oh my! If it's too many questions*" [Focus group participant]

Participants also indicated that they feel frustrated when asked for information that may not be available at hand (e.g., social security number), especially if this disrupts or blocks task completion.

- "*When they ask for information you don't know. Something that wouldn't be straight forward*" [Focus group participant]

Participants also pointed at complex security requirements (e.g., strong password policies or in-person identity validation) as another major cause for concern, having a direct bearing on their decision to enrol.

- "[Focus group participant] *Another thing I hate about registration is the password. The format.*

  [General agreement]

  [Focus group participant] *And unique numbers, and special characters...*

  [Further general consensus]

  [Focus group participant] *I just add two numbers at the end just to make it different.*"

7.3.3.5 Theme 5: Usage patterns and lifestyles

Participants reported that if the service has a positive impact on their lifestyle they would be encouraged to enrol. This includes productivity as well as the services' entertainment value. Also, when the expected frequency of use is high, participants feel more compelled to enrol since the services' perceived value increases.

- "*Again it's just one site [Play.com] and you can buy several times using. It's not like you give these details to an Ebay seller from whom you buy only once. There you have to give information each and every time.*" [Focus group participant]

7.3.3.6 Theme 6: Convenience

Participants highlighted several aspects that contribute towards a higher level of perceived convenience in online services (both commercial and governmental). These include service personalisation, access to (free) resources, reuse of identity information (through identity federations), better communication facilities (with service provider or peers) and improved productivity (e.g., paying fees online rather than

visiting brick and mortar offices). Participants pointed out that they are willing to accept more workload and less privacy in order to gain additional long-term convenience.

- "*I would go to Gmail, because Gmail groups emails together, forming a thread and that's more neat.*" [Focus group participant]
- "*I would use PayPal, because you input the data once and use it across different sites*" [Focus group participant]
- "*I would use to Gmail because it's linked to YouTube also. And then it's simple to use also*" [Focus group participant]

7.3.4 Framework contributions and modifications

This study provided additional insights on this group of users and how they interact with online services. The following aspects were considered to be important contributors to the development of *Sentire*.

7.3.4.1 Adding meta-data to quantitative behavioural models

The calibration process would be significantly richer if participants are given the opportunity to speak up and share their thoughts on aspects related to enrolment and workload. This highly depends on the available resources, mainly time, man-power and availability of participants. The cheapest and quickest way to generate this qualitative information is to encourage participants to think-aloud during the calibration session, however other techniques exist, including post-task semi-structured discussions on a one-to-one basis, focus-groups or post-task text capturing for remote calibration. This additional corpus of qualitative data provides a deeper understanding on users' attitudes, and if analysed well, could be used to supplement the simulations generated via the quantitative behavioural models. This information should be added to the respective user groups in the user group library for future reference and reuse (i.e., within *Sentire*'s CASE tool).

7.3.4.2 One-size does not fit all

This study has shown that major differences exist between the user groups considered in the first two case studies (*young urban professionals (30–40)* and *undergraduate students (non-IT)*). These differences have been mainly highlighted through the constructed quantitative behavioural models, and have shown that different groups of users react in a significantly different way to the same set of design factors. The adoption of Calibrated Personas during the requirements development process as well as the calibration process itself have confirmed the hypothesis that a one-size-fits-all enrolment strategy is not aligned with a user-centric design stance.

Through the adoption of NASA-TLX as well as thematic analysis on qualitative data it was observed that a user's decision not to complete a task is highly influenced by the level of perceived workload, however this perception may exist for different reasons altogether (across different users). High levels of perceived workload may arise due to the task's cognitive demands for one group, but due to physical demands or feelings of helplessness for other groups of users.

Following the main premise of *Sentire*, the user population needs to be analysed (i.e., different user groups within a region or nation) by building models that can be reused in future projects to inform the requirements development process through simulated user feedback ('listening' to users in absentia). Simulated feedback mitigates the risk of over-shooting budget constraints by reducing the dependence on frequent user evaluation exercises (following each design decision) while also mitigating the risk of expensive late-changes by flagging critical design issues earlier on, specifically at the requirements stage.

7.3.5 Plans for next study

Following the evaluation of this group of users it was observed that NASA-TLX was not always comprehended by younger calibration participants and a significant level of hand-holding was at times necessary (e.g., explaining workload dimensions several times, even though on-screen tips and examples were provided). It was decided to tackle this issue through the subsequent study, adopting the following objectives:

1. Carry out a systematic evaluation of NASA-TLX's suitability with younger audiences, in terms of
   (a) **understandability** and
   (b) **sensitivity**

   The government of Malta rolled out an exam registration e-service for A-level students sitting for their examinations in 2013. This was considered to be a unique opportunity to:

2. (a) Build behavioural models for this particular user group (i.e., digital natives)
   (b) Conduct an ex post facto evaluation for this national e-service and on the enrolment policy adopted, and
   (c) Assess the impact of this national policy on the digital natives' lived experience

7.4 Summary

This chapter outlined a second study wherein *Sentire* was evaluated with another user group and outside the e-government context. Changes arising from the previous study, mainly to the user group calibration process were applied with positive results and confirmed by a series of statistical tests on the models produced. The strategy adopted for data collection and model evaluation was similar to the one used in the first study, however in this case user group calibration was supplemented with qualitative insights generated using focus-groups as part of the exercise. These were then analysed thematically and six themes emerged (see Section 7.3.3). These themes revolve around aspects that influence this group's decision making process during enrolment, highlighting a number of factors that may encourage and discourage such users from adopting online services. Although this study was conducted in the commercial domain these factors may also apply in a governmental context.

A number of insights emerged from this study informing the evolution of the framework as well as its supporting CASE tools (see Section 7.3.4). Plans and goals for the subsequent study were outlined in Section 7.3.5.

Chapter 8

Case Study 3: Assessing NASA-TLX With Younger Users — Evaluating a Compulsory E-Service for Digital Natives

In 2013 Malta launched a new e-service for students aged 16–18 who were applying for their A-Level exams. Adoption was compulsory and students also needed to enrol for a national e-ID to gain access to the service. Governments have been using rhetoric of 'transformative' and 'citizen-centric' e-government services, but to deliver on this promise, they need to pay more attention to the design of identity-related processes [11, 127]. Enrolment is a pivotal part of the user experience, and without proper considerations, it can be the major hurdle which stops users from transacting online. Axelsson and Melin [11] note that while there has been ample technical research into e-ID, little research has been done on social and organisational dimensions. This study will give particular attention to *digital natives* – people who have grown up with, and are highly accustomed to digital technology [130]. This user group is contrasted with digital immigrants – people who have adopted technology later on in life by choice or necessity.

8.1 Defining the Context

The examinations department stipulated that students are to use a new e-service to register for their A-level examinations. Unless there were exceptional circumstances, students could not apply via the traditional method (i.e., visiting the examinations department in person). A 'Click Here to Apply' button was made available on a clean and easy to follow landing page at https://exams.gov.mt. Once clicked, students were asked to login using their e-ID credentials. No immediate information is given on how to obtain an e-ID. Instructions on how to enrol for an e-ID are provided in another e-government site[1], as follows:

1. Visit the registration office in Valletta in person (on average it takes 30 minutes each way by bus)

2. Go through a short enrolment process (on average it takes 5 minutes to complete and students need to present their national ID card and a valid email address). Queues are possible since this is a

[1] Instructions for e-ID enrolment are provided at https://mygov.mt/PORTAL/webforms/howdoigetaccesstomygov.aspx, (accessed June 2014)

central-government office.

3. A security PIN is sent by post to the address given at enrolment

4. Students are to activate the e-ID account using the PIN received by post and a password received by email (provided in step 2)

5. Finally students have to create a new password (following a strict password policy)

Once students are successfully enrolled on the National Identity Register, they could proceed to register and pay for their A-level examinations through the e-service at https://exams.gov.mt/.

8.2 Aims

This study aims to develop an understanding on how enrolment-specific workload, as a multidimensional measure, impacts the digital native's experience with online services. Qualitative techniques are used to capture this citizen group's perceptions of, expectations from and reactions to identity related tasks. This study also aims to determine whether NASA-TLX (1) is easy to understand and follow for younger (and untrained) participants, (2) whether it is applicable within this particular context (e-government) and (3) whether it is sensitive enough to significantly detect changes in workload parameters.

8.3 Method

8.3.1 Participants

Two sets of participants were involved in this study, one for each of the two phases discussed below, namely the (1) collection and analysis of users' experiences (online questionnaire) and the (2) follow-up workshops to verify and validate NASA-TLX ratings. Details for each group of participants are given in Sections 8.4.1.1 and 8.4.2.1 respectively.

8.3.2 Process

The author's goal was to capture as much feedback as possible from the pool of students sitting for their 2013 exams. An online questionnaire was opted for since it would help (1) reach as many students as possible while (2) minimising disruptions to their studies. A number of interesting insights and recommendations emerged during this exercise. It was also felt that this study would benefit highly from a second intervention through which the initial results could be validated and substantiated. This was the motivation for the second part of the study which offered the opportunity to assess the applicability and understandability of NASA-TLX with digital natives and to investigate its sensitivity towards workload induced by enrolment-specific factors. By this time all students had completed their exams and a series of follow-up workshops were organised in small groups of two to five students. Students who indicated that they would be willing to participate in follow-up meetings were contacted and a series of five workshops were scheduled between August and November 2013.

All ethical considerations recommended by the research ethics committees at UCL and University of Malta were observed for both phases of the study.

8.3.3 Sequential overview of study activities

Two major post-secondary schools in Malta were contacted in order to kick off the investigation. The questionnaire was designed to capture users' perceptions of, and reactions to, the new e-service for which an e-ID was required. Two pilot sessions were conducted to improve the questionnaire's layout and comprehensibility. A NASA-TLX evaluation sheet was also provided to take ex post facto workload measurements. This questionnaire was sent out to over 1000 students who were sitting for the 2013 A-Level examination sessions. Following this exercise a group of students were invited to participate in a series of focus groups during which their experience was investigated in more depth and additional measurements taken.

8.3.3.1 Online questionnaire

An online questionnaire was designed over a number of cycles, to capture several elements including workload perceptions, expectations and demographics. An online version of the TLX evaluation process was also presented and students were asked to consider their experience, including e-ID enrolment (if applicable), and then rate the six workload scales on a simplified scale (0 to 10 in the data preparation stage this was then scaled up to the original 0-100 range). Following the rating exercise, students were presented with 15 pairs containing all the possible combinations of the six workload dimensions which were rated earlier. For each pair they were asked to select the scale that contributed most towards workload. This produced a weighting for each scale which would then be used to generate a weighted overall workload score. An open-source survey engine (*Limesurvey*) was used to build the questionnaire and its logic. Two rounds of reviews were then conducted with fellow researchers and an initial version was established. A pilot session was held with a small sample of students from the target audience. A think-aloud session was held at their school within which six students voiced their thoughts while filling in the questionnaire. The students' feedback was instrumental and modifications were carried out accordingly to capture further information which was crucial for the generation of meaningful insights. A second pilot was then conducted with a medium sized higher secondary school through which the nature of the data and the best strategies to analyse it were determined (28 students participated in this pilot study). The final version of the questionnaire was then sent to students sitting for their 2013 examination sessions via the schools' official channels.

8.3.3.2 Processing quantitative questionnaire data

Responses from the online questionnaire were collected over a period of one month. Participant responses were segregated, distinguishing those who registered for their exams in person from those who adopted the e-service. The responses were then further categorised, splitting the latter group into two: (1) *students who already had an e-ID* and (2) *students who had to enrol for an e-ID*.

The next step was to calculate the weighted overall Task Load Index (or Mean Weighted Workload). Because of the multi-dimensional nature of NASA-TLX the specific dimensions that contributed most to the overall workload figure can be determined. For each participant the weighting for each scale (a number between 0 and 5) was multiplied by the rating provided for that scale (a number between 0 and 100). This figure (adjusted rating) was then divided by 15 to get a weighted score for workload (a number

between 0 and 100). See Section 2.3.5.1 for a description of the mechanics behind NASA-TLX. Each participant now has an overall mean weighted workload (*MWW*) figure for the exam registration task. This is considered to be the overall task load index, weighted according to each participant's perception as to what contributes most to workload. Given the structure of NASA-TLX data one could track back to the specific dimensions that each participant perceived as major contributors towards overall workload. The six workload scales were processed and both mean and median values were recorded across all participants (i.e., per group) to get a general understanding of the dimensions that contributed most to the overall workload rating.

8.3.3.3 Processing qualitative questionnaire data

The questionnaire was organised in three sections. The first section captured details about the students' experience with both the e-ID enrolment process and the e-service in general. The second section measured workload based on a NASA-TLX workload rating process and finally the last section captured basic demographics. Thematic analysis was performed on the first section in order to provide further insights on participants' ratings.

8.3.3.4 Follow-up workshops

Five follow-up sessions were organised (see Section 8.4.2), allowing the author to dig deeper into the students' experiences while eliciting further meta-data to explain the quantitative dataset generated via the questionnaire. During these sessions students had the opportunity to discuss, debate and agree on various issues raised by the researcher. These sessions were recorded, transcribed and analysed thematically. The sessions were of one hour each, and in the last 20 minutes students were asked to go through the user group calibration (UGC) exercise. This gave the author the opportunity to (1) build behavioural models for this user group, (2) validate workload feedback given via the questionnaire and (3) assess the students' understanding of the original, albeit digitised NASA-TLX 'pen-and-paper' process.

8.3.3.5 Analysing and synthesising results

Finally, results were analysed and synthesised. These are presented in Section 8.4 followed by a discussion on the main findings in Section 8.5. A set of recommendations arising from this study are provided in Section 8.6 .

8.4 Results

Three data sets were generated following this case study: (1) qualitative results from the questionnaire outlining experiences for the various subgroups, (2) quantitative workload data obtained from the questionnaire's NASA-TLX assessment and (3) data from follow-up sessions which includes both qualitative and TLX related information.

8.4.1 Online questionnaire

8.4.1.1 Participants

The questionnaire was sent to over 1000 students who were sitting for the 2013 A-Level examination sessions. A total of 134 valid responses were received (13% response rate). 62% of the participants

were female, 21% male and 17% decided not to disclose their gender. 81% of students declared that they fall within the 16–18 age-group while 15% chose not to disclose their age. Four participants stated that they are aged 19-24 and one was over 25 years of age. Only those falling within the 16–18 bracket were considered in the analysis stage. Furthermore, around 10% (13) of the respondents accepted the invitation to participate in one of a series of follow-up workshops held in the following months.

8.4.1.2 Students' reported experience

Given the compulsion to register for A-level examinations using the online service (and thus the need to enrol for an e-ID), around 93% of the respondents managed to complete the process successfully while the remaining 7% resorted to the manual process. Of those who completed the online process 13% had already enrolled for their e-ID in previous interactions with a government entity (e.g., passport renewal) while 87% had to go through the e-ID enrolment process. 34% of students who adopted the online method, and who had to enrol for an e-ID, stated that – given a choice – they would not use the online process again.

Through a thematic analysis of the main reasons given for this, a number of themes emerged:

- Hassle to register for an e-ID (57%)
  - "*Too much hassle for the e-ID*"
  - "*...the queues were extremely long for such a short process. It is highly time wasting for serious students*"
- Lack of trust in online systems (10%)
  - "*I feel safer if I had to go personally. I don't really trust the internet.*"
- Preference to traditional means (13%)
  - "*I would have preferred doing all the process by visiting the Examinations Department so that if a problem arises one could find help.*"
- Lack of process clarity (13%)
  - "*I don't know why the registration process had to involve an e-ID. To apply for a University course, it was different and we didn't need an e-ID. We just submitted normal information about us [via an online form]*"

7% gave no reason for their decision. The remaining 66% (who stated that they would register online even though they had to enrol for an e-ID) gave various reasons for this decision, represented by the following themes:

- Additional convenience disregarding the e-ID enrolment process (64%)
  - Given the e-ID enrolment process's reported difficulty, a state of cognitive dissonance amongst this group of students might have been created [56, 106]. To reduce cognitive

dissonance, people generally re-evaluate their attitude towards the action and this can also lead people to truly believe that the action was a positive one (refer to the peg experiment in Festinger and Carlsmith's paper [56]). Private opinion would generally change in order to "*bring it into closer correspondence with the overt behaviour the person was forced to perform*" [56].

The following are some comments from students who stated "*convenience*" as the main factor behind their positive intent to voluntarily make use of the e-service. These students **also had** to personally visit the e-ID offices in Valletta to enrol for an e-ID account:

* "*Saves me a lot of time instead of queuing up at the [examinations] department.*"
* "*More comfortable and easy to do it from the comfort of your home than to go to the [examinations] department.*"

The author argues that these students blocked out (ignored) the e-ID enrolment process and considered the exam registration process in isolation. If this group of students took the enrolment process into account the actual number of students who – given a choice – would make use of the e-service may be significantly lower.

- Additional convenience taking the e-ID enrolment process into consideration (16%)
  - Only 16% of students who had to enrol for an e-ID and who stated that they would still make use of the e-service explicitly took the e-ID enrolment process into account. In particular, some students even considered the e-ID enrolment process as an investment for future interactions (since the e-ID can be used for other e-government services). Some statements include:
    * "*I chose yes since if I took all the trouble to register for an e-ID account, I might as well use it again, even though I cannot pay online.*"
    * "*If I will be using the service regularly it would be worth waiting 3 days for activation.*"
- Enhanced security (3%)
  - "*Sure you have to register for e-ID once more but I feel it is more comfortable and secure applying for your exams online*"

17% did not give any reason for their decision. Interestingly, 31% of students who adopted the online method, and who already had an e-ID, stated that – given a choice – they would still not use the online process once again.

8.4.1.3 Students' reported workload

The second part of the questionnaire was an online version of the TLX workload assessment procedure. Initially students were asked to rate the six sub-scales (or workload dimensions) for the exam registration task (including e-ID enrolment if applicable), followed by the pairwise comparison to get a weighted overall workload measure (mean of weighted ratings). The six workload dimensions are Mental Demand

(*MD*), Physical Demand (*PD*), Temporal Demand (*TD*), Own Performance (*P*), Effort (*E*) and Frustration (*F*). The overall task load index (*MWW*) was calculated for each participant, and averaged across the various student subgroups (see Figure 8.1).

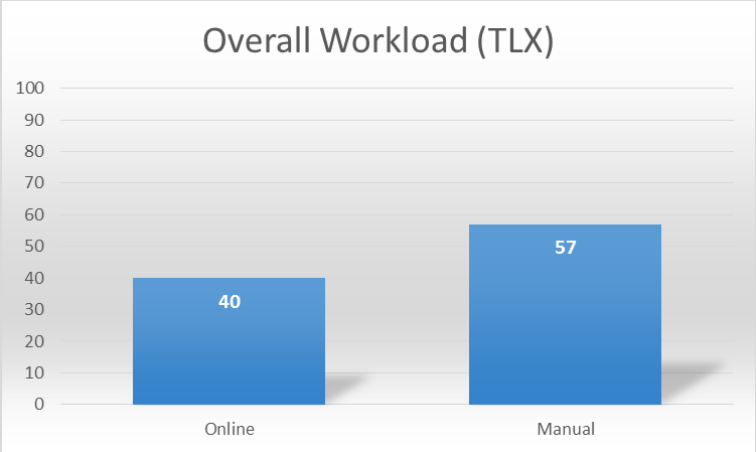


Figure 8.1: Mean weighted workload (*MWW*) for e-service users (online) and for those who adopted the offline exam registration process (at the exams registration department)

The average rating for the online method takes into consideration the ratings given by students who already had an e-ID and also by those who had to enrol for one. Table 8.1 shows how students who already owned an e-ID weighted the different workload dimensions.

Table 8.1: Workload dimension weighting by students who used the e-service and who already owned an e-ID

	MD[1]	**PD[2]**	**TD[3]**	**OP[4]**	**E[5]**	**F[6]**
Mean	3.7	0.4	1.9	2.6	2.2	4.3
Median	4	0	2	3	2	4
Standard Deviation	1.2	0.9	1.1	1.4	0.9	0.7

[1] Mental Demand
[2] Physical Demand
[3] Temporal Demand
[4] Own Performance
[5] Effort
[6] Frustration

Table 8.2: Workload dimension weighting by students who used the e-service but had to enrol for an e-ID

	MD[1]	**PD[2]**	**TD[3]**	**OP[4]**	**E[5]**	**F[6]**
Mean	2.8	1.4	3	2.1	2.2	3.7
Median	3	1	3	2	2	4
Standard Deviation	1.4	1.5	1.4	1.4	1.1	1.3

[1] Mental Demand
[2] Physical Demand
[3] Temporal Demand
[4] Own Performance
[5] Effort
[6] Frustration

Adjusted ratings are obtained by combining these weighted dimension values with raw ratings, as shown in Figure 8.2. In this case, Physical Demand is the lowest contributor to workload (adjusted rating = 2.5) however Frustration has an adjusted rating of 247, making it the highest contributor. Mental Demand follows Frustration, and thus these have a great influence on the average overall mean weighted workload (*MWW*). On the other hand, Table 8.2 shows how students who had to enrol for an e-ID weighted the different workload dimensions (out of 5). Figure 8.3 shows the respective adjusted ratings for this group. At a glance it is evident that this group of students had a different experience than the previous group and reported an increase in Physical and Temporal Demand. Frustration is still the highest contributor to workload, given an average weighting of 3.7, followed by Temporal Demand (3).

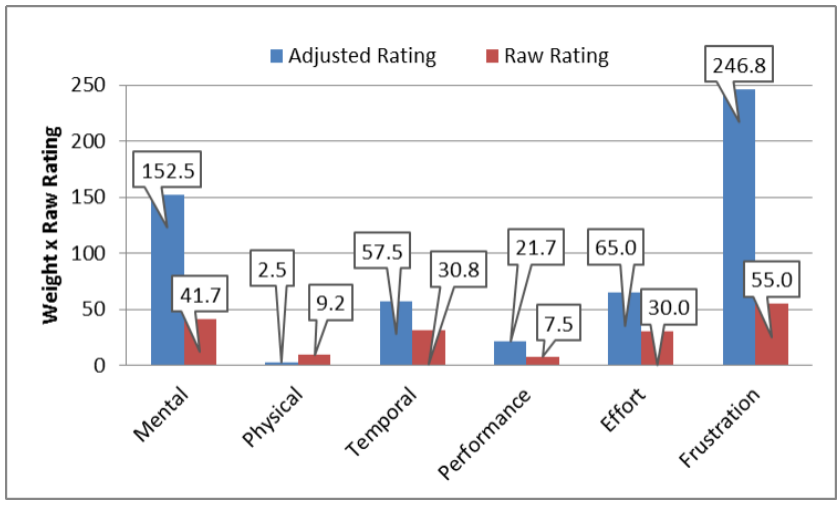


Figure 8.2: Adjusted rating for e-service users who already owned an e-ID (adjust rating = weighting x raw rating)

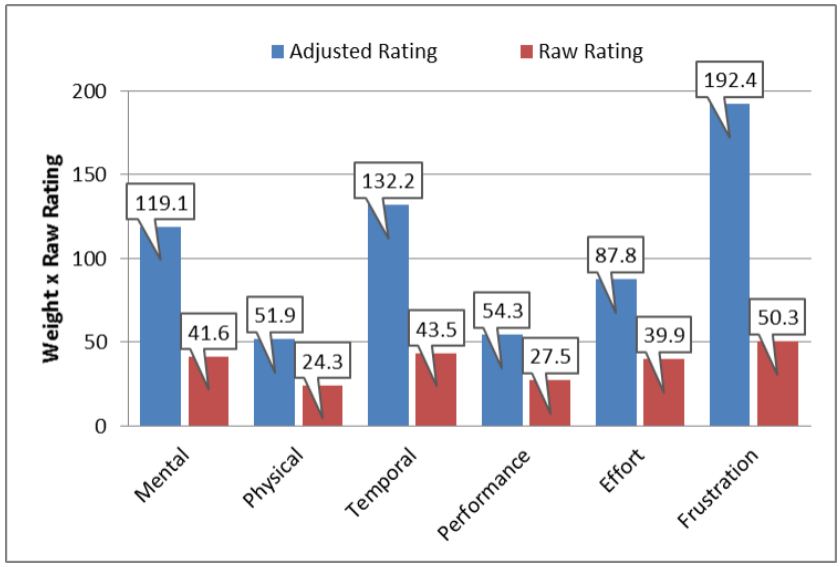


Figure 8.3: Adjusted rating for e-service users who had to enrol for an e-ID (adjust rating = weight x raw rating)

Given this information, the author argues that both groups of students (those who already had an e-ID and those who had to enrol for one) exhibited high levels of workload, albeit, for different reasons:

Those who had an e-ID Overall Task Load Index (TLX) was high mainly due to Frustration and Mental

Demand. Causes for this outcome were various, including lack of process clarity, preference for traditional means, lack of trust in online systems and site performance.

