



How representative is the Swedish PISA sample? A comparison of PISA and register data

Linda Borger¹ · Stefan Johansson¹ · Rolf Strietholt² 

Received: 7 February 2024 / Accepted: 4 June 2024 / Published online: 6 July 2024
© The Author(s) 2024

Abstract

PISA aims to serve as a “global yardstick” for educational success, as measured by student performance. For comparisons to be meaningful across countries or over time, PISA samples must be representative of the population of 15-year-old students in each country. Exclusions and non-response can undermine this representativeness and potentially bias estimates of student performance. Unfortunately, testing the representativeness of PISA samples is typically infeasible due to unknown population parameters. To address this issue, we integrate PISA 2018 data with comprehensive Swedish registry data, which includes all students in Sweden. Utilizing various achievement measures, we find that the Swedish PISA sample significantly overestimates the achievement levels in Sweden. The observed difference equates to standardized effect sizes ranging from $d = .19$ to $.28$, corresponding to approximately 25 points on the PISA scale or an additional year of schooling. The paper concludes with a plea for more rigorous quality standards and their strict enforcement.

Keywords Global yardstick · PISA 2018 assessment · Quality standards

1 Introduction

The Programme for International Student Assessment (PISA) is a prominent global study that aims to evaluate the math, reading, and science proficiency of students every 3 years, with the goal of comparing countries and analyzing trends within them.

Linda Borger, Stefan Johansson, and Rolf Strietholt contributed equally to this paper and are listed in alphabetic order.

✉ Rolf Strietholt
rolf.strietholt@tu-dortmund.de

¹ Department of Education and Special Education, University of Gothenburg, PO Box 300, 405 30 Gothenburg, Sweden

² Center for Research On Education and School Development (IFS), TU Dortmund University, Vogelpothsweg 78, 44227 Dortmund, Germany

The study receives considerable academic, policy, and media attention (Hopfenbeck et al., 2018; Meyer et al., 2017, Strietholt et al., 2014; Strietholt & Scherer, 2018). Ensuring that the samples drawn in each country accurately reflect the underlying population of 15-year-olds is essential for the study's quality. Comparisons across countries or over time are only meaningful if the samples are unbiased. This article focuses on the issue of exclusions and non-response as a potential source of bias, using Sweden as a case study, where a controversial debate about Sweden's score in PISA emerged after the cycle in 2018.

When the Swedish PISA 2018 results were first presented in December 2019, the Director-General of the Governmental agency Skolverket (Swedish National Agency for Education), responsible for PISA, happily announced that Sweden had significantly improved from the past PISA cycles (OECD, 2020; Skolverket, 2019). While the positive message initially gained attention, the discussion soon shifted to focus on the robustness of the Swedish results. The debate was initiated by an intense public discussion in the Swedish media, where the journalist Hellberg (e.g., 2020, 2021) heavily criticized that a large proportion of students were excluded; he argued that this exclusion led to a biased sample which was not representative for the underlying population of all 15-year-olds. The results were also questioned by researchers who contended that serious mistakes had been made (e.g., Munther, 2020).

The criticism of PISA was taken up by various parties: While Skolverket and the OECD defended the published results (OECD, 2020; Skolverket, 2021), the national audit agency supported the critique on the high level of exclusions in Sweden (Swedish National Audit Office, 2021; Andersson and Sandgren Massih, 2023). The audit agency interviewed approximately 10% of the school coordinators involved in PISA and concluded that many of them had misunderstood the exclusion criteria, leading to an excessive number of students being improperly excluded. Overall, the audit agency's report and the corresponding publication in an academic journal constitute a harsh critique of Skolverket's procedures and practices.

To date, deliberations on the representativeness of the Swedish PISA 2018 cohort have largely been theoretical, with positions ranging from those minimizing the impact of student exclusions to those asserting their substantial influence. The aim of this study is to elucidate this debate using empirical evidence. To this end, we use registry data of all Swedish students and compare the distribution of student characteristics and their performance with the distribution in the PISA sample.

2 Literature review

The essence of sampling is to facilitate inferences about a population from a subset thereof (Cochran, 1977). When a sample is drawn randomly, it allows for extrapolation of population parameters. Sophisticated sampling methodologies are commonplace in international large-scale assessments (Meinck & Vandenplas, 2021; Strietholt and Johansson, 2023); nonetheless, the random selection of students remains a critical element in all sampling plans and is crucial for the overall quality of such studies. The extent to which PISA samples accurately reflect the underlying student population is a subject of both past and current scholarly inquiry.

Anders et al. (2021) discuss that non-participation of students is a threat to the representativeness of the Canadian PISA sample. They not only demonstrated the limited coverage of the PISA 2015 sample for Canada, encompassing only about half of the 15-year-old cohort, but also conducted simulation studies to assess the impact of this limitation. Their simulations revealed that even slight differences between students who participated in PISA and those who did not can significantly distort the average student performance. In a similar vein, Andersson and Sandgren Massih (2023) conducted simulations for Sweden, exploring various scenarios to demonstrate how the exclusion or non-participation of students with low performance can quickly lead to a biased sample that significantly overestimates the performance level in Sweden.

A limitation of the studies by Anders et al. and Anderson and Sandgren is that the performance of students who should have, but did not, participate in PISA remains unknown (but is based on untestable assumptions). Consequently, any conclusions about the extent of potential biases are, to some degree, speculative. Jerrim (2021) addressed this issue by utilizing data on the performance distribution from the GCSE (General Certificate of Secondary Education; a set of national tests) in the UK. This allows for a comparison between the PISA sample's performance distribution and that of the full population, offering insights into the representativeness of the PISA sample and the magnitude of any biases. Based on the performance differences observed in the GCSE, Jerrim projected that the PISA study overestimates the performance level in England and Wales by 11 and 14 points, respectively, on the PISA scale. This assessment suggests a significant upward bias in PISA's estimation of student achievement within these regions. These findings confirm earlier research from England using PISA data combined with national tests (Durrant & Schnepf, 2018; Micklewright et al., 2012).

The comparability of PISA data over time has been a subject of scrutiny in various countries, with the primary issue being that different student groups are considered eligible to take the PISA test in different years. This eligibility variance can lead to inconsistencies in trend comparisons within countries over time. For instance, Spaul (2018) has raised questions regarding the comparability of PISA data for Turkey, while Pereira (2011) has expressed similar concerns for Portugal. Jerrim (2013) questions the comparability across PISA cycles criticizing the comparability of the data quality for England. These critiques suggest that changes in the populations deemed eligible for PISA across assessment cycles may affect the interpretation of educational progress or decline within these countries.

The paper proceeds as follows: First, we delve into the PISA sampling model and discuss different threats to the representativity of the sample. Subsequently, we utilize Sweden as a case study to examine the sampling outcomes of PISA 2018. Our empirical analysis involves contrasting PISA data with Swedish registry data across diverse measures of achievement. We conclude with a discussion of our empirical findings and advocate for a more rigorous sampling approach.

3 The PISA sampling model

PISA does not test all 15-year-old students, but a sample in each of the participating countries. The PISA sampling model involves multiple stages, including the selection of schools and students within schools. In the first stage, a sample of schools is selected from a list of all eligible schools in each country. In the second stage, a sample of students is selected from each selected school. If the sampling at each stage is at random, it is possible to infer from the distribution of the sample parameters the corresponding characteristics in the population of all 15-year-olds. The process of random sampling is thoroughly documented by the OECD including in PISA technical standards, sampling guidelines, and the technical report chapters on the sampling design and the sampling outcomes (the documents can be found on OECD's PISA website: oecd.org/pisa/publications/). However, there are two sets of threats to the representativeness of the PISA sampling, namely exclusion and non-response, we will elaborate on these issues in the following.

3.1 Exclusions

Exclusions of entire schools or individual students pose a significant threat to the representativeness of samples and make it difficult to compare data across countries or over time. If a country systematically excludes low-performing students from the PISA test, the average performance observed in the sample may overestimate the actual performance level of all 15-year-olds in the country. To mitigate this risk, the OECD permits only a few exceptions for countries to exclude individual schools or students from PISA. The PISA sampling standards permit countries to exclude up to 5% of the relevant population of 15-year-old students, either by excluding entire schools or individual students within schools (OECD 2015, 2019a, p 159). However, these exclusions must meet specific criteria:

- At the school level, exclusions are limited to schools that are geographically inaccessible or where the administration of the PISA assessment is not considered feasible and schools that only provide teaching for specific student categories, such as schools for the blind. The percentage of 15-year-olds enrolled in such schools must be less than 2.5% of the nationally desired target population, with a maximum of 0.5% for the former group and 2% for the latter.
- Exclusions at the student level are also restricted. Students with intellectual or functional disabilities that would prevent them from performing in the PISA testing environment, and those with limited assessment-language proficiency, who are unable to read or speak any of the languages of assessment in the country at a sufficient level,¹ may be excluded. However, exclusions cannot be made solely for low proficiency or disciplinary issues. The percentage of 15-year-olds excluded within schools must also be less than 2.5% of the national desired target population.

¹ Exclusions of students are limited to students that are not proficient in the assessment language as native speakers (immigrants) and have limited proficiency in the assessment language and who have received less than one year of instruction in the assessment language.

The PISA standards, by design, aim to ensure ambitious sampling practices that mitigate risks to the representativeness of PISA samples. However, adherence to these standards has been inconsistent. Annex A2 of the 2018 PISA international report reveals that 16 participating countries, 14 of which are OECD members including Sweden, exceeded the prescribed maximum overall exclusion rate of 5% (OECD, 2019b).

3.2 Non-response

School or student non-response poses another threat to the representativeness of samples and makes it difficult to compare data across countries or over time. To mitigate this risk, PISA quality standards required minimum participation rates for both schools and students to minimize potential bias from non-response:

- At the school level, at least 85% of the schools initially selected were required to agree to conduct the test. If the response rates of schools were between 65 and 85%, replacement schools were used to improve participation rates and reach the 85% threshold.
- At the student level, a total participation rate of 80% of the students within participating schools was required. Student participation rates were calculated over both original and replacement schools.

In the PISA 2018 assessment, five countries did not meet the 85% threshold but met the 65% threshold at the school level, and three of these countries still failed to reach an acceptable school participation rate even after replacement. The Netherlands did not even meet the 65% school response rate threshold but reached an 87% response rate upon replacement. The student level 80% threshold was met in every country except Portugal, where only 76% of students who were sampled actually participated.

3.3 Consequences of non-compliance with quality standards

Quality standards can only secure quality when they are genuinely followed. However, as demonstrated by Anders et al. (2021), certain countries have been included in PISA reports despite not meeting the criteria established by the OECD since the inception of the PISA cycles. In other words, the self-imposed quality criteria by the OECD are often disregarded and this raises doubts about whether the samples accurately reflect the performance of the population. This observation was also evident in PISA 2018, where all the countries mentioned in the preceding paragraphs, despite failing to meet the quality standards, were still incorporated into the reports (e.g., OECD, 2019a).

To legitimize reporting results for countries that do not meet pre-established standards, the OECD requires the performance of so-called non-response bias analyses (NRBA) for these countries. It is crucial to recognize that the representativeness

of a sample is only impacted by exclusions or non-responses if the non-participating or excluded students exhibit systematic differences from the students in the sample. As such, the primary objective of NRBA is to determine whether this is the case. A fundamental problem with NRBA, however, is the lack of observable key variables of interest, such as student performance in the PISA test. Consequently, the quality of NRBA is inherently contingent on the availability and adequacy of reference data. However, as Anders et al. (2021) note, the OECD provides little guidance for conducting NRBA, does not release them, and the technical reports contain only very brief statements in which countries are certified that the NRBA supports the representativeness of the samples. Consequently, assessing the representativeness of PISA samples is often challenging. Next, we will examine this issue more closely using the Swedish PISA sample from 2018 as an example.

4 A case study of PISA 2018: exclusions and non-response in Sweden

Table 1 overviews the exclusion and non-response rates for Sweden for all previous PISA cycles. In 2018, the highest combined exclusion rates across all countries were observed in Sweden. The national team that implemented the study excluded 1.38% of the school, which is well in accordance with PISA technical standards that permit up to 2.5%. However, the within-school exclusions were 9.84%, which was higher than in any other country and not in accordance with PISA technical standards that permit up to 2.5%. The school and student exclusions accumulate to a total exclusion rate of 11.08%,² which is well above the 5% threshold (OECD, 2019a, Annex A2, Table I.A2.1). In terms of participation rates, Sweden achieved a participation rate of 99% for schools and 86% for students (OECD, 2019b, Annex A2, Table I.A2.6), exceeding the thresholds set by the OECD of 85% and 80%, respectively.

The high within-school exclusion rates observed in 2018 are significantly higher than in all previous PISA cycles. Although Sweden has consistently exceeded the 2.5% threshold since 2000, the within-school exclusions were notably lower at 3.43 (2000), 2.80 (2003), 2.67 (2006), 2.89 (2009), 3.84 (2012), 5.71 (2015), and 6.26 (2022) compared to the 9.84% (2018) observed in the 2018 cycle. These varying exclusion rates raise the question of whether the PISA 2018 results can provide meaningful insights into trends in Sweden or not. On the other hand, the school-level exclusion rates as well as the school and student non-response rate are more stable over time and in all cycles below the thresholds set by the OECD.

5 The present study

We investigate the representativeness of the Swedish PISA sample focusing on student achievement. To this end, we draw on register data provided by Statistics Sweden. Specifically, we examine if the distribution of performance measures by GPA

² The total exclusion rate is calculated as follows: $0.0138 + (1 - 0.0138) \times 0.0984 = 0.1108$.

Table 1 Exclusion and non-response rates for Sweden in PISA 2000–2018 (all in %)

	2000	2003	2006	2009	2012	2015	2018	2022
Exclusions								
School-level exclusion rate	1.35	1.44	1.83	1.91	1.67	1.25	1.38	1.20
Within-school exclusion rate	3.43	2.80	4.46	2.89	3.84	4.51	9.84	6.26
Non-response								
Weighted school participation rate before replacement	99.96	99.08	99.59	99.91	98.92	99.70	98.75	98
Weighted student participation rate after replacement	87.96	92.61	91.37	92.97	92.17	90.67	86.46	85

Sources are PISA Technical Reports (Adams & Wu, 2002; OECD, 2005, 2009, 2012, 2014, 2017, n.d.). The PISA report for 2022 does not provide decimals for the school and student participation rates

and national tests differs from that in the full population. Further, we use three key measures of social background: parental education as a proxy for socioeconomic status, immigrant background, and student gender. These variables are central to the reporting and research based on PISA data.

6 Methods

6.1 Combining PISA with register data

In this study, we investigated the Swedish PISA 2018 data, consisting of 5504 students, which was utilized for both national and international reports on PISA 2018 (OECD, 2019a, 2019b; Skolverket, 2019). PISA assessed students who were aged between 15 years and 3 months and 16 years and 2 months at the time of the assessment and who were enrolled in schools in grade 7 or higher. In Sweden, this target sample definition corresponds to all students born in 2002. To enhance the dataset, additional variables from Swedish registers, such as various measures of student social background and achievement, were added to the PISA dataset through the unique personal ID assigned to all residents in Sweden, known as the personal number. This extended PISA dataset was made available by the Swedish National Agency for Education (Skolverket), which was responsible for implementing PISA 2018 in Sweden. Out of the initial 5504 students in the PISA sample, the large majority of the students granted permission for linking their PISA data with registry data using their personal number, resulting in having both PISA and registry data for those 5432 students (98.7%), while PISA data alone is available for 72 students (1.3%).

In order to investigate the representativeness of a sample, it is necessary to have knowledge of the population parameters of interest. For this purpose, we use register data that contain data of all students in Sweden who were born in 2002. The register data we used in the present study contains 106,653 students. As the register data contains information of the entire population, there is no sampling variance and population parameters can be calculated without any sampling error. Access to the register data was available through the Gothenburg Educational Longitudinal

Database (GOLD). This database includes education-related information such as grades and national test results, and the data is stored at Statistics Sweden, the government agency responsible for the collection and processing of register data.

6.2 Measures

6.2.1 Student achievement

To compare the performance distribution in the PISA sample with the entire population, we use different performance measures: the results of standardized national tests in Swedish and English and the grades assigned by teachers.

Grade point average (GPA) We measure student performance using a GPA score, which is the sum of all student grades and a measure of overall academic performance. The score is based on the sum of the 16 final subject grades a student receives in grade 9. Sweden has a criterion-referenced grading system where teachers evaluate students' knowledge against specific grading criteria using a six-step scale (A–F), with A representing the highest passing grade and F denoting a failing grade. The letter grades correspond to the following numeric values: $A = 20$, $B = 17.5$, $C = 15$, $D = 12.5$, $E = 10$, $F = 0$. The GPA can range from 0 to 320, accounting for students with the lowest or highest grades across all 16 subjects.

Teachers have exclusive responsibility for assessment and grading in the Swedish system and assign subject grades at the end of a course or the semester. Grading is centrally regulated by the national curriculum, but local autonomy exists in the selection of assessment methods and strategies for assigning grades. Externally developed national tests for core subjects such as Swedish, mathematics, and English are provided to support comparable and fair grading. The GPA is used as a selection criterion for students when they apply to upper secondary school, adding a high-stakes dimension to grades in the Swedish school system.

National tests Results from mandatory national tests in grade 9 for Swedish and English were additionally employed as metrics of student achievement.³ The primary aim of these national tests is to support equitable grading. The tests, focusing on the subjects of Swedish and English, are based on the respective curricula and syllabi. The Swedish test includes speaking, reading, and writing sections, while the English test includes speaking, receptive skills (listening and reading), and writing. Test scores from the three components are aggregated and reported on the same six-step grading scale ($A = 20$ to $F = 0$) that is used for subject grades.

National tests have high stakes for students because teachers are expected to use and, as per a 2018 regulation, give special consideration to the results from these assessments when assigning grades. It is worth noting that in the current system,

³ The national tests in mathematics were excluded from this analysis due to a security breach in 2018 that resulted in their leakage.

no central marking occurs at the national level. Teachers are responsible for grading these tests, with a strong recommendation for collaborative marking but no formal requirement. The national tests are administered in spring, which coincides with the administration of PISA tests.

6.2.2 Social background

We investigated whether the PISA sample is representative of the population by examining three key student background characteristics available in Swedish register data.

Immigration The variable immigrant status combines register information on the birth country of the child and their parents and has three categories (Swedish background, born in Sweden with two foreign-born parents, foreign-born).

Gender The gender variable in this study is binary, distinguishing between boys and girls. There were no missing values for gender in the register data. In the PISA data, student gender was obtained from the PISA tracking form, which lists the gender of all students included in the data set with no missing data.

Parental education The variable parental education was formed on the basis of the highest educational attainment of both parents as reported in Swedish registers. This variable can be categorized into seven values, with the first four indicating no higher education and the remaining three representing higher education: (1) pre-secondary education less than 3 years, (2) pre-secondary education 9 [10] years (equivalent), (3) secondary education up to 2 years, (4) secondary education longer than 2 years but not more than 3 years, (5) post-secondary education less than 3 years, (6) post-secondary education 3 years or more, (7) postgraduate education).

6.3 Missing data

There are two forms of missing data in the present study. First, students in the PISA study had to consent to have their data merged with the registry data. Of the $n=5,504$ students in the PISA sample, $n=72$ refused to consent. Second, there are missing values in the variables from the registry data, such as when students were sick on the day the national tests were administered or when their parents were unknown. Table 2 shows that the proportions of missing data by item for PISA and register data range between 0 and 21%.

In two separate analyses, we used multiple imputation to impute missing values ten times using the R package mice (Buuren & Groothuis-Oudshoorn, 2011). The imputation model for the register data includes all variables listed in Table 2. For the PISA data, we added further test and survey data from PISA, which we considered particularly valuable for students who did not consent to have their PISA data merged with the registry data. Specifically, we used the PISA variable immigration

Table 2 Missing values in PISA and register datasets

	PISA (sample)			Register (full population)		
	Valid	Missing	% Missing	Valid	Missing	% Missing
GPA	5.418	86	1%	106.653	0	0%
National test Swedish	4.672	832	15%	86.408	20.245	19%
National test English	4.347	1.157	21%	84.566	22.087	21%
Immigration	5.432	72	1%	106.653	0	0%
Gender	5.504	0	0%	106.653	0	0%
Parental education	5.393	111	2%	106.058	602	1%
	<i>N</i> = 5.504			<i>N</i> = 106.653		

Unweighted sample size

status (variable name in the PISA dataset: IMMIG), occupational status (HISEI), parental education (PARED), socioeconomic status (HOMEPOS), math achievement (PV1MATH), reading achievement (PV1READ), and science achievement (PV1SCIE). The predictive mean matching approach was applied as the imputation method for all variables.

6.4 Analysis

To assess whether the 2018 PISA sample accurately represents Sweden, we conducted comparisons between the PISA sample and population parameters obtained from registry data. We examined the distribution by considering both the mean and the standard deviation.

In the case of PISA data, we utilized sampling weights to estimate the parameters. We applied the BRR (Balanced Repeated Replication) method to calculate the sampling variance, performed the analyses across all 10 plausible values to address imputation variance, and then amalgamated this information to compute the standard error using Rubin's rules (1976).

In the case of registry data, as we were dealing with data encompassing the entire population, there was no need to compute the sampling variance. However, to account for the imputation of missing data (if any), we conducted analyses across all 10 imputed datasets to quantify the imputation variance. In this context, the standard deviation among the results for the 10 imputed datasets effectively represents the standard error. Consequently, variables that contain no missing data exhibit a standard error of zero.

To compare the PISA sample with register data, we computed the mean differences. To compute the standard error of these mean differences, we utilize the square root of the sum of the squared standard errors for PISA and register data. Partially, this equates to the standard error for PISA, since the standard error for the registry data is zero for variables without missing data.

Table 3 Comparing the achievement distribution in PISA with register data

	PISA		Register		$M_{\text{PISA}} - M_{\text{Register}}$	d
	M	SD	M	SD		
GPA	232.80 (0.77)	53.17 (1.29)	217.97 (-)	68.56 (-)	14.85*** (0.77)	0.28
National test Swedish	13.68 (0.06)	4.26 (0.11)	12.80 (0.01)	5.03 (0.01)	0.88*** (0.06)	0.21
National test English	15.45 (0.05)	4.12 (0.06)	14.65 (0.01)	4.96 (0.01)	0.80*** (0.05)	0.19
N	5.504		106.653			

Weighted statistics for PISA. Standard errors in parentheses (the hyphen symbol “-” indicates a standard error of zero because there is no imputation error). The formula $d = (M_{\text{PISA}} - M_{\text{Register}}) / \text{SD}_{\text{PISA}}$ was applied to compute Cohen’s d . Significance levels * < 0.050 , ** < 0.010 , *** < 0.001

7 Results

7.1 The PISA sample overestimates the achievement level in Sweden

To determine if the PISA sample provides unbiased estimates of the achievement distribution of all 15-year-old Swedish students, we compared it with the statistics of the full population using register data. Our key finding is that the PISA sample is not representative for the full population of 15-year-old students in terms of student achievement but overestimates the actual achievement level in Sweden. Specifically, we find that the average GPA in the PISA sample is 15.64 points higher than in the full population of all 15-year-olds and this difference corresponds to a standardized effect size of Cohen’s $d=0.28$ (see Table 3).⁴ In further analyses, we used the scores from national tests in Swedish and English to compare the achievement level in PISA with the full population. These analyses confirm the finding as the average performance in the PISA sample is considerably higher than the performance of all students in the Swedish population with standardized effect sizes of $d=0.21$ and $d=0.19$, respectively. All observed mean differences are statistically significantly different from zero.

How much does the PISA sample overestimate the performance level of all 15-year-olds in Sweden on the PISA metric? We cannot answer this question directly since not all children in Sweden participated in the PISA test—only those within the PISA sample did. However, we can make an estimation based on the observed differences between grades (GPA) and the results of standardized national tests.⁵ Sweden scored an average of 506 points with a standard deviation of 107 points on the

⁴ We used the standard deviation observed in the PISA sample for the standardization as it allows us to project the observed differences based on GPA and national test scores to the PISA sample.

⁵ In the Swedish PISA sample ($n=5504$), we observe that the PISA reading scores correlate at $r=0.61$ with the GPA scores, at $r=0.63$ with the national test in Swedish, and at $r=0.58$ with the national English language test.

PISA reading scale (OECD, 2019a, p. 62). Effect sizes ranging from $d=0.19$ to 0.28 correspond to 20 to 30 points on the PISA reading scale.⁶ According to Woessmann (2016, p. 6; see also Avvisati & Givord, 2023), “As a rule of thumb, learning gains on most national and international tests during one year are equivalent to between one-quarter and one-third of a standard deviation, which amounts to 25–30 points on the PISA scale.” Using this guideline, we can infer that the PISA results overestimate the performance of Swedish students by approximately one academic year.

7.2 The PISA sample underestimates the dispersion in achievement in Sweden

The standard deviation of the achievement scores is a measure of inequality that quantifies the variation of test scores between low- and high-performing students. Table 2 shows that the standard deviation of the GPA in the PISA sample is considerably smaller than the dispersion in the full population of all 15-year-olds. The variance ratio is 0.78 (52.96/69.55), suggesting that PISA underestimates the actual dispersion in student achievement by 22%. A similar pattern can be observed for national test scores. The variance ratios are 0.85 (4.26/5.02) and 0.83 (4.12/4.96) for Swedish and English, respectively.

We observed that the variation in performance within the PISA sample is considerably smaller than in the full student population. This discrepancy raises a pertinent question: does the PISA sample lack representation from students who perform at either the lower or higher ends of the academic spectrum? Considering the previously presented data regarding average student performance, a direct answer can be formulated. When examining the entire distribution of student abilities, it becomes evident that the PISA sample disproportionately excludes students with weaker academic performance.

7.3 Children of low-educated parents and with immigrant backgrounds are underrepresented in the PISA sample

In addition, we compared the social composition of the PISA sample to that of the entire population. Table 4 illustrates that students with an immigrant background and those with less-educated parents are underrepresented in the PISA sample. Conversely, native students and children of highly educated parents are overrepresented. Regarding student gender, our observations indicate an equal distribution of boys and girls in the PISA sample, whereas boys are slightly overrepresented in the full population. Nonetheless, the observed differences in social background variables are minor in comparison to the more substantial differences in achievement.

7.4 Robustness analyses: is recent immigration an issue?

The registry data includes all children who were in Sweden in 2018, including those who immigrated after the administration of the PISA tests in the spring of 2018. Obviously, they could not be sampled and tested in PISA. Do the results of our

⁶ Calculation: $107 \times 0.19 = 20$ and $107 \times 0.28 = 30$.

Table 4 Comparison of PISA and register data

	PISA	Register
Immigration		
Swedish background	0.770 (0.007)	0.737 (-)
Born in Sweden with two foreign-born parents	0.107 (0.004)	0.111 (-)
Foreign-born	0.123 (0.004)	0.151 (-)
Gender		
Boy	0.500 (0.003)	0.5146 (-)
Girl	0.500 (0.003)	0.485 (-)
Parental education		
Pre-secondary education less than 3 years	0.022 (0.002)	0.005 (0.000)
Pre-secondary education 9 (10) years (equivalent)	0.022 (0.001)	0.104 (0.000)
Secondary education up to 2 years	0.135 (0.003)	0.132 (0.000)
Secondary education longer than 2 years but not more than 3 years	0.224 (0.004)	0.176 (0.000)
Post-secondary education less than 3 years	0.185 (0.003)	0.179 (0.000)
Post-secondary education 3 years or more	0.383 (0.005)	0.374 (0.000)
Postgraduate education	0.029 (0.002)	0.029 (0.000)
<i>N</i>	5.504	106.653

comparison with the PISA sample change when we exclude these newly immigrated students? To test this, we excluded all 383 children who immigrated in 2018 to replicate our analyses with the remaining $n = 106,205$ students. We observed that the mean GPA for these students stands at 218.54, with a mean score of 12.85 on the National Test for Swedish and 14.71 on the National Test for English. These scores demonstrated a slight improvement when compared to those in the entire dataset, yet the differences are so minimal as to be inconsequential.

