Educational Assessment, Evaluation and Accountability
https://doi.org/10.1007/s11092-020-09329-5

# Is Canada really an education superpower? The impact of non-participation on results from PISA 2015

Jake Anders[1] · Silvan Has[1] · John Jerrim[1] · Nikki Shure[1] · Laura Zieger[1]

Received: 19 November 2019 / Accepted: 13 July 2020/Published online: 25 July 2020


## Abstract

The purpose of large-scale international assessments is to compare educational achievement across countries. For such cross-national comparisons to be meaningful, the participating students must be representative of the target population. In this paper, we consider whether this is the case for Canada, a country widely recognised as high performing in the Programme for International Student Assessment (PISA). Our analysis illustrates how the PISA 2015 sample for Canada only covers around half of the 15-year-old population, compared to over 90% in countries like Finland, Estonia, Japan and South Korea. We discuss how this emerges from differences in how children with special educational needs are defined and rules for their inclusion in the study, variation in school participation rates and the comparatively high rates of pupils' absence in Canada during the PISA study. The paper concludes by investigating how Canada's PISA 2015 rank would change under different assumptions about how the non-participating students would have performed were they to have taken the PISA test.

**Keywords** PISA · Non-response

## 1 Introduction

The Programme for International Student Assessment (PISA) is an important international study of 15-year-olds' achievement in reading, science and mathematics. It is

Jake Anders, Silvan Has, John Jerrim, Nikki Shure and Laura Zieger contributed equally to this work.

Jake Anders, Silvan Has, John Jerrim, Nikki Shure and Laura Zieger are joint first authors.

**Electronic supplementary material** The online version of this article (https://doi.org/10.1007/s11092-020-09329-5) contains supplementary material, which is available to authorized users.

✉ John Jerrim
J.Jerrim@ucl.ac.uk

Extended author information available on the last page of the article

conducted every three years by the Organisation for Economic Cooperation and Development (OECD) and receives substantial attention from policymakers, the media, academics and the wider education community. Particular attention is often paid to the top-performing nations in PISA and these often inspire policy development in other countries (Raffe 2011). Although Finland (Hendrickson 2012; Takayama et al. 2013) and the high-performing East Asian nations (Feniger and Lefstein 2014) have often taken the limelight, a North American country, Canada, has also received significant attention. Indeed, despite its cultural, linguistic and historical similarities to many other Western nations, Canada achieves much higher average PISA scores than most OECD countries, while also apparently having a more equitable distribution of educational achievement. This is illustrated in Table 1, which benchmarks Canada's PISA 2015 reading scores against key comparators (OECD 2016). Based upon these results, Canada has been described as an 'education super-power' in the press (Coughlan 2017), with Andreas Schleicher—who led the development of the OECD's PISA programme—suggesting that this is driven by its strong commitment to equity.

Such international comparisons of countries—of the type routinely undertaken through PISA—requires strict criteria to ensure one is comparing like-with-like. A long and extensive literature has discussed the importance of translation (e.g. Masri et al. 2016), cross-cultural comparability of the test instruments (e.g. Kankaraš and Moors 2014) and the importance of establishing measurement invariance across countries (Rutkowski and Rutkowski 2016). Yet issues surrounding population definitions, school enrolment rates, sample exclusions, school participation and pupil participation are also important. For instance, if country A systematically excludes many of its low-achieving students (e.g. by deeming them ineligible for the study or because they are absent on the day of the test), then the data and results generated may not be comparable with country B (where a truly representative cross-section of the student population participated). Consequently, comparisons of educational achievement across these two archetypal countries will not be meaningful. As this paper will describe, the PISA 2015 data for Canada seem to have some of the characteristics of country A, which clearly have the potential to cause comparability issues with countries that are more like country B. This, in turn, undermines Canada's apparently strong performance in the PISA study in terms of both equity and efficiency.

This is not the first paper to discuss issues of population coverage and non-response bias in the context of PISA. Similar concerns have previously been raised about the quality of data available from other countries. For instance, using trends in the PISA scores of Turkey as an example, Spaull (2018) highlights how limitations with the eligibility criteria used in PISA can lead to overestimation of academic achievement and underestimation of educational inequality. A similar analysis conducted by Education Datalab (2017) highlights how issues with differential school enrolment rates across countries can partially explain the strong PISA performance of Vietnam. Pereira (2011) focuses upon changes to the PISA sampling method used in Portugal over time, suggesting that this can help explain recent trends in this country's performance. Furthermore, Micklewright et al. (2012) and Durrant and Schnepf (2018) tackle the issue of school and student non-response in England. They find that low-achieving schools, and schools with a large proportion of disadvantaged pupils, are more likely to refuse to take part in PISA, which may bias estimates of educational achievement compared to countries that do not allow such refusals. Similarly, Jerrim (2013)

**Table 1** PISA 2015 reading scores compared across OECD countries

| Country | Mean | Confidence interval | P10 | P90 | P90–P10 |
|---|---|---|---|---|---|
| Canada | 527 | 522–531 | 404 | 642 | 238 |
| Finland | 526 | 521–531 | 401 | 640 | 239 |
| Ireland | 521 | 516–526 | 406 | 629 | 223 |
| Estonia | 519 | 515–523 | 404 | 630 | 226 |
| South Korea | 517 | 511–524 | 386 | 637 | 251 |
| Japan | 516 | 510–522 | 391 | 629 | 238 |
| Norway | 513 | 508–518 | 381 | 636 | 255 |
| New Zealand | 509 | 505–514 | 368 | 643 | 275 |
| Germany | 509 | 503–515 | 375 | 634 | 259 |
| Poland | 506 | 501–511 | 386 | 617 | 231 |
| Slovenia | 505 | 502–508 | 382 | 621 | 239 |
| Netherlands | 503 | 498–508 | 368 | 630 | 262 |
| Australia | 503 | 500–506 | 365 | 631 | 266 |
| Sweden | 500 | 493–507 | 364 | 625 | 261 |
| Denmark | 500 | 495–505 | 383 | 608 | 225 |
| France | 499 | 494–504 | 344 | 637 | 293 |
| Belgium | 499 | 494–503 | 360 | 623 | 263 |
| Portugal | 498 | 493–503 | 374 | 614 | 240 |
| UK | 498 | 493–503 | 372 | 621 | 249 |
| USA | 497 | 490–504 | 364 | 624 | 260 |
| Spain | 496 | 491–500 | 379 | 603 | 224 |
| Switzerland | 492 | 486–498 | 360 | 614 | 254 |
| Latvia | 488 | 484–491 | 374 | 595 | 221 |
| Czech Republic | 487 | 482–492 | 352 | 614 | 262 |
| Austria | 485 | 479–490 | 347 | 611 | 264 |
| Italy | 485 | 480–490 | 359 | 602 | 243 |
| Iceland | 482 | 478–485 | 350 | 607 | 257 |
| Luxemburg | 481 | 479–484 | 336 | 616 | 280 |
| Israel | 479 | 472–486 | 326 | 621 | 295 |
| Hungary | 470 | 464–475 | 338 | 593 | 255 |
| Greece | 467 | 459–476 | 334 | 590 | 256 |
| Chile | 459 | 454–464 | 342 | 572 | 230 |
| Slovak Republic | 453 | 447–458 | 312 | 583 | 271 |
| Turkey | 428 | 421–436 | 322 | 535 | 213 |
| Mexico | 423 | 418–428 | 321 | 523 | 202 |

illustrates how a combination of non-response bias, changes to the target population and test month led policymakers to reach erroneous conclusions about changes to PISA test scores in England.