Those who did not have an e-ID Overall TLX was high due to Frustration, Temporal and Mental Demand. Causes for this outcome were various, mostly due to the hassle involved to get an e-ID (e.g., waiting time at the e-ID enrolment office). Physical Demand was also significantly higher than that reported by the previous subgroup.

8.4.2 Follow-up workshops

Students who accepted to participate in follow-up sessions were first asked to discuss their experience with the exam registration process and compulsory e-ID enrolment. Following this they were asked to compare and rate the perceived effort required to enrol for various online services including social networks, e-learning tools, payment gateways, email services and e-commerce sites. Each group had to reach a consensus for each rating decision and their interaction was observed. Following this, students were asked to go through the user group calibration exercise.

8.4.2.1 Participants

13 students accepted to participate in a series of follow-up sessions in small groups, eight of whom were female and five were male. Their median age was 17 years old. All participants had just finished their A-level examinations.

8.4.2.2 Perceived workload by consensus

Each group of students was presented with a list of online services that they might have used at any point in time (e.g., *Gmail*, Facebook, *Skype*, *PayPal* and *Hotmail* amongst others). The most commonly used services (for each group) were then listed on a white board next to a rating scale indicating the level of perceived effort required to enrol for that specific service (i.e., easy, medium, difficult/annoying).

Figure 8.4: Participants had to agree on the level of perceived enrolment-specific workload (from personal experience) for several online services

Students had to agree on the level of perceived workload for each services' enrolment process and their feedback was recorded. Both mean and median values for the most commonly used services

(across all groups) are presented in Figure 8.5. Feedback provided by different groups was normalised according to each group's rating patterns (e.g., some groups always rated high, while others were more conservative). This made it possible to generate high-level cross-group observations. Table 8.3 adds some context to these scores, providing annotations for the respective services' enrolment processes.

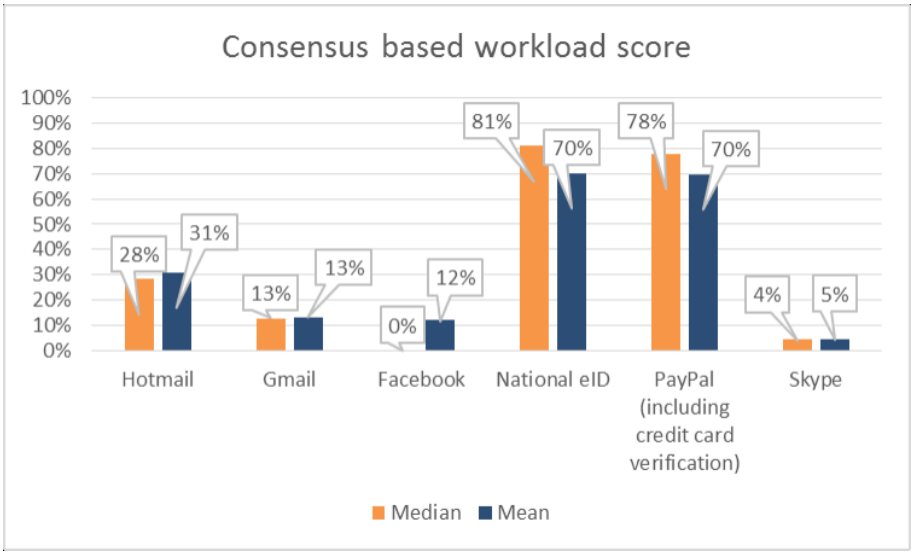


Figure 8.5: Perceived enrolment-specific workload for the most common online services

Table 8.3: Various services' enrolment processes, their design factors and consensus based perceived workload

Service	ItR[1]	ItG[2]	I[3]	D[4]	Perceived Workload (by consensus)
Hotmail	10	2	No	No	28%
Gmail	8	2	No	No	13%
Facebook	6	1	No	No	0%
National e-ID	NA	3	Yes	Yes	81%
PayPal [5]	13	1	Yes [6]	Yes [7]	78%
Skype	11	2	No	No	5%

[1] *ItR*: Items to Recall
[2] *ItG*: Items to Generate
[3] *I*: Interruptions to daily routines
[4] *D*: Delays
[5] Including credit card verification
[6] User needs to get hold of a bank statement
[7] Hours or days until transaction is visible in a credit card/bank statement

8.4.2.3 Sensitivity of NASA-TLX

During the follow-up sessions, students were also asked to go through the user group calibration (UGC) process (individually). One of the outputs from this process is a set of NASA-TLX scores for the various workload dimensions. It was decided to maintain the final NASA-TLX pairwise rating and thus generating a weighted workload rather than a raw TLX score (see Table 8.4 for weighting values).

It is evident from the weighting exercise that digital natives consider Frustration (*F*), Physical Demand (*PD*), Temporal Demand (*TD*) and Effort (*E*) as the major sources of workload (in this order). Frustration (*F*) was presented as a measure of irritation, stress and annoyance during the task while

Table 8.4: Workload dimension weighting by students following the final pairwise comparison

	MD[1]	PD[2]	TD[3]	OP[4]	E[5]	F[6]
Mean	0.7	3	2.7	1.7	2.5	4.4
Median	1	3	3	1	3	5
Standard Deviation	0.8	1.4	1.1	1.2	1	1

[1] Mental Demand
[2] Physical Demand
[3] Temporal Demand
[4] Own Performance
[5] Effort
[6] Frustration

Effort (*E*) was explained to be the level of mental and physical work required to accomplish the task. This corroborates with the consensus based perceived workload levels shown in Table 8.3 whereby the highest workload scores were given to those enrolment processes that interrupted the primary task. In the National e-ID case students had to visit an office in Valletta, while in *PayPal*'s case participants had to wait a couple of hours or days until a small *PayPal* transaction was processed and made visible on the credit card statement. The transaction details on the statement contain an activation code which is required to complete the verification process (i.e., card ownership).

The participants' overall weighted workload values for each of the nine fictitious enrolment processes presented during calibration are shown in Table 8.5. In Section 8.5.2 this data is compared and contrasted with results obtained through the consensus-based perceived workload exercise, shown in Table 8.3.

Table 8.5: Median value for the mean weighted workload (*MWW*) score across all participants for the nine fictitious enrolment processes presented during the user group calibration (UGC) exercise

Task	ItR	ItG	I	D	MWW
A	1	0	No	No	0%
B	2	1	No	No	0%
C	5	1	No [1]	Minor [2]	18%
D	4	2	No	Major [3]	11%
E	5	2	Yes [4]	Major [4]	32%
F	6	3	No	Minor [5]	14%
G	6	4	No	No	12%
H	9	3	No	Minor [6]	21%
I	NA	3	Yes [7]	Major [8]	81%

[1] Credit card details are required
[2] Wait a few minutes for activation email
[3] Wait three days before account is activated
[4] Visit closest outlet to confirm identity
[5] Upload recent photo
[6] Call free-phone to activate account
[7] Visit enrolment office during specific opening hours
[8] Three day waiting period till PIN is received

8.5 Discussion

8.5.1 Digital natives and NASA-TLX

The use of NASA-TLX to measure perceived workload in the exam registration process and e-ID enrolment (whenever applicable), provided the author with very useful insights. This, together with data from follow-up sessions helped to understand how students related to NASA-TLX's terminology and processes originally introduced by Hart and Staveland in [70], with the aim to maximise NASA-TLX's validity and utility for use with this group of users and within this context.

8.5.1.1 Workload manifests itself in different ways

Students who have used the exam registration e-service but had to go through the e-ID enrolment process were expected to give significantly higher overall workload ratings than those who already had an e-ID, mainly due to additional physical and temporal workload involved in travelling and queuing. This was not the case, with a negligible difference in overall *MWW* between the two groups. By drilling down into NASA-TLX's multi-dimensional results it was noticed that sources of workload were significantly different for the two groups. Both presented a high measure of overall workload, albeit for different reasons. In principle those who had to enrol for an e-ID were concerned with delays and interruptions to their primary task, however they indicated that the exam registration process was – in comparison – acceptable. On the other hand students who already had an e-ID based their feedback mainly on the non-functional aspects of the exam registration process, such as lack of clarity in the process and site loading speed, resulting in a high level of frustration. Uni-dimensional workload measurement techniques do not explain the user experience in its entirety. Issues in design and performance can cause frustration, and this can be an equally important contributor to perceived workload, together with the more traditionally accepted sources of workload (i.e,, physical and cognitive demand). The author recommends the adoption of a multi-dimensional workload assessment tool in order to understand the various sources of workload for different service alternatives. Future governments depend on the trust of younger citizens, and the interaction with government institutions is formative for trust perceptions. Riegelsberger and Sasse [137] point out that trust depends on the users' perception of motivation and competence – so being confronted with less than competently designed e-government services will undermine young people's trust in government.

8.5.1.2 Demystifying workload dimensions

Although provided with on-screen guidelines, participants in follow-up sessions were at times confused while rating certain dimensions, especially Own Performance (*P*), Effort (*E*) and Temporal Demand (*TD*). In particular Temporal Demand (*TD*) caused a level of confusion in its interpretation.

- "*What is Temporal Demand? Is it how long it took me to complete or how long it should have taken me?*"

Temporal Demand (*TD*) was originally introduced in NASA-TLX as a measure of time related pressure during a task, specifically on the pace at which tasks occurred. This is a very context specific dimension especially suited for critical scenarios such as an emergency landing of an aircraft in bad weather. As

is, this dimensions may not be adequate for non-critical and mundane tasks. Further to this, some participants also voiced their concern on the similarity of certain workload dimensions:

- "*The main problem is that some of them are really similar. And you wouldn't know what to choose*"

It was a non-trivial task to help participants understand the difference between the more abstract workload dimensions (e.g., Frustration (*F*) and Own Performance (*P*) or Effort (*E*) and Mental Demand (*MD*)). Students were given the opportunity to think aloud and clarify their doubts throughout the exercise by asking questions.

- "*The only thing which struck me was the 'own performance' rating. Sometimes it is a bit hard to figure out what you did right or wrong so it's kind of hard to assess own performance*"

Another comment related to how participants felt while conducting the final pairwise comparison, especially when they were asked to choose between Physical (*PD*) and Mental Demand (*MD*):

- Participant A: "*I also feel lazy with my choices*" Participant B: "*True true, same here*"

In this case both participants felt uncomfortable disclosing the fact that they preferred mental demand rather than physical demand, and it therefore cannot be excluded that lack of anonymity may influence feedback. This ties in with Malheiros's [101] observations on disclosure, whereby participants are less likely to disclose information (comfortably and honestly) if this portrays them in a bad light.

8.5.1.3 Keep out of reach of digital natives?

Consider Tables 8.1, 8.2 and 8.4. The weighting values for some of the workload dimensions provided via the online questionnaire (unsupervised) are considerably different from those provided for the same dimensions during the follow-up sessions (supervised) – see Table 8.6. This presents the possibility that participants who had no immediate support, as opposed to the supervised group, may have interpreted the rating scales differently from the supervised group, thus affecting the validity of results collected. If this is the case, the unmodified (original) NASA-TLX process would not be suitable in an unsupervised environment and with untrained participants. A set of tests are presented below to assess this hypothesis.

Table 8.6: Workload dimension weighting (median) varied when students were supervised as opposed to unsupervised responses (i.e., no immediate help was available)

	MD[1]	**PD**[2]	**TD**[3]	**OP**[4]	**E**[5]	**F**[6]
Unsupervised (online)	4	0	2	3	2	4
Unsupervised (online without e-ID)	3	1	3	2	2	4
Supervised (follow-up sessions)	1	3	3	1	3	5

[1] Mental Demand
[2] Physical Demand
[3] Temporal Demand
[4] Own Performance
[5] Effort
[6] Frustration

Given a non-normal distribution for the workload dimensions' weighting, a set of non-parametric tests were conducted (using the Related Samples Wilcoxon Signed Rank test) to determine whether there

is a statistically significant difference between the unsupervised and supervised sets of weighting values. The following null hypothesis was therefore adopted: the median of differences between each pair of data sets (e.g., Supervised MD and Unsupervised MD) is equal to 0 (i.e., no statistically significant difference exists between the two).

Table 8.7: Tests to determine whether there is a statistically significant difference between an Unsupervised and a Supervised TLX weighting exercise (i.e., pairwise comparison)

Null Hypothesis [1]	Significance (.025)[2]	Decision
Supervised MD and Unsupervised MD	.000	Reject the NH
Supervised PD and Unsupervised PD	.000	Reject the NH
Supervised TD and Unsupervised TD	.304	Retain the NH
Supervised OP and Unsupervised OP	.021	Reject the NH
Supervised E and Unsupervised E	.011	Reject the NH
Supervised F and Unsupervised F	.000	Reject the NH

[1] **Null Hypothesis (NH):** The median of differences between each pair of data sets (e.g., Supervised MD and Unsupervised MD) is equal to 0 (i.e., no statistically significant difference exists between the two)

[2] A comparison of two tests under different conditions is being presented using a Bonferroni adjusted alpha level $(0.05/2 = 0.025)$

Most tests in Table 8.7 indicate that a supervised TLX exercise will yield a significantly different result in the way the six workload dimensions are weighted by digital natives. In the follow-up sessions the facilitator explained each and every workload dimension before going through the different tasks. This might have contributed towards the variance in interpretation, and thus in weighting outcomes, between online and workshop participants. Table 8.6 clearly indicates the differences in the interpretation of rating scales with and without supervision and hand-holding.

It was noticed that this group of users did not fully understand the official NASA-TLX descriptions for the various workload dimensions, in particular those for Mental Demand (*MD*), Effort (*E*) and Own Performance (*P*). Specific and age-appropriate examples were found to be helpful.

8.5.2 NASA-TLX, e-government enrolment and digital natives — does it really work?

Can this technique be used to measure workload confidently with digital natives? This section will tackle a subset of tasks from the user group calibration exercise carried out in the follow-up sessions and their respective workload ratings across the six dimensions. Statistical tests show that there is a significant correlation between the resulting ratings and the demands imposed by the task.

Figure 8.6 represents the overall mean adjusted ratings for the selected tasks across the six workload dimensions. A pattern emerged across the three services and their respective workload ratings. Service D had no major workload issues, however Temporal Demand (*TD*) and Frustration (*F*) were rated as being considerably high (requires three days for account activation). Service G had low levels of workload across all dimensions, however Mental Demand (*MD*) was the highest rated dimension. This can be explained by the fact that participants had to come up with a new password, a password hint and a call-in-PIN (used to authenticate themselves in case they need to call a help-desk). Finally, service I had the

Table 8.8: This table shows three different tasks from the user group calibration exercise denoting the participants' perceived mean weighted workload (*MWW*)

Task	ItR[1]	ItG[2]	I[3]	D[4]	MWW
D	4	2	No	Major [5]	11%
G	6	4	No	No	12%
I	NA	3	Yes [6]	Major [7]	81%

[1] Items to Recall
[2] Items to Generate
[3] Interruptions to daily routines
[4] Delays
[5] Wait three days before account is activated
[6] Visit registration office during specific opening hours
[7] Three day waiting period till PIN is received

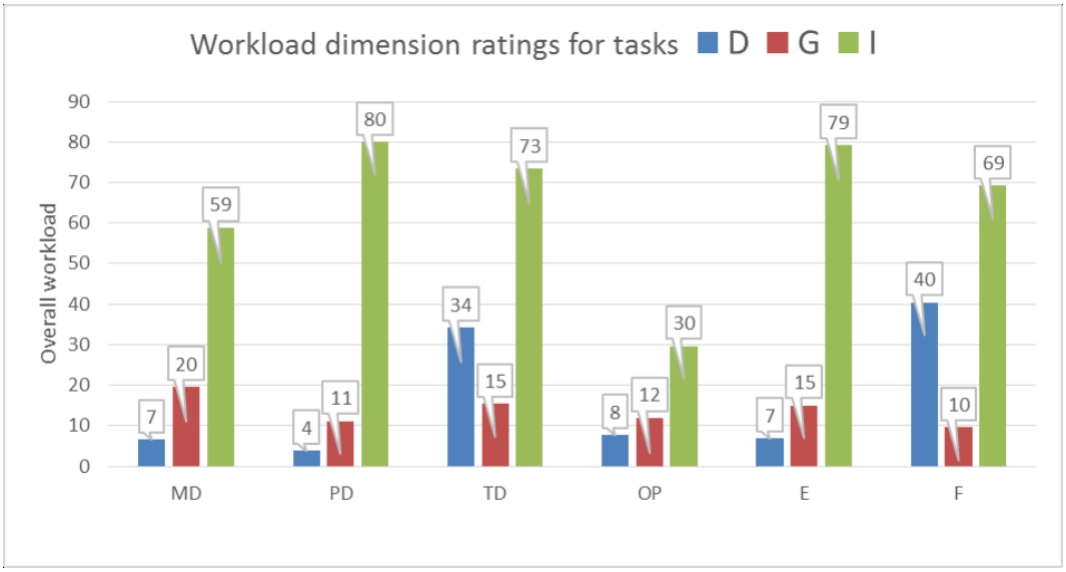


Figure 8.6: This chart shows the overall mean workload for the three tasks listed in Table 8.8

highest ratings across all dimensions, particularly those for Physical Demand (*PD*), Temporal Demand (*TD*), Effort (*E*) and Frustration (*F*). Half a day of travelling and queuing is required to complete the identity verification process as well as a three day period until the activation PIN is received by post.

A degree of consistency was observed between the perceived workload values (by consensus) for enrolment processes used on popular online services (see Table 8.3) and median weighted workload values for the nine user group calibration tasks (see Table 8.5). Some noticeable examples are provided in the Table 8.9. Although the two sets of results are close, one cannot exclude the possibility of other design factors placing significant influence on workload especially on dimensions such as Frustration (*F*) and Effort (*E*) (e.g., multi-page enrolment processes).

Table 8.9: Contrasting perceived enrolment workload (*PEW*) derived by consensus from actual enrolment processes with TLX-based Mean Weighted Workload (*MWW*) values for similar, but fictitious tasks

Real Service	PWL	Fictitious Service	MWW
Hotmail	28%	Task H	21%
Gmail	13%	Task G	12%
National e-ID	81%	Task I	81%

Even though the nine calibration (fictitious) tasks were presented in a random order, on average participants reported statistically significant differences in perceived workload for tasks designed to be theoretically more demanding (see Table 8.5). These task-workload comparisons were tested using the Related-Samples Wilcoxon Signed Rank non-parametric test for non-normally distributed data, as shown in Table 8.10. The null hypothesis set for these tests is that no statistically significant increase in perceived workload exists for subsequent incrementally (theoretical) demanding task.

Table 8.10: Tests to determine whether there is a statistically significant difference between reported workload levels for subsequent incrementally (theoretical) demanding tasks

Null Hypothesis [1]	**Significance (.05)**	**Decision**
PEW[*] for Task C over Task B	.001	Reject the NH
PEW for Task H over Task G	.033	Reject the NH
PEW for Task I over Task H	.003	Reject the NH
PEW for Task C over Task A	.001	Reject the NH
PEW for Task F over Task E	.039	Reject the NH
PEW for Task H over Task B	.001	Reject the NH
PEW for Task G over Task C	.039	Reject the NH

[1] **Null Hypothesis (NH):** The median of differences between each pair of data sets is equal to 0 (i.e., there is no statistically significant increase in perceived workload for subsequent incrementally demanding tasks).
[*] *PEW*: Perceived Enrolment Workload

In some cases, although the task was intended to be theoretically less demanding than the subsequent one, it turned out that digital natives perceived it as more demanding (although fairly low statistical significance is reported). Cases in point are tasks C and D as well as tasks F and G whereby the null hypothesis was retained. This can be explained by referring to the participants' supervised workload dimensions' weighting values (see Table 8.6) wherein Physical Demand (*PD*) and Temporal Demand (*TD*) (both given a weight of 3) are considered to be two major contributors to workload, as opposed to Mental Demand (*MD*) (weight of 1). Although tasks C and F are theoretically less demanding than their subsequent tasks (D and G respectively) with lower levels of mental demand (*MD*), they present users with more physical (*PD*) and temporal demands (*TD*) (i.e., travelling, looking up information and waiting for account activation).

A final set of tests sheds more light on the need to retain the pairwise comparison exercise that is used to produce weighted workload values for each participant. The following table shows the medians for Mean Weighted Workload (*MWW*) and Raw TLX workload (*RTLX*) together with their respective deviations from the mean. *RTLX* does not take workload dimensions' weighting into consideration and is calculated by dividing the sum of all workload dimensions' raw ratings (for each task/participant) by six (total number of dimensions).

The results shown in Table 8.11 do not highlight any major advantages for the adoption of weighted workload (*MWW*) values over raw workload (*RTLX*), given the additional effort required from participants to complete the final pairwise comparison. Eliminating this final step may in turn simplify the TLX process even further. To test this hypothesis a Spearman's rho correlation was run on the non-normally distributed values for *MWW* and *RTLX*. Two tests were carried out, one on the data collected during

Table 8.11: This table shows the set of nine calibration tasks together with their respective median *MWW* values alongside the median *RTLX* values

Task	MWW	St. Dev.	RTLX	St. Dev.
A	0%	2.2	0%	2.6
B	0%	9.7	0%	8.1
C	18%	15.6	16%	13.3
D	11%	21.8	8%	17.7
E	32%	28	33%	22.3
F	14%	18.6	13%	17.1
G	12%	9.1	13%	8.8
H	21%	13.3	21%	13
I	81%	27.3	72%	24.4

the follow-up workshops (117 observations from 13 participants reporting on nine fictitious tasks) and another test on values reported through the online questionnaire (94 students who had to enrol for an e-ID before using the e-service). In both cases the Spearman's rho revealed a positive and statistically significant relationship between *MWW* and *RTLX* ($r_s[117] = .989, p < .001$ and $r_s[94] = .937, p < .001$ respectively). In line with these observations, Cao et al. observed [27] that *RTLX* is more commonly adopted over *MWW*, citing the high correlation between weighted and unweighted workload scores as the main determining factor.

8.6 Recommendations

In this section the results obtained are revisited and a number of recommendations for both practitioners (see Sections 8.6.1 and 8.6.3) and researchers (see Section 8.6.2) are proposed.

8.6.1 Encourage secure behaviour

Although the following recommendations stem from this study it is believed that they are generic enough to be well suited for natives and non-natives alike.

8.6.1.1 Measure

By systematically measuring workload for proposed enrolment processes, designers are able to visualise the effort required to complete the task and contrast this with the benefits that users could obtain. A balance must be struck while keeping the required identity assurances in check for safe operation of the e-service. Is in-person identity verification really required at first contact? Can verification be deferred, eliminated or replaced with additional non-invasive and non-blocking requests for identifying information?

8.6.1.2 Simplify and guide

Security processes should be broken down into distinguishable bite-sized steps, preferably adopting a wizard layout using short, clear yet concise age-appropriate messages. E-service designers may find inspiration from popular online services while considering open-sourced user interface frameworks (e.g., Twitter's *Bootstrap* and Zurb's *Foundation* frontend frameworks) and widely accepted design patterns. In principle, use less (but more age appropriate) text prompts, adopt visual cues and move towards

responsive interfaces that reorganise themselves to preserve the intended flow and action clarity across different devices and display resolutions.

8.6.1.3 Integrate and defer

Interruptions and delays are considered to be two major sources of workload, contributing towards Physical Demand (*PD*), Temporal Demand (*TD*) and Frustration (*F*). Integrating security tasks with primary tasks, either implicitly (security by designation – refer to [175]) or asynchronously, may reduce the perception of a hurdle while allowing users to complete their primary task (or partially) at first contact. Integrating enrolment tasks within the primary task may give the impression of a non-enrolment service whereby users are not explicitly coerced to create an account. Nonetheless, following task completion users could be given the option to store their details for future interactions. Whenever integration is not possible (e.g., requiring manual verification of identity information before activating account) one may consider asynchronous (or deferred) enrolment. This can be achieved by deferring the disruptive step (e.g., manual identity verification) while allowing for the completion of the primary task. Backend verification can then occur in batch, by (1) verifying the identity of the user and (2) process the required transaction (following successful verification). If issues in the enrolment or transactional information are discovered the service provider (automatically or assisted) would then ask for clarifications or further details, but only as an exception to the rule. Splitting the primary task in two (i.e., enrolment (secondary task) and actual transaction (primary task)) might add to the perceived costs of using the service. The idea of having to go back to the primary task, especially if delays are present, may discourage users to initiate the transaction in the first place. This may also lead them to find alternative channels through which they could complete the transaction at one go (e.g., in person or by post).

If in-person verification is required (as in this case study), one may consider deferring the in-person portion of the process while still allowing for the completion of the online task. Eventually, in-person enrolment would then validate the data provided online without acting as a blocking step at first interaction. One should also consider the real need for in-person verification at some official government office. Can a trusted institution (e.g., employer) provide the necessary information instead (with the user's consent)? Clear guidelines should be published on the various identity mechanisms and processes that can be used to deliver the various levels of identity assurance requirements (see Table 2.2). Would a driver's license suffice instead of a birth certificate to provide a specific level of assurance for a given e-service? Both documents can be obtained, however the probability of having a driver's license readily available is higher than that of a birth certificate (which would require more effort to obtain). This is also based on the assumption that target users have a drivers license.

Separating enrolment from the primary task adds to the perceived cost of the transaction (i.e., breaking the flow with delays or interruptions) and this may in turn make goal-oriented users abandon the task altogether. Users may feel that security measures interfere "*with their ability to deliver their work on time*" [12], justifying their circumvention or non-adoption. Delays or disruptions could result from various technical or political decisions; from downloading a browser plug-in to waiting in line to enrol.

8.6.1.4 Communicate utility and value

A small percentage of participants understood that the e-ID could be used for future government interactions and with this in mind they were happy to go through the whole enrolment process. Outlining the benefits gained through enrolment might help reduce the level of resentment and frustration, especially when compulsion exists. In this case the e-ID enrolment process was a pre-requisite for the e-service itself, rather than part of it. If there is no other way to eliminate or improve the enrolment process (through technical or procedural re-engineering) additional effort should then be made to communicate the potential benefits one would obtain following enrolment.

8.6.2 Modifying NASA-TLX for use in enrolment

The NASA-TLX rating process may require some minor changes to make it more palatable and understandable by this particular group of users (within an e-service enrolment context). This suggestion might also be well suited for other user groups.

The meaning of Temporal Demand (*TD*) needs to be modified to fit within an e-government context. "*Feeling rushed*" is not an appropriate measure for enrolment processes (as opposed to other situations such as engaging landing gears during an emergency landing). In the follow-up sessions Temporal Demand (*TD*) was expressed as a measure of the time required to complete the task. The associated hint should read: "*How much time did you require to complete this task?*". This represents the perceived amount of time taken-up by the enrolment portion of an e-service, rather than the pressure exerted from time limitations.

Simpler definitions and context specific examples are needed for most of the rating scales:

- *Own Performance:* How confident were you during the enrolment process? Was the process easy to follow?

  The inverted labels for *Own Performance* (Good to Poor rather than Low to High) did not seem to be problematic.

- *Physical Demand:* How much physical effort did the process involve? Did you have to reach for some documents? Did you need to go somewhere in-person to complete the transaction?

- *Mental Demand:* How much thought was required during this process? Did you have to come up with new secrets, such as usernames, passwords or PINs? Did you have to provide a lot of information to complete the form(s)?

- *Effort:* Considering both mental and physical demand, did you require a lot of effort to perform the process?

- *Frustration:* How irritating or annoying was this enrolment process?

If possible provide a channel for immediate feedback during the TLX rating process using voice over IP (VoIP) if physical proximity is not possible. Finally, Raw TLX was found to be a suitable measure to inform designers on perceived workload for this group of users (digital natives), while also

simplifying the overall rating process. This was mainly due to the fact that a high level of correlation was found between Raw TLX and Mean Weighted Workload values, making the additional effort required to generate *MWW* values unjustifiable.

8.6.3 Privacy and frustration

Frustration may also result from the request for personal information which goes beyond the acceptable parameters of information requirements by the e-service or service provider. Users may feel a disproportionate level of identity exposure [131]. This may have a severe impact on user perceptions and decision making. Other factors exist (e.g., invasive data requests and fairness) that would have an impact on the Frustration (*F*) workload dimension. The impact of privacy on process completion is discussed thoroughly in [83, 102, 129].

8.7 Conclusions

Following a rigorous empirical exercise, this chapter offers two perspectives on security and the user experience:

1. The impact that security-related policies can have on the users' lived experience (ULX), in particular for digital natives, and
2. The applicability of NASA-TLX as a highly cited human factors technique to measure such impact

In this particular scenario, enrolment involved travelling and waiting in a line, and it comes as no surprise that physical and temporal demand got high scores. Nonetheless NASA-TLX is a multidimensional workload measurement technique and this chapter assesses the impact of this particular security policy across the six workload dimensions. Observations are also confirmed using statistical tests.

The findings are revisited in light of the aims set for this study:

- *What effect, if any, did this national policy have on the digital natives' lived experience?*

34% of those who had to enrol for an e-ID stated that given a choice, they would not use the e-service. A number of insights were produced from the feedback generated. Thematic analysis on focus group discussions helped to establish recurring themes that emphasise the sentiments expressed by participants (see Section 8.4.1.2)

- *Can the original NASA-TLX technique be used as-is with this group of users?*

With minor modifications NASA-TLX could be improved to serve its purpose better within this particular context and with this user group. This also includes additional guidance on the meaning and implications of the various workload dimensions (see Sections 8.5.1 and 8.6.2).

- *Is NASA-TLX sensitive enough to measure enrolment related workload in this context?*

NASA-TLX provided interesting insights into the possible sources of workload for this group of users, and it was found to be fairly sensitive to changes in workload parameters, informing the researcher on

possible actions to reduce workload perceptions, improve adoption and if compulsion exists, minimise resentment (see Sections 8.4.1.3, 8.4.2 and 8.5.2).

8.8 Process Evaluation

8.8.1 Framework contributions and modifications

This study gave the author some interesting insights into the sensitivity and applicability of NASA-TLX with younger audiences. A number of minor modification were necessary to improve its understandability, and these were reflected in the user group calibration (UGC) exercise (e.g., providing age and group appropriate descriptions to explain the various calibration stages, including NASA-TLX workload dimensions).

8.8.2 Plans for next study

The first three studies contributed towards the evolution of *Sentire*, its underlying theory as well as supporting CASE tools. This motivated the planning of a final study in which *Sentire*'s performance could be evaluated in a fully fledged public facing e-government service project.

8.9 Summary

This study offered further insights on the sensitivity and applicability of NASA-TLX as a subjective and multi-dimensional workload rating technique and its use within the user group calibration process. User experience considerations should be an integral part of policy making, rather than a non-functional (or secondary) exercise within the development process. NASA-TLX is a suitable starting point to measure workload and assess potential impact on users. *Sentire* offers the necessary framework and supporting CASE tools to include NASA-TLX within a systematic process of discovery, design and development.

Chapter 9

Case Study 4: Building a National Consumer Affairs E-Service

This case study aims to assess the latest iteration of *Sentire* within a fully-fledged e-service project, from inception to launch. This will encompass the entire process and determine the impact that *Sentire* affords at each stage. The requirements development process (project blast-off) started in November 2013 and the first high-fidelity prototype was delivered in May 2014.

9.1 Defining the Context

The Malta Competition and Consumer Affairs Authority (MCCAA) are aiming to improve the way citizens interact with the authority while streamlining internal processes for increased efficiency. For this reason a series of meetings were carried out with top management to discuss their needs as well as the author's objectives, and eventually a collaborative agreement was formalised for the development of the Consumer Advice Portal (CAP) – a public facing e-service that acts as the first point of contact for consumers. CAP was planned to offer an advice and complaints wizard, a publication repository, frequently asked questions as well as a knowledge base on past cases. This e-service will also act as an internal knowledge base giving easier access to information and improve knowledge transfer and reuse for current and future case officers. Based on this it was agreed to adopt *Sentire* to develop CAP's requirements as well as design and build the actual portal using an agile development methodology. CAP can be categorised as a *Rabbit project* in *Volere* terminology.

9.2 Aims

9.2.1 Research objectives

This case study aims to review the latest iteration of *Sentire* in its entirety as well as the corresponding CASE tools within a fully-fledged public facing and potentially enrolment based e-government service. This will act as a confirmatory study targeting a wide variety of user groups providing the author with the opportunity to reuse existing Calibrated Personas to guide the requirements development process. Simulated user feedback will iteratively inform the project team on the right balance between perceived workload and identity assurance levels for the different user groups under consideration.

9.2.2 Practical objectives

MCCAA need to develop and launch a consumer advice e-service (CAP) to improve its interaction with the public while optimising internal resource utilisation (e.g., encouraging case officers to reuse information via an internal knowledge base, providing frequently asked questions online, directing users to the right authority/channels, providing a publicly available knowledge base and so forth).

CAP will also serve as a first point of contact for consumers (and potentially traders) through which they can seek advice, file complaints and subscribe for market-related updates and alerts.

9.3 Method

Figure 9.1 outlines the workflow adopted for this project, broken down by work unit. Each work unit will be tackled in the subsequent discussion, highlighting implications arising from the adoption of *Sentire*. In the initial stages (**T1**, **T2**) high level goals were specified and approved by top-management. A project team was designated, which in turn initiated discussions with various stakeholders who may have a direct or indirect influence on the e-service (**T3**). Stakeholder identification and short-listing was based on a simplified version of Ian Alexander's onion-ring model [8]. Several pre-determined categories (from the *Volere* template) were also used to inform the discovery of people, entities and systems that can have an impact or are impacted by the authority's work. By the end of the requirements development process around 47 stakeholders were discovered, together with their inter-relationships (wherever applicable). These included domain experts, internal specialists and other government entities.

A discussion on business events was then initiated (**T4**). These are events that occur at the authority, including externally initiated events (e.g., consumer calls the help-desk) and time-triggered events (e.g., bi-weekly batch processing of complaints). By creating an outline of business events the project team

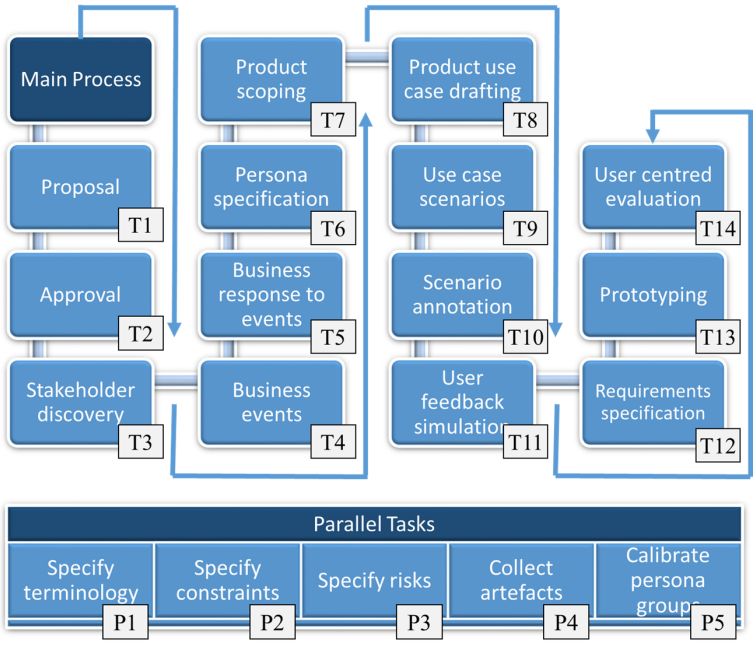


Figure 9.1: Consumer Advice Portal project workflow – *T* labels represent sequential tasks while *P* labels indicate parallel tasks

was preparing a solid foundation to (1) rigorously identify business use cases (how the authority responds to business events) and (2) discover new areas that may need further investigation. These are foundational steps enabling a fine grained scoping exercise for the system-to-be. By the end of the requirements development process 12 externally initiated events were identified along with six time triggered events (18 in all). The project team's understanding of how the different sections within the authority operate was quickly developing (i.e., directorates, divisions, institutes and entities). This knowledge also included the way different internal stakeholders respond to both external and time triggered events. This helped formulate a first set of technology-agnostic business use cases (**T5**) explaining the various work processes and exceptions thereof. Relevant artefacts, such as physical forms, documents, leaflets, meeting recordings, transcripts, photos and correspondence were stored within the project's workspace in *Sentire*'s CASE tool (**P4**).

So far, all discussions revolved around the authority's work, and there was no mention of any e-service related features. Initially the team wanted to draw a picture of the authority's world and work as well as any form of interaction with consumers and traders. This serves as empirical evidence into which well-scoped decisions may be grounded. This is in contrast to a situation in which requirements are drawn up around unknown problem(s) within an undefined domain. Scoping becomes much easier and robust when based on broad knowledge. At this point, workflows started to emerge and various (sometimes unexpected) domain experts were involved in order to walk the project team through the various scenarios. Business use cases were specified within the CASE tool reflecting how the authority responds to the various events. Information held included use case pre-conditions, business rules and outcomes as well as normal, alternative, exception, misuse and negative use case scenarios. Active and interested stakeholders were associated with use cases. Several domain specific terms and acronyms started to emerge and these were recorded in the terminology library within *Sentire*'s CASE tool (**P1**). The terminology library facilitates communication across the entire team (including future team members) by providing a common dictionary of domain and project specific terms. Project constraints (**P2**) and risks (**P3**) were discussed in tasks **T1** and **T2** and also stored in their respective repositories. Furthermore, identified risks can also be linked to other aspects of the project (e.g., Use Cases) that can help to mitigate their occurrence. An online, centralised and structured repository makes it easier for team-members to monitor progress without the need to request and go through any ancillary documentation.

Following the exploration of business events and associated responses, the project team now had a clearer picture of the main citizen groups who are presently interacting with the authority. A set of hypothetical personas was outlined – flagging those personas who might eventually make use of the consumer advice portal (**T6**). At this point a parallel task was spawned – persona elaboration, grounding and calibration (**P5**). Persona hypotheses were stored in the CASE tool and this encouraged the project team to collect more empirical data that would shed more light on the authority's clients, including actual case data, statistics and past experience. This enabled the team to evolve these personas even further while grounding aspects such as user activities, aptitudes, attitudes, motivations and skills in empirical evidence. Personas were elaborated even further during the persona calibration exercise. This

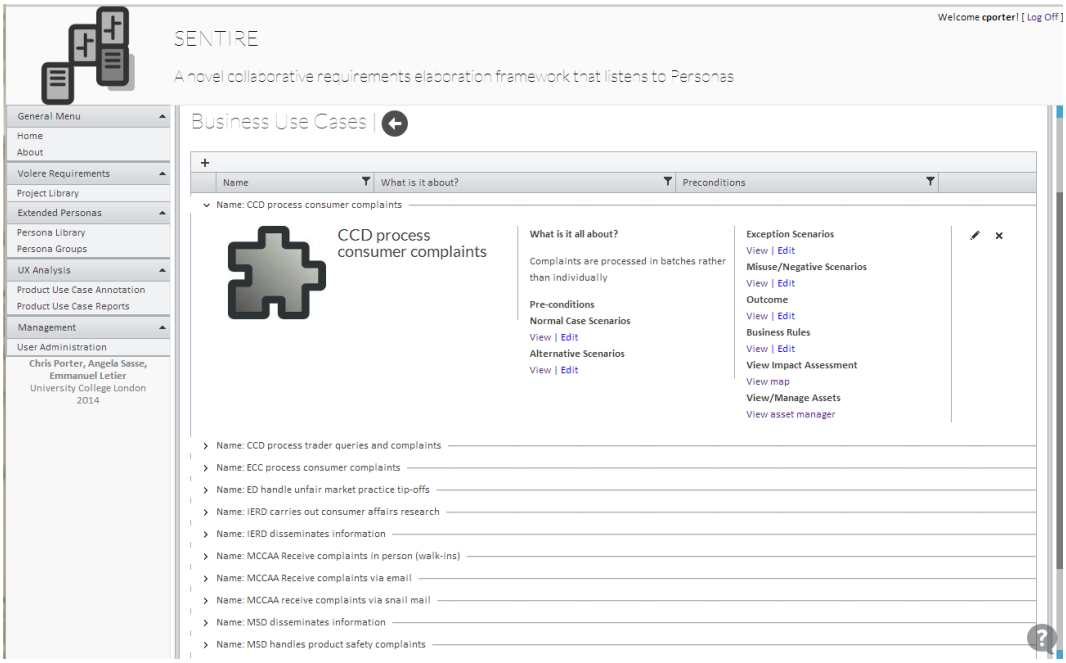


Figure 9.2: Business use case screen in *Sentire*'s CASE tool

task ran in parallel with the main project requirements workflow.

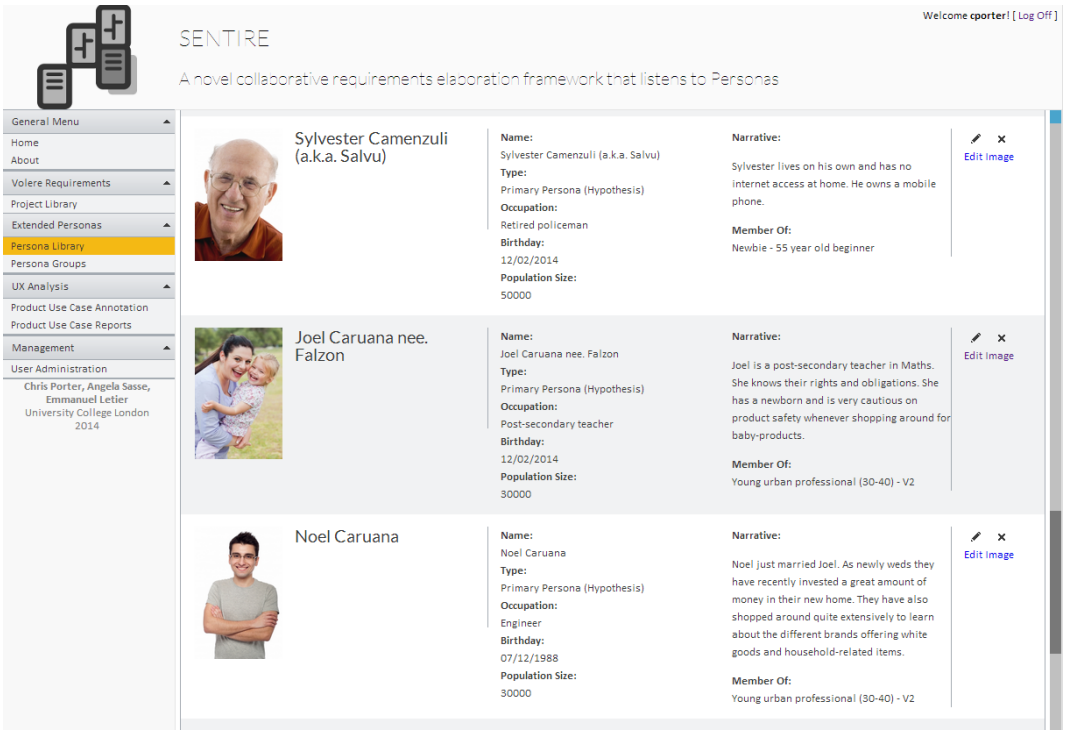


Figure 9.3: Persona library in *Sentire*'s CASE tool

Initially four primary and four secondary personas were specified:

- *Mary Piscopo* — 55–65 year old technology-newbie
- ***Joel Caruana*** — 30–40 year old teacher and mother

- ***Noel Caruana*** — 30–40 year old engineer
- ***Shanya Borg*** — 16–18 year old student

Secondary personas (will not be directly affected by the public facing e-service)

- *Joe Grech* — 55–65 year old trader (not a target user)
- *Sylvester Camenzuli* — 70+ year old retired policeman (non-adopter)
- ***Joanne Bonnici*** — 18–25 year old portal content manager
- ***Joseph Zammit Borda*** — 30–40 year old case officer

A number of user models were already available, generated for use in previous case studies (for similar project personas to those shown in bold). No models were available for Mary Piscopo. User group calibration (UGC) sessions were organised with participants falling under this user archetype. Seven individuals accepted to participate and in-context calibration was conducted at each participant's home. Each session took around 50 minutes to complete. All ethical considerations recommended by UCL's Research Ethics Committee were observed.

The online and self-administered version of the calibration process was not practical with this group of people mainly due to a low level of confidence in using online services for the first time. A new approach for data collection was devised whereby the facilitator explained each task using visual and verbal cues while calibration feedback was recorded on their behalf. This approach might pose a risk to validity due to potential bias (as opposed to going through the calibration process independently), however a large degree of openness and honesty was noticed in their responses. This served as re-assurance on the authenticity and quality of the data generated. Leading questions were also avoided to mitigate this risk. This technique also helped to uncover several unexpected insights on this user group's attitudes towards enrolment. The calibration data from each session was consolidated and prepared for processing using a statistical package (SPSS). Throughout the process it was noticed that some participants behaved in a significantly different manner, even though they were theoretically accurate representatives of the persona under consideration. This led the project team to believe that the initial persona (Mary Piscopo) may have been an over-generalisation of that particular user group. This was also considered to be possible empirical evidence for an emerging persona.

Additional insights on participants denoted the emergence of a new and distinctive set of attitudes, suggesting the existence of two rather than one user groups arising from the original persona hypothesis (Mary Piscopo). This new group of people appeared to be extremely afraid of "*breaking something*" even though they are willing to adopt new technologies. Another attitude was that they prefer enduring physical demand (i.e., going somewhere) than feeling frustrated (i.e., fear of doing something wrong or lack of understanding). This was related to the issue of confidence, specifically wanting to get the job done with high confidence in the outcome.

It was decided to create a new persona and associated user models to cater for this variant on the original persona. This new persona was called Doris Piscopo and its attributes were based on the emerg-

Figure 9.4: Facilitated in-context calibration session with a participant

ing insights. Qualitative data was also being generated thanks to the adopted data collection mechanism (affording a higher degree of dialogue), providing more insights into aptitudes, attitudes and other important factors that contribute towards the construction of a well grounded persona.