Another challenge related to immigration is that PISA allows countries to exclude students who have been in the test country for less than a year if they do not yet speak the test language. Therefore, the high exclusion rates in Sweden may be partially “legitimate.” Recently immigrated children who speak Swedish, on the other hand, should not be excluded (e.g., Norwegian is similar to Swedish, there is a Swedish-speaking minority in Finland, and children raised bilingually with

a Swedish-speaking parent). To examine this more closely, we excluded 734 students from the registry data who immigrated to Sweden in 2017 or 2018 and did not receive a GPA score in 2018 but one year later. The rationale behind this approach is that newly arrived students in the Swedish school system receive a GPA score for grade 9 once they have achieved a sufficient level of proficiency in Swedish. We observe that the mean GPA for the remaining $n = 105,919$ students stands at 218.72, with a mean score of 12.88 on the National Test for Swedish and 14.73 on the National Test for English. Once more, these scores show minimal variation when compared to the results presented for the entire dataset.

8 Discussion

In this study, we explored the representativeness of the Swedish PISA sample by integrating PISA 2018 data with comprehensive Swedish registry data encompassing all students born in 2002 in Sweden. Our analysis focused on different achievement measures, revealing a significant overestimation of achievement levels by the Swedish PISA sample. We found standardized effect sizes ranging from $d = 0.19$ to 0.28, translating to approximately 25 points on the PISA scale or the equivalent of an additional year of schooling. These insights point to potential biases in PISA estimates, underscoring the need for more stringent sampling quality standards and enforcement to ensure an accurate representation of student performance.

Firstly, while PISA's target population does not include all children but rather those enrolled in school, one might hypothesize that the PISA sample may lack children who have left school early, potentially those with lower academic performance. However, in Sweden, school attendance is compulsory for all 15-year-olds. Thus, the absence of early school leavers cannot account for the observed discrepancies in the PISA results.

Secondly, PISA's definition of the target population excludes 15-year-old children who have not yet reached the 7th grade. In contrast, the register data encompasses all children born in 2002, regardless of their grade. It is reasonable to conjecture that older students in lower grades would generally perform poorly; this might justify the observed differences when compared to the entire cohort of 15-year-olds, including those who were in the 7th grade when PISA was administered. However, the OECD (2019a, 2019b, p. 174) indicates that in 2018, all 15-year-old students were in grades 8 (2.1%), 9 (96.3%), or 10 (1.6%). Consequently, this factor cannot account for the overestimated performance in the PISA sample.

Secondly, PISA's definition of the target population excludes 15-year-olds who have not yet reached the 7th grade (e.g., due to late school entry or class repetitions). It is plausible to assume that older students who are in lower grades would generally perform poorly; this could legitimize the differences observed when compared to the entire cohort of 15-year-olds including those who were in 7th grade when PISA was administered (e.g., Strietholt et al., 2013). However, the OECD shows that 2.1% of 15-year-olds were in the 8th grade, 93.3% were in the 9th grade, and 1.6% were in the 10th grade. Since there were no students in the 7th grade, it is implausible to suggest the presence of children in lower grades, thus this cannot explain the overestimation of performance in the PISA sample.

Thirdly, PISA allows for the exclusion of children who have recently immigrated and have not yet mastered the test language Swedish. Yet, there is no obligation to exclude these children, which aligns with the educational goal of teaching newcomers the national language. Furthermore, the exclusion is permitted only if the children have been in the country for less than a year. However, we conducted different robustness checks where we excluded recent immigrants from the register data and replicated our analyses. The differences between the PISA sample and the new population parameters remained largely unchanged, albeit slightly reduced.

The findings indicate a need for more rigorous scrutiny of PISA sampling methods. While PISA offers valuable insights into global education standards and outcomes, its methodologies must be continually evaluated and refined to ensure accuracy and fairness in its representational scope. This is especially crucial as educational policymakers and stakeholders increasingly rely on PISA data to shape educational reforms and initiatives. Therefore, understanding and addressing the limitations and potential biases in PISA is essential for its continued relevance and effectiveness as a tool for educational assessment and improvement.

One question arises: what are the reasons for the discrepancies between the PISA sample and the population parameters? It is very unlikely that the sampling process for PISA selected only high-performing students at random. Rather, as we have noted, in Sweden, about 10% of students were excluded from the sample—that is, roughly one in every ten students. In contrast, many countries participating in PISA reported exclusion rates of less than 1% (for instance, Belgium, Mexico, Korea, Colombia, France, Japan, Czech Republic; cf. OECD, 2019a, p. 164). The subjectivity inherent in the legitimacy of school or student exclusions is highlighted by Aursand and Rutkowski (2021), who noted the varied interpretations of the OECD criteria by Norwegian school principals. Another contributing factor could be non-participation rates. In Sweden, only 86% of students who were supposed to participate in PISA actually did so. While low participation rates represent an international phenomenon that can also be observed in other countries (e.g., Jerrim, 2021), there are many countries that achieved participation rates of 95% or higher (for example, Greece, Japan, Korea, Luxembourg, Turkey; cf. OECD, 2019a, p. 172). It appears likely to us that the lack of representativeness in the PISA sample is a product of both exclusions and non-participation—further research on this matter is essential.