In this paper, we add to this literature by explaining how data from one of the top PISA performers, Canada, potentially suffers from similar issues. We begin by

discussing the rules that the OECD set for inclusion in the PISA study and investigate whether Canada meets each of these criteria. We find that it either fails to meet them, or meets them only marginally, in all cases. It is then demonstrated how this has a significant cumulative impact upon the PISA 2015 Canadian sample. Our empirical analysis then moves on to sensitivity analysis of the Canadian PISA results, focusing upon how it compares to two genuinely high-performing countries (Japan and South Korea) where student exclusion and school/student non-response rates are much lower. These sensitivity analyses estimate the scores that excluded and non-responding students would need to have achieved in order to 'disturb' a finding (Gorard and Gorard 2016); in other words, to make the difference between countries disappear. We argue that this is a more important reflection of uncertainty in the Canadian PISA results than the standard forms of statistical inference (confidence intervals and statistical significance tests) that are routinely reported by the OECD (as it captures different forms of bias rather than just sampling variation alone). Our results illustrate how Canada's PISA results could change in non-trivial ways, relative to other countries, under plausible assumptions about how excluded/non-responding students would have performed on the test. It is, hence, concluded that the OECD should do more to communicate the uncertainty in PISA results due to sample exclusions and missing data.

The paper now proceeds as follows. Section 2 describes the criteria set by the OECD to try to ensure the PISA data are of high quality and illustrate how the data for Canada performs relative to these benchmarks. Our empirical approach is set out in Section 3, along with the sensitivity analyses we have conducted around Canada's PISA results. Conclusions and implications for communication and interpretation of the PISA results follow in Section 4.

## 2 Key elements of the design of the PISA study

### 2.1 Target population and exclusions

The target population of PISA is 15-year-olds attending educational institutions in seventh grade[1] or higher (OECD 2017: Chapter 4). This definition has some subtle, but important, implications. In particular, note that young people not enrolled in education (due to, for instance, permanent exclusion, home schooling or having surpassed the minimum school leaving age) are excluded. This is consistent with the aim of PISA to measure outcomes of school systems towards the end of compulsory education. Yet, as previous research has suggested, this definition means many 15-year-olds are excluded from PISA in low- and middle-income countries (Spaull 2018; Education Datalab 2017) upwardly biasing results compared to those that would be expected if the target population were all 15-year-olds. Yet, as Table 2 illustrates, it is also not a trivial issue in some OECD countries. In Canada, around 4% of 15-year-olds are excluded from PISA due to non-enrolment at school. This is greater than in the other high-performing OECD nations of Estonia, Finland, Japan and South Korea, where between 98 and 100% of 15-year-olds are enrolled in an educational institution. Yet there are also some

[1] Based upon the education system in the USA.

other OECD countries where this is clearly a very important issue, most notable Turkey (83% of 15-year-olds are enrolled in schools) and Mexico (62% of 15-year-olds enrolled in school).

**Table 2** Enrolments, exclusions and response rates in PISA 2015 (OECD countries)

| | (a) % of 15-year-olds enrolled in school | (b) % Excluded | (c) School response % before replacement | (d) School response % after replacement | (e) Student participation rate (%) | (f) % of 15-year-old covered by test |
|---|---|---|---|---|---|---|
| Australia | 100 | 5.3 | 91 (94) | 92 (95) | 81 | 71.7 |
| Austria | 94 | 2.1 | 99 (100) | 99 (100) | 71 | 62.9 |
| Belgium | 99 | 1.7 | 81 (83) | 95 (95) | 91 | 83.8 |
| Canada | 96 | 7.5 | 70 (75) | 72 (79) | 81 | 52.5 |
| Chile | 96 | 1.8 | 89 (92) | 97 (99) | 94 | 87.3 |
| Czech Republic | 100 | 2.4 | 99 (98) | 99 (98) | 89 | 84.8 |
| Denmark | 99 | 5.0 | 88 (90) | 89 (92) | 87 | 74.0 |
| Estonia | 98 | 5.5 | 100 (100) | 100 (100) | 93 | 85.5 |
| Finland | 100 | 2.8 | 99 (100) | 100 (100) | 93 | 90.2 |
| France | 96 | 4.2 | 91 (91) | 95 (94) | 88 | 74.5 |
| Germany | 100 | 2.1 | 96 (96) | 99 (99) | 93 | 90.0 |
| Greece | 100 | 1.9 | 90 (92) | 99 (98) | 94 | 90.2 |
| Hungary | 95 | 3.3 | 92 (93) | 97 (99) | 92 | 82.8 |
| Iceland | 99 | 3.6 | 95 (99) | 95 (99) | 86 | 80.5 |
| Ireland | 98 | 3.1 | 99 (99) | 99 (99) | 89 | 83.0 |
| Israel | 95 | 3.4 | 89 (91) | 91 (93) | 91 | 76.2 |
| Italy | 92 | 3.8 | 78 (74) | 87 (88) | 89 | 66.5 |
| Japan | 98 | 2.4 | 95 (94) | 99 (99) | 97 | 91.6 |
| Latvia | 98 | 5.1 | 86 (86) | 92 (93) | 90 | 76.6 |
| Luxemburg | 96 | 8.2 | 100 (100) | 100 (100) | 96 | 83.8 |
| Mexico | 62 | 0.9 | 95 (95) | 97 (98) | 95 | 54.2 |
| Netherlands | 100 | 3.7 | 62 (63) | 92 (93) | 85 | 75.4 |
| New Zealand | 95 | 6.5 | 69 (71) | 84 (85) | 80 | 56.5 |
| Norway | 100 | 6.8 | 95 (95) | 95 (95) | 91 | 79.7 |
| Poland | 95 | 2.4 | 89 (88) | 99 (99) | 87 | 78.7 |
| Portugal | 91 | 1.3 | 84 (86) | 94 (95) | 82 | 67.6 |
| Slovak Republic | 99 | 4.3 | 92 (93) | 98 (99) | 91 | 84.8 |
| Slovenia | 98 | 3.1 | 95 (98) | 95 (98) | 91 | 84.1 |
| South Korea | 100 | 0.9 | 99 (100) | 99 (100) | 99 | 98.1 |
| Spain | 94 | 3.2 | 99 (99) | 100 (100) | 89 | 79.8 |
| Sweden | 99 | 5.7 | 99 (100) | 99 (100) | 91 | 84.3 |
| Switzerland | 98 | 4.4 | 91 (93) | 97 (98) | 93 | 84.7 |
| Turkey | 83 | 1.1 | 90 (97) | 96 (99) | 95 | 76.0 |
| UK | 100 | 8.2 | 85 (84) | 91 (93) | 88 | 73.6 |
| USA | 95 | 3.3 | 67 (67) | 83 (83) | 90 | 66.4 |
| **OECD average** | **96** | **3.7** | **90 (91)** | **95 (96)** | **89** | **78.1** |
| **OECD median** | **98** | **3.3** | **91 (93)** | **97 (98)** | **91** | **79.8** |

Both weighted and unweighted school response rates are provided (the former appear in brackets). Student response rate refers to weighted figures. Shading should be read vertically, with darker green shading indicating a 'worse' outcome on the measure (e.g. lower response rate, higher level of exclusions). The coverage rate of 15-year-olds (f) is computed as follows: 100%−[[100%−(a)] + (b) + [100%−weighted (d)×(e)]]

Countries are also allowed to exclude some schools or students within the defined target population from the PISA study. This is usually due to severe Special Educational Needs (SEN) limiting the opportunity for some young people to take part. The criteria set by the OECD is that a maximum of 5% of students can be excluded from PISA within any given country. As noted by Rutkowski and Rutkowski (2016), this maximum of 5% should 'ensure that any distortions in national mean scores due to omitted schools or students would be no more than ± 5 score points on the PISA scale'.

Yet the second column of Table 2 illustrates how several countries breached this 5% threshold for exclusions in PISA 2015, but were still included within the study. This includes Canada, which has one of the highest rates of student exclusions (7.5%)—double the OECD average (3.7%). Further inspection of the PISA 2015 national report for Canada (O'Grady et al. 2016) indicates that the excluded students were mainly those with intellectual disabilities (5%), with a further 1.5% of students removed due to limited language skills and 0.5% for physical disabilities. As Table 2 illustrates, the percentage of excluded students differs across countries—with many more excluded in Canada than in some of the other high-performing OECD countries (e.g. Japan and South Korea).[2] This has the potential to bias comparisons between these nations, which the OECD recognises, if certain groups we would not expect to perform well on the PISA test are routinely excluded in some nations (e.g. students with intellectual disabilities in Canada) but not in others (e.g. Japan and South Korea).

### 2.2 Sample design

PISA utilises a stratified, clustered sampling approach. The purpose of stratification is to boost the efficiency of the sample (i.e. increase power to narrow confidence intervals) and to ensure there is adequate representation of important sub-groups.

To begin, each country selects a set of 'explicit stratification' variables, which should be closely related to PISA achievement.[3] These are essentially used to form different sub-groups (strata). Although these differ across countries, geographic region and school type are common choices. In Canada, province, language and school size are used. Within each of these explicit strata, schools are then ranked by a variable (or set of variables) that are likely to be strongly associated with PISA test scores. This is known as implicit stratification. Unfortunately, the implicit stratification variables used in Canada (level of urbanisation, source of school funding and the ISCED level taught) are only relatively weakly associated with academic achievement. For instance, when regression PISA reading scores upon the stratification variables, the R-squared statistic (i.e. the variance explained by the stratification variable) is only 0.01 for level of urbanisation and source of school funding and 0.02 for ISCED level taught. This creates a potential issue if replacement schools need to be targeted, which we discuss below.

[2] This in part demonstrates an inherent challenge in cross-national comparative research; the definition, identification and treatment of students with special educational needs is likely to differ significantly across different national settings.

[3] The ideal stratification variables are usually based upon school/student performance in national examinations. Although standardised tests are conducted within Canada, they differ across provinces, potentially explaining why they are not used.

Schools are then randomly selected, with probability proportional to size, from within each of the explicit strata. The minimum number of schools to be selected is 150, although some countries oversample in order achieve sufficient statistical power to produce meaningful sub-group estimates. This is the case in Canada, where results are reported nationally and at the province level. Hence, in total, 1008 Canadian schools were approached to take part.

Not all schools approached agree to participate in the PISA study. In Canada, 305 (30%) of schools initially approached refused to participate in PISA 2015. In this situation, PISA allows countries to approach up to two 'replacement schools' to take the place of the originally sampled schools. These replacement schools are those that are adjacent to the originally sampled school on the sampling frame. The intuition behind this approach is that the replacement schools will be 'similar' to the originally sampled school that they replace. It is hence a form of 'near neighbour' donor imputation; however, this is only effective at reducing non-response bias if the stratification variables used in the sampling are strongly correlated with the outcome of interest (PISA scores). As noted above, this is questionable in the case of Canada, where only quite modest predictors of academic achievement were used as stratification variables.

After including these replacement schools, a total of 726 Canadian schools (72% of those approached) took part. The school response rate was particularly low in Quebec (40% before replacement), where there were teacher strikes later in 2015 (though we are unclear whether the low school response rates for PISA were related to these strikes). Even the overall figure is much lower than in most other OECD countries (OECD average = 95%), including the other high-performing nations of Estonia (100%), Finland (100%), Japan (99%) and South Korea (99%), as illustrated by Table 1.

### 2.3 School response rate criteria

To encourage countries to achieve adequate school response rates, the OECD have set criteria linked to these that are meant to result in exclusion from the PISA study if they are not met. The criteria are depicted by Fig. 1 and can be summarised as follows:

- Acceptable 1 (light-blue region). More than 85% of originally sampled schools participated. The PISA sample for countries in this category is assumed to be unbiased and automatically included in the results.
- Acceptable 2 (light-blue region). Between 65 and 85% of originally sampled schools participated, with this percentage increasing substantially[4] once replacement schools are added. The PISA sample for countries in this category is assumed to be unbiased and automatically included in the results.
- Intermediate (blue region). Between 65 and 85% of originally sampled schools participated, with this percentage not increasing sufficiently even when replacement

[4] A sliding scale is used, where a higher 'after-replacement' response rate is required if a lower 'before-replacement' response rate was obtained. For instance, a country that had a 65% response rate amongst initially sampled schools would require this to increase to around 95% once replacement schools are added.

schools are added. Countries that fall into this category are required to undertake a non-response bias analysis (NRBA) as discussed in the following sub-section.
- Not acceptable (dark-blue region). Less than 65% of the originally sampled schools participated in the study.[5] Countries that fall in this category should be automatically excluded from the PISA results.

Figure 1 illustrates how, in PISA 2015, four OECD countries (Italy, New Zealand, USA and Canada) fell into the intermediate category where a NRBA was required—with Canada the furthest of these from the 'acceptable' zone after replacement. The data for one other OECD country (the Netherlands) appears in the 'not acceptable' zone and, as such, should have been automatically excluded.[6] Nevertheless, all five countries were included in the final PISA 2015 rankings without any explicit warning given about their results.

Comparing Canada with other high-performing countries over several years, Fig. 2 illustrates how Estonia, Finland and South Korea have generally had high school response rates in all PISA cycles (after allowing replacement schools, they often reach close to 100% response rates). The situation is rather different for Canada, however, where the unweighted school response rate fell dramatically in 2015 to around 72% (from around 90% in the previous PISA cycles).

### 2.4 Non-response bias analyses

Non-response does not necessarily introduce bias into the sample—it only does so if the non-response is not random. However, some previous research does indeed suggest that school non-response in PISA is selective (Heine et al. 2017). One way of investigating whether certain 'types' of schools choose not to participate in PISA is to compare the observable characteristics of participating and non-participating schools.[7] Ideally, the variables used to compare responding and non-responding schools should, like stratification variables, be strongly associated with the outcome of interest (i.e. PISA scores)—such as national measures of school achievement. The intuition behind this approach is that, if responding and non-responding schools differ in terms of (for instance) national measures of achievement (e.g. scores on a national mathematics exam), then they are also likely to differ in terms of their likely performance on PISA.

Unfortunately, few details about what the OECD deems to be an acceptable NRBA are published. The only details available come from a small section in the technical report (OECD 2017: Chapter 14). However, some more description of what is required

[5] One way to think of this criterion is that a maximum of 35% of the PISA data is allowed to be 'imputed' from the donor (substitute) schools.

[6] The PISA 2015 technical report (OECD 2017) notes that the Netherlands was permitted to conduct a non-response bias analysis (NRBA) rather than face automatic exclusion. Based upon the findings from the NRBA, the OECD deemed the Dutch sample to be of acceptable quality and hence could be included in the PISA 2015 results. We have obtained a copy of the NRBA for the Netherlands using freedom of information laws and provide this for interested readers in the online Appendix C.

[7] Borrowing terminology from the literature on missing data, such investigations attempt to establish whether the missing data are Missing at Random (MAR) rather than Missing Completely at Random (MCAR).

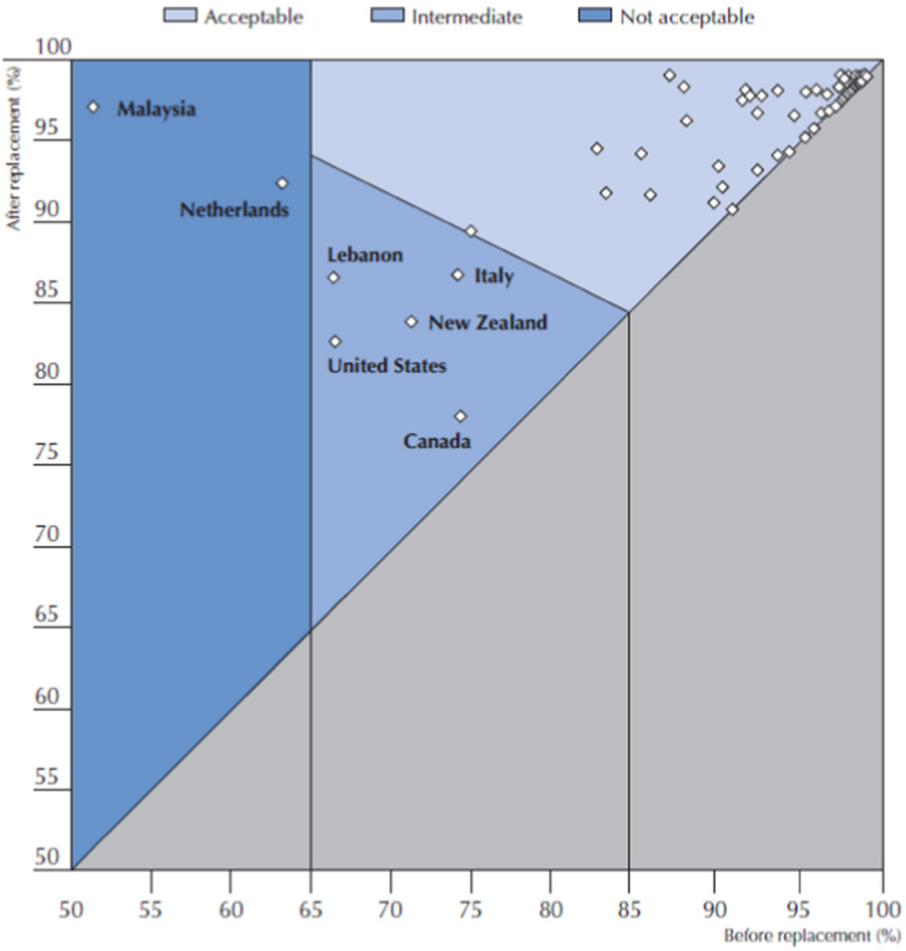


**Fig. 1** School-response rates in PISA 2015. Source: PISA (2015) technical report. Fig. 14.1.

is provided by some countries where NRBA have previously been conducted, such as Kastberg et al. (2017) for the USA. In summary, the characteristics of responding schools are compared to non-responding schools in terms of a small set of observable characteristics (usually the stratification variables included on the sampling frame plus, occasionally, some additional auxiliary variables).

The key criterion then used to determine evidence of bias seems to be whether or not any of the differences between participating and non-participating schools, in terms of the observable school-level characteristics available, were statistically significant. If there are no statistically significant differences, this seems to be treated as an indication of a lack of bias and, hence, reason for inclusion in the PISA results. Critically, full results from these NRBA are *not* routinely published by the OECD (bar a nebulous paragraph included in the depths of the technical report—OECD 2017: Chapter 14), with the information eventually provided largely left to the discretion of individual countries within their national reports.

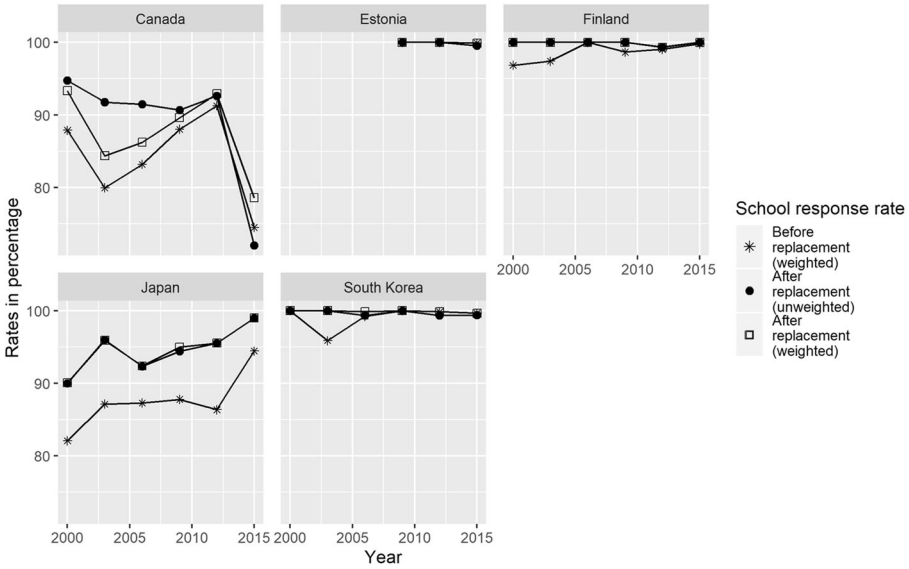


**Fig. 2** School response rates over time in selected high-performing countries

The only publicly available details about the NRBA conducted for the PISA 2015 Canada sample are provided within O'Grady et al. (2016: Appendix A). This explains that a NRBA was not conducted for Canada as a whole, but for just three provinces where school response rates were particularly low (Québec, Ontario and Alberta). The report goes on to explain how the characteristics of participating schools was compared to the characteristics of all originally sampled schools (i.e. both participating and non-participating) in terms of school funding source, language, size and recent results in provincial assessments. All analyses were conducted separately for the three provinces with schools (rather than students) being the unit of analysis.

Unfortunately, very little detail is provided about the specific analyses undertaken within the information made publicly available. Likewise, little formal detail is provided about the results (e.g. there are no tables illustrating the results of the NRBA conducted). Instead, it is simply offered that 'non-response analysis revealed no potential bias' in Ontario and that 'very few statistically significant differences were observed between the non-response adjusted estimates and the population parameter estimates' in Alberta (O'Grady et al. 2016: Appendix A). On the other hand, in Québec (a reasonably large province that accounts for approximately a fifth of Canada's population), statistically significant differences were observed and it is reported that the NRBA 'revealed potential bias'. Yet, despite this, it was concluded that 'the PISA international consortium judged that the Canadian data overall were of suitable quality to be included fully in the PISA data sets without restrictions'.

There are, however, at least two significant problems with the current approach to NRBA used within PISA, and the analysis for Canada shows:

1. Only a limited selection of variables is investigated, with the choice of these variables to some extent at the discretion of individual countries. Lack of evidence of a difference on this small handful of variables is then taken as indicating a lack

of bias. Yet it could simply reflect that countries had not looked for bias very hard. Alternatively, countries may lack the resources (i.e. staff and data) to perform a valid and truly informative NRBA, especially given the extremely tight timelines of producing the PISA reports.

2. Whether there are any *statistically significant* differences in characteristics between responding and non-responding schools seems to be the main criterion for evidence of bias. Yet with only a very limited number of schools (e.g. just 114 in the case of the NRBA conducted in Alberta)—and with the NRBA conducted at school-level—such significance tests are likely to be underpowered. In other words, the small sample size will make it extremely difficult to detect 'significant' differences between participating and non-participating schools. In fact, the *magnitude* (such as standardised differences) are of much more use and interest (Imbens and Rubin 2015). Yet, as in the example of Canada, such crucial information is not generally made publicly available.

The main consequence of the discussion above is that it is far from clear that the PISA data is indeed representative for countries that 'passed' the NRBA. Not enough detail has been published by the OECD and countries themselves (including Canada) to allow proper independent scrutiny of the matter. It is for this reason that we have used freedom of information laws to obtain and publish—for the first time—the full school-level NRBA that was conducted for Canada in PISA 2015. This is provided in the Appendix B, illustrating that, in some Canadian provinces (e.g. Ontario), there was a big difference in response rates between public and private schools. The information available within the NRBAs that are publicly available is limited, and we believe have been designed to support the inclusion of a country's data wherever possible. Indeed, in Table 3, we document all occasions where a country has been required to complete a NRBA between 2000 and 2015, noting that on 21 out of 24 occasions (88%), they have come through the process unscathed.

### 2.5 Pupil response rates

The OECD stipulates that at least 80% of students from within participating schools complete the PISA assessment. Pupils who are selected to participate may end up not participating if they are absent from school on the day of the test, they (or their parents) do not consent to participation in the study, or there were issues with how the study was conducted (e.g. as a computer-based assessment, non-participation could have been the result of a computer 'crash'). In 2015, Canada narrowly met the pupil participation threshold (81%) but, as Table 2 illustrates, this is one of the lowest rates of student response across the OECD (OECD average = 89%). Yet, as the official student response rate criteria was met, no further evidence is available about the characteristics of non-participating pupils. This is despite analysis within previous PISA cycles suggesting that students who were absent from the PISA test tend to achieve lower scores on Canadian provincial assessments (Knighton et al. 2010) and that low student participation rates might be more problematic (in terms of introducing bias into the sample) than low school participation rates (Durrant and Schnepf 2018). The fact that almost a fifth of sampled Canadian pupils within participating schools did not take the PISA test is therefore a concern (although we do not know the extent to which such non-response

**Table 3** Countries having to do a non-response bias analysis in PISA between 2000 and 2015

| Country/year | % school response | Included in report? |
|---|---|---|
| 2000 | | |
| Netherlands | 27% | Excluded |
| USA | 56% | Included |
| UK | 61% | Included |
| Belgium | 69% | Included |
| New Zealand | 77% | Included |
| Poland | 79% | Included |
| 2003 | | |
| UK | 64% | Excluded |
| USA | 65% | Included |
| Canada | 80% | Included |
| 2006 | | |
| USA | 69% | Included |
| Scotland | 64% | Included |
| UK | 76% | Included |
| 2009 | | |
| Panama | 84% | Included |
| UK | 70% | Included |
| USA | 67% | Included |
| 2012 | | |
| Netherlands | 74% | Included |
| USA | 67% | Included |
| 2015 | | |
| Malaysia | 51% | Excluded |
| Netherlands | 63% | Included |
| Lebanon | 67% | Included |
| USA | 67% | Included |
| Canada | 74% | Included |
| New Zealand | 71% | Included |
| Italy | 74% | Included |

School response rate reported before replacement

is 'selective' in Canada). Figure 3 illustrates how this is not a new problem facing the PISA sample for Canada; it has historically had both high rates of student exclusions and low student participation rates relative to other high-performing countries over the 2000–2015 PISA cycles. This could introduce bias into comparisons of educational achievement across these countries if less able students are most likely not to participate in PISA (as indicated by Heine et al. 2017) and if groups with certain characteristics (e.g. those with learning disabilities) are more likely to be excluded.

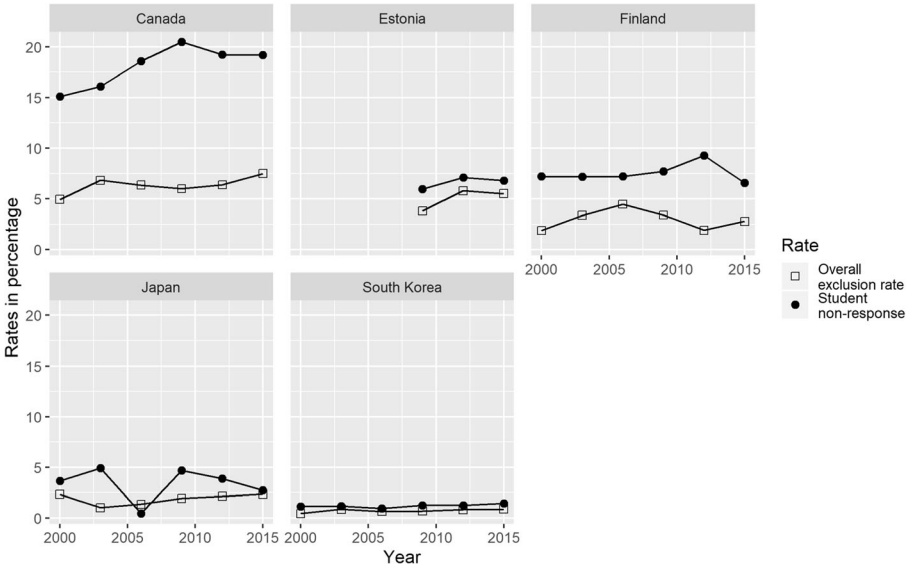


**Fig. 3** Student exclusion and non-response rates over time in selected high-performing countries

## 2.6 Weighting for non-response

The PISA database includes a set of response weights which attempt to adjust estimates for non-random non-response (amongst other functions). These weights are only as effective in reducing non-response bias as the variables used in their construction. In an ideal world, they would be both (a) predictive of non-participation and (b) strongly associated with the outcome of interest (PISA scores). This is likely to be the case in, for example, the PISA data for England, where prior school achievement in high-stakes national examinations is included in the non-response adjustment at school level.

Unfortunately, this is unlikely to hold true in the case of Canada (and potentially many other counties as well). Only the implicit and explicit stratification variables are used to adjust for non-participation at the school level. These are level of urbanisation, source of school funding and the ISCED level taught (OECD 2017: Chapter 4) which are only moderately related to PISA outcomes. For instance, when regression PISA reading scores upon the stratification variables, the R-squared statistic (i.e. the variance explained by the stratification variable) is only 0.01 for level of urbanisation and source of school funding and 0.02 for ISCED level taught. Then, at the student level, essentially, no correction for non-response has been made; as noted within the PISA technical report: 'in most cases, this student non-response factor reduces to the ratio of the number of students who should have been assessed to the number who were assessed' (OECD 2017: 122). This means that the non-response adjustment model in PISA is 'non-informative'—based upon an assumption that missing pupil data effectively occurs at random within schools (i.e. within schools data are Missing Completely At Random—MCAR). The main implication of this is that the application of the weights supplied as part of the Canadian PISA data should do little to mollify concerns over school and student non-response.