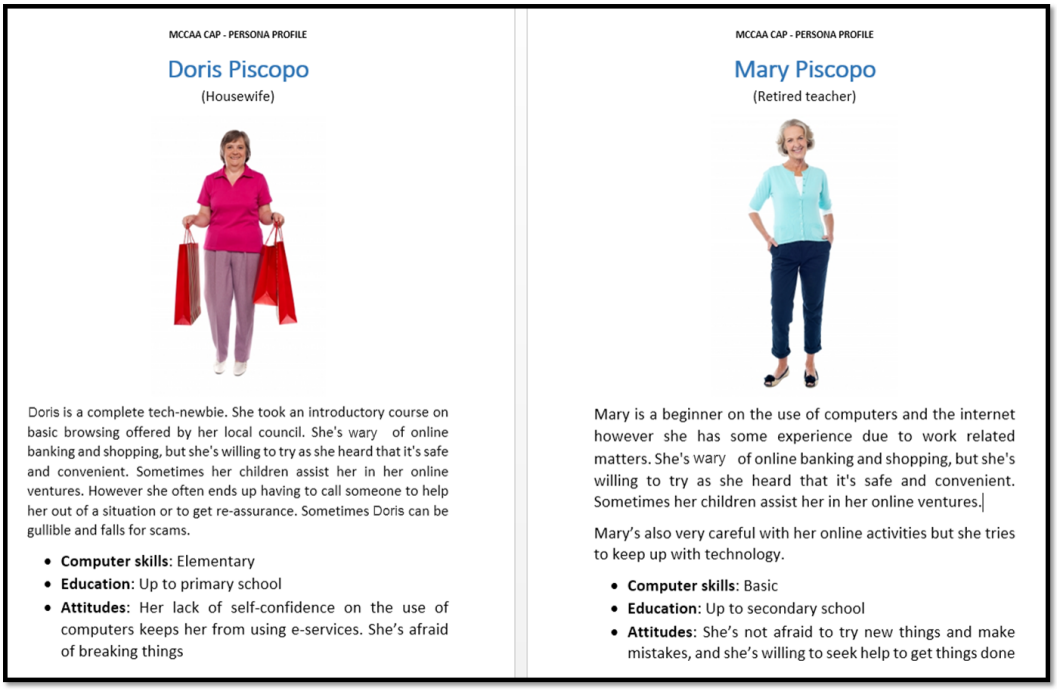


Figure 9.5: Creation of a new persona hypothesis to reflect an emerging user archetype. Persona posters (shown here) were used during project meetings.

A qualitative analysis of the calibration sessions confirmed the author's hypothesis for this new persona. The following quotations sum this up:

- "*When I see an enrolment page I stop as I'm afraid of breaking something [the computer]*" [Participant 2a (retired couple)]
- "*I depend on my daughter who's still at home when she leaves, then I'll make an effort to overcome my fear*" [Participant 1]

- "*I prefer to go out and finish tasks in person, its part of life and its relaxing*" [Participant 1]
- "*Time and physical effort are not an issue*" [Participant 2b (retired couple)]

"*Lack of self-confidence*" and "*fear of breaking things*" are two main attitudes associated with this new project persona. These observations are grounded in feedback obtained throughout the calibration process. In the final part of the calibration process (comparison of NASA-TLX workload dimensions) it became extremely clear that this group of people are more willing to endure physical and temporal demand as long as they feel confident that they completed the job, they did it "*the right way*" and that they're not making any mistake that could cause harm to anyone or damage anything. One participant noted that she has a "*fear of breaking the national network [laughing]*". The calibration exercise helped the author learn more about this group of users, beyond their attitudes towards design factors. This exercise also allowed the researcher to model their attitudes for future reuse.

Two regression models were created for the *confident newbies' (55+)* user group: a linear regression model to explain perceived workload and a binary logistic regression model to predict the users' willingness to adopt the e-service.

Table 9.1: This table shows regression coefficients generated for the *confident newbies (55+)* user group (represented by Mary Piscopo)

	Regression coefficients	
	Task completion (see Figure 9.6)	Perceived workload (see Figure 9.7)
B-Coefficient	40.642	15.328
Items to Generate	NA	4.614
Items to Recall	NA	NA
No Delays	2.425	-5.116
Minor Delays	.734	34.486
Major Delays	0	0
Interruption	-20.043	34.700
Type of Service 1	-41.871	NA
Type of Service 2	-40.882	NA
Type of Service 3	NA	NA
Type of Service 4	NA	NA

Considering the *confident newbies (55+)* user group (represented by Mary Piscopo) the backward stepwise entry method was adopted to conduct a binary logistic regression analysis for the participants' willingness to complete the task (*WCT* model outcome is binary – yes/no). Three significant predictors were found (see Table 9.3) explaining over 74% of the total variation in the yes/no response (Negelkerke = 0.746 – see Table 9.2).

Table 9.2: *Confident newbies' (55+) WCT* model – testing fitness to the data

Pseudo R-Square	
Cox and Snell	.509
Nagelkerke	.746
McFadden	.621

No significant predictors were found for the perceived enrolment workload (PEW) regression model, and resource constraints limited the possibility for further calibration sessions. It was nonetheless

Figure 9.6: Complete results for the task completion (*WCT*) regression coefficients generated for the *confident newbies (55+)* user group

Parameter Estimates

Decision[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
Yes	Intercept	40.642	.939	1874.618	1	.000			
	[NatureOfService=0]	-41.871	.724	3345.192	1	.000	6.544E-19	1.584E-19	2.704E-18
	[NatureOfService=1]	-40.882	.000	.	1	.	1.758E-18	1.758E-18	1.758E-18
	[NatureOfService=2]	.000	.000	.	1	.	1.000	1.000	1.000
	[NatureOfService=3]	0[b]	.	.	0	.	.	.	.
	[Delays=0]	2.425	1.111	4.759	1	.029	11.298	1.279	99.785
	[Delays=1]	.734	1.015	.523	1	.469	2.084	.285	15.241
	[Delays=2]	0[b]	.	.	0	.	.	.	.
	[Interrupts=1]	-20.043	6157.799	.000	1	.997	1.973E-9	.000	.[c]
	[Interrupts=2]	0[b]	.	.	0	.	.	.	.

a. The reference category is: No.

b. This parameter is set to zero because it is redundant.

c. Floating point overflow occurred while computing this statistic. Its value is therefore set to system missing.

Figure 9.7: Complete results for the perceived workload (*PEW*) regression coefficients generated for the *confident newbies (55+)* user group

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	15.328	12.8162	-9.791	40.447	1.430	1	.232
[Delays=0]	-5.116	12.9980	-30.592	20.360	.155	1	.694
[Delays=1]	34.486	20.1137	-4.937	73.908	2.940	1	.086
[Delays=2]	0[a]	.	.	.	.	.	.
[Interrupts=1]	34.700	23.8460	-12.037	81.438	2.118	1	.146
[Interrupts=2]	0[a]	.	.	.	.	.	.
NewFacts	4.614	2.6258	-.533	9.760	3.087	1	.079
(Scale)	2.720[b]	.2972	2.196	3.370			

Dependent Variable: Workload

Model: (Intercept), Delays, Interrupts, NewFacts

a. Set to zero because this parameter is redundant.

b. Maximum likelihood estimate.

Table 9.3: *Confident newbies' (55+) WCT* model – likelihood ratio tests

Effect	*Chi-Square*	*df*	*Sig*
ToS	62.824	3	.000
D	7.328	2	.026
I	5.016	1	.025

decided to go ahead with the data at hand. A Gamma Regression model identifies three predictors for the PEW model, whereby Delays (*D*, $p > 0.05$) is the best predictor followed by Interruptions to daily routines (*I*, $p > 0.05$) and Items to Generate (*ItG*, $p > 0.05$).

On the other hand, the participants categorised under the newly discovered user group (represented by Doris Piscopo – *complete newbies (55+)*) were processed separately, and a separate set of user models was created. Given the small number of participants available for this user group (i.e., a retired couple

in one UGC session and an individual participant in a second session), the generated models were not sufficiently significant in statistical terms. Nonetheless the models generated reflected the team's initial hypotheses with respect to this particular group of user: *Doris Piscopo won't be willing to use any online service*. For instance, perceived workload is consistently high even for the simplest enrolment processes.

No calibration was required for the other groups of users (represented by primary personas) since these were calibrated during previous studies.

In the meantime tasks **T3**, **T4**, **T5** as well as **P1** provided the team with a solid foundation to scope the new e-service (**T7**) while evaluating possible product use cases, or rather functions that the new e-service may afford to stakeholders, primarily consumers (**T8**). Following *Volere*'s templates, each product use case was discussed and scenarios (normal and alternative) were drafted (**T9**). All of the project personas were present in a visible location during these sessions (see Figure 9.8).

Figure 9.8: Persona posters were highly visible during meetings

Given the required agility for this project, a high-fidelity prototype was used to guide the team in the requirements development process. The author believes that projects for which high agility is required, product use case and scenario development are in themselves a design activity rather than just requirement elicitation and specification techniques. Product use cases are considered as a painting's outline while the different types of requirements act as colour guides for developers to follow. In similar cases, the requirements development process can be supported through the design of product use cases, which are in turn informed through the adoption of hi or low-fidelity prototypes. Prototyping was considered to be an important element in this project, since it informed the requirements development process itself. Prototypes, or parts thereof, might eventually find themselves in the final deliverable. Eye-tracking, card-sorting and other user-centred design tools are valuable tools to assess, monitor and verify usability requirements.

As part of the product use case specification exercise a card-sorting exercise was conducted to determine an initial take on the e-service's information architecture. A set of yellow post-it notes representing

functional and informational pages were provided together with a set of blank green ones. The project team was asked to use the green notes to create categories under which the other post-it notes would be placed. This would in turn translate into a consensus based hierarchical representation of information and service pages. Following a team effort and based on experience and expectations the team managed to come up with new informational and functional categories for the e-service, while existing ones were also removed or consolidated. The card-sorting exercise was crucial to reflect on specific requirements, answering questions such as: *How will a user look for advice?*, *Will consumers follow a hierarchical navigation pattern through FAQ drill-downs?*. This clarified potential user flows and an initial design started to emerge. Participants also had to agree on naming conventions for groups of concepts, and this informed the creation of menus, menu items and layouts (e.g., *News* vs *Alerts*). The idea of closed card sorting was used (with a predetermined set of categories) however the team opted to create new categories and concepts whenever necessary. This gave way to discussions about user-journeys, defining the path a user would take to complete a task (e.g., seek advice or file a complaint). Using a marker the team joined various post-it notes with arrows, indicating flow. By referring to the project personas throughout the discussion, the team constantly double checked whether any target user group was being excluded from the design process. The guiding principle was to design for the lowest-denominator in terms of skills. Whenever a design decision was taken the team referred back to the various personas and asked questions such as: *Would it be intuitive to Joe? Would it be too complex to Mary? How would Shanya react to this?*.

This was however based on subjective interpretation. Following Faily and Fléchais' recommendations [55], the team backed up project specific personas with as many facts as possible, also derived from first hand experience with real consumers who have previously interacted with the authority.

Figure 9.9: Card sorting exercise following an initial iteration of product use case designs

User journeys convey important information to the design team as they uncover flaws in assumed workflows and allow for an early rethink on the steps required to complete a primary task. This exercise should be conducted with both the users' goals as well as the authority's goals in mind.

Figure 9.10: Second stage of the card-sorting exercise was to determine user journeys

Following a systematic yet well scoped investigation of the authority's work, a set of product use cases was finally produced. Certain requirements started to emerge, and a note was taken (although no formal specifications were made at this point). The team then proceeded to test the current product use cases to determine their potential impact on users (**T10**, **T11**). Various advantages and disadvantages of introducing obligatory enrolment were discussed, both from a users' point of view but also from MCCAA's point of view. Two enrolment approaches were considered: (1) either via the national e-ID infrastructure or (2) via a custom-built and internal MCCAA user management facility. These scenarios placed different demands on users and the impact from each scenario was discussed. From a users' perspective the team considered and discussed both physical workload (i.e., national e-ID requires users to visit an enrolment office in person) as well as cognitive workload (i.e., custom MCCAA user accounts would require users to create yet another set of credentials). *Sentire* was adopted to simulate and visualise the impact that enrolment might have on users. Normal case scenarios were created for both enrolment options and these were then annotated (**T10**). Each use case was specified using scenarios, which were in turn specified as consecutive steps. Enrolment-specific steps were then annotated with measurements for each of the design factors identified in Chapter 4 (see Table 4.1).

Primary project personas were associated with their respective user group models (from previous UGC exercises), turning these into Calibrated Personas. User feedback was then simulated for all of the active actors and across the different enrolment scenarios (**T11**). *Sentire*'s CASE tool was used to automate this process.

The simulated feedback confirmed the team's concerns and also strengthened their conviction that

adopting an enrolment process based on the national e-ID for basic and infrequent transactions is an overkill, which would then result in major adoption issues (see Figure 9.11). The team was operating on the assumption that the number of active e-ID accounts in Malta is relatively low (i.e., around 10%). Using an MCCAA specific account would result in generally improved process completion rates (sign-ups) however this would have a negative impact on users represented by Mary Piscopo and Doris Piscopo, while discouraging many younger users represented by Shanya Borg (16–18 year old users). *Sentire* confirmed the team's gut feelings with objective data and decisions could be taken with higher confidence.

For each group of users *Sentire* reported the amount of perceived workload (histogram) as well as their willingness to enrol (pie-chart) for the e-service and complete the task online. The alternative methods to file complaints and seek advice would be more conventional, including phone, email, snail-mail or in person – or give up on the process altogether (see Figure 9.11). Following a couple of iterations (changing enrolment process parameters) it was decided to leave the e-service open across all use cases without the need for any compulsory enrolment. Nonetheless an optional MCCAA account was considered to be a reasonable offering for those consumers who would wish to track their complaints and other interactions. Although not necessarily precise, simulated feedback challenges designers to re-assess their assumptions and decisions following a systematic and repeatable process.

At this point a set of product use cases were formulated and tested for critical enrolment-related issues. Atomic requirements were specified for these use cases adopting a modified version of *Volere*'s requirements Snow Card template (**T12**). These low-level requirements covered various categories including functional, usability and humanity, look and feel and maintainability and support. As outlined in Chapter 5.4, testable and measurable fit-criteria can also be specified for aspects such as willingness to complete task and perceived workload.

Each requirement was assigned to specific product use case(s) or marked as global. This offers various levels of requirements granularity, with the product use cases being a high level view of what the system shall do, and atomic requirements specifying low level detail denoting how a system shall achieve such functionality. *Sentire*'s CASE tool offers a project visualisation map which allows (see Figure 5.24) the team to view requirements at various levels of granularity as well as their inter-relationship with higher level groupings (e.g., product use cases, business use cases and events). Other project elements are also displayed (and linked), including stakeholders, personas/actors, events, risks, use cases and requirements. This can serve as a visual impact assessment utility for regression testing following modifications to requirements or use cases.

Based on the current level of coverage and detail, a prototype started to emerge (**T13**) (see Figure 9.13). This high-fidelity prototype was tested at the University of Malta's Interaction Design Lab (**T14**) (see Figure 9.14). Using Tobii's eye-tracking and analysis studio (see Figure 9.14), a set of predetermined tasks (goals) were provided during which eye-gaze data was captured for a deeper assessment on findability, navigability and explicit pain-points (e.g., using heat-maps to uncover points of failure). Retrospective Think Aloud (RTA) sessions provided deeper and invaluable insights and knowledge on

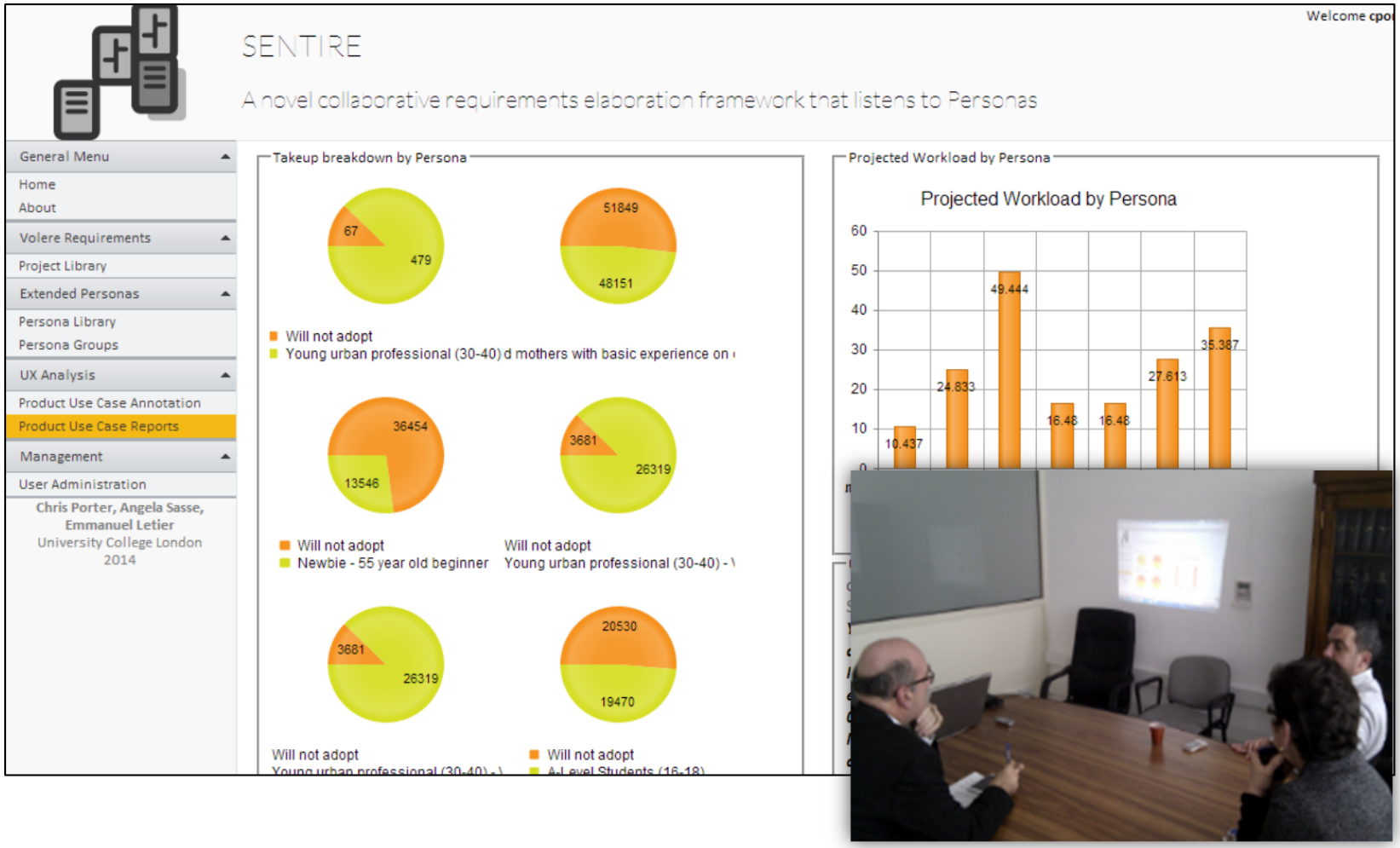


Figure 9.11: Simulated feedback for the different user groups represented by the various Calibrated Personas used throughout the design process. The feedback shown above was generated using *Sentire*'s CASE tool on the '*enrol with an MCCAA account*' product use case (Inset: *Sentire* workshop participants)

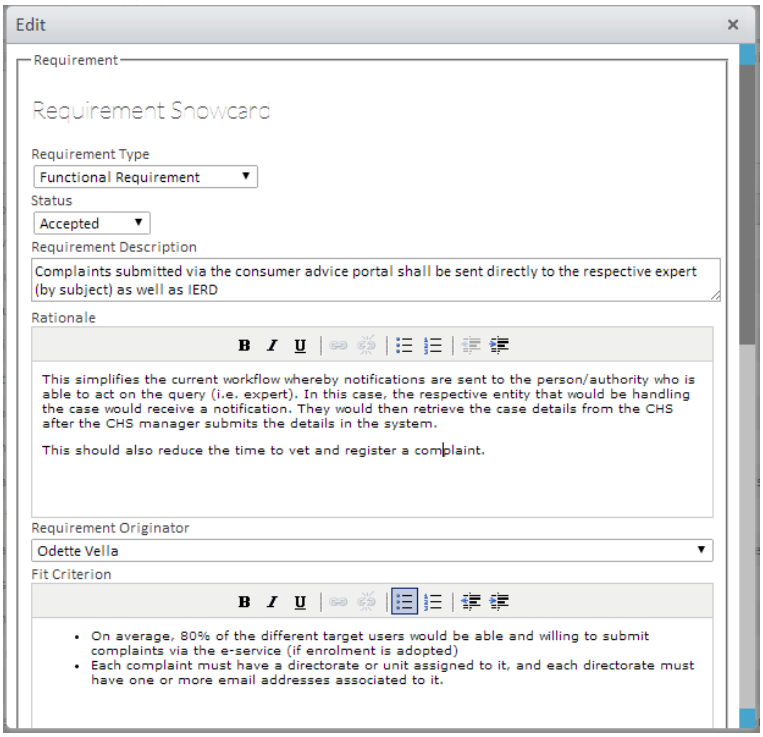


Figure 9.12: Atomic requirements were specified using a modified version of *Volere*'s requirements Snow Card template

what users expect, what they look for and the rationale behind their decisions. This data supplemented the eye-tracking data. A number of severe usability issues were uncovered during the first few sessions, and corrected prior to subsequent sessions. Following this iterative process the authority can then plan case officers' training, data migration and e-service release.

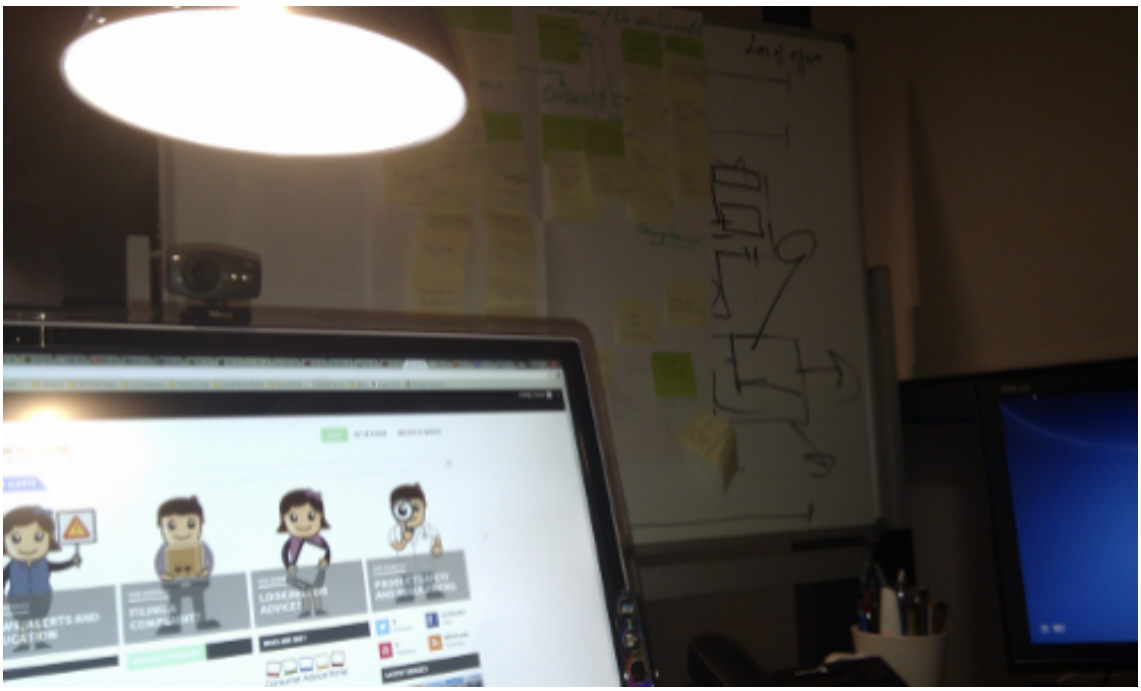


Figure 9.13: Prototyping the e-service based on the initial information architecture session

Figure 9.14: An eye-tracking session participant

9.4 Evaluation and Findings

The following is a discussion on the learning outcomes from the case study presented in this chapter.

9.4.1 Theoretical evaluation

9.4.1.1 Persona evolution through calibration

An interesting side-effect of the persona calibration exercise was the identification of new personas, stemming from behavioural information. Mary Piscopo, one of the project's primary personas, was built on assumptions, primary data, observations, opinions and experience provided by the stakeholders. When representatives of this project persona were calibrated, it was immediately clear that there were several distinctive behavioural clusters present within participants' data, who should have been theoretically similar in demographic terms. This informed the evolution of Mary Piscopo, but more importantly to the creation of a new persona Doris Piscopo which reflected the contrasting attitudes towards specific enrolment-related design factors. Although a similar result could have emerged through ethnographic methods, it is believe that the systematic, structured and objective nature of the calibration process highlighted marked differences between expected and actual behaviour across participants. When distinct clusters of behavioural information emerge from the data, then this would be indicative of a phenomenon that requires further investigation (e.g., out of seven participants, three generated commonly divergent results with respect to perceived workload and willingness to complete a task).

A small number of participants may not yield enough information to generate statistically significant results (i.e., models), however this data provides important insights that could drive the team to consider different angles of the same problem while being critical of their own actions and conclusions. This also promotes a structured and actively reflective design process. An unstructured and qualitative discussion

with participants might not have the same effect as the systematic, directed and repeatable user group calibration (UGC) process.

9.4.1.2 Partial user models

Sometimes user models can only partially capture user behaviour and reactions towards specific design factors. This is mainly because certain predictors used during the calibration process would not be statistically significant for a specific group of users (e.g., for a given user group and in a given context, an increasing amount of fields to fill may have an impact on workload but not on the willingness to complete the task). Other major causes of partial user models include instances when data collection is not extensive enough to perform good model fitting, and when the influence exerted by certain predictors would degrade the model's overall predictive power. Statistical tests would highlight these predictors which are in turn excluded from the final model to (1) strengthen the effect of the remaining predictors and (2) improve overall predictive power, albeit for specific design factors only. However the presence of partial models must be made explicitly clear to avoid misinterpretation. Figure 9.15 shows a conceptual representation of how this can be communicated to project teams.

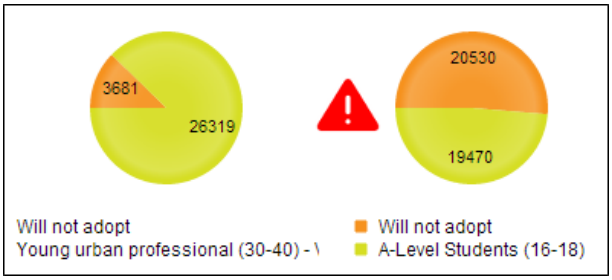


Figure 9.15: A conceptual representation of alerts shown when simulations are generated via partial user models

9.4.1.3 Indicative can be as good as precise

Simulated feedback provides a solid and objective grounding for discussion. From the experience gained throughout this project it was clear that the team was not interested in precise predictions but considered general trends to be sufficiently useful to inform decision making. The introduction of a new Calibrated Persona (Doris Piscopo) was a case in point wherein behavioural simulations were only indicative and based on weak statistical user models. This was mainly because data for the underlying user group (*complete newbies (55+)*) was based on readings from just two UGC sessions (with two and one participants respectively). Nonetheless this Calibrated Persona was still useful as it gave an objective indication of what kind of reaction to expect, in comparison to the other Calibrated Personas (even if not statistically accurate). Additional calibration sessions would strengthen the user model underlying this project persona. Nonetheless the current, albeit weak model still contributed towards a design decision, that of eliminating all enrolment processes from the e-service.

9.4.1.4 Contextual feedback is possible with in-context calibration

Persona group calibration can be carried out either in a lab environment or within the users' natural environment from where the e-service might eventually be used. In an earlier case study, calibration

for participants representing the *young urban professionals (30-40)* user group was conducted at each individual's workplace. In this particular case it was evident that decisions were highly influenced by contextual nuances.

With in-context calibration, participants could refer to their physical surroundings and work-conditions before submitting their feedback on perceived workload and task completion (e.g., "*how hard would it be to scan a utility bill over here [at home]?*"). Behaviours and attitudes might change in different contexts, however in-context calibration may mitigate this risk by creating behavioural models influenced by contextual nuances.

9.4.2 Evaluation of task completion predictions

A quantitative user evaluation was conducted to study users' attitudes towards this e-service, and their willingness to use it given different enrolment process options (scenarios). An online questionnaire was distributed via the social media and 119 full responses were received. Four basic enrolment scenarios were provided together with screenshots of enrolment page mock-ups. Respondents were asked to state whether they would be willing to use the e-service given each enrolment process. Basic demographic data was also requested in order to be able to group responses. The primary task presented was to *obtain information on consumer rights and file a complaint against a trader*. A free-phone number was provided as an alternative channel to complete the primary task while bypassing the e-service and the need to enrol. Tables 9.4, 9.5, 9.6 and 9.7 provide the annotations for the four scenarios presented in the questionnaire.

Table 9.4: CAP – enrolment process alternative 1

Scenario 1 – Low workload	
Design element	**Measurement**
Items to Generate	1
Items to Recall	1
Delays	False
Interruptions to daily routines	False
Type of Service	1

Table 9.5: CAP – enrolment process alternative 2

Scenario 2 – Low workload	
Design element	**Measurement**
Items to Generate	1
Items to Recall	5
Delays	None
Interruptions to daily routines	False
Type of Service	1

9.4.2.1 Participants

A total of 119 participants completed the questionnaire. Based on the provided demographic data (i.e., age range, level of education and IT proficiency) around 11% of participants could be categorised as *undergraduate students (18–25)*, 57% as *young urban professionals (30–40)* and around 14% as *confident*

Table 9.6: CAP – enrolment process alternative 3

Scenario 3 – Medium workload	
Design element	**Measurement**
Items to Generate	2
Items to Recall	9
Delays	None
Interruptions to daily routines	False
Type of Service	1

Table 9.7: CAP – enrolment process alternative 4

Scenario 4 – High workload	
Design element	**Measurement**
Items to Generate	3
Items to Recall	2
Delays	Major
Interruptions to daily routines	True
Type of Service	1

newbies (55+). There were no respondents from the *A-level students (16–18)* user group.

9.4.2.2 Results

Table 9.8: Predictions for perceived workload and the willingness to complete the task, generated for the four enrolment scenarios and user groups

	Simulated user feedback							
	Group A[1]		**Group B[2]**		**Group C[3]**		**Group D[4]**	
	PEW	*WCT*	*PEW*	*WCT*	*PEW*	*WCT*	*PEW*	*WCT*
Scenario 1	11%	73%	13%	60%	6%	94%	15%	77%
Scenario 2	2%	73%	23%	60%	15%	94%	15%	77%
Scenario 3	0%	73%	34%	53%	24%	88%	19%	77%
Scenario 4	64%	12%	61%	14%	67%	10%	64%	0%

[1] A-level students students (16–18)
[2] Undergraduate students (18–25)
[3] Young urban professionals (30–40)
[4] Confident newbies (55+)

Table 9.9: Actual user feedback (on *WCT*) for the four scenarios. Respondents are grouped based on demographic similarities to project personas

	Willingness to complete the task			
	Group A[1]	**Group B[2]**	**Group C[3]**	**Group D[4]**
Scenario 1	NA	58%	80%	64%
Scenario 2	NA	50%	70%	58%
Scenario 3	NA	33%	49%	52%
Scenario 4	NA	33%	30%	45%

[1] A-level students students (16–18)
[2] Undergraduate students (18–25)
[3] Young urban professionals (30–40)
[4] Confident newbies (55+)

Simulated user feedback generated through *Sentire* is provided in Table 9.8 while actual user feedback from questionnaire respondents (categorised under three groups that are demographically similar

to user groups underpinning CAP's Calibrated Personas) is provided in Table 9.9. No feedback was received from respondents who are demographically similar to the *A-level students (16–18) user group*.

Figures 9.16, 9.17 and 9.18 provide a visual perspective of the results obtained.

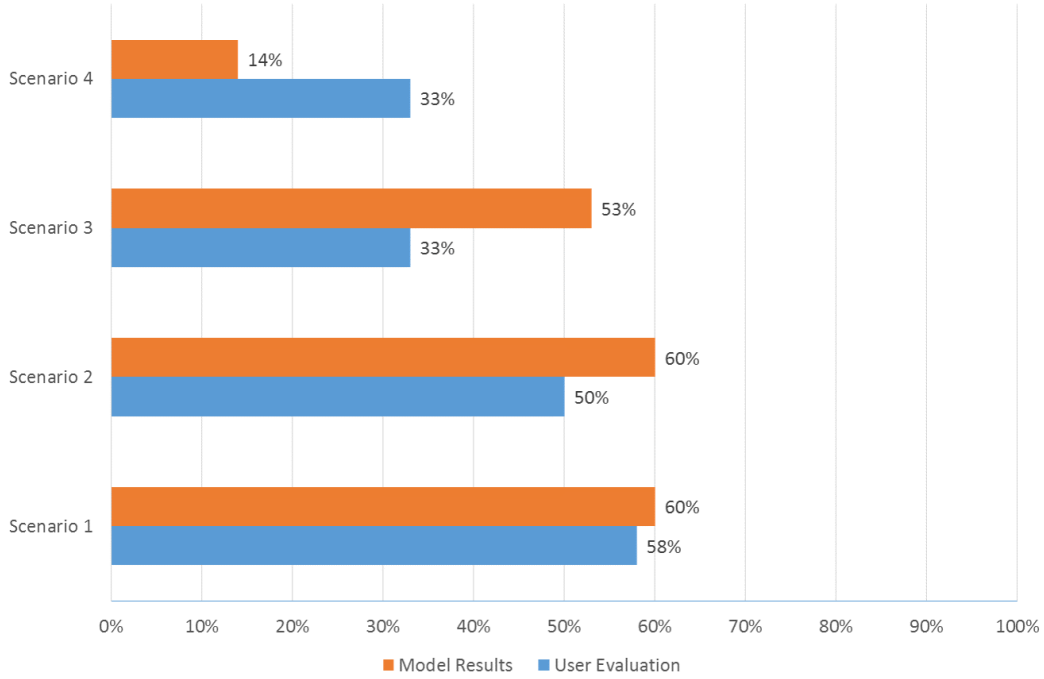


Figure 9.16: *Undergraduate students (18–25)* – feedback from actual users and predictions generated via *Sentire* for the willingness to adopt the e-service and complete the primary task online across the four enrolment scenarios

9.4.3 Practical modifications to the *Sentire* CASE tool

During the course of this case study a number of possible improvements to the CASE tool emerged:

- Stakeholders at MCCAA were too busy to actively engage using the online CASE tool. This required email-based updates and calls for feedback. All correspondence was stored in the tool's project repository (as HTML documents). *Multi-channel notifications* might encourage engagement. Stakeholders could be given the option to select their preferred way(s) by which to receive alerts on project updates (e.g., SMS notifications for critical updates and email notifications for less important changes).

- Meeting notes were taken in an online notepad made available in the project workspace, however following a number of meetings this got too large and impractical to refer to and update. A project *meeting management area* could be introduced, offering meeting recording facilities (i.e., meeting notes and meta-information) as well as a meeting-specific repository for artefacts collected during each meeting (e.g., photos, recordings, and documents).

- Use case annotation and feedback simulation should be made available on a *sidebar* within the use case formulation page. At the moment user feedback simulations are accessible from a separate

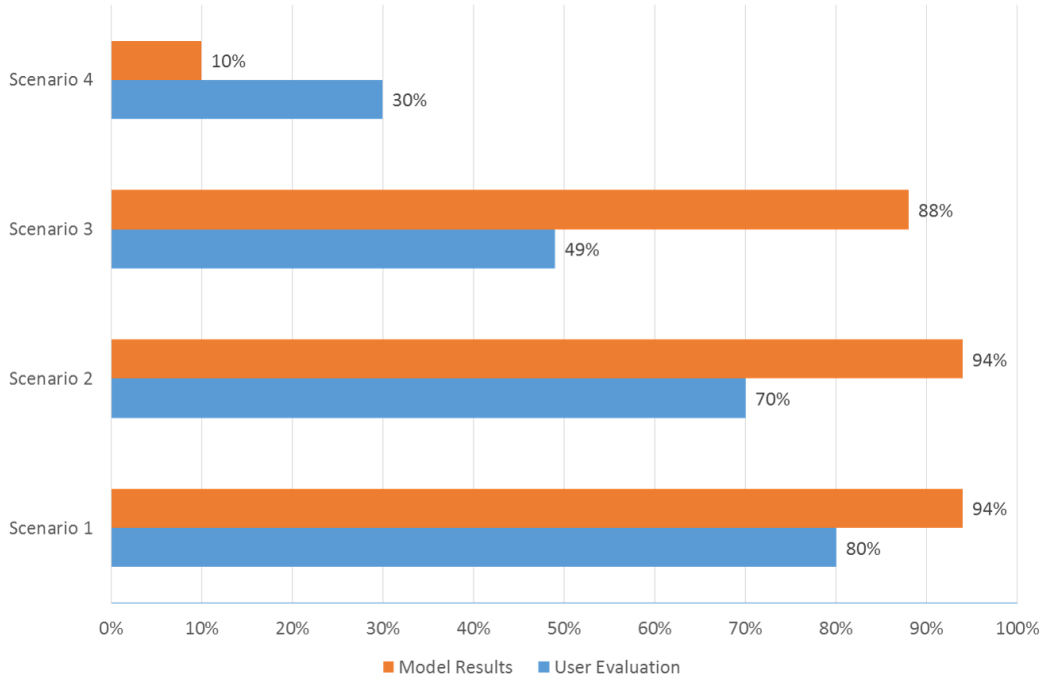


Figure 9.17: *Young urban professionals (30–40)* – feedback from actual users and predictions generated via *Sentire* for the willingness to adopt the e-service and complete the primary task online across the four enrolment scenarios

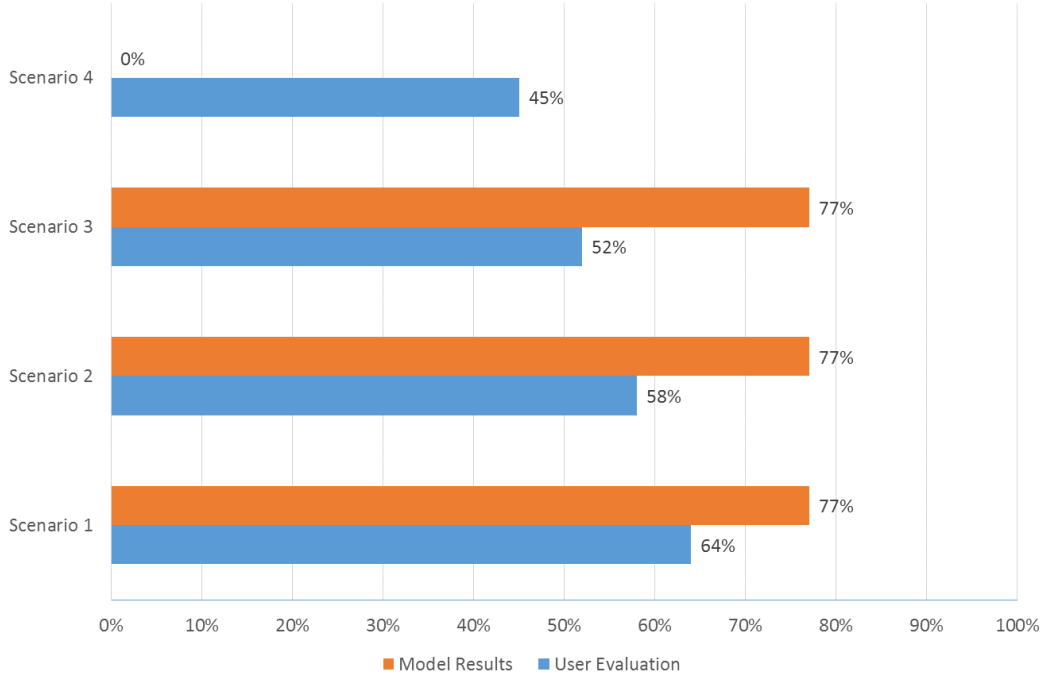


Figure 9.18: *Confident newbies (55+)* – feedback from actual users and predictions generated via *Sentire* for the willingness to adopt the e-service and complete the primary task online across the four enrolment scenarios

area within the CASE tool (i.e., UX-analytics dashboard), which could potentially disrupt the workflow.

- Realtime *messaging* facilities would offer a centralised space for teamwork, allowing stakeholders to collaborate through instant messaging (IM) and also asynchronously via email messages sent to project-specific email addresses (e.g., mccaa-p1@devbell.com) which would be automatically transformed into instant messages.

- Wireframes are currently created using external tools and images are then manually imported into the CASE tool. At the time of writing the author was communicating with a wire-framing application vendor to create an *integration API to offer in-app collaborative wireframing capabilities*.

9.5 Summary

Sentire has been adopted in a fully-fledged e-government service project with positive results. The project team described the experience as intuitive and non-threatening to the non-technical person. *Sentire*'s user feedback simulations together with the centrality of personas throughout the process (see Figures 9.8 and 9.11) provided a rich user-centred environment in which the different target user groups were virtually present from the first product use case design meeting. New and unexpected user groups were also discovered through the calibration process itself (see Section 9.4.1.1). Discussions were tailored around a set of project personas which were in turn grounded in empirical data. Furthermore, management felt more comfortable taking decisions based on quantitative and comparable views (i.e., simulations) of the impact that certain design decisions could have on users (see Section 9.4.1.3). Knowledge gained throughout the project (i.e., CAP), including terminology, stakeholders, project personas, statistical behavioural models and requirements have been stored in *Sentire*'s knowledge base for reuse in future projects, within MCCAA and potentially across government entities. Finally, a set of practitioner-oriented improvements to the online collaborative CASE tool were identified and discussed in Section 9.4.3.

Chapter 10

Conclusions

Developing studies suggest that the disconnect between software engineering and behavioural science persists, even in the private sector. Caputo et al. [28] observed that developers rarely expressed an understanding of user characteristics, the impact of security activities on the primary task (i.e., performance constraints), the users' context of use and the impact of security solutions on individuals and organisations. In their findings the authors reported the existence of a cultural gap between the two disciplines, wherein developers are reluctant to get guidance from usability specialists who are viewed as "*people who simply cannot code*" [28]. On the other hand Seffah et al. [147] noted that whenever a usability specialist is also a "*strong programmer and analyst*", it is more likely that user-centred techniques and methods would be accepted by software engineers and integrated within the development process.

"*Developers simply do not feel the pain of bad usability*" – Caputo et al. [28]

In their findings Caputo et al. [28] also noticed a "*strong belief*" that developers know best about a product's usability. Developers "*viewed themselves as users*", strengthening the belief that there is no need to reach out and get feedback from [real] potential users. This is not to say that developers do not care about the end user, but it shows the need for better alignment between the two mindsets. On one hand, software developers' thinking patterns are riddled with technical considerations to make systems more secure, maintainable, scalable and robust, while on the other hand a usability expert may overlook technical complexities in favour of seemingly-obvious user-centred solutions. Developers are *not* trained in usability either through academic programmes or through work-sponsored training [28]. The authors believe that one way to increase usable security is to increase training on usability and usable security for developers before and on-the-job, so as to introduce them to "*the complexities of human judgment and perception*" [28]. Seffah et al. [147] also identified the shortage of training on usability engineering as a major obstacle for the effective integration of software development practices and usability.

This thesis commenced by asking this question: *How can user behavioural modelling support the requirements process to encourage takeup in enrolment based and public facing e-government services?* Personas, as a user-centred design technique, was considered to be a primary candidate in supporting the requirements development process, however this technique requires specialist knowledge for proper development and use. Also, project teams may not have any UX specialists on board at the earliest stages, or at worst, at no point in time throughout the project. Sole reliance on hypothesis personas may

introduce threats to the process with decisions taken on highly subjective and unscientific user profiles (e.g., stereotypes). Challenges in e-service procurement and development processes (see Section 2.4) as well as the knowledged and culture gaps found within e-service project teams (see Section 2.5) backup these claims.

This thesis presents a user group calibration protocol through which predictive behavioural user models, targeting specific aspects of e-service design, can be generated for re-use across e-government projects. These user group models are embedded within traditional personas (sharing similar attributes) and the resulting construct is referred to as a Calibrated Persona. This design technique narrows the gap between traditional software engineering and user experience design practices. Calibrated Personas, as part of *Sentire*, situate simulated user feedback at the centre of a practitioner-centric requirements development process, turning user-centric rhetoric into a set of practical and systematic techniques for use by both practitioners (to build new services) as well as researchers (to build a knowledge base on user behaviour).

Equipping developers with tools such as Calibrated Personas, in a form that is readily available across the various stages of development would be highly beneficial. *Sentire* together with Calibrated Personas informs developers on the potential impact that critical design decisions may have on users, through unobtrusive yet actionable user experience insights. This does not replace usability experts or diminish the need for training, but providing such insights in a form that is usable by both developers and policy makers would reinforce the importance of behavioural science in software activities, especially within e-government projects. This information at the requirements development stage could help developers and policy makers avoid the most egregious kind of design flaws, which would otherwise only be discovered when users and usability experts are introduced to the project – generally at a later stage, if at all.

Feedback at the earliest stages of a system's lifecycle has the highest value in terms of risk-mitigation potential (e.g., costs for rectification) however this depreciates with time. Getting user feedback as early as possible at least on critical design decisions is important, however it is also time consuming. Furthermore, for effective and meaningful user feedback a prototype might be required, yet this may only be available at an advanced stage of the requirements development process. Statistical models fill this gap by providing objective, measurable and testable insights (simulated) on critical requirements and design decisions. This thesis has shown that such feedback is sensitive enough to changes in design parameters and can be used to inform the design team on possible risks arising from design decisions. Furthermore, introducing measurable experience-related fit-criteria improves the chances that non-functional experience requirements are honoured.

Table 10.1 reviews the sub-research questions posed in Chapter 1.2.

Table 10.1: Revisiting the sub-research questions

Code	Question	Section	Summary
SRQ1	What is the relationship between different levels of enrolment-specific friction and the adoption, security and cost of e-government services?	Section 2.3	High friction affects the user's experience negatively and therefore lowers adoption. This question was tackled through a review of existing literature, evidence from the field (lessons learnt within both the public and private sectors) and statistical reports. E-service take-up issues were linked directly to e-service over-protection, causing severe inconvenience to potential users.
SRQ2	Which enrolment-specific design factors contribute to friction?	Section 4.1.1	A qualitative study helped the author uncover a number of themes representing general enrolment-related frustrations amongst regular internet users. Thematic analysis was applied on a series of semi-structured interview transcripts during which the codebook's progress was monitored for saturation. A set of design factors were then drawn from these themes (grounded in empirical evidence) and operationalised for modelling purposes.
SRQ3	How can user behaviour be modelled to simulate reactions towards friction in new enrolment processes – given different e-service contexts and across a rich diversity of user groups?	Sections 4.2 and 4.3	A number of statistical modelling techniques were explored. These can be used to explain users' behaviour given a series of design parameters as well as predict their reactions when these parameters are modified. Users' attitudes towards enrolment design factors are believed to vary from one user group to another as well as from one context to the next, and user feedback simulation is meaningful only if these nuances are captured. For this purpose the user group calibration (UGC) process was presented, which is used to build contextual user models based on observed behaviour and reported attitudes.
SRQ4	How can UX simulations help non-HCI practitioners design better e-services?	Sections 5.1, 5.2 and 5.3 as well as Chapters 6 and 9	User group specific predictions (based on annotated product use cases and Calibrated Personas) allow the project team to understand the impact that each design alternative might have on both the end users (perceived workload) as well as on the e-service itself (adoption), in a pragmatic, granular and actionable way. Also such simulations allow project teams to maintain a positive lived experience (for users) while delivering the required level of identity assurance (for product owners). *Sentire* adopts and extends the *Volere* process (an industry strength requirements development process and specification template) with Calibrated Personas and simulated user feedback. User experience simulations (or UX-analytics) were pivotal to the introduction of *testable experience requirements* which are specified using *measurable fit-criteria* (e.g., 80% of *young urban professionals* shall be willing to use the service with a perceived level of workload of 20% or less). While designing product use cases, the project team should determine whether their designs respect the specified UX-requirements, by cross-checking simulated user feedback with the respective requirements' fit-criteria (using *Sentire*'s CASE tool). This technique can help highlight egregious user experience issues at the requirements and design stage that may arise from assumptions about potential user groups or due to a lack of knowledge of users' attitudes and perceptions. *Sentire* was evaluated and refined through two major interventions (see Chapters 6 and 9).

10.1 Contributions

10.1.1 Calibrated Persona – a technique to model and predict user reactions to and perceptions of e-service enrolment processes

*Code: **C1**; Type: **Main**; Contributing to: **HCI(Sec) research***

Measuring user experience (UX) is not a trivial task and specifying testable and verifiable experience-related requirements is even more challenging. Unless UX activities and requirements are not specified together with explicit verification mechanisms, it is more likely that developers would shortchange user experience considerations to (1) qualify as the lowest bidder (before development) and (2) cut costs (during development). This thesis presents a technique to understand and model users' reactions to and perceptions of enrolment processes within different types of e-government services and contexts of use (see Chapter 4). This provides a first step towards the development of predictive HCI modelling techniques for practitioner-centric decision support systems, reflecting user behaviour with respect to the service under consideration and its context of use. The enrolment process is considered to be a major hurdle for e-service adoption, and for each group of potential users this technique explains their willingness to enrol and complete a primary task online (as opposed to traditional service channels) as well as their levels of perceived workload (see Sections 4.2.3 and 4.2.4). Figure 10.1 outlines the main aspects of this contribution. Models are built following a standardised test, referred to as the user group calibration (UGC) exercise (see Sections 4.2 and 4.3). Calibration requires participants, representing a specific user group, to complete a set of fictitious tasks which were designed following an empirical study on e-service enrolment processes and user attitudes (see Section 4.1). During calibration, feedback and measurements are recorded for each task. This data is processed to generate two statistical models for each user group (i.e., perceived workload and willingness to complete the task online), which models can then be used to explain why users behave the way they do in an existing scenario and also to predict how users might react towards a new enrolment-centric e-service. Contextual calibration adds to the validity of the resulting models and a discussion on this is provided in Section 4.3.3.

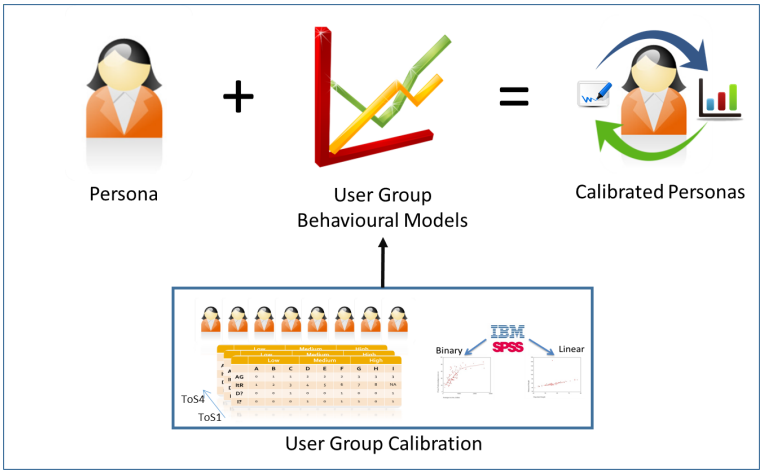


Figure 10.1: Calibrated Personas – embedding user behavioural models (generated via UGC sessions) within project personas. This adds a 'voice' to traditional personas via simulated user feedback.

This thesis embeds the resulting models within the persona construct – referred to as Calibrated Personas – for use within a user-centric requirements process. By introducing Calibrated Personas into the requirements and design process, the project team would be able to assess design alternatives through simulated user feedback (see Section 10.1.2). This technique can act as an early warning system informing the project team on the impact that specific design alternatives may have on the various target user groups. This also provides more guidance by which project teams could comply with the required levels of identity assurance for a specific e-service while respecting end users' limitations, perceptions and attitudes.

10.1.2 *Sentire* – a requirements framework based on simulated user feedback

Code: ***C2****; Type:* ***Main****; Contributing to:* ***RE/HCI research and practice***

This thesis presents *Sentire* – a requirements framework developed to assist project teams elicit and specify user-centred requirements for public facing and enrolment-based e-government services (see Chapter 5). As outlined in Figure 10.2, *Sentire* extends the *Volere* requirements process and introduces Calibrated Personas and simulated user feedback as a central activity within the quality assurance stage (i.e., *Volere*'s quality gateway) (see Section 5.3). By simulating user feedback the project team can ensure that any proposed use case respects the various user groups' limitations, perceptions and attitudes. Furthermore, *testable fit-criteria* for user experience requirements can be specified for product use cases, especially if the design team opts to stop short from specifying low-level product use case designs (i.e., step by step). User feedback simulations would then be computed on proposed designs produced by the entity that actually develops the e-service (i.e., internal development team or external contractor) as part of an iterative design process. Embedding measurable and verifiable UX fit-criteria within the requirements development process adds a new dimension to this discipline, by reducing the level of subjectivity in specifying and evaluating non-functional requirements, in this case related to enrolment processes. Based on empirical evidence and following a number of case studies it is believed that *Sentire* can help project teams diminish e-service uptake hurdles while complying with identity assurance level requirements.

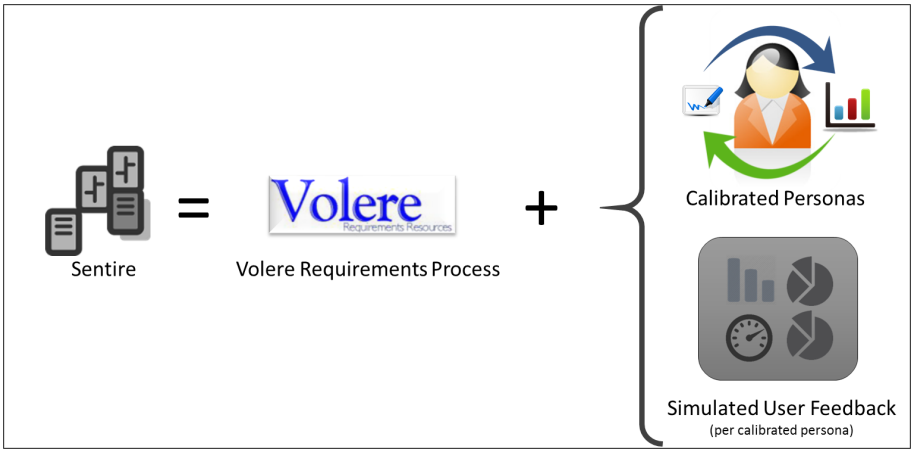


Figure 10.2: *Sentire* – embedding Calibrated Personas within *Volere*'s Quality Gateway. This introduces user feedback simulations to the requirements development process.

Calibrated Personas can be extended to other areas of requirements and design (as discussed in Section 10.3.1), giving the user a say, albeit through simulations, even before prototypes have been developed. Embedding this technique within the requirement development workflow ensures a truly user-centric philosophy based on objective UX simulations within a systematic process of discovery. This also mitigates the risk of unexpected late-stage rework on critical design aspects wherein risks of budget-overruns are significantly higher.

Sentire was applied in a series of real-world case studies and a collaborative CASE tool was also developed to facilitate *Sentire* activities.

10.1.3 Collaborative tool support for *Sentire*

*Code: **C3**; Type: **Main**; Contributing to: **RE practice***

This thesis also presents a computer-aided software engineering (CASE) tool for *Sentire* that supports and manages the entire requirements development process through an online collaborative workspace, currently hosted at *http://devbell.com* (see Section 5.7). Project team members should be an integral part of the process (i.e., collaborators) rather than mere sources of information.

Based on the premise of delivering user-centric e-services, the framework and respective toolset are practitioner-centric themselves, considering the practitioners' limitations and impact on their own lived experience (ULX).

User feedback simulations are automated, eliminating the need for laborious calculations while visualising the impact of design decisions on users. Process standardisation and simplification, efficiency, improved dependency-mapping, traceability, reusability and consistent reporting are some of the benefits this CASE tool provides. *Sentire*'s CASE tool offers exportable *Volere* compliant requirements specifications, as well as an artefacts library for project teams to collect and organise use case and project specific artefacts (e.g., scanned forms, photographs, recordings and emails).

This is a primary contribution to practice, based on the theory developed throughout this thesis, extending industry-strength practices and refined through a number of real-world studies.

10.1.4 Other contributions

10.1.4.1 A study on user attitudes towards enrolment processes

*Code: **C4**; Type: **Minor**; Contributing to: **HCI(Sec) research and practice***

This thesis presents a set of enrolment-process related design considerations. These considerations were derived from an empirical study planned and executed to learn more about the impact of enrolment processes on the users' experience (UX and ULX) and how people relate and react to this aspect of e-service design. The study revolved around reported experiences of, and attitudes towards enrolment processes from 20 participants (see Section 4.1). Coding was continuously monitored for codebook saturation across multiple rounds of interviews, and Katy Charmaz's [31] coding recommendations were also followed. This study resulted in a set of enrolment-specific design factors that are considered to be primary contributors to friction (i.e., increase perceived workload and the probability of end users abandoning the e-service in favour of alternative service channels). The impact level of these factors

(on end users) may also be influenced by the type of service under consideration and by the number of competing providers for that service (this is generally not applicable in an e-government context). These factors are discussed in Section 4.1.1 and outlined in Table 4.1.

Also, Sections 7.3.3 and 8.6 presents a set of insights and recommendations on the design and development of enrolment-based e-services for two groups of users (undergraduate students and digital natives – people who have grown up with and are accustomed to digital technology). These insights are collateral contributions arising from the second and third case studies, presented in Chapters 7 and 8 respectively.

10.1.4.2 Assessment of NASA-TLX's sensitivity for enrolment-specific perceived workload on younger audiences

*Code: **C5**; Type: **Minor**; Contributing to: **HCI(Sec) research***

This thesis assessed the applicability, understandability and sensitivity of NASA-TLX with younger participants, particularly 16 to 18 year old A-level students (see Chapter 8). This group of users are referred to as digital natives, people who have grown up with and are accustomed to digital technology. The context for this investigation was a new national e-service specifically built to allow students to register for their 2013 A-level examinations.

The study consisted of two main phases: (1) an online questionnaire and (2) a series of follow-up sessions in small groups. The questionnaire was essential to synthesise students' experiences with the e-service which also included an ex post facto NASA-TLX evaluation using a digitised version of the pen-and-paper process. This led to the second phase of the study in which a systematic assessment was conducted on the sensitivity, understandability and general applicability of NASA-TLX within this specific domain and with this group of users. Minor modifications were proposed to this technique (see Section 8.6.2) which could improve its validity in similar scenarios especially when used with younger participants. Section 8.6 provides an evaluation of results together with a set of learning outcomes.

10.1.4.3 Testable fit-criteria for experience related (non-functional) requirements

*Code: **C6**; Type: **Minor**; Contributing to: **RE practice***

Sentire adopts Calibrated Personas as a technique that allows for the specification of measurable (and thus testable) UX-oriented fit-criteria. Leading industry experts (see Section 10.4) have suggested that Calibrated Personas could be adopted as a standard industry technique (i.e., pattern) to specify non-functional requirements' fit-criteria (e.g., "*65% of Shanya Borg's [Calibrated Persona] shall be willing to complete the primary task online*" or "*George Smith's [Calibrated Persona] level of perceived workload shall be lower than 30% for all enrolment-related scenarios*").

Although enrolment-specific design issues were central to this work, it is believed that this pattern can be extended for use in other critical areas of e-service design (e.g., "*Shanya Borg [Calibrated Persona] shall not feel that her privacy is threatened when submitting a police report online (level of perceived threat must be less than 30%)*"). See Section 10.3.1 for a discussion on potential future work.

10.1.4.4 User-group knowledge base for reuse across government projects

*Code: **C7**; Type: **Minor**; Contributing to: **RE practice***

This thesis proposes a technique by which user behaviour could be captured, analysed, modelled and reused to simulate user feedback (see Chapter 4). This presents an opportunity to build an inventory of behavioural models for the different groups of users generally found within a given geographic area for use during e-service development at a regional or national government level. This knowledge base can evolve over time through the introduction of new models, re-calibration of existing ones and retirement of 'expired' models (i.e., UGC participants used to construct a set of models for a specific user group may have been replaced by a new generation of people carrying different attitudes and behavioural traits).

Throughout this thesis a number of user groups were calibrated, including *young urban professionals (30–40 years old)*, *digital natives (students sitting for their A-level exams)*, *55+ technology newbies* and *undergraduate students*. User group models created during the first case study were stored within the knowledge base and in turn reused with (re-embedded in) project-specific personas (i.e., Calibrated Personas) during subsequent studies, whenever applicable. Applicability is determined by the user group's demographics and context within which calibration took place. If a new e-service targets a user group for which no models exist (or models exist but for another context of use) then the project team should conduct a user group calibration (UGC) exercise for that group of users within the desired context of use. Resulting models are then added to the knowledge base for use in future projects.

Based on this mechanism, this thesis proposes the creation of national and regional user-group knowledge bases for use across government projects and entities. Calibrated Personas are central to *Sentire*, however the underlying user group models and associated meta-data (e.g., NASA-TLX workload subscale weightings) are agnostic to *Sentire* and *Sentire*'s CASE tool. These can be used by project teams adopting any other requirements development processes (see Section 4.4). Section 10.3.6 discusses potential operational models for such knowledge bases, including commercial opportunities.

10.2 Critical Reflection

This section outlines a number of personal observations regarding the research process itself and its deliverables. It primarily answers the following questions: *If someone else takes this research further, what should they look out for? What aspects need to be revisited? And how should these be tackled?*

10.2.1 Empirical validation limitations

10.2.1.1 Evaluating quantitative results in case studies

The author would have preferred to evaluate both HRIU (see Chapter 6) and CAP (see Chapter 9) in a live environment (i.e., post launch) to determine actual adoption rates. However this was not possible for two reasons: (1) lack of access to actual end users (e.g., CAP will be launched without compulsory enrolment – thus no or little adoption information will be available), and (2) no involvement in the development and launch of HRIU (i.e., collaboration on HRIU ended as soon as the tender document was published, and the author could not monitor live usage statistics). Nonetheless other techniques to elicit user feedback were adopted, albeit from potential users, through focus groups, interviews and

questionnaires. However it would be highly insightful to gain access to actual and accurate user and usage data following the launch of an e-service (e.g., using tools such as *Google Analytics* or *UserStats* which provide custom analytics on e-service usage streams). Through engagement analysis (e.g., flow or funnel reports) backed with demographic data (e.g., through ad-network supported analytics providers) one could determine the drop-off rates at enrolment for the different user groups. A/B testing could also provide interesting results for comparative exercises whereby multiple enrolment processes are rolled out in parallel and tested in a live environment. Users would be presented with any of the enrolment process variants based on specific experimental parameters (e.g., IP geolocation or demographics) and performance outcomes are then compared.

10.2.1.2 Evaluating qualitative studies

A final critique is on the validation of qualitative analysis processes (e.g., thematic analysis to uncover the design factors presented in Section 4.1). Ideally such processes are delegated to another two (or more) experienced qualitative researchers. This would then be followed up with a discussion on findings to produce a consensus-based set of themes. This technique would cross-validate the researchers' findings while ensuring the quality of the process itself.

10.2.2 Limitations of the calibration process

The user group calibration (UGC) process must be broad enough to cover edge cases without making the calibration process unbearably (and unnecessarily) lengthy. In scenarios for which use case annotations go beyond the values used during calibration (i.e., exceed lower and upper bounds used in calibration tasks – as shown in Table 4.8) the project team should be notified that the underlying models may not produce reliable results, while denoting the boundaries used during calibration. These notifications could be similar to the ones used for partial models (see Section 9.4.1.2).

10 to 15 representative participants for the calibration process were deemed to be sufficient to construct indicative behavioural models (see Section 4.3.1 for a discussion on model saturation), however if sub-groups are identified during calibration more participants might be required to compensate for the split in the underlying datasets. This thesis acknowledges the complexity of the domain whereby user perceptions, attitudes and behaviours are affected by numerous factors, including the context of use, user attitudes, skills, aptitudes, experiences, culture, choice of technology and interface design amongst others. It would be presumptuous to assert that calibration captures all of these aspects, however by adopting a divide and conquer approach this complexity can be abstracted into smaller and more manageable problems (i.e., scoped user behavioural models).

This work is inter-disciplinary in nature and input is required from the HCI, behavioural science and software engineering domains. Support from statistical scientists is also highly recommended. A significant amount of time was spent gaining the necessary skills and insights to build the theory presented in this thesis. In Section 10.3.3 the author suggests possible research directions that can help improve the modelling and prediction aspects of Calibrated Personas and *Sentire* in general.

10.2.3 Final remark on *Sentire*

The goal of *Sentire* is not to replace HCI-specialists and associated user centred design techniques, but it is a way forward to help bridge the gap between software engineering, behavioural sciences, HCI, policy making and general practice. *Sentire* helps to:

1. Flag potentially critical design issues at the requirements stage, before e-service prototypes and HCI specialists are introduced. This also affords low-cost and quick design iterations without the need to invite actual users for evaluation purposes following each design decision.

2. Define testable experience-related fit-criteria when non-prescriptive requirements documents are produced (e.g., calls for tender). The winning bidder must then produce low-level e-service designs (e.g., through use cases and scenarios) that meet the fit-criteria specified within the provided requirements. In turn, these designs are tested and signed off by the contracting authority during the quality assurance stage through the adoption of user feedback simulations.

3. Provide practitioners with an integrated framework whereby users are placed at the centre of the design process while encouraging user-knowledge accumulation and re-use.

10.3 Future Work

10.3.1 Adoption of *Sentire* for other critical e-service design aspects

As well as enrolment, the current user models may be well suited for other aspects of identity management, including authentication (single or *n*-factor) and account recovery processes. Enrolment processes that result in the provision of hard-tokens (e.g., one-time pin (OTP) device) may cause delays and possibly interruptions. These factors are already handled by the enrolment-specific behavioural models built for this thesis. On the other hand during authentication, the use of an OTP token may cause a minor delay, which again is handled by the current enrolment-specific user models. Automated account recovery processes can be considered to be a critical aspect of e-service design, which can also impact the users' experience in a negative way (e.g., complex recovery processes leading to task abandonment). Also, a badly designed (or non-existent) account recovery process adds to the cost of handling user requests manually (i.e., help-desk visits/calls). The enrolment-related design factors identified for this thesis may also be applicable in this scenario, nonetheless further empirical evidence is required to confirm this hypothesis.

Another possible research avenue could be to extend *Sentire* for use beyond enrolment-oriented e-government services, and adapted to model and simulate user feedback for other critical areas of e-service design (e.g., trust, quality and privacy perceptions as well as pricing strategies). Researchers would need to devise new calibration processes following the method outlined in Section 4.1. This may result in multiple or multi-purpose calibration processes, producing a variety of behavioural models for different user groups.

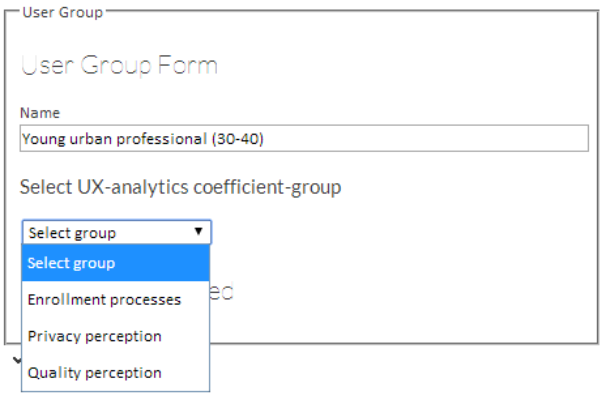


Figure 10.3: Calibrated Personas can be adopted to model user behaviour for other critical e-service design aspects

10.3.2 Simplifying simulated feedback

It might be sufficient to adopt and present simpler heuristics for user experience feedback rather than attempting to obtain significantly precise predictions. Such figures may result to be counter-productive for practitioners, who may only require a confirmation to move on with a decision, rather than an exact prediction. This resulted from the final case study (see Chapter 9) wherein the author observed that the project team wasn't interested in exact workload or take-up predictions but simply considered the dominant colour in the simulation report charts to guide decision making. This would be the result of further abstraction of user feedback simulations to provide coarsely grained recommendations rather than percentage based predictions. Additional information could in reality be counterproductive, since it creates the perception of more work which might eventually lead to the circumvention of the supporting tools themselves. Figure 10.4 shows the current feedback reporting mechanism together with an alternative yet simpler reporting technique. The hypothesis is that the simpler option (right) may be equally (or more) effective than its current counterpart (left), however there is no empirical evidence to support this claim.

10.3.3 Collaboration with statistical sciences research groups

A discussion with Gordon Ross (from UCL's Department of Statistical Sciences) exposed a number of potential avenues for future work and improvement. The following list provides a summary of potential paths of exploration:

- **On identifying user groups for calibration**: Analysing data with potentially overlapping categories (i.e., demographics) to identify clusters (user groups) is intrinsically hard. However, sophisticated techniques exist to automate this process (rather than relying solely on heuristics). Formal clustering techniques can be used to learn about the number and composition of clusters (user groups) directly from a large dataset, without splitting the data collection process and resulting datasets in advance based on assumptions about user groups. Such techniques could also be used to determine whether an existing calibrated user group has been previously overgeneralised (i.e., check for the presence of distinct clusters within the user group's data). The *Cluster Analysis* pack-

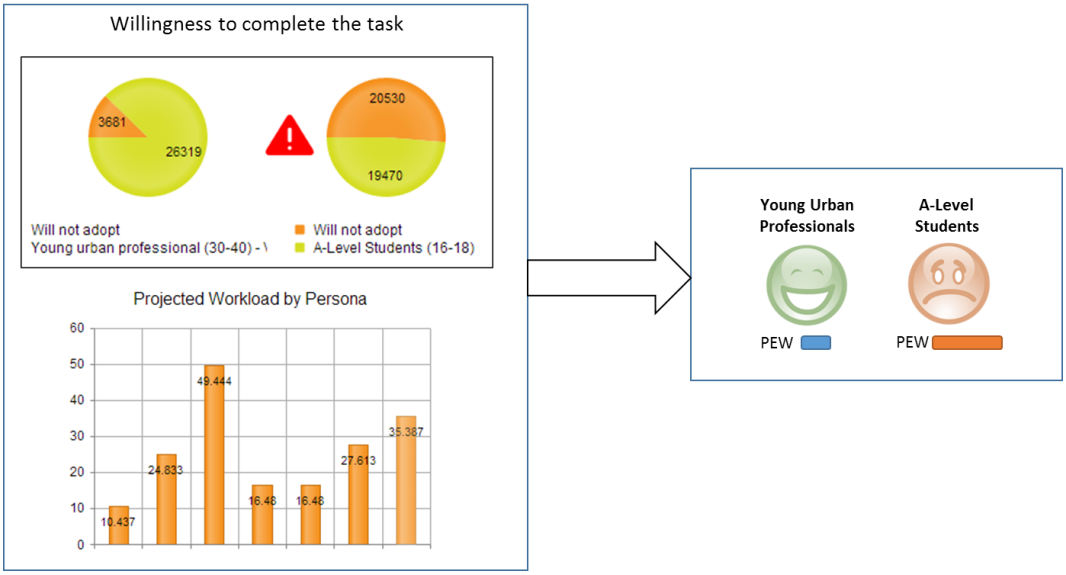


Figure 10.4: Providing coarser grained results (right) might be more effective for practitioners (as observed during the final case study), whereby the original reporting format (left) might create a perception of more work – thus discouraging adoption or proper use of the tool (*Note: PEW = Perceived Enrolment Workload*)

age in R[1] may be considered for this type of analysis. This technique can help to systematically confirm or reject hypotheses about the existence of distinct user groups within a specific context. In turn this adds validity to project specific Calibrated Personas and resulting predictions.

- **On sampling for calibration**: Sampling presents numerous challenges that may introduce threats to validity (e.g., selection bias and randomisation). This problem is compounded even further by the fact that user groups are currently specified manually and generally in advance (i.e., user group clustering). This may present a severe risk whereby researchers generalise for a whole cluster (user group) based on models built from a biased sample. Models created for the *young urban professionals (30–40)* user group based on City workers may not be generalisable to Scottish school teachers, even though they are both considered to be young professionals. For this reason this thesis argues that user groups should be built for (and are valid to) specific regions, especially in the case of larger countries (e.g., London should manage its own user group library, distinct from the one built for and used in North Yorkshire). On the other hand, smaller nations (e.g., Luxembourg) could build and maintain one national user group library (see Section 10.1.4.4). Future work should also consider techniques for sampling bias corrections.

- **On data collection for calibration**: Alternative data collection processes (for calibration) should be explored (e.g., including more variants of the current nine tasks) to determine whether the quality of data collected and eventually the models produced could be improved even further. Nine calibration tasks may not be exhaustive enough to cover all the possible conditions (treatments) necessary to produce enough observations for robust modelling. However data collection is expensive and time consuming, and thus a balance between the number (and configuration) of calibration

[1]Quick-R: Cluster Analysis, http://www.statmethods.net/advstats/cluster.html, (accessed December 2014)

tasks, the number of participants and model accuracy is required.

- **On black-box machine learning techniques for prediction models**: The adoption of black-box methods may help to construct a better understanding of the underlying data while mitigating risks arising from human error and bias during the modelling stage. Machine learning techniques can handle variable selection, interaction terms and non-linearities automatically. R provides several libraries for sophisticated black-box methods that have been engineered to work well on most problems. Shifting to machine learning techniques could have an impact on the user group calibration process as well as on the design of the Calibrated Persona construct itself (i.e., in theory but also in terms of how this could be represented and used in *Sentire*'s CASE tool to generate immediate user feedback simulations as part of the requirements development process). Researchers adopting regression-based techniques are encouraged to compare and contrast their results with those produced through machine learning methods. This would help determine whether there are any substantial variances in outcomes, highlighting potential issues with the modelling strategy adopted.

- **On user group maintenance and recurring calibration**: This thesis suggests that user group models should be maintained through re-calibration at predefined time-intervals. However tests must be made to ensure that new calibration participants for a specific user group are still representative of the original group, mainly because new generation of users may differ in terms of expectations, perceptions and general behavioural patterns. The multiple comparison problem (multiplicity) should be considered when proposing user group maintenance protocols.

- **On out-of-sample analysis**: To evaluate the constructed models' accuracy one may adopt out-of-sample analysis, whereby the models generated are evaluated by comparing forecasts to actual real-world outcomes. It was generally difficult to obtain actual e-service usage statistics to evaluate the models' prediction accuracy (i.e., WCT). In this case pseudo out-of-sample analysis could be adopted to determine a model's out-of-sample accuracy by assessing model forecasts against a set of historical observations (used solely for cross-validation and not to construct the model itself). One must be careful when interpreting R^2 as a measure of model accuracy especially when in-sample analysis is used (i.e., risking over-fitting). This risk can be mitigated with an out-of-sample procedure.

- **On standardised predictors**: Standardised predictors could help project teams understand the impact (i.e., importance) of each predictor on the outcome (dependent variable) irrespective of the units of measurement used for each predictor (e.g., one unit of age vs one unit of length).

10.3.4 Calibrating non-technical users

More work is required to study the effectiveness of the calibration process with non-technical participants (e.g., people with low fluency in the use of technology). This was particularly challenging when dealing with older participants who were not able to complete the calibration tasks on their own (see Section 9.3

– step P5). This was mainly due to a low level of self-confidence, especially when using new systems. Can calibration be conducted without having to go through a number of enrolment tasks? Would a questionnaire-styled process be as effective as having to perform the tasks themselves?

Empirical evidence is required to answer these questions, while tackling other aspects such as computer literacy as well as confidence.

10.3.5 Remote and large-scale calibration

During the first intervention (see Chapter 6) a one-to-one in-context calibration process was opted for. This can be an expensive and lengthy process especially when it involves a large number of calibration participants, potentially requiring the services of additional facilitators. Studies on how to conduct an in-context user group calibration process at scale without affecting response validity would be extremely beneficial. This could in turn enable more frequent calibration exercises at a fraction of the cost (compared to on-to-one in-context calibration). The use of social media or human-based computation services[2] can help to reach more calibration participants, which would in turn provide more data with which to generate and maintain user behavioural models. However, empirical data on remote calibration (and especially on the use of human-based computation services) is required, specifically on issues related to data quality, response validity (due to lack of direct support) and contextual validity.

Knowledge on statistical modelling techniques is required to construct statistically significant behavioural models following a calibration exercise. By automating the calibration and model generation processes one would be abstracting even more from the underlying intricacies, allowing non-technical stakeholders to focus on the core requirements development activities.

10.3.6 Calibrated user group marketplace and cold start issues

During discussions with peers it has been argued that the expense involved to calibrate personas could be avoided and resources should be used to test e-services directly with end users. Calibrated Personas (as part of *Sentire*) are not intended to replace user testing but their purpose is to enable knowledge accumulation (about users) through a systematic process of discovery, which knowledge could then be reused in future projects. By adopting *Sentire*, project teams make better informed decisions at the earliest stages of an e-services' lifecycle through simulated UX-analytics. The value of the techniques presented in this thesis increase as more user-specific knowledge is accumulated and maintained for use in future projects.

The calibration techniques proposed in this thesis (see Section 4.2) can be taken up by (or outsourced to) commercial and research entities who would in turn participate in and contribute to a *user-group model marketplace*. Project teams could then import (or purchase) the required user group models (e.g., *models for London-based teachers in a school environment (20-35 years old)*) for use within specific projects (e.g., e-learning platform for London-based schools). These models are then associated with project specific personas – thus turning them into Calibrated Personas. Models can also be maintained through this marketplace approach. This marketplace should address or at least acknowledge regional differences, clearly indicating the origin of the model's underlying participants (e.g., geographic

[2] Such as Amazon's Mechanical Turk, https://requester.mturk.com/developer, (accessed June 2014)

and cultural context) as well as the environment in which calibration took place (e.g., at the participant's home, in transit or at the office). This meta-data is important, helping the project team pick better-fitting user-group models to a set of project-specific personas in order to generate contextually robust simulations. This marketplace is a commercial opportunity for entities involved in model-provision, however further assessment is required to explore alternative modus-operandi, including an open-source ecosystem as well as a freemium model[3].

Sentire depends on calibrated user groups to generate simulated user feedback and for this reason early adopters of the framework (and CASE tool) may face *cold start* issues. This would require the project team to factor in estimation of effort required for user group calibration as part of their project plans. This may not be an attractive proposition for one-off projects. *Sentire*'s value increases over time and with subsequent projects, whereby user group models would then be immediately available for reuse, thus avoiding the upfront costs of calibration (although maintenance may still be required). However by encouraging the market-driven user group modelling approach (with user group models made available by, or outsourced to third parties), the *cold start* problem may be further diminished, potentially also for first-time and one-time users. Nonetheless, the author argues that in the e-government domain, a user group behavioural knowledge base would be a justifiable and comparatively minimal investment. This is mainly because knowledge and process reuse is critical given the amount of government bodies investing time and money to build better public facing e-government services, within specific regional and national contexts and serving subsets of the same population.

10.4 Expert Evaluation

Sentire was presented to James and Suzanne Robertson, principals at The Atlantic Systems Guild (together with Tom DeMarco, Steve McMenamin, Peter Hruschka and Tom Lister) and also creators of the *Volere* requirements development process (including templates and techniques for the validation and specification of requirements). They co-authored 'Mastering the Requirements Process' [138] – a complete guide for practitioners on discovering and communicating requirements effectively using the *Volere* process and associated templates. Suzanne is also the founding editor of the requirements column in IEEE Software[4]. James Robertson (personal communication, May 5, 2014) stated that *Sentire* is "*refreshingly useful*" and the approach is both "*sensible and useful in that it will in all probability deliver good results*". He continued to state that the use of '*all probability*' was intentional, referring to the fact that "*there are always practitioners that can misuse even the most sensible of tools*". Robertson concludes his evaluation by stating that *Sentire* is "*a valuable piece of work*" and points at the utility of Calibrated Persona as a standard method to specify requirements' fit-criteria (e.g., "*65% of Shanya Borg's are able to complete some task*").

and user experience practices be integrated to support the design of acceptable

[3] Fremium.org define freemium as a term "*coined using two words 'Free' and 'Premium'. It describes a business model wherein you give away a core product for free and then generate revenue by selling premium products to a small percentage of free users.*" – Freemium.org, (accessed October 2014)

[4] The Atlantic Systems Guild, http://www.systemsguild.com/sqr.htm, (accessed June 2014)

Appendix A

Research Artefacts

A.1 Initial Theory – Interview Guide

A.1.1 Experiences in registration and authentication processes

1. **Identifying respondents**: Respondents for this exercise are regular internet users who are currently in the workforce (aged 16-60) with at least a secondary level of education.

2. **Number of respondents**: Data collection stops as soon as codebook saturation is identified.

3. **Interview**

   - *Objectives*

     The main objective of this interview is to explore the various experiences you might have had (both positive and negative) with registration and authentication processes across different online services. The discussion will span across various online services, including emailing, social networking, e-banking, e-government and so forth.

     Online Service is defined as: Any online activity requiring users to register for an account, including but not restricted to, email, social networks, e-banking, e-government and so forth. Registration processes could range from the use of a username and password and up to physical visits to some official location (e.g. Registration Authority).

   - *Plan*

     – Introduction

       * Interviewer introduces him/herself
       * Interviewer explains that the respondent has been selected because he/she is representative of the group of people under investigation
       * Why are we doing this interview? (refer to objectives)
       * Interview will take approximately 30 minutes (or less)
       * Repeat the conditions for this interview (consent section below) and ask for their consent to go on.

- General and open ended questions. These are useful to kick off the discussion and to contextualize the interview
  - What kind of online services do you use? (refer to the definition of Online Service)
  - What services have you registered for lately?
  - Does the need for a new username/password annoy you? How do you come up with passwords? (No specific details required)
- To encourage discussion, the following questions may be adapted according to feedback provided by previous respondents. This can be done by providing anonymous statements from previous interviews.
  - Can you recall any negative experiences you had (or heard of) during the registration process of any online service? This is your chance to vent out any frustrations you might have had with specific services (or in general).
  - What are your best experiences with online services registration processes?
  - What do you hate/like most in registration processes?
    - What's your view on account activation processes? (e.g. email activation, manual checks before being able to access an account – such as requesting a bank statement)
  - What is an acceptable registration process in your opinion?
    - Would you apply this process to an e-Banking or e-Government service?
  - Do you own an electronic identity (eID) issued by the Government? Try to encourage discussion around these points
    - How did you register for it?
    - Do you think that the process is acceptable?
    - What do you like and dislike in the eID registration process?
    - How would you change it?
    - Would this process apply for any other online service?
  - Do you have an e-Banking account? (same questions for eID apply here)
  - Would you consider sharing identities across online services? (e.g. Using your Facebook ID for a Yahoo mail account). Why?
  - There is an online service that allows you to submit the annual tax returns once a year. But to use it you have to register for an identity by visiting an authority in Valletta between 9am and 1pm, sign some documents, and wait for the username and password to arrive by post.
    - Would you do it?
    - What other services may be well suited for this kind of registration process?

      (1) How about a service that allows you to receive general notifications about planned electricity cuts? (2) How about a service that allows you to apply for

a birth/marriage certificate? (3) How about a service that allows you to pay for your utilities online? (4) How about a service that allows you to receive weather forecasts? (5) How about a service that allows you to vote online?

- Conclude by asking if there's anything else he/she would like to add
- Ask whether the respondent could be contacted later on just in case we have additional questions.

4. *Consent required*

- Interview is confidential and no names will be divulged
- You can terminate the interview at any point in time
- Interview will be recorded (audio only) for post-interview analysis. Interview will be transcribed in full.
- Interview will (at most) take 30 minutes

5. *References*

1. Gouvernement du Québec (2009), Guide to Organizing Semi-Structured Interviews With Key Informant, Institut National de Santé publique du Québec, retrieved from http://www.crpspc.qc.ca/Guide_entretien_versionWEB_eng.pdf, (accessed, February 2011)

2. Ted Zorn, Designing and Conducting Semi-Structured Interviews for Research, Waikato Management School, retrieved from http://home.utah.edu/~u0326119/Comm4170-01/resources/Interviewguidelines.pdf, (accessed, February 2011)

Appendix B

Colophon

This document was typeset using LaTeX and *BaKoMa TeX* (v.10.30). All styling is based on the `UCL LaTeX Thesis` template, developed and maintained by Ian Kirker. You can download this template from `https://github.com/ucl/ucl-latex-thesis-templates`.

The image used for the introductory vignette is a royalty free image hosted at `http://www.morguefile.com/archive/display/97496`

Bibliography

[1] Anne Adams and M. Angela Sasse. Users are not the enemy. *Communications of the ACM*, 42(12):40–46, 12 1999.

[2] T. Adlin and J. Pruitt. *The Essential Persona Lifecycle: Your Guide to Building and Using Personas*. Morgan Kaufmann, 2010. 2011292293.

[3] GARTEUR Action Group FM AG13. Garteur handbook of mental workload measurement. Technical report, GARTEUR, 2003.

[4] Michel C. Desmarais Ahmed Seffah, Jan Gulliksen. *Human-Centered Software Engineering - Integrating Usability in the Software Development Lifecycle*. Springer, 2005.

[5] Georg Aichholzer and Stefan Strauß. The Austrian case: multi-card concept and the relationship between citizen ID and social security cards. *Identity in The Information Society*, 3:65–85, 2010.

[6] Stephen Kwamena Aikins. *Managing e-government projects concepts, issues and best practices.* IGI Global, Pennsylvania, 2012.

[7] Ian F. Alexander. A taxonomy of stakeholders: Human roles in system development. *International Journal of Technology and Human Interaction (IJTHI)*, 1(1):23–59, 2005.

[8] Ian F. Alexander and Suzanne Robertson. Understanding project sociology by modelling stakeholders. http://www.scenarioplus.org.uk/papers/stakeholders_without_tears/stakeholders_without_tears.htm, 2013. Accessed November 10, 2013.

[9] Ian F. Alexander, Suzanne Robertson, and Neil Maiden. What influences the requirements process in industry? a report on industrial practice. In *Requirements Engineering, 2005. Proceedings. 13th IEEE International Conference on*, pages 411–415, 2005. ID: 1.

[10] Susan N. Allen. Formative evaluation. http://www.beyondintractability.org/essay/formative-evaluation, 12 2003. Accessed November, 2013.

[11] Karin Axelsson and Ulf Melin. Citizens' attitudes towards electronic identification in a public e-service context an essential perspective in the eid development process. In Hans J. Scholl, Marijn Janssen, Maria A. Wimmer, Carl Erik Moe, and Leif Skiftenes Flak, editors, *Electronic Government*, volume 7443 of *Lecture Notes in Computer Science*, pages 260–272. Springer Berlin Heidelberg, 2012.

[12] Adam Beautement, M. Angela Sasse, and Mike Wonham. The compliance budget: Managing security behaviour in organisations. In *Proceedings of the 2008 Workshop on New Security Paradigms*, NSPW '08, pages 47–58. ACM, 2008.

[13] Hugh Beyer and Karen A. Holtzblatt. *Contextual Design: Defining Customer-centered Systems*. Morgan Kaufmann, 1998. 97035927.

[14] Ann Blandford. Semi-structured qualitative studies. In *The Encyclopedia of Human-Computer Interaction*. The Interaction Design Foundation, Aarhus, Denmark, second edition, 2013.

[15] Bodil Stilling Blichfeldt and Jesper Rank Andersen. Creating a wider audience for action research: Learning from case-study research. *Journal of Research Practice*, Volume 2, Issue 1, Article D2, 2006, 2006.

[16] Barry William Boehm and Philip N. Papaccio. Understanding and controlling software costs. *Software Engineering, IEEE Transactions on*, 14(10):1462–1477, 1988. ID: 1.

[17] Joseph Bonneau, Cormac Herley, Paul C. van Oorschot, and Frank Stajano. The quest to replace passwords: a framework for comparative evaluation of web authentication schemes. Technical report, University of Cambridge, Computer Laboratory, 3 2012.

[18] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006. doi: 10.1191/1478088706qp063oa; M3: doi: 10.1191/1478088706qp063oa; 23.

[19] Leo Breiman. Statistical modelling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 8 2001.

[20] Judy Brewer and Shawn Henry. Policies relating to web accessibility. `http://www.w3.org/WAI/Policy/`, 8 2006. Accessed Novemberm, 2013.

[21] Sacha Brostoff and M. Angela Sasse. Are passfaces more usable than passwords? a field trial investigation. In S. McDonald, editor, *Proceedings of HCI 2000*, pages 405–424, Sunderland, 9 2000. Springer.

[22] Joe Bruin. Introduction to SAS. `http://www.ats.ucla.edu/stat/stata/ado/analysis/`, 2 2011. Accessed April, 2013.

[23] A. Bryman. *Social Research Methods*. OUP Oxford, 2012. 2011938966.

[24] Elizabeth Buie and Dianne Murray. *Usability in Government Systems - User Experience Design for Citizens and Public Servants*. Elsevier, Waltham, 2012.

[25] Brad Cain. A review of the mental workload literature. Technical report, Defence Research and Development Canada Toronto, 2007.

[26] Kim Cameron. The laws of identity. http://msdn.microsoft.com/en-us/library/ms996456.aspx, 5 2005. Accessed April, 2013.

[27] Alex Cao, Keshav K. Chintamani, Abhilash K. Pandya, and R. Darin Ellis. Nasa tlx: Software for assessing subjective mental workload. *Behavior Research Methods*, 41(1):113–117, 2009.

[28] Deanna Caputo, Shari Lawrence Pfleeger, M. Angela Sasse, Paul Amman, Jeff Offutt, and Laura McNamara. More than just pretty interfaces: Three case studies of usable security. Under review, oct 2014.

[29] Stuart K. Card, Thomas P. Moran, and Allen Newell. The keystroke-level model for user performance time with interactive systems. *Commun.ACM*, 23(7):396–410, 7 1980.

[30] Jaelson Castro, Manuel Kolp, and John Mylopoulos. Towards requirements-driven information systems engineering: the tropos project. *Information Systems*, 27(6):365–389, 2002.

[31] Kathy Charmaz. *Constructing Grounded Theory: A practical guide through qualitative analysis*. SAGE Publications Ltd, London, 2006.

[32] Alistair Cockburn. Use cases, ten years later. http://alistair.cockburn.us/Use+cases,+ten+years+later, 2002. Accessed 2012.

[33] Louis Cohen, Lawrence Manion, and Keith Morrison. *Research Methods in Education*. Taylor & Francis, 2007.

[34] Alan Cooper. *The Inmates Are Running the Asylum*. Macmillan Publishing Co., Inc, Indianapolis, IN, USA, 1999.

[35] Alan Cooper. *The Inmates are Running the Asylum*. Sams, 2004. 99060546.

[36] Alan Cooper. *Inmates Are Running the Asylum, The: Why High-Tech Products Drive Us Crazy and How to Restore the Sanity*. SAMS, Indiana, 2004.

[37] Alan Cooper. The pipeline to your corporate soul. http://www.cooper.com/journal/2011/09/the_window_to_your_corporate_s.html, 9 2011. Accessed August, 2013.

[38] Alan Cooper, Robert Reimann, and David Cronin. *About Face 3: The Essentials of Interaction Design*. Wiley, 2007.

[39] IBM Corp. Ibm spss statistics. http://www-01.ibm.com/support/knowledgecenter/SSLVMB_20.0.0/com.ibm.spss.statistics.help/idh_cpca.htm, 2011. Accessed 2013.

[40] Terry Crooks. The validity of formative assessments. In *British Educational Research Association Annual Conference*, 9 2001.

[41] Ian Dey. *Grounding Grounded Theory: Guidelines for Qualitative Inquiry*. Academic Press, 1999. 98083122.

[42] Damon Dimmick. Design spikes - fitting big-picture ux into agile development. http://uxdesign.smashingmagazine.com/2012/11/06/design-spikes-fit-big-picture-ux-agile-development, 11 2012. Accessed December, 2013.

[43] Paolo Donzelli and Paolo Bresciani. Goal-oriented requirements engineering: A case study in e-government. In *CAiSE*, pages 601–616, 2003. DBLP:conf/caise/2003.

[44] Amir Dotan, Neil Maiden, Valentina Lichtner, and Lola Germanovich. Designing with only four people in mind? — a case study of using personas to redesign a work-integrated learning support system. In *Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction: Part II*, INTERACT '09, pages 497–509. Springer-Verlag, 2009.

[45] John Dowell and John Long. Towards a conception for an engineering discipline of human factors. *Ergonomics*, 32(11):1513–1535, 1989.

[46] Stephan Dreiseitl and Lucila Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*, 35(5-6):352 – 359, 2002.

[47] Jonathan Earthy, Brian Sherwood Jones, and Nigel Bevan. The improvement of human-centred processes - facing the challenge and reaping the benefit of iso 13407. *International Journal of Human-Computer Studies*, 55(4):553–585, 10 2001.

[48] Entrust. Real-time fraud detection with entrust transactionguard. http://www.entrust.com/products/entrust-transactionguard/, 2012. Accessed August, 2014.

[49] Rebecca Eynon. Breaking barriers to egovernment: Overcoming obstacles to european public services. http://www.oii.ox.ac.uk/research/projects/?id=14, 2007. Accessed August, 2014.

[50] Rebecca Eynon, William H. Dutton, and Helen Margetts. A legal and institutional analysis of barriers to egovernment. deliverable 1b for the ec-funded project 'breaking barriers to egovernment'. http://www.oii.ox.ac.uk/research/projects/?id=14, 2007. Accessed August, 2014.

[51] Rebecca Eynon, William H. Dutton, and Helen Margetts. Solutions for egovernment. deliverable 3 for the ec-funded project 'breaking barriers to egovernment'. http://www.oii.ox.ac.uk/research/projects/?id=14, 2007. Accessed August, 2014.

[52] Shamal Faily. *A framework for usable and secure system design*. PhD thesis, University of Oxford, 2011.

[53] Shamal Faily and Ivan Fléchais. Barry is not the weakest link: Eliciting secure system. In *BCS '10 Proceedings of the 24th BCS Interaction Specialist Group Conference*, Swinton, 2010. BCS.

[54] Shamal Faily and Ivan Fléchais. The secret lives of assumptions: developing and refining assumption personas for secure system design. In *HCSE'10 Proceedings of the Third international conference on Human-centred software engineering*, Reykjavik, Iceland, 2010. Springer-Verlag.

[55] Shamal Faily and Ivan Fléchais. Persona cases: A technique for grounding personas. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2267–2270. ACM, 2011.

[56] Leon Festinger and James M. Carlsmith. Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology*, 58:203–210, 1959.

[57] Andy Field. *Discovering Statistics Using SPSS*. Sage Publications Ltd, 2009.

[58] Ivan Fléchais, M. Angela Sasse, and Stephen Hailes. Bringing security home: A process for developing secure and usable systems. In *NSPW '03 Proceedings of the 2003 workshop on New security paradigms*, pages 49–57. ACM, 2003.

[59] Bent Flyvbjerg. Five misunderstandings about case-study research. *Qualitative Inquiry*, 12(2):219–245, 4 2006.

[60] Helge Fredheim. Why user experience cannot be designed. http://www.smashingmagazine.com/2011/03/15/why-user-experience-cannot-be-designed/, 5 2011. Accessed November, 2013.

[61] Britta Glade. Identity assurance framework: Assurance levels. Technical report, Kantara Initiative, 2010.

[62] Barney G. Glaser and Anselm L. Strauss. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine, 1967. 66028314.

[63] Martin Glinz. Improving the quality of requirements with scenarios. In *Proceedings of the Second World Congress on Software Quality*, 2000 2000.

[64] Joseph A Goguen and Charlotte Linde. Techniques for requirements elicitation. In *Requirements Engineering, 1993., Proceedings of IEEE International Symposium on*, pages 152–164, 1 1993.

[65] GOV.UK. Gov.uk verify guidance: For government service providers. https://www.gov.uk/service-manual/identity-assurance, 2014. Accessed October, 2014.

[66] Megan Grocki and Jamie Thomson. Illustrating the big picture: Journeys, experiences and interactions. http://uxmag.com/articles/illustrating-the-big-picture, 6 2012. Accessed August, 2014.

[67] Greg Guest, Arwen Bunce, and Laura Johnson. How many interviews are enough? an experiment with data saturation and variability. *Field Methods*, 18(1):59–82, 2 2006.

[68] Fraser Hamilton, Pete Pavan, and Kevin McHale. Designing usable e-government services for the citizen - success within user centred design. *International Journal of Public Information Systems*, 7:159–167, 2011.

[69] Sandra G. Hart. Nasa-task load index (nasa-tlx); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9):904–908, 2006.

[70] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in Psychology - Human Mental Workload*, 52:139–183, 1988.

[71] Shawn Henry and Andrew Arch. Social factors in developing a web accessibility business case for your organization. `http://www.w3.org/WAI/bcase/soc.html`, 9 2012. Accessed November, 2013.

[72] Cormac Herley and Paul C. Van Oorschot. A research agenda acknowledging the persistence of passwords. *Security and Privacy, IEEE*, 10(1):28–36, 2012. ID: 1.

[73] Susan G. Hill, Helene P. Iavecchia, James C. Byers, Alvah C. Bittner, Allen L. Zaklad, and Richard E. Christ. Comparison of four subjective workload rating scales, 1992.

[74] Graeme D. Hutcheson and Nick Sofroniou. *The Multivariate Social Scientist: Introductory Statistics Using Generalized Linear Models*. SAGE Publications Ltd, 1999.

[75] Philip Inglesant and M. Angela Sasse. Usability is the best policy: Public policy and the lived experience of transport systems in london. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI...But Not As We Know It - Volume 1*, BCS-HCI '07, pages 35–44. British Computer Society, 2007.

[76] ISO. Iso dis 9241-210: Ergonomics of human-system interaction. part 210: Human-centred design process for interactive systems., 2008.

[77] Michael Jackson. The meaning of requirements. *Annals of Software Engineering*, 3:5–21, 1 1997.

[78] Michael Jackson. The meaning of requirements. *Annals of Software Engineering*, 3(1):5–21, 1997.

[79] Tomasz Janowski, Elsa Estevez, and Adegboyega Ojo. A project framework for e-government. In *Proceedings of the 4th International Conference on E-Government*. Trauner Verlag, 8 2005.

[80] Matthias Jarke and Reino Kurki-Suonio. Introduction to the special issue. *Software Engineering, IEEE Transactions on*, 24(12):1033–1035, 1998. ID: 1.

[81] Homa Javahery, Alexander Deichman, Ahmed Seffah, and Thiruvengadam Radhakrishnan. Incorporating human experiences into the design process of a visualization tool: A case study from bioinformatics. In *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on*, pages 1517–1523, Oct 2007.

[82] Homa Javahery, Alexander Deichman, Ahmed Seffah, and Mohamed Taleb. A user-centered framework for deriving a conceptual design from user experiences: Leveraging personas and patterns to create usable designs. In Ahmed Seffah, Jean Vanderdonckt, and Michel C. Desmarais, editors, *Human-Centered Software Engineering*, Human-Computer Interaction Series, pages 53–81. Springer London, 2009.

[83] Charlene Jennett, Miguel Malheiros, Sacha Brostoff, and M. Angela Sasse. Privacy for loan applicants versus predictive power for loan providers: Is it possible to bridge the gap? In Serge Gutwirth, Ronald Leenes, Paul De Hert, and Yves Poullet, editors, *European Data Protection: In Good Health?*, pages 35–51. Springer Netherlands, 2012.

[84] Timo Jokela and Elizabeth Buie. Getting ux into the contract. In *Usability in Government Systems: User Experience Design for Citizens and Public Servants*, pages 251–266. MK - Elsevier, 2012.

[85] Sara Jones and Neil Maiden. Rescue: An integrated method for specifying requirements for complex sociotechnical systems. In *Requirements Engineering for Sociotechnical Systems*, pages 245–265. IGI Global, Hershey, PA, USA, 2005. ID: 28413.

[86] Sara Jones, Neil Maiden, Sharon Manning, and John Greenwood. Informing the specification of a large-scale socio-technical system with models of human activity. In Pete Sawyer, Barbara Paech, and Patrick Heymans, editors, *Requirements Engineering: Foundation for Software Quality*, volume 4542 of *Lecture Notes in Computer Science*, pages 175–189. Springer Berlin Heidelberg, 2007.

[87] Christos Katsanos, Nikos Karousos, Nikolaos Tselios, Michalis Xenos, and Nikolaos Avouris. Klm form analyzer: Automated evaluation of web form filling tasks using human performance models. In *Human-Computer Interaction - INTERACT 2013*, volume 8118 of *Lecture Notes in Computer Science*, pages 530–537. Springer Berlin Heidelberg, 2013.

[88] Axel Van Lamsweerde. Requirements engineering in the year 00: A research perspective. In *Proceedings of the 22nd international conference on Software engineering*, pages 5–19. ACM, 2000.

[89] Axel Van Lamsweerde. Goal-oriented requirements engineering: A guided tour. In *Requirements Engineering, 2001. Proceedings. Fifth IEEE International Symposium on*, pages 249–262. IEEE, 2001.

[90] Axel Van Lamsweerde. *Requirements engineering : from system goals to UML models to software specifications*. John Wiley, Chichester, England Hoboken, NJ, 2009.

[91] Axel Van Lamsweerde. *Requirements Engineering: From System Goals to UML Models to Software Specifications*. Wiley, 2009.

[92] Axel Van Lamsweerde and Emmanuel Letier. From object orientation to goal orientation: A paradigm shift for requirements engineering. In *Radical Innovations of Software and System Engineering, Montery'02 Workshop, Venice(Italy), LNCS*, pages 4–8. Springer-Verlag, 2003.

[93] Alexei Lapouchnian. Goal-oriented requirements engineering: An overview of the current research. *University of Toronto*, 2005.

[94] Effie Lai-Chong Law, Virpi Roto, Marc Hassenzahl, Arnold P. O. S. Vermeeren, and Joke Kort. Understanding, scoping and defining user experience: A survey approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 719–728. ACM, 2009.

[95] Nancy G. Leveson. Intent specifications: An approach to building human-centered specifications. *IEEE Trans. Softw. Eng.*, 26(1):15–35, January 2000.

[96] Kurt Lewin. The dynamics of group action. *Educational Leadership*, 1:195, 1944.

[97] Bob Lohfeld. Will low-price contracting make us all losers? http://washingtontechnology.com/articles/2012/06/04/lohfeld-low-price-technically-acceptable.aspx, 2012. Accessed November, 2013.

[98] Atlantic Systems Guild Ltd. Experiences of *Volere* users. http://www.volere.co.uk/experience.htm, 2014. Accessed August, 2014.

[99] Neil Maiden. Improve your requirements: Quantify them. *Software, IEEE*, 23(6):68–69, 2006. ID: 1.

[100] Neil Maiden and Suzanne Robertson. Developing use cases and scenarios in the requirements process. In *Proceedings of the 27th International Conference on Software Engineering*, ICSE '05, pages 561–570. ACM, 2005.

[101] Miguel Malheiros. *User behaviour in personal data disclosure*. PhD thesis, UCL (University College London), 2014.

[102] Miguel Malheiros and Sören Preibusch. Sign-up or give-up: Exploring user drop-out in web service registration. *Symposium on Usable Privacy and Security (SOUPS)*, 2013.

[103] Tarvi Martens. Electronic identity management in estonia between market and state governance. *Identity in the Information Society*, 3(1):213–233, 2010.

[104] Mark Mason. Forum: Qualitative social research - sample size and saturation in phd studies using qualitative interviews. http://www.qualitative-research.net/index.php/fqs/article/view/1428/3027, 2010. Accessed 2013.

[105] John McCarthy and Peter Wright. Technology as experience. *interactions*, 11(5):42–43, 2004.

[106] Saul A. McLeod. Cognitive dissonance theory. http://www.simplypsychology.org/cognitive-dissonance.htm, 2008. Accessed August, 2013.

[107] Scott Menard. *Logistic Regression: From Introductory to Advanced Concepts and Applications*. SAGE Publications, 2009. 20084993.

[108] Glenn Micallef and Chris Porter. Web personalisation through mouse motion analytics. In *Workshops in ICT 2013*, 2013.

[109] Peter Morville. User experience design. http://semanticstudios.com/user_experience_design/, 6 2004. Accessed August, 2014.

[110] Peter Morville. User experience deliverables. http://semanticstudios.com/user_experience_deliverables/, 1 2009. Accessed August, 2014.

[111] In Jae Myung. Tutorial on maximum likelihood estimation. *Journal of mathematical psychology*, 47(1):90–100, 2 2003.

[112] Jakob Nielsen. 1994 Design of SunWeb: Sun Microsystems' Intranet. http://www.nngroup.com/articles/1994-design-sunweb-sun-microsystems-intranet/, 1994. Accessed April, 2014.

[113] Jakob Nielsen. Usability metrics. http://www.nngroup.com/articles/usability-metrics/, 1 2001. Accessed April, 2014.

[114] Jakob Nielsen. Usability 101: Introduction to usability. http://www.nngroup.com/articles/usability-101-introduction-to-usability/, 1 2012. Accessed November, 2013.

[115] Jakob Nielsen and Thomas Landauer K. A mathematical model of the finding of usability problems. In *CHI '93 Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, pages 206–213. ACM, 1993.

[116] Jakob Nielsen and Don Norman. The definition of user experience. http://www.nngroup.com/articles/definition-user-experience/, 2012. Accessed November, 2014.

[117] Lene Nielsen. Engaging personas and narrative scenarios. Technical report, Department of Informatics, Copenhagen Business School, 2004.

[118] Bashar Nuseibeh and Steve Easterbrook. Requirements engineering: A roadmap. In *Proceedings of the Conference on The Future of Software Engineering*, ICSE '00, pages 35–46. ACM, 2000.

[119] OECD. Electronic authentication and oecd guidance for electronic authentication. http://www.oecd.org/internet/ieconomy/38921342.pdf, 6 2007. Accessed 2012.

[120] OECD. Security and authentication issues in the delivery of electronic services to taxpayers. http://www.oecd.org/site/ctpfta/49428035.pdf, 1 2012. Accessed November, 2012.

[121] Treasury Board of Canada Secretariat. Standard on identity and credential assurance. https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=26776&section=text, 2013. Accessed October, 2014.

[122] Harvard Graduate School of Education. Thematic analysis. http://isites.harvard.edu/icb/icb.do?keyword=qualitative&pageid=icb.page340897, 2008. Accessed February, 2013.

[123] George Olsen. Persona creation and usage toolkit. http://www.interactionbydesign.com/presentations/olsen_persona_toolkit.pdf, 3 2004. Accessed February, 2014.

[124] Shari Lawrence Pfleeger. Blending in instead of bolting on: Harmonizing security and usability in software development. In *The 1st Software And Usable Security Aligned For Good Engineering (SAUSAGE) Workshop*, National Institute of Standards and Technology, Gaithersburg, MD USA; I3P, apr 2011. Institute for Information Infrastructure Protection.

[125] Shari Lawrence Pfleeger and Deanna D. Caputo. Leveraging behavioral science to mitigate cyber security risk. *Computers Security*, 31(4):597–611, 6 2012.

[126] Shari Lawrence Pfleeger, M. Angela Sasse, and Deanna Caputo. Studying usable security: How to design and conduct case studies. Under review, oct 2014.

[127] Chris Porter, M. Angela Sasse, and Emmanuel Letier. Designing acceptable user registration processes for e-services. In *Proceedings of HCI 2012 - The 26th BCS Conference on Human Computer Interaction*. BCS, 9 2012.

[128] Chris Porter, M. Angela Sasse, and Emmanuel Letier. Giving a voice to personas in the design of e-government identity processes. In *Research to Design: Challenges of Qualitative Data Representation and Interpretation in HCI - in BCS HCI 2013*. BCS, 9 2013.

[129] Sören Preibusch, Kat Krol, and Alastair R. Beresford. The privacy economics of voluntary overdisclosure in web forms. In Rainer Böhme, editor, *The Economics of Information Security and Privacy*, pages 183–209. Springer Berlin Heidelberg, 2013.

[130] Marc Prensky. Digital natives, digital immigrants. *On the horizon*, 9(5):1–6, 2001.

[131] Adrian Rahaman and M. Angela Sasse. A framework for the lived experience of identity. *Identity in the Information Society*, 3(3):605–638, 2010.

[132] Trygve Mikjel H. Reenskaug. *The original MVC reports*. University Of Oslo, 1979.

[133] Karen Renaud. User issues in security. In *Usability in Government Systems: User Experience Design for Citizens and Public Servants*, pages 217–229. MK - Elsevier, 2012.

[134] HM Revenue and Customs. Hmrc online services: Increasing use of online filing and electronic payment. http://www.legislation.gov.uk/ukia/2007/37/pdfs/ukia_20070037_en.pdf, 2007. Accessed August, 2014.

[135] HM Revenue and Customs. Departmental report 2009. Technical report, HM Revenue and Customs, 2009. Accessed August, 2014.

[136] HM Revenue and Customs. Online services: Digital certificates. http://webarchive.nationalarchives.gov.uk/+/http://www.hmrc.gov.uk/ebu/digital-certs.htm, 2012. Accessed August, 2014.

[137] Jens Riegelsberger and M. Angela Sasse. Ignore these at your peril: Ten principles for trust design. In *Trust 2010. 3rd International Conference on Trust and Trustworthy Computing*, 2010.

[138] Suzanne Robertson and James Robertson. *Mastering the Requirements Process: Getting Requirements Right*. Addison-Wesley, 2012.

[139] Yvonne Rogers, Helen Sharp, and Jenny Preece. *Interaction Design - Beyond Human-Computer Interaction, 3rd Edition*. Wiley, 2011.

[140] Susana Rubio, Díaz Eva, Martín Jesús, and José M. Puente. Evaluation of subjective mental workload: A comparison of swat, nasa-tlx, and workload profile methods. *Applied Psychology*, 53(1):61–86, 2004.

[141] M. Angela Sasse. Usability and trust in information systems. Technical report, Cyber Trust and Crime Prevention Project, 2004.

[142] M. Angela Sasse. People and security. Lecture Notes from COMPGA10 - People and Security (UCL), 2014.

[143] M. Angela Sasse, Sacha Brostoff, and D. Weirich. Transforming the 'weakest link' - a human/computer interaction approach to usable and effective security. *BT Technology Journal*, 19(3):122–131, 7 2001.

[144] M. Angela Sasse and Ivan Fléchais. Usable security: Why do we need it? how do we get it? In *Security and Usability*, pages 13–30. O'Reilly, 2005.

[145] M. Angela Sasse, Michelle Steves, Kat Kron, and Dana Chisnell. The great authentication fatigue - and how to overcome it. In P.L.Patrick Rau, editor, *Cross-Cultural Design*, volume 8528 of *Lecture Notes in Computer Science*, pages 228–239. Springer International Publishing, 2014.

[146] Jeff Sauro. *Quantifying the user experience : practical statistics for user research*. Elsevier/Morgan Kaufmann, Amsterdam Boston, 2012.

[147] Ahmed Seffah, Jan Gulliksen, and Michel C. Desmarais. An introduction to human-centered software engineering. In Ahmed Seffah, Jan Gulliksen, and Michel C. Desmarais, editors, *Human-Centered Software Engineering - Integrating Usability in the Software Development Lifecycle*, volume 8 of *Human-Computer Interaction Series*, pages 3–14. Springer Netherlands, 2005.

[148] Ahmed Seffah and Eduard Metzker. The obstacles and myths of usability and software engineering. *Communications of the ACM*, 47(12):71–76, 12 2004.

[149] Philip Seltsikas and Nikolaos Papas. Developing user requirements for trans-national government information systems. In *System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on*, pages 1–6, 1 2009.

[150] Norbert Seyff, Florian Graf, and Neil Maiden. End-user requirements blogging with irequire. In *Software Engineering, 2010 ACM/IEEE 32nd International Conference on*, volume 2, pages 285–288, 2010. ID: 1.

[151] Norbert Seyff, Florian Graf, and Neil Maiden. Using mobile re tools to give end-users their own voice. In *Requirements Engineering Conference (RE), 2010 18th IEEE International*, pages 37–46, 2010. ID: 1.

[152] Helen Sharp, Anthony Finkelstein, and Galal Galal. Stakeholder identification in the requirements engineering process. In *Database and Expert Systems Applications, 1999. Proceedings. Tenth International Workshop on*, pages 387–391. Ieee, 1999.

[153] Marten D. Shipman. *The limitations of social research*. Longman, London New York, 1997.

[154] Joel Sklar. *Web design principles*. Course Technology Cengage Learning, Australia, 2012.

[155] DG Information Society and Media. A legal and institutional analysis of barriers to egovernment. Technical report, DG Information Society and Media, 2007.

[156] Frank Stajano. Pico: No more passwords! In *Proceedings of the 19th International Conference on Security Protocols*, SP'11, pages 49–81. Springer-Verlag, 2011.

[157] Inc StatSoft. Electronic statistics textbook. *StatSoft, Tulsa, OK*, 2007.

[158] Michelle Steves, Dana Chisnell, M. Angela Sasse, Kat Krol, Mary Theofanos, and Hannah Wald. Report: Authentication diary study. nistir 7983. Technical Report NISTIR 7983, 2014.

[159] Anselm L. Strauss. *Qualitative Analysis for Social Scientists*. Cambridge University Press, 1987. lc86021608.

[160] Anselm L. Strauss and Juliet M. Corbin. *Basics of qualitative research: grounded theory procedures and techniques*. Sage Publications, 1990. lc90039609.

[161] Alistair Sutcliffe. Requirements engineering from an hci perspective. In *The Encyclopedia of Human-Computer Interaction*. The Interaction Design Foundation, Aarhus, Denmark, second edition, 2013.

[162] Symantec. Endpoint, cloud, mobile and virtual security solutions. http://www.symantec.com/index.jsp, 8 2012. Accessed August, 2014.

[163] Gary Thomas and David James. Re-inventing grounded theory: some questions about theory, ground and discovery. *British Educational Research Journal*, 32, 6:767–795, 2006.

[164] Stephen E. Toulmin. *The Uses of Argument*. Cambridge University Press, 2003.

[165] UsabilityNet. Relevant international standards in usability and user-centred design. http://www.usabilitynet.org/management/b_standards.htm, 2006. Accessed November, 2013.

[166] Lex van Velsen, Thea van der Geest, Marc ter Hedde, and Wijnand Derks. Requirements engineering for e-government services: A citizen-centric approach and case study. *Government Information Quarterly*, 26(3):477–486, 7 2009.

[167] Gianluigi Viscusi, Carlo Batini, and Massimo Mecella. *Information Systems for eGovernment: A Quality-of-Service Perspective*. SpringerLink: Springer e-Books. Springer, 2010.

[168] Constantin von Saucken, Ioanna Michailidou, and Udo Lindemann. How to design experiences: Macro ux versus micro ux approach. In Aaron Marcus, editor, *Design, User Experience, and Usability. Web, Mobile, and Product Design*, volume 8015 of *Lecture Notes in Computer Science*, pages 130–139. Springer Berlin Heidelberg, 2013.

[169] Diane Walker and Florence Myrick. Grounded theory: An exploration of process and procedure. *Qualitative health research*, 16(4):547–559, 4 2006.

[170] Léonie Watson. Gov.uk accessibility: beyond box-ticking. https://gds.blog.gov.uk/2013/02/11/beyond-box-ticking/, 2 2013. Accessed November, 2013.

[171] Susan Weinschenk. *Neuro web design : what makes them click*. New Riders, Berkeley, CA, 2009.

[172] Karl E. Wiegers. When telepathy won't do: Requirements engineering key practices. *Cutter IT Journal*, 13(5):9–15, 2000.

[173] Karl E. Wiegers. *Software Requirements*. Pro-Best Practices. Microsoft Press, 2009.

[174] Graham Williamson, Ilan Sharoni, David Yip, and Kent. E. Spaulding. *Identity Management: A Primer*. MC Press Online, 2009.

[175] Ka-Ping Yee. Guidelines and strategies for secure interaction design. In *Security and Usability: Designing Secure Systems That People Can Use*, pages 247–273. O'Reilly, 2005.

[176] Robert K. Yin. *Case Study Research: Design and Methods*. SAGE Publications, Inc., 4th edition, 2009.

[177] Konstantinos Zachos and Neil Maiden. Art-scene: enhancing scenario walkthroughs with multi-media scenarios. In *Requirements Engineering Conference, 2004. Proceedings. 12th IEEE International*, pages 360–361, 2004. ID: 1.

[178] Konstantinos Zachos, Neil Maiden, and Amit Tosar. Rich-media scenarios for discovering requirements. *Software, IEEE*, 22(5):89–97, 2005. ID: 1.

[179] Pamela Zave. Classification of research efforts in requirements engineering. In *Requirements Engineering, 1995., Proceedings of the Second IEEE International Symposium on*, pages 214–216. IEEE, 1995.

[180] Mary Ellen Zurko and Richard T. Simon. User-centered security. In *Proceedings of the 1996 Workshop on New Security Paradigms*, NSPW '96, pages 27–33. ACM, 1996.