9 Conclusions

Our study revealed a consistent overestimation of student performance in the PISA sample, with effect sizes suggesting a discrepancy equivalent to about a year of schooling. This discrepancy raises critical questions about the sampling process, which does not seem to occur by chance but is indicative of systematic exclusions and non-participation. It is essential to emphasize that our findings go beyond the scope of the present study. We strongly believe that exclusions and non-participation likely affect representativeness in other countries, across different years, and in various studies. Subsequently, we will pinpoint several strategies to address this issue.

9.1 (More) rigorous quality standards

Exclusions and non-participation are not problems specific to PISA. Nevertheless, it seems crucial to critically evaluate the quality standards of PISA. In some instances, the exclusion of certain schools or students may be necessary, and achieving 100% participation from all schools and students can be challenging. PISA quality standards allow countries to exclude up to 5% of schools or students under certain circumstances (in other words, 95% must be included). Furthermore, school-level participation rates of 85% and student-level participation rates of 80% are deemed acceptable. Multiplying the specified values results in 65% ($0.95 \times 0.85 \times 0.80$). In other words, the OECD deems it acceptable if one-third of the students who should participate in PISA do not or were excluded. From our point of view, these thresholds are too lenient. We advocate for more rigorous quality standards. Various countries around the world achieve very low exclusion rates and high participation rates, which demonstrates what is possible in the realm of surveys with dedication and effort. For example, while participating in PISA is compulsory for schools and students in some countries, other countries give them the opportunity to opt-out. Making participation obligatory appears to be a straightforward measure to minimize bias in the sample. Once again, PISA serves as an illustration of the issue, but it is worth noting that other studies encounter the same problem.

9.2 Adherence to quality standards

Standards that are not followed are not worth the paper they are written on. As we have reported in the section *The PISA Sampling Model*, the quality standards—which we consider too lenient—are not being met by many countries, and this is a recurring issue. Countries such as the USA and UK are notorious for their poor participation rates (Anders et al., 2021), which is a situation we find untenable. A draft of the technical report of PISA 2018 has commented for several countries “in consideration of the fact that the nature and amount of exclusions in 2015 was similar (...), data were included in the final database.”⁷ This viewpoint suggests that a uniform lack of data quality over successive cycles could be deemed acceptable. We dissent from this opinion and maintain that such a practice may lead to biased comparisons across study cycles.

9.3 Not representative, but comparable over time?

A particular value of international assessments is its ability to describe trends, that is, the development of a country's performance over time (e.g., Johansson & Strietholt, 2019; Strietholt & Rosén, 2016). One might argue that exclusions and non-participation do not distort such trends (we presume that this assumption underlies

⁷ <https://www.oecd.org/pisa/data/pisa2018technicalreport/PISA2018-TecReport-Ch-14-Data-Adjudication.pdf>

the quote referenced in the previous paragraph). However, this would only be the case if both the proportion of exclusions and non-participations, the position of the excluded or non-participating schools and students within the performance distribution as well as the spread of the distribution remained constant. Whether both conditions are met is questionable. In Sweden, we observed (see Table 2) that on a school level, the rates of exclusion and participation are nearly constant. However, at the student level, there are significant differences between the years. More importantly, we do not know which schools and students were excluded or did not participate in each year. In the current study, we have analyzed data from 2018 because, in that year, Swedish registry data was linked to PISA for the first time. To what extent exclusions and non-participation have distorted performance estimates in previous years cannot be empirically determined by us. However, for PISA 2022, a linkage with registry data was possible again, allowing us to at least investigate prospectively whether PISA provides unbiased estimates of performance development.

In conclusion, our research offers compelling evidence of the overestimation of student performance within the PISA sample for Sweden, raising profound concerns about the representativeness of this influential international assessment in Sweden. The magnitude of discrepancy—notably equivalent to a year’s worth of educational progress—necessitates a reevaluation of PISA’s methodological rigor and adherence to quality standards. It is imperative that the global education community and policymakers critically address these biases to safeguard the integrity of educational assessments. As we continue to shape the future of education based on such data scientific rigor is essential.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was financially supported by Riksbankens Jubileumsfond under grant number P20-0095.

Data availability The authors are willing to provide the code used for the analyses upon request. The data supporting the findings of this study are not publicly available due to sensitivity concerns. Swedish register data from the GOLD database are stored in controlled-access data storage at Statistics Sweden. The PISA data, including the personal IDs, were provided by Skolverket.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Adams, R. & Wu, M. (2002). *PISA 2000 Technical Report*. OECD Publishing.
- Anders, J., Has, S., Jerrim, J., Shure, N., & Zieger, L. (2021). Is Canada really an education superpower? The impact of non-participation on results from PISA 2015. *Educational Assessment, Evaluation, and Accountability*, 33, 229–249. <https://doi.org/10.1007/s11092-020-09329-5>
- Andersson, C., & Sandgren Massih, S. (2023). PISA 2018: Did Sweden exclude students according to the rules? *Assessment in Education: Principles, Policy & Practice*, 30(1), 33–52.
- Aursand, L., & Rutkowski, D. (2021). Exemption or exclusion? A study of student exclusion in PISA in Norway. *Nordic Journal of Studies in Educational Policy*, 7(1), 16–29. <https://doi.org/10.1080/20020317.2020.1856314>
- Avvisati, F., & Givord, P. (2023). The learning gain over one school year among 15-year-olds: An international comparison based on PISA. *Labour Economics*, 84, 102365.
- Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). Wiley.
- 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), 1057–1075.
- Hellberg, L. (2020). *Debatt: Dagens ETC har fel om min Pisa-granskning*. ETC. Retrieved January 30, 2024, from <https://www.expressen.se/nyheter/pisa-skandalen-dag-for-dag/>
- Hellberg, L. (2021). *PISA-skandalen–dag för dag*. Expressen. Retrieved January 30, 2024, from <https://www.expressen.se/nyheter/pisa-skandalen-dag-for-dag/>
- Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J. A. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the programme for international student assessment. *Scandinavian Journal of Educational Research*, 62(3), 333–353.
- 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.
- Jerrim, J. (2021). PISA 2018 in England, Northern Ireland, Scotland and Wales: Is the data really representative of all four corners of the UK? *Review of Education*, 9(3), 1–41. <https://doi.org/10.1002/rev3.3270>
- Johansson, S., & Strietholt, R. (2019). Globalised student achievement? A longitudinal and cross-country analysis of convergence in mathematics performance. *Comparative Education*, 55(4), 536–556.
- Meinck, S., & Vandenplas, C. (2021). Sampling design in ILSA: Methods and implications. *International Handbook of Comparative Large-Scale Studies in Education: Perspectives, Methods and Findings* (pp. 1–25). Springer International Publishing.
- Meyer, H. D., Strietholt, R., & Epstein, D. Y. (2017). Three models of global education quality: the emerging democratic deficit in global education governance. In *International Handbook of Teacher Quality and Policy* (pp. 132–150). Routledge.
- 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.
- Munther, V. (2020). *Svenska Pisa-resultaten ifrågasätts: "Grava metodfel"*. svf. Retrieved January 30, 2024, from <https://www.expressen.se/nyheter/pisa-skandalen-dag-for-dag/>
- OECD. (n.d.). PISA 2018 technical report. Retrieved January 30, 2024, from <https://www.oecd.org/pisa/data/pisa2018technicalreport/>
- OECD. (2005). *PISA 2003 Technical Report*. OECD Publishing.
- OECD. (2009). *PISA 2006 Technical Report*. OECD Publishing.
- OECD. (2012). *PISA 2009 Technical Report*. OECD Publishing.
- OECD (2014). *PISA 2012 Technical Report*. OECD Publishing.
- OECD. (2017). *PISA 2015 Technical Report*. OECD Publishing.
- OECD. (2015). PISA 2018 technical standards. Retrieved January 30, 2024, from <https://www.oecd.org/pisa/pisaproducts/PISA-2018-Technical-Standards.pdf>
- OECD. (2019b). *PISA 2018 results (volume II): where all students can succeed*, PISA, OECD Publishing, Paris. <https://doi.org/10.1787/b5fd1b8f-en>
- OECD. (2020). A review of the PISA 2018 technical standards in regard to exclusion due to insufficient language experience. Retrieved January, 2024, from https://www.skolverket.se/download/18.2f324c2517825909a16357f/1628070064497/Revised_note_on_SWE_PISA_2018_20210310-2.pdf

- OECD. (n.d.). PISA 2022 technical report. Retrieved January 30, 2024, from <https://www.oecd.org/pisa/data/pisa2022technicalreport>
- Pereira, M. (2011). An analysis of Portuguese students' performance in the OECD programme for international student assessment (PISA). Retrieved January 30, 2024, from https://www.bportugal.pt/sites/default/files/anexos/papers/ab201111_e.pdf
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592.
- Skolverket. (2019). *PISA 2018, 15-åringars kunskaper i läsförståelse, matematik och naturvetenskap*. Skolverket, Stockholm.
- Skolverket. (2021). *PISA 2018: Undersökningens syfte, genomförande och representativitet*. Retrieved January 30, 2024, from <https://www.skolverket.se/skolutveckling/forskning-och-utvarderingar/internationella-jamforande-studier-pa-utbildningsomradet/pisa-internationell-studie-om-15-aringsars-kunskaper-i-matematik-naturvetenskap-och-lasforstaelse/pisa-2018-undersokningens-syfte-genomforande-och-representativitet>
- 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>
- Strietholt, R., & Scherer, R. (2018). The contribution of international large-scale assessments to educational research: Combining individual and institutional data sources. *Scandinavian Journal of Educational Research*, 62(3), 368–385.
- Strietholt, R., Rosén, M., & Bos, W. (2013). A correction model for differences in the sample compositions: The degree of comparability as a function of age and schooling. *Large-Scale Assessments in Education*, 1, 1–20.
- Strietholt, R., & Johansson, S. (2023). Challenges for the design of international assessments: Sampling, measurement, and causality. *On Education. Journal for Research and Debate*, 6(18). https://doi.org/10.17899/on_ed.2023.18.2
- Strietholt, R., & Rosén, M. (2016). Linking large-scale reading assessments: Measuring international trends over 40 years. *Measurement: Interdisciplinary Research and Perspectives*, 14(1), 1–26.
- Strietholt, R., Bos, W., Gustafsson, J. E., & Rosén, M. (Eds.). (2014). *Educational policy evaluation through international comparative assessments*. Waxmann Verlag.
- Swedish National Audit Office. (2021). Pisa-undersökningen 2018–arbetet med att säkerställa ett tillförlitligt elevdeltagande. Retrieved January 30, 2024, from https://www.riksrevisionen.se/download/18.4cb651881790d82c1eb2111a/1619525207429/RiR%202021_12%20Anpassad.pdf
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67. <https://doi.org/10.18637/jss.v045.i03>
- Woessmann, L. (2016). The importance of school systems: Evidence from international differences in student achievement. *Journal of Economic Perspectives*, 30(3), 3–32. <https://doi.org/10.1257/jep.30.3.3>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.