### 2.7 Summary

Table 2 also provides a summary of the combined impact of these issues upon the Canadian PISA sample in the last column. In total, the OECD estimated there to be almost 400,000 15-year-olds in Canada. Yet, through a combination of some young people not being enrolled in school, exclusions from the sample, schools refusing to participate and student absence, only around 53% of the students are assessed. This is quite some distance below the OECD average (78%) and especially far from the other high-performing OECD nations of Finland (90%), Estonia (86%), Japan (92%) and South Korea (98%).

Returning to Table 1, we note that the mean score for Canada has one of the narrowest confidence intervals. As this is the only measure of uncertainty routinely reported within PISA, it would be easy for many of its consumers to conclude that the Canadian PISA results are amongst the most secure and robust. The reality is, of course, rather different—with uncertainty due to missing data particularly acute. This highlights the need for more sensitivity analyses of the PISA results and for there to be clearer articulation by the OECD about the various different uncertainties that surround them (see Schnepf 2018). We illustrate this point in the following section.

## 3 Sensitivity analyses conducted for the Canadian results

The issues raised above mean it is important to consider the potential cumulative effect of missing data upon Canada's PISA results. Our approach to doing so consists of two sets of simulations, the first of which can be summarised as follows. We assume that students not enrolled in schools, students excluded from the study (due to, for instance, special educational needs), students in non-responding schools and non-responding pupils (within responding schools) have a different distribution of PISA achievement scores than those covered in the PISA data. As we know little about the characteristics about these students (i.e. we have no micro data about them), we make some assumptions about the distribution of the likely PISA scores for these individuals.

Our starting point is that the average PISA scores of 'non-participants' (those 15-year-olds not in school, those who were excluded from the study, those whose school chose not participate and student who chose not to participate) would be lower than those of the students who actually sat the test. For instance, students having intellectual disabilities was the main reason for student exclusion in Canada—a group defined as having low levels of academic achievement. Similarly, previous research has shown how pupil absence is more common amongst low academic achievers (e.g. Gottfried 2009), including in the context of PISA tests conducted in Canada in previous cycles (Knighton et al. 2010). It has also been shown that weaker schools are more likely to opt out of PISA (Micklewright and Schnepf 2006). We hence believe that our assumption of non-participating students being weaker academically (on average) than participating students is likely to hold. This is also consistent with what was found for Quebec in PISA 2012, where 'on average, PISA non-respondents did not perform as well as PISA respondents on the provincial test of French administered to students in Quebec' (Brochu et al. 2013).

However, one does not know how much lower non-participating students in Canada would have scored on the PISA assessment. Consequently, our sensitivity analysis essentially investigates how Canada's PISA scores would change under different assumptions made about the achievement of not enrolled/excluded/non-participating students. We are particularly interested in comparisons with four other high-performing OECD nations (Estonia, Finland, Japan and South Korea) where student exclusions and school/student non-response are much lower (see Section 2 for further details). This approach is similar in spirit to investigations of the number needed to disturb a finding (Gorard and Gorard 2016): what would the average score of non-participants need to be in order for Canada and (for example) South Korea to be equally ranked on the PISA test?

This approach is implemented as follows. First, we take the total number of 15-year-olds in Canada from the PISA 2015 technical report (396,966) and divide them into two groups: the number of participants weighted by the final student weight (210,476) and the weighted number of non-participants (186,490).[8] For the participants, we simply use their PISA scores[9] as recorded within the international database, but deflate the final student weight so that it totals 210,476. Then, for unobserved excluded/non-participating students, we randomly draw 186,490 scores from a normal distribution, assuming different values for the mean (detailed below), with the standard deviation taking the same value as for participants (e.g. 93 points in the case of reading). The values we use for the mean of this normal distribution correspond to different percentiles of the observed PISA score distribution for Canada. Specifically, we report results when assuming the mean score of excluded/non-participating students is equal to:

- The observed 45th percentile (assumed mean of non-participants = 519 in reading).
- The observed 40th percentile (assumed mean of non-participants = 507 in reading).
- The observed 35th percentile (assumed mean of non-participants = 494 in reading).
- The observed 30th percentile (assumed mean of non-participants = 480 in reading).
- The observed 25th percentile (assumed mean of non-participants = 465 in reading).
- The observed 20th percentile (assumed mean of non-participants = 449 in reading).
- The observed 15th percentile (assumed mean of non-participants = 429 in reading).
- The observed 10th percentile (assumed mean of non-participants = 402 in reading).

For each of these different scenarios, the randomly drawn scores of the 186,490 unobserved excluded/non-participating students (all of whom are assigned a weight of one) are appended to the database with the observed data for the 20,058 participants (who, when the weight is applied, total up to represent 210,476 Canadian 15-year-olds). Key results for Canada—most notably mean scores and inequality as measured by the gap between the 90th and 10th percentiles—are re-estimated, incorporating the simulated effect of exclusion/non-participants. This, in turn, allows us to consider how Canada's performance in PISA would change, particularly in comparison to other high-performing countries with much lower exclusion/non-response rates, under a set of

[8] From this point forward, we use the terms "participants" and "non-participants" for brevity. We note that "participants" are technically those who actually took the test. But, throughout this section, we use the term "participants" to mean the weighted total of those who actually took the PISA test.

[9] We use 'scores' in this context to refer to plausible values. To simplify the process, our analysis focuses upon the first plausible value only.

different plausible scenarios. We do not argue that any of our alternative scenarios are 'correct', but that some of them are at least as plausible as the results used to construct the PISA rankings, while resulting in quite different conclusions.

As Canada was the highest-performing OECD country in reading in 2015, we focus upon the robustness of scores within this domain when reporting our results. Key findings are presented in Table 4. Columns (1) and (2) provide information about the simulated average PISA reading scores of non-participants, while columns (3) to (6) illustrate revised estimates of the mean, 10th and 90th percentile of PISA reading scores in Canada following the simulated inclusion of the not-enrolled/excluded/non-participating students.

Even under the most moderate of our assumed performance distributions of excluded/non-participating students, reading scores in Canada decline dramatically with their simulated inclusion. For instance, if we assume non-participants have only slightly lower levels of achievement than participants (i.e. they would achieve the same score as those at the 40th percentile of participants), then the mean score in Canada falls to 517. This is below the average for Finland (526) and now level with South Korea (517) and Japan (516). Hence, the scores of non-participants in Canada do not need to be particularly low (only 507, which is still substantially above the OECD average of 493) to eliminate any difference between Canada and these other high-achieving nations. If we alter the assumption so that non-participants score at (on average) the 30th percentile of participants (480 points), the average score for Canada would decline to 505, which is similar to Germany (509), Poland (506) and Slovenia (505). Indeed, under the scenario that non-participants would have achieved an average score of 465 points (equivalent to the 25th percentile amongst participants), the mean score for Canada (497) would be similar to the OECD average (493).

**Table 4** Simulated PISA reading scores under differing assumptions about the likely average scores of non-participants

| 1. Non-participants achievement as a percentile of observed Canadian distribution | 2. Assumed average score of non-participants | Revised PISA scores | | | |
|---|---|---|---|---|---|
| | | 3. Mean | 4. P10 | 5. P90 | 6. P90–P10 |
| Original | 527 | 527 | 404 | 642 | 238 |
| 45 | 519 | 523 | 402 | 640 | 238 |
| 40 | 507 | 517 | 395 | 635 | 240 |
| 35 | 494 | 512 | 388 | 631 | 243 |
| 30 | 480 | 505 | 380 | 626 | 247 |
| 25 | 465 | 497 | 370 | 622 | 252 |
| 20 | 449 | 490 | 358 | 619 | 260 |
| 15 | 429 | 481 | 343 | 615 | 272 |
| 10 | 402 | 468 | 321 | 612 | 291 |

Column 1 refers to the percentile of the Canadian PISA reading score distribution that the average non-participant would have achieved had they sat the test (column 2 illustrates the actual PISA score this corresponds to). Columns 3 to 6 then illustrate how PISA reading scores for Canada would change under the different scenarios

A similar finding emerges with respect to inequality in reading achievement, as measured by the gap between the 90th and 10th percentiles. Using the data from participants only, inequality in reading achievement in Canada (238 points) is around 11 points lower than in the average OECD country (249 points). Yet, using plausible assumptions about the likely scores of non-participants, there is potentially no difference between Canada and the OECD average at all. For instance, were non-participants in Canada to achieve reading scores that were (on average) around the 30th percentile (480 points) then inequality in reading scores in Canada (247) and across the OECD (249) would effectively be the same.[10]

While our first simulations assume lower PISA scores of non-participating students, our second set of simulations use information about student background characteristics. The extensive background questionnaires give us a detailed view on participating students, their homes and the schools they attend. We now assume that the low overall participation rate in Canada makes it more likely for students of certain background characteristics, such as low SES, not to participate. For instance, if we assume that pupils who have repeated a class in the past to be twice as likely not to be covered by the PISA study than those who have never repeated a class, our simulations see the Canadian reading score drop by 10 points to 517. Similarly, the gap between the 90th and 10th percentiles increases to 250 points, indicating a higher inequality in reading achievement. To obtain these results, we manipulate the original student weights used for in the computation of the PISA scores to account for hypothesised differences in non-response. Further results and a detailed description of how we conducted this second set of simulations can be found in the Appendix A.

What do these sensitivity analyses imply for how one should interpret the Canadian PISA 2015 results? Our interpretation is that, although it remains plausible that average reading scores in Canada are above the OECD average (and inequality in achievement below the OECD average), there is not the strength of evidence to classify this country as an 'education super-power'. We think it is just as plausible that Canada's average PISA scores fall below those of four other genuinely high-performing OECD countries (Finland, Canada, Japan and South Korea) in all three core PISA domains. Likewise, it is plausible that inequality in educational achievement in Canada is quite similar to the average across OECD countries.

## 4 Conclusions

PISA is an influential large-scale study of 15-year-olds achievement in reading, science and mathematics which is now conducted in more than 70 countries and economies across the world. Results from PISA are widely reported by international media and have had a significant influence upon policymakers (and policymaking). High-performing PISA countries have received much attention, with Finland and a group of high-performing East Asian nations (e.g. Japan, South Korea, Singapore, Hong

[10] Note that results referring to inequality are also sensitive to the standard deviation used for non-participants used in the simulations. For reading, we have kept this at 93 points throughout (i.e. the same standard deviation as for participants) as it is not clear whether the reading score of non-participants would be more or less equal.

Kong) being prominent examples. Canada has also performed extremely well on the PISA tests, being lauded for its high average scores and low levels of inequality in achievement. This is striking because—given its similar language, culture, economy, political system and population size—it is a more natural comparator for many Western education systems than some of the high-performing East Asian nations. Canada has, hence, been held up as an example of a high-quality, equitable education system which leads the Western world (Coughlan 2017).

Yet are Canada's PISA results really as strong as they first seem? This issue has been explored in this paper, considering critical elements of the quality of the Canadian PISA data. We have highlighted how Canada only just meets the minimum threshold the OECD sets for several criteria, with the PISA data for this country suffering from a comparatively high student exclusion rate, low levels of school participation and high rates of student absence. The combination of these factors leads us to believe that there are serious problems with comparing the PISA 2015 data for Canada to other countries with higher response rates. It is, hence, suggested that additional sensitivity analyses should be applied to the Canadian PISA results, particularly if it is going to be compared to other high-performing OECD countries where exclusion rates are much lower and participation rates are much higher (most notably Estonia, Finland, Japan and South Korea). Our analysis shows that, under plausible scenarios, average PISA scores in Canada drop below those of these other world-leading systems, while inequality in achievement draws close to the OECD average. We hence conclude that, although it remains plausible that educational achievement in Canada is higher than in the average OECD country, there is not the strength of evidence to put it in the same class as the world's genuine top performers.

This case study of Canada has wider implications for how the PISA results are reported by the OECD. Three particular issues arise. First, the criteria the OECD sets for a country's inclusion in the results need to be tightened and how they apply these rules needs to be more transparent. In our opinion, the minimum student response rate should be raised from 80 to 90% and the 5% criterion for student exclusions much more strictly applied. Likewise, given that a school response rate below 65% is labelled 'unacceptable', countries with school participation below this level (such as the Netherlands in 2015) should be excluded. We also believe that the OECD should introduce a new criterion for the overall coverage rate to be above some minimum level, in order to avoid the situation that has emerged for Canada (where we believe the cumulative impact of being on the border line for several of the OECD's rules has led to problems).

Second, non-response bias analyses (NRBA) need to become much more thorough and transparent. We wholeheartedly believe that comparisons of respondents to non-respondents in terms of observable characteristics is a sensible and insightful approach and that such analyses should be undertaken by all countries as a matter of course (i.e. not just for countries that fall into the 'intermediate' or 'unacceptable' response zones). Although such an exercise might seem futile for those countries with very high response rates (e.g. above 95%), it would allow these nations to illustrate clearly to non-specialist audiences how the PISA data for that country really are representative of the national population. This should be done at both the school- and student-levels wherever possible, given that non-response amongst either could generate bias in the results. A 'gold-standard' example can be found in Micklewright et al. (2012), using

pupil-level administrative linked data to interrogate thoroughly the bias in England's PISA data from 2000 and 2003. We also advocate an increased focus on the magnitude of differences between participants and non-participants (e.g. standardised differences or probability differences), rather than on statistical significance. However, most importantly, full details and results from the NRBA *must* be routinely published by the OECD as part of their technical report. The current brief, nebulous summaries provided within the technical report and individual country reports are not fit for purpose. The only way the OECD (and individual countries) will inspire greater confidence in their data is by becoming more transparent about such issues. In an effort to push the OECD in this direction, we have used freedom of information laws to gain access to selected school-level NRBAs that have been produced by England (for PISA 2009), New Zealand,[11] the Netherlands and Canada for PISA 2015. We provide a selection of these in online Appendices B-D for readers to inspect. This is (to our knowledge) the first time such evidence has been made available in the public domain, and hence will help readers reach their own conclusion about potential bias in the PISA sample due to school non-response.

Finally, it is unreasonable to expect non-specialist audiences to go through the same level of detail as this paper, or to have the necessary technical understanding (and time) to decipher details that can only be found in the depths of the PISA technical report. Therefore, for each country, the OECD should provide a 'security rating' for the quality of the data that is presented in the same table as the headline PISA results. These could be based upon existing information collected (e.g. exclusion rates, school response rates, student response rates, population coverage) and be presented on a simple scale (e.g. 1* to 5*). A similar system is already being used by some organisations devoted to research use amongst the education community, such as the Education Endowment Foundation in England, and have generally been well-received.[12] Given the importance and wide interest and influence of PISA, we believe that the introduction of such a system would significantly improve understanding about the uncertainties surrounding the results.



## References

Brochu, P., Deussing, M.-A., Houme, K. and Chuy, M. (2013). Measuring up: Canadian results of the OECD PISA study. The performance of Canada's youth in mathematics, reading and science. 2012 first results for Canadians

[11] We are unable to publish the NRBA for New Zealand that was received, due to a confidentiality agreement that had to be signed in order to obtain the relevant documentation.

[12] The EEF use a 1* to 5* security rating system to help non-specialist audiences understand the 'quality' of the randomised controlled trials (RCTs) that they conduct.

aged 15. Accessed 01/05/2020 from http://cmec.ca/Publications/Lists/Publications/Attachments/318/PISA2012_CanadianReport_EN_Web.pdf.

Coughlan, S. (2017). How Canada became an education superpower. BBC news website. Accessed 08/04/2019 from https://www.bbc.co.uk/news/business-40708421.

Durrant, G., & Schnepf, S. (2018). Which schools and pupils respond to educational achievement surveys? A focus on the English Programme for international student assessment sample. *Journal of the Royal Statistical Society, Series A, 181*(4), 1,057–1,075.

Education Datalab. (2017). Why does Vietnam do so well in PISA? An example of why naive interpretation of international rankings is such a bad idea Accessed 08/04/2019 from https://ffteducationdatalab.org.uk/2017/07/why-does-vietnam-do-so-well-in-pisa-an-example-of-why-naive-interpretation-of-international-rankings-is-such-a-bad-idea/.

Feniger, Y., & Lefstein, A. (2014). How *not* to reason with PISA data: an ironic investigation. *Journal of Education Policy, 29*(6), 845–855.

Gorard, S., & Gorard, J. (2016). What to do instead of significance testing? Calculating the 'number of counterfactual cases needed to disturb a finding'. *International Journal of Social Research Methodology, 19*(4), 481–490.

Gottfried, M. (2009). Excused versus unexcused: how student absences in elementary school affect academic achievement. *Educational Evaluation and Policy Analysis, 31*(4), 392–415.

Heine, J. H., Nagy, G., Meinck, S., Zühlke, O., & Mang, J. (2017). Empirische Grundlage, Stichprobenausfall und Adjustierung im PISA-Längsschnitt 2012-2013. *Zeitschrift für Erziehungswissenschaft, 20*, 287–306. https://doi.org/10.1007/s11618-017-0756-0.

Hendrickson, K. A. (2012). Learning from Finland: formative assessment. *The Mathematics Teacher, 105*(7), 488–489.

Imbens, G. M., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: an introduction*. New York: NY, Cambridge University Press.

Jerrim, J. (2013). The Reliability of Trends over Time in International Education Test Scores: Is the Performance of England's Secondary School Pupils Really in Relative Decline? *Journal of Social Policy, 42*(2), 259–279.

Kankaraš, M., & Moors, G. (2014). Analysis of cross-cultural comparability of PISA 2009 scores. *Journal of Cross-Cultural Psychology, 45*(3), 381–399.

Kastberg, D.; Lemanski, N.; Murray, G.; Niemi, E. and Ferraro, S. (2017). Technical Report and User Guide for the 2015 Program for International Student Assessment (PISA). Data Files and Database with U.S.-Specific Variables. *National Center for Education Statistics report 095*. Accessed 08/04/2019 from https://nces.ed.gov/pubs2017/2017095.pdf.

Knighton, T.; Brochu, P. and Gluszynski, T. (2010). Measuring up. Canadian results of the OECD PISA study. Accessed 08/04/2019 from https://www.cmec.ca/Publications/Lists/Publications/Attachments/254/PISA2009-can-report.pdf

Masri, Y., Baird, J., & Graesser, A. (2016). Language effects in international testing: the case of PISA 2006 science items. *Assessment in Education: Principles, Policy & Practice, 23*(4), 427–455.

Micklewright, J. and Schnepf, S. (2006). Response bias in England in PISA 2000 and 2003. Department for Education and skills research report 771. Accessed 08/04/2019 from https://webarchive.nationalarchives.gov.uk/20130323024553/https://www.education.gov.uk/publications/eOrderingDownload/RR771.pdf.

Micklewright, J., Schnepf, S. V., & Skinner, C. J. (2012). Non-response biases in surveys of school children: the case of the English PISA samples. *Journal of the Royal Statistical Society. Series A (General), 175*, 915–938.

O'Grady, K., Deussing, M., Scerbina, T. Fung, K., & Muhe, N. (2016). *Measuring up: Canadian results of the OECD PISA study*. 2015 first results for Canadians aged 15. Accessed 08/04/2019 from https://www.cmec.ca/publications/lists/publications/attachments/365/pisa2015-cdnreport-en.pdf.

OECD. (2016). *PISA 2015 results (volume I): excellence and equity in education*. PISA: OECD Publishing, Paris. https://doi.org/10.1787/9789264266490-en.

OECD (2017). PISA 2015 technical report. OECD: Paris. Accessed 08/04/2019 from http://www.oecd.org/pisa/data/2015-technical-report/.

Pereira, M. (2011). An analysis of Portuguese students' performance in the OECD programme for international student assessment (PISA). Accessed 08/04/2019 from https://www.bportugal.pt/sites/default/files/anexos/papers/ab201111_e.pdf.

Raffe, D. (2011). Policy borrowing or policy learning? How (not) to improve education systems. CES briefing, 57. Accessed 24/04/2019 from http://www.ces.ed.ac.uk/PDF%20Files/Brief057.pdf.

Rutkowski, L., & Rutkowski, D. (2016). A call for a more measured approach to reporting and interpreting PISA results. *Educational Researcher, 45*(4), 252–257.

Schnepf, S. (2018). "Insights into survey errors of large scale educational achievement surveys" Working Papers 2018–05, Joint Research Centre, European Commission (Ispra site).

Spaull, N. (2018). Who makes it into PISA? Understanding the impact of PISA sample eligibility using Turkey as a case study (PISA 2003–PISA 2012). *Assessment in Education: Principles, Policy & Practice, 26*, 397–421. https://doi.org/10.1080/0969594X.2018.1504742.

Takayama, K., Waldow, F., & Sung, Y.-K. (2013). Finland has it all? Examining the media accentuation of 'Finnish education' in Australia, Germany and South Korea. *Research in Comparative and International Education, 8*(3), 307–325.



## Affiliations

**Jake Anders[1] · Silvan Has[1] · John Jerrim[1] · Nikki Shure[1] · Laura Zieger[1]**

Jake Anders
Jake.Anders@ucl.ac.uk

Silvan Has
S.Has@ucl.ac.uk

Nikki Shure
Nikki.Shure@ucl.ac.uk

Laura Zieger
L.Zieger@ucl.ac.uk

[1] UCL Institute of Education, 20 Bedford Way, Bloomsbury, London, WC1H 0AL, UK