Educational Assessment, Evaluation and Accountability
https://doi.org/10.1007/s11092-023-09416-3

Exclusion rates from international large-scale assessments: an analysis of 20 years of IEA data

Umut Atasever[1] · John Jerrim[2] · Sabine Tieck[1]

Received: 17 August 2022 / Accepted: 31 August 2023
© The Author(s) 2023

Abstract

Cross-national comparisons of educational achievement rely upon each participating country collecting nationally representative data. While obtaining high response rates is a key part of reaching this goal, other potentially important factors may also be at play. This paper focuses on one such issue—exclusion rates—which has received relatively little attention in the academic literature. Using data from 20 years of international large-scale assessment data, we find there to be modest variation in exclusion rates across countries and that there has been a relatively small increase in exclusion rates in some over time. We also demonstrate how exclusion rates tend to be higher in studies of primary students than in studies of secondary students. Finally, while there seems to be little relationship between exclusion rates and response rates, there is a weak negative association between the level of exclusions and test performance. We conclude by discussing how information about exclusions—and other similar issues—might be more clearly communicated to non-specialist audiences.

Keywords TIMSS · PIRLS · ICCS · ICILS · IEA · Exclusion rates · International large-scale assessments

Umut Atasever, John Jerrim and Sabine Tieck are joint first authors.

✉ John Jerrim
J.Jerrim@ucl.ac.uk

Umut Atasever
Umut.Atasever@iea-hamburg.de

Sabine Tieck
Sabine.Tieck@iea-hamburg.de

[1] IEA Hamburg, Hamburg, Germany

[2] UCL Social Research Institute, London, UK

Published online: 21 September 2023

1 Introduction

International large-scale assessments (ILSAs)—such as the Trends in Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA)—have become high-profile investigations of how educational achievement compares across the world. Having been conducted in their present form for over 20 years, the results from such studies receive significant global attention from policymakers, educationalists, academics and journalists. When well-executed, findings from ILSAs can provide countries with a robust, independent measure of educational standards in key subject areas such as reading, science and mathematics, while also capturing whether these standards are improving or declining over time. The rich information gathered within the background questionnaires can also deepen our understanding of the correlates and potential causes of educational achievement, providing clues to policymakers as to what they might change in their education system based upon the results. It is for such reasons that many policymakers hold ILSAs in such high regard.

Yet, given the new prominent role of ISLAs across the world, it is vital that the data and evidence they provide are as robust as possible. Given their central aim of comparing educational achievement across countries and over time, obtaining high-quality, nationally representative samples is vital. Indeed, if this is not achieved, then one cannot be sure that such comparisons (either over time or to other countries) are unbiased and that we might not be comparing like with like. When considering such issues, many within the education research community focus upon response rates—the extent to which schools and students who were randomly sampled to participate are willing and able to take part. Yet, as noted by Jerrim (2021) and Anders et al. (2021), response rates are only part of the story. Other factors—such as the precise definition of the target population and the decision about how many schools/students to exclude—also have an impact as well. When taken together, this can result in the data collected having sub-optimal levels of population coverage, jeopardising a key assumption underpinning cross-country comparisons—that the data for each nation is nationally representative.

A collection of previous studies has considered factors that may impact the representativeness of ILSAs. While most have investigated one specific issue in isolation such as response rates, others have attempted to understand their joint impact when taken together. For instance, Durrant and Schnepf (2018) and Micklewright et al. (2012) investigated bias induced into the PISA data for England due to school and student non-response. Both studies report findings of selective participation, likely leading to an upward bias in this country's results. This is supported by recent evidence presented by Jerrim (2021) who demonstrates how the national examination grades of students who participated in PISA differ significantly from population-level figures. Pereira (2011) makes a similar argument in the case of Portugal, suggesting that changes to the sample design in this country might have had an impact upon the trend observed in this country's PISA results. In Sweden, the National Audit Office (2021) focused specifically upon the issue of the high Swedish exclusion rates in PISA. They concluded that

the exclusion rate reported in Sweden in 2018 did not follow OECD criteria and that consequently the results could have been significantly affected. Anders et al. (2021) reach a similar conclusion for Canada, noting how a combination of high exclusion rates and relatively low response rates is likely to lead to bias in the Canadian PISA sample. Spaull (2018) focuses upon the related issue of eligibility criteria, noting how ILSAs focus on the within-school population is likely to lead to estimates of academic achievement being overestimated, particularly within lower- and middle-income settings. Education Datalab (2017) has then demonstrated how this issue may have impacted the PISA results for Vietnam.

Although insightful, there remain gaps in our understanding and knowledge about some of the factors that potentially jeopardise the representativeness of ILSAs. In particular, some issues (most notably school and student response rates and patterns) have received much more scrutiny than others. The overarching aim of this paper is to hence provide new descriptive evidence on one specific component—exclusion rates— to which less independent academic attention has been paid. Indeed, many consumers of ILSA data and results may not realise that the technical standards allow countries to exclude up to 5% of students from such studies—and that the actual proportion of exclusions varies both across jurisdictions and over time.[1] Our hope is that this paper will foster a better understanding of this issue across the education research and policy community by providing a comprehensive analysis of data on exclusion rates across ILSAs conducted by the International Association for the Evaluation of Educational Achievement (IEA) since 1999.

To do so, we attempt to answer the following six research questions. First, our analysis will provide a descriptive overview of how exclusion rates compare across countries. Although such information is routinely reported in ILSA technical reports (e.g. Martin et al., 2020), they typically focus on a single wave of data in isolation and are rarely given much academic attention. In contrast, we will investigate country-level exclusion rates pooling across a large number of studies (e.g. TIMSS, The Progress in International Reading Literacy Studies and International Civic and Citizenship Education Study) and timepoints to better reflect which countries generally have comparatively high or low exclusion rates. We will pay particular attention to countries where the average exclusion rate is greater than the maximum 5% level stipulated by many ILSAs (see Section 2 for further details). Our first research question is thus:

1.1 Research question 1. How do exclusion rates compare across countries? How common is it for countries to exceed the maximum 5% exclusion rate?

Next, we turn to patterns in exclusion rates. In particular, do these differ between fourth-grade (approximately age 9/10) and eighth-grade (approximately age

[1] Up until 2003, the technical standards for IEA studies—while aiming at a maximum of 5%—allowed countries to exclude up to 10% of students without any annotation. This was reduced to 5% from 2006 onwards. Note also that, although 5% is the widely reported and understood maximum limit now used in IEA studies, it is effectively 5.5% in practise (due to rounding to the nearest whole number).

13/14) students? While exclusions occur for many reasons, some are related to test accessibility. It may be that this particularly affects younger students, who may be more likely than older students to get excluded on such grounds. Similarly, older students may be more likely to be diagnosed with a severe special education, with the test then being deemed inappropriate. Yet, there is currently little evidence explicitly comparing how exclusion rates vary across student age groups. Our second research question is thus:

1.2 Research question 2. Are exclusion rates higher in studies of primary (grade 4) or secondary (grade 8) school students?

Relatedly, exclusion rates—and, in particular, student exclusion rates—may vary according to test subject. This is because schools may deem a greater proportion of students to be unable to access the content of a reading test (as compared to—for instance—a mathematics test) due to a learning disability or limited language skills. Our third research question provides an exploration of this issue by asking:

1.3 Research question 3. Do exclusion rates amongst primary school students differ depending on the subject of the test (reading versus mathematics/science)?

Of course, excluding some students from ILSAs is not new—it has been a recurring issue over a long period of time. But are such exclusions becoming more common as ILSAs have increased in prominence and importance, or have they been on the decline? We provide new descriptive evidence on this matter when addressing research question 4:

1.4 Research question 4. How have exclusion rates from IEA studies changed over the last 20 years?

As noted above, exclusion rates are only one part of the picture with respect to obtaining nationally representative samples. Other factors—such as response rates—are a key factor as well. Yet relatively little consideration has been given to the interplay between these issues, and to the extent that they trade off against one another. For instance, is the "counterfactual" to excluding a school or student from a study that they get asked to participate but then refuse (or are unable) to take part? If so, then relaxing exclusion rates is unlikely to help in improving the representativeness of the sample collected, as the problem (selective non-participation) is simply shifted from one category (exclusions) to another

(non-response). Our fifth research question provides some descriptive, cross-country correlational evidence on this issue by asking:

1.5 Research question 5. Is there a trade-off between exclusions and response rates? Do countries achieve higher response rates when the exclusion rate is higher?

Finally, selectively excluding schools/students could potentially impact certain ILSA results, particularly if the decision of which and how many to exclude varies across countries or over time. Moreover, some key statistics are more likely to be impacted than others. For instance, one of the major drivers of student exclusion is special educational needs—students who would likely score below average on the test were it accessible to them. But if some countries are more likely to exclude such students than others, this has the potential to impact cross-national comparisons of the percentage of the population who lacks basic skills. We examine this relationship in our final research question:

1.6 Research question 6. Is there an association between exclusion rates and the proportion of students who fail to reach (a) the low international proficiency benchmark and (b) the intermediate international proficiency benchmark?

To summarise our key findings, we find a small number of countries (e.g. Israel, the USA and Denmark) to have exclusion rates persistently around or above the 5% maximum stipulated within ILSA technical standards. While exclusion rates tend to be higher in studies of primary (fourth grade) students than secondary (eighth grade) students, there is little clear evidence of a difference across subject areas (though with significant cross-national heterogeneity in these results). There has been a modest increase in overall exclusion rates across IEA studies over time, though with a greater increase in some (e.g. Germany and Sweden) than others. Little evidence emerges of a trade-off between response rates and exclusion rates, with higher levels of exclusions not resulting in lower levels of survey non-response. On the other hand, we do find some evidence of a modest association between exclusion rates and test scores, where the exclusion of more students is associated with higher levels of overall achievement (as measured by fewer students failing to reach key international benchmarks).

The paper now proceeds as follows. In Section 2, we describe the exclusion rate criteria used in ILSAs, the data we use to address our six research questions, and provide an overview of our empirical methodology. Results are then provided in Section 3, with conclusions following in Section 4.

2 Data and methodology

2.1 What exclusion criteria are used in IEA studies?

Within ILSAs, countries are permitted to exclude a small percentage of schools and students from the target population. With respect to school-level exclusions, this essentially means that the school is removed from the sampling frame (i.e. it has zero probability of being selected into the sample). Using TIMSS as an example, countries are permitted to exclude a small number of schools on the following grounds (Martin et al., 2020):

- Geographical inaccessibility
- Extremely small size (e.g. fewer than four students[2] in the target grade)
- The school's grade structure of curriculum is radically different from mainstream education
- The school provides instruction solely to those in the student exclusion categories (e.g. students with special educational needs only)

Note also that the technical standards also specify that the number of students excluded due to attending very small schools should be kept below two percent of the target population.

In contrast, student-level exclusions occur after sampling has taken place. Specifically, after schools and classes have been randomly selected to take part in the study, school coordinators can choose to exempt a small number of students from taking the test and completing the background questionnaires. Such exclusions are permitted for three reasons:

- Functional (physical) disabilities mean the student is unable to take the test.
- Intellectual disabilities.
- Students unable to read or speak the test language (and who usually have received less than 1 year of instruction in the test language).

Note that as disability criteria vary, countries are asked to translate the international criteria developed by the IEA into a local equivalent and apply it within schools.

To keep exclusion rates to a minimum, the IEA stipulates that countries should:

Keep the overall exclusion rate (the combination of school and student exclusions) below 5% of the national target population. However, note that in practise, the maximum limit to the exclusion rate is effectively 5.5% (due to the IEA rounding figures to the nearest whole number).

In addition, after the field trial, feedback is provided to countries about the level of within-school exclusions they are making and whether more students should be

[2] The number of students used to define "small size" is not fixed and varies across countries.

included. These criteria are used to try and ensure that each country's sample can—within reason—maintain the representativity of the target population. It is noteworthy, however, that high within-sample exclusion rates may be of greater concern than high school-level exclusion rates—given the select nature of this group[3]—and are something that is harder to control.

Note that there is also a related issue—not covered in this paper—of "reduced coverage". This is where a country is explicitly clear that they are not interested in a part of their population, and hence it is purposefully excluded from the study. ("Exclusions" are different, as the group in question is of interest, but too difficult to access.) Some information on this issue is reported however in Appendix A and Appendix D. This includes how often there has been "reduced coverage" of a country and the reasons why (it is usually due to a focus on public schools only, a focus upon only specific regions of a country, or a focus upon particular language groups). Amongst those that have taken part in at least five IEA studies, reduced coverage has most commonly affected Georgia, Lithuania, Canada and Florida.

2.2 Data

The data we analyse about such exclusion rates are drawn from all IEA studies conducted between 1999 and 2019. This incorporates six rounds of TIMSS, four rounds of PIRLS, two rounds of ICCS and two rounds of ICILS. All countries that took part in any of these studies/cycles are included. The total sample size is hence 807 country-by-study cycle data points, encompassing information from across 101 jurisdictions[4]. Country-level sample sizes vary from just a single data point (Iran) through to a maximum of 18 in Lithuania, Italy and the Russian Federation (mean = 8 studies; median = 7).

For each country-by-study cycle, our data file includes the following variables:

1. Percent school-level exclusions
2. Percent within-sample exclusions
3. Percent total exclusions
4. School response rates (before and after replacement, unweighted and weighted)
5. Student response rates
6. Coverage rates
7. The percentage of students who took the test but failed to reach the "low" international benchmark
8. The percentage of students who took the test but failed to reach the "intermediate" international benchmark

[3] It seems reasonable to assume that most within-sample exclusions are likely to be below-average academic achievers according to the criteria set. On the other hand, school-level exclusions are likely to include a mix of higher and lower achievers, assuming that such exclusions include small and/or geographically inaccessible schools.

[4] Although we use the term "country" throughout, sub-national jurisdictions such as Madrid and Buenos Aires are included as separate data points in our analysis.

Note that the fourth and fifth variables listed above are of particular use in addressing research question 5—studying the link between exclusion rates and response rates. The final two variables are used to answer research question 6, exploring the link between exclusion rates and the percent of the population in each country classified as a low academic achiever. Specifically, each IEA study uses cut-offs on the assessment scores to define the international benchmark groups ("low", "intermediate" and "advanced"). Our primary interest is in the percent who fail to reach (a) the low benchmark and (b) the intermediate benchmark.

2.3 Research question 1

To address research question 1, we compute the average exclusion rate for each country using information from all IEA studies they have participated in since 1999. This has the benefit of maximising the sample size for each country, though with the limitation that some will have participated in a more diverse set of studies than others. It will nevertheless provide an overarching overview of how well each country has done in minimising exclusion rates when they have participated in ILSAs over the last 20 years, including if they have generally managed to keep these below the maximum 5% threshold.

2.4 Research question 2

To compare exclusion rates in primary school (grade 4) and secondary school (grade 8), we restrict our focus to TIMSS. This is because it is the only IEA study to include both grade 4 and grade 8 students in most countries each time it is conducted. We will then use descriptive statistics—in the form of scatterplots, correlations and estimating international averages—to illustrate how exclusion rates from primary and secondary ILSAs differ.

2.5 Research question 3

To compare overall exclusion rates from tests covering different subjects, we use a similar approach to research question 2. To begin, we pair together PIRLS and TIMSS cycles by their proximity in time. This leads us to match together:

- PIRLS 2001 with TIMSS 2003.
- PIRLS 2006 with TIMSS 2007.
- PIRLS 2016 with TIMSS 2015.

Note that we exclude PIRLS/TIMSS 2011 as, in many countries, these studies were conducted simultaneously with the same students and schools (thus resulting

in identical exclusion rates).[5] We then further restrict the sample to only those countries that participated in both of these study-cycle pairs (e.g. countries that participated in both PIRLS 2001 and grade 4 TIMSS 2003). Descriptive statistics are again used to investigate how exclusion rates compare across studies focused on reading (PIRLS) versus science/mathematics (TIMSS) amongst primary school students.

2.6 Research question 4

To explore whether overall exclusion rates have changed over time, we estimate the following ordinary least squares regression model using the full sample (including all 807 country-by-study cycle observations in a single model):

$$E_{ij} = \alpha + \beta.T_{ij} + \gamma.G_{ij} + \mu_j + \varepsilon_{ij} \quad (1)$$

where E_{ij} is the overall exclusion rate from the study. T_{ij} is a variable capturing the year of the study, running from 0 (referring to 1999) to 19 (referring to 2019). G_{ij} is a binary variable capturing whether the study includes primary or secondary students. μ_j is the country-fixed effects. i is the study i. j is the country j. ε_{ij} is the random error term.

The parameter of interest from this model is β—the estimated change in the exclusion rate over time. Estimates will be reported in terms of the change in the exclusion rate per each 10-year increase in time. Country-fixed effects are included as controls, meaning all between-country variation in exclusions is stripped out, with our focus hence upon within-country changes over time. We also control for whether the study includes primary/secondary school students to control for potential changes in whether countries have changed participation in primary/secondary ILSAs over our 20-year time horizon. Standard errors are clustered at the country level. Together, estimates from this model will reveal whether there has been a general tendency for exclusion rates to have increased over time (across all participating countries).

Our analysis then turns to trends for individual countries. To begin, we restrict our focus to only those countries that have participated in at least ten IEA studies between 1999 and 2019 (although in Appendix B we present an extended set of results restricting the set of countries to those that have participated in at least five studies rather than ten). The model presented in Eq. (1) is then re-estimated separately for each country, excluding the country-fixed effects. A separate time trend (β) estimate is thus produced for each country, providing a single-figure summary statistic of whether overall exclusion rates have changed in each. Finally, for each country with a statistically significant increase or decrease in exclusion rates

[5] In 2011, 33 countries conducted PIRLS and TIMSS with the same students, and thus the total level of exclusions was identical. In 12 countries, PIRLS and TIMSS were conducted in different schools or with different classes in the same schools, and hence, the student-level and overall exclusion rates differ. Given the limited number of observations this would add, we have decided to exclude TIMSS/PIRLS 2011 from our analysis.

according to this estimated time trend, we produce a scatterplot to examine whether this seems to be driven by outliers or otherwise unusual data points.

2.7 Research question 5

Our analysis for research question 5 begins by presenting a scatterplot of school-level/within-sample exclusion rates against school/student response rates. We hypothesise there to be a positive relationship, where countries that exclude more students on a school-level/within-sample level are less likely to suffer from problems with survey non-response. The intuition is that, if countries choose to not exclude many students upfront, then they will simply drop out of the study in another way (e.g. by not being able or willing to take part).[6] We then test this relationship formally by estimating OLS regression models of the form:

$$R_{ij} = \alpha + \beta.Ex_{ij} + \delta.T_{ij} + \theta.P_{ij}\mu_j + \varepsilon_{ij} \quad (2)$$

where R_{ij} is the school (before replacement) or student response rate. Ex_{ij} is the school-level or within-school exclusion rate from the study. T_{ij} is a linear time trend. P_{ij} is a dummy variable capturing whether the study focuses upon primary or secondary school students. μ_j is the country-fixed effects. ε_{ij} is the random error term. i is the data point i. j is the country j.

The parameter of interest from this model is β. This captures the estimated percentage point change in the response rate associated with each percentage point increase in the exclusion rate. Note that we estimate the model presented in Eq. (2) twice—once for school exclusions/response rates and once for within-school/student response rates. As the model includes country-fixed effects, our estimates focus upon within-country differences across studies and cycles.[7]

2.8 Research question 6

To address research question 6, we begin by presenting a scatterplot of overall exclusion rates against the percent of students failing to achieve the low and intermediate benchmarks on relevant IEA achievement tests. This will provide an initial descriptive overview as to whether there is any relationship between exclusion rates and test performance at the country level. A set of OLS regression models will then be estimated in the form:

$$B_{ij} = \alpha + \beta.Ex_{ij} + \gamma.S_{ij} + \delta.T_{ij} + \theta.P_{ij} + \tau.R_{ij} + \mu_j + \varepsilon_{ij} \quad (3)$$

[6] Take a student with a severe functional disability. If one did not exclude this student—and hence tried to get them to take the assessment—it is likely they would not be able to—or that their school/teacher/parent would not agree to them—taking part (hence being classified as non-response).

[7] Note that throughout this process, we remove a small number of outliers with very high exclusion rates (greater than 12%) from the sample. This does not have a substantive impact upon our key results.

where B_{ij} is the percentage of students who do not achieve the low/intermediate international benchmark. Ex_{ij} is the overall exclusion rate from the study. S_{ij} is a vector of dummy variables capturing differences across IEA studies (e.g. PIRLS and TIMSS). T_{ij} is a linear time trend. P_{ij} is a dummy variable capturing whether the study focuses upon primary or secondary students. R_{ij} is the overall study response rate. μ_j is the country-fixed effects. ε_{ij} is the random error term. i is the data point i. j is the country j.

The parameter of interest is β. This captures the change in the percentage of students not achieving the relevant international benchmark per each percentage point increase in the exclusion rate. We test the hypothesis that β is negative that higher exclusion rates are associated with fewer students failing to achieve the low/intermediate international benchmark (i.e. a higher level of overall test performance).

Five specifications of the model outlined in Eq. (2) are estimated. This allows us to see how the addition of various control variables impacts our substantive results. Model M1 estimates the bivariate association with no controls included. Country-fixed effects (μ_j) are then added to model M2. Hence, from this point forward, our estimates will capture within-country variation only—i.e. the association between the change in the exclusion rate with the change in performance within each country. A time trend (T_{ij}) and a set of IEA study dummy variables (S_{ij}) are then added to model M3—our preferred model specification. These control for the fact that both exclusion rates and the percent failing to reach key international benchmarks may differ across subjects (e.g. reading, mathematics and civic education) and may have simultaneously changed over time. Likewise, a dummy variable capturing whether the study focuses upon primary- or secondary-aged students is included in specification M4, allowing for potential differences in exclusion and achievement rates across students of different ages. Finally, the overall survey response rate is added to the model in M6 to account for any potential trade-off between exclusion and response rates confounding the results. Note that, for each specification, standard errors are clustered at the country level using a sandwich estimator.

3 Results

3.1 Research question 1. How do exclusion rates compare across countries? How common is it for countries to exceed the maximum 5% exclusion rate?

Table 1 begins by presenting exclusion rates across 38 jurisdictions that have participated in at least 10 IEA studies since 1999 (see Appendix A for an analogous table including all 101 jurisdictions). The shading on this table should be read vertically, with red (green) cells indicating where exclusion rates are comparatively high (low) relative to other countries. The two dashed horizontal lines in Table 1 illustrate the stated 5% maximum exclusion rate permitted by the IEA technical standards and the 5.5% exclusion rate that is used in practise (due to the IEA rounding figures to the nearest whole number). These results are complemented by Appendix F, where we provide an interactive map (as an HTML file) providing an overview of the results.

Table 1 Exclusion rates across countries. Pooled estimates across all IEA studies

Country name	Number of studies	School-level exclusion	Within-sample exclusion	Total exclusions
Israel	10	16.5%	6.2%	22.4%
United States	17	0.2%	5.7%	5.9%
Denmark	11	1.8%	4.1%	5.9%
Hong Kong SAR	13	4.0%	1.4%	5.4%
Quebec, Canada	12	2.7%	2.7%	5.3%
Ontario, Canada	13	1.3%	3.7%	5.0%
Lithuania	18	2.3%	2.7%	5.0%
Singapore	15	4.8%	0.1%	4.9%
Dubai, UAE	10	2.3%	2.6%	4.9%
Russian Federation	18	2.4%	2.5%	4.9%
Georgia	11	2.0%	3.0%	4.9%
Sweden	15	1.8%	2.9%	4.7%
Hungary	15	2.8%	1.9%	4.7%
Italy	18	0.5%	4.1%	4.7%
Czech Republic	11	3.8%	0.6%	4.4%
Slovak Republic	12	3.1%	1.3%	4.4%
New Zealand	15	1.9%	2.3%	4.2%
Latvia	10	2.9%	1.0%	3.9%
Norway	15	1.3%	2.6%	3.9%
Netherlands, The	14	3.1%	0.8%	3.9%
Belgium (Flemish)	10	2.6%	1.1%	3.6%
Australia	14	1.4%	2.1%	3.6%
Iran, Islamic Rep. of	14	2.9%	0.5%	3.4%
Qatar	11	2.1%	1.2%	3.3%
Saudi Arabia	10	3.0%	0.3%	3.2%
England	16	2.0%	1.3%	3.2%
Finland	11	2.0%	1.2%	3.2%
Bulgaria	11	1.8%	1.4%	3.2%
Turkey	10	1.6%	1.5%	3.1%
Chile	13	1.4%	1.5%	2.9%
South Africa	11	1.8%	1.0%	2.8%
Korea, Rep. of	13	1.2%	1.3%	2.5%
Germany	10	1.1%	1.3%	2.4%
Slovenia	16	1.5%	0.9%	2.4%
Chinese Taipei	16	0.4%	1.9%	2.3%
Japan	11	0.9%	1.1%	2.0%
Morocco	14	1.2%	0.0%	1.2%
Kuwait	10	0.8%	0.4%	1.2%

In almost a fifth (7) of the 38 jurisdictions, the average exclusion rate sits on or exceeds the 5% maximum allowed within the ILSA's technical standards. These seven jurisdictions are a heterogeneous group, encompassing those within Europe (e.g. Denmark), North America (e.g. the USA) and East Asia (e.g. Hong Kong). Israel is a notable outlier, where both the school-level and within-sample exclusion rates are particularly high. This may be due to a combination of excluding ultra-orthodox schools and those schools/classes specifically focused upon serving the needs of students with special educational needs.

It is interesting to note that there are also significant differences across countries in whether school-level or within-sample exclusion rates are higher. There are some—such as Israel, Singapore, Hong Kong and the Czech Republic—where school-level exclusion rates are by far larger than within-sample exclusion rates. Yet, the opposite holds true in the USA, Ontario and Italy. In other words, there is no single pattern that seems to hold across countries, with quite a different mix of school and student exclusions determining how the overall exclusion rate is formed. It is noteworthy, however, that there is a negative association between the two; once Israel is excluded as an outlier, countries with higher school-level exclusion rates tend to have lower within-sample exclusion rates (Pearson correlation = −0.43).

There are also some countries with quite low levels of student exclusion. While one would expect that students with special education needs would be part of the student population, we suspect that in some countries, these pupils are not involved in the respective school system.

3.2 Research question 2. Are exclusion rates higher in studies of primary (grade 4) or secondary (grade 8) school students?

Figure 1 compares the overall exclusion rates across primary (horizontal axis) and secondary (vertical axis) studies using data gathered from TIMSS. Each data point refers to a country within a given survey cycle. The diagonal 45° line on this plot demonstrates where the grade 4 and grade 8 overall exclusion rates from TIMSS are equal.

From this plot, there are three points to note. First, most points tend to sit below the 45° line. This illustrates how exclusion rates tend to be slightly higher for studies of primary school students (median = 4.0%) compared to those for secondary school students (3.3%). Second, there is nevertheless a strong correlation between the two (Pearson correlation = 0.83). In other words, countries that excluded a higher proportion of students/schools from the fourth grade TIMSS generally had higher exclusion rates from the eighth grade TIMSS as well. Finally, there are a handful of notable outliers. The most obvious was in Quebec in TIMSS 2007, where the overall exclusion rate for the eighth-grade sample (13.6%) was more than double that for the fourth-grade sample (6.4%). Yet other examples include Florida (12% exclusion in grade 4 versus 7% in grade 8) and Hong Kong (8.5% in grade 4 versus 5.3% in grade 8) in TIMSS 2011. Nevertheless, with respect to research question 2, Fig. 1 provides clear evidence that exclusion rates tend to be higher for primary schools than secondary schools, although this is not always the case.

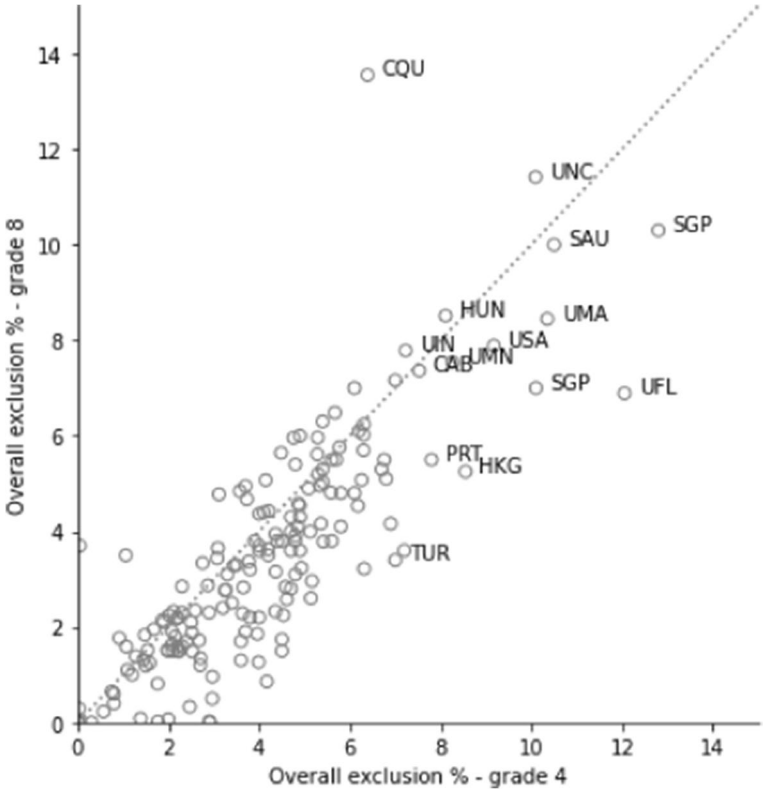

Fig. 1 A comparison of overall exclusion rates across the grade 4 and grade 8 TIMSS samples. Notes: Each data point refers to a participating country within a given TIMSS cycle. The sample was restricted to countries that participated in both the grade 4 and grade 8 studies within a given year (2003, 2007, 2011, 2015 and 2019). The diagonal 45° line illustrates where the grade 4 and grade 8 exclusion rates are equal. The median exclusion rate is 4% in grade 4 and 3.3% in grade 8. Pearson correlation = 0.83. Countries identified by three-letter country code (CQU, Quebec; UNC, North Carolina; UMA, Massachusetts; UIN, Indiana; UFL, Florida)

3.3 Research question 3. Do exclusion rates amongst primary school students differ depending upon the subject of the test (reading versus mathematics/science)?

The results presented in Fig. 2 are similar to those in Fig. 1, but now focus upon differences between subjects—reading from PIRLS (horizontal axis) and mathematics/science from fourth-grade TIMSS (vertical axis). To directly answer research question 3, there is no clear pattern of the data points tending to fall above or below the 45° line—i.e. there is no evidence that exclusion rates vary by subject. This is further borne out by the international medians, which are similar across fourth-grade TIMSS (4.0%) and PIRLS (3.7%).

There is again evidence of a positive correlation between the PIRLS/TIMSS exclusion rates, indicating once more the tendency for countries that have high

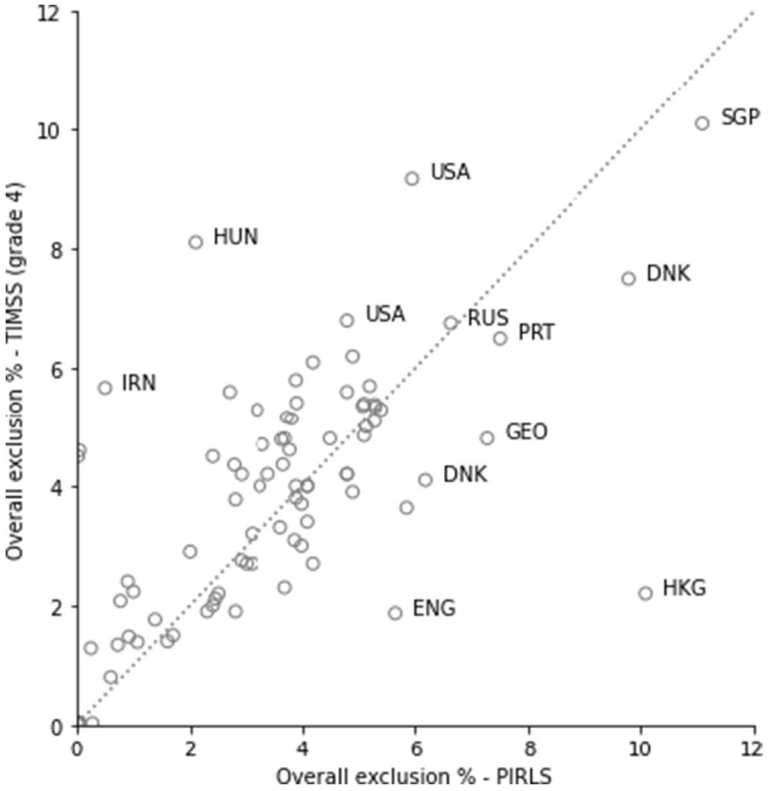

Fig. 2 A comparison of overall exclusion rates across the PIRLS and TIMSS grade 4 samples. Notes: PIRLS/TIMSS cycles have been matched as follows: PIRLS 2001-TIMSS 2003; PIRLS 2006-TIMSS 2007; PIRLS 2016-TIMSS 2015. Each data point refers to a participating country within a given TIMSS/PIRLS matched cycle. The diagonal 45° line illustrates where matched PIRLS and TIMSS exclusion rates are equal. The median exclusion rate is 3.7% in PIRLS and 4.0% in TIMSS. Outlying data points have been labelled with their two-letter country code (some countries appear twice due to their exclusion rates being high across more than one cycle). Pearson correlation = 0.64

exclusion rates on one ILSA to also have high exclusion rates on others. Yet the correlation here (0.64) is smaller than in Fig. 1 (0.83) suggesting that there is a weaker association in exclusion rates across studies (PIRLS and fourth-grade TIMSS) than across different grades within a single study (fourth-grade and eighth-grade TIMSS). This is to some extent reflected by the fact that there are some more extreme outliers in Fig. 2 than in Fig. 1. For instance, both Hungary (HU) and Iran (IR) had much higher exclusion rates in TIMSS 2003 than PIRLS 2001. On the other hand, exclusions in Hong Kong (HK) were significantly smaller in TIMSS 2015 than in PIRLS 2016.

Hence, while there is no general pattern that exclusion rates tend to be higher for ILSAs focusing upon reading (PIRLS) than science/mathematics (TIMSS), there are also some clear instances of big differences even when these studies have been conducted only a short time apart.

Table 2 Estimated increase in exclusion rates in IEA studies over time

	M1		M2		M3	
	Change	SE	Change	SE	Change	SE
Change in exclusion rate per 10 years						
Overall exclusions	0.55%*	0.22%	0.51%*	0.23%	0.47%*	0.23%
School-level exclusion	0.01%	0.18%	0.18%	0.23%	0.16%	0.23%
Within-sample exclusion	0.50%*	0.14%	0.31%*	0.12%	0.28%*	0.11%
Controls						
Country fixed effects	–		Y		Y	
Grade 4 or 8	–		–		Y	

Figures refer to the estimated change in exclusion rates for a 10-year increase in time. Estimates based upon data from 807 country-study data points between 1999 and 2019

*The increase in exclusion rates is statistically significant at the 5% level. Standard errors have been clustered by country. Estimates based upon OLS regression models. M1 includes year as the only covariate, M2 adds country-fixed effects and M3 whether the study involved grade 4 or grade 8 students

3.4 Research question 4. How have exclusion rates from IEA studies changed over the last 20 years?

We begin by Table 2 by presenting estimates from the OLS regression model specified in Eq. (1). Estimates refer to the percentage point change in exclusion rates per 10-year increase in time. This provides a simple summary of the aggregate picture of how exclusion rates in ILSAs have changed.

Table 2 points towards two key results. The first is that there seems to have been a modest increase in overall exclusion rates over time, even after accounting for the changing pool of countries participating in ILSAs and changing participation rates in grade 4 and grade 8 studies. Specifically, overall exclusion rates have increased on average by around 0.5 percentage points every 10 years, or by one percentage point over our 20-year time horizon. Although this is relatively modest, any further increase is likely to be uncomfortable, given how many countries now sit close to the 5% limit (see Table 1).

The second point of note from Table 2 is that the evidence is strongest for within-sample exclusion rates driving this increase. Across all model specifications, the increase in within-school exclusions is statistically significant. In contrast, the trend for school-level exclusion rates is less clear, with none of the three estimates reaching statistical significance at the 5% level.

This aggregate pattern does of course mask potential heterogeneity in trends across countries. Table 3 hence presents country-level estimates of the trend in overall exclusion rates over time, with the parameters having the same interpretation as in Table 2 above. Note that Table 3 only includes 38 jurisdictions that participated in at least 10 IEA studies between 1999 and 2019. An extended set of results including a wider pool of countries can be found in Appendix B.

Out of the 38 countries included in Table 3, overall exclusion rates have seen a statistically significant increase in 11 countries, with a clear decrease in just one

Table 3 Estimated change in overall exclusion rate by country per 10-year increase in time

	N	M0		M1	
		Change	SE	Change	SE
Singapore	15	6.9*	0.6	6.8*	0.6
Saudi Arabia	10	5.5*	1.7	6.0*	1.9
Israel	10	2.4*	1.0	2.4*	0.8
Latvia	10	1.8*	0.3	1.8*	0.3
Germany	10	1.8*	0.5	1.8*	0.6
Kuwait	10	1.7*	0.6	1.8*	0.6
Sweden	15	1.5*	0.4	1.5*	0.4
Norway	15	1.0*	0.5	1.3*	0.5
Denmark	11	0.8	1.5	1.2	1.6
Australia	14	1.3*	0.4	1.2*	0.3
Slovenia	16	1.2*	0.5	1.2*	0.5
Lithuania	18	1.1*	0.3	1.1*	0.3
Turkey	10	0.9	0.8	0.8	0.7
Japan	11	0.7	0.4	0.7	0.5
Qatar	11	0.6	1.3	0.7	1.3
New Zealand	15	0.7	0.5	0.6	0.4
Slovak Republic	12	0.2	0.7	0.6	0.8
Netherlands, The	14	0.4	0.7	0.4	0.8
Russian Federation	18	0.3	0.6	0.3	0.6
England	16	0.2	0.6	0.2	0.6
USA	17	0.0	0.6	0.1	0.6
Dubai, UAE	10	0.1	0.6	0.1	0.7
Italy	18	0.0	0.4	0.0	0.4
Chile	13	0.2	0.5	−0.1	0.4
Finland	11	−0.3	0.3	−0.2	0.3
Morocco	14	0.0	0.3	−0.2	0.2
Hungary	15	−0.4	0.7	−0.3	0.7
Bulgaria	11	0.0	0.8	−0.4	0.8
Iran, Islamic Rep. of	14	−0.5	0.8	−0.6	0.8
Hong Kong SAR	13	−1.1	2.1	−0.6	2.0
Georgia	11	−0.7	0.7	−0.7	0.7
Chinese Taipei	16	−0.7	0.4	−0.7	0.4
Belgium (Flemish)	10	−0.4	1.1	−0.9	1.2
Korea, Rep. of	13	−0.9	0.5	−1.0	0.6
Czech Republic	11	−0.7	0.5	−1.1*	0.5
Quebec, Canada	12	−1.3	1.5	−1.2	1.5
South Africa	11	−0.5	1.5	−1.3	1.4
Ontario, Canada	13	−1.4*	0.7	−1.4*	0.7

The sample is restricted to countries with at least 10 observations (participated in at least 10 IEA studies between 1999 and 2019). Estimates based upon OLS regression models. M0 includes year as the only covariate, while M1 controls for whether the study involved grade 4 or grade 8 students

*Change statistically significant at the 5% level

(Ontario). There is a geographical spread amongst these countries, including East Asia (Singapore), the Middle East (Saudi Arabia and Kuwait) and a host of coutries throughout Europe (Germany, Sweden and Slovenia). In each of these 11 countries, overall exclusion rates have increased on average by at least one percentage point every 10 years, or two percentage points over our 20-year time horizon. Yet there are also examples of countries with a much more extreme increase in exclusions. This includes Singapore, where private schools (and the students that attend them) have been increasingly excluded, and Saudi Arabia, where school-level exclusions reached 8% in TIMSS 2019 due to some schools being located in a war zone.

Appendix C provides some further details by presenting scatterplots of school and student exclusion rates over time for each jurisdiction with a statistically significant increase or decrease. This reveals how the countries with a statistically significant change can be broadly divided into four sub-groups:

(a) Change driven by within-sample exclusions (Germany, Latvia, Norway, Sweden and Ontario).
(b) Increase driven by school-level exclusions (Israel[8] and Singapore).
(c) Increase driven by a mix of school-level and within-sample exclusions (Australia, Denmark, Kuwait, Lithuania and Slovenia).
(d) Change heavily influenced by an outlier (Saudi Arabia, where the exclusion rate was exceptionally high—9.1%—in grade 8 TIMSS 2019. This was driven by the exclusion of schools located in a war zone, encompassing around 8% of the student population).

Moreover, Appendix C also makes clear how these countries started from very different bases. For instance, Israel has always had very high levels of school exclusions, and these have risen further over the last 20 years. On the other hand, Germany historically had low overall exclusion rates (e.g. around one percent in PIRLS 2006 and TIMSS 2007), but these have recently increased (e.g. up to around four percent in PIRLS 2016 and TIMSS 2019).

3.5 Research question 5. Is there a trade-off between exclusions and response rates? Do countries achieve higher response rates when the exclusion rate is higher?

Figure 3 documents the bivariate association between exclusion rates (horizontal axis) and response rates (vertical axis) at both the school (Fig. 3a) and within-sample (Fig. 3b) levels. In contrast to our hypothesis, no clear relationship between the variables seems to exist. There is no clear pattern to the scatter of data points, with the plotted locally weighted scatterplot smoothing (LOWESS)

[8] The increasing school level exclusion rate in Israel is also to some extent driven by TIMSS 1999 as an outlier, where the level of exclusions was much lower than in subsequent years. However, even if the TIMSS 1999 data is removed from the analysis, the increase in school exclusions for Israel over time remains significant.

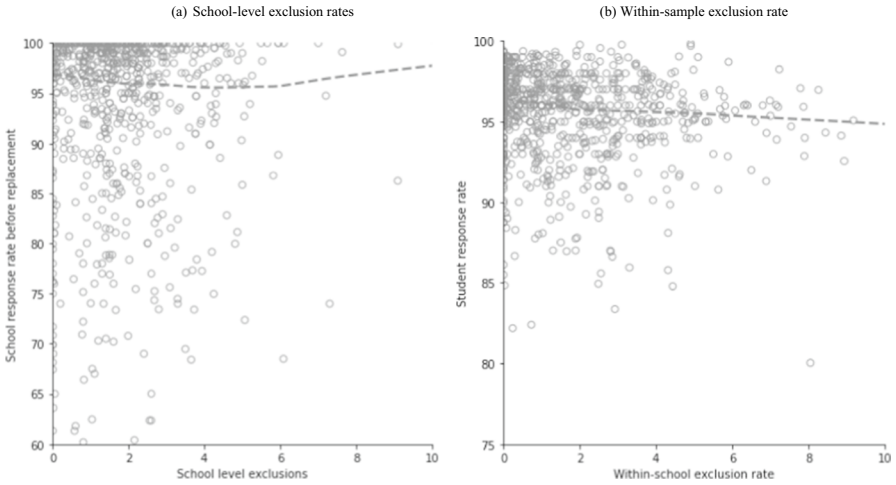

Fig. 3 The relationship between exclusion rates and response rates at the country level. Notes: We have removed a small number of outliers from the sample where the overall exclusion rate is greater than 12%. Dashed grey line generated by locally weighted scatterplot smoothing (LOWESS). Graphs based upon 789 observations

Table 4 OLS estimates of the relationship between response rates and exclusion rates at the country level

	School level		Student level	
	% change	SE	% change	SE
Change in response rate per 1% increase in exclusion rate	−0.0475%	0.238%	−0.0611%	0.112%
Controls				
Country fixed effects	Y		Y	
Year (continuous)	Y		Y	
Grade 4 or 8 study	Y		Y	

Data points were removed when the respective overall exclusion rate was greater than 12%. Estimates based upon 787 observations. Figures refer to the change in the response rate (in percentage points) per each percentage point increase in response rates. Models control for country fixed effects, a continuous year variable and whether the study involves grade 4 or grade 8 pupils

lines essentially being flat. This is further confirmed in Table 4, where we present estimates from our regression models. The estimated effect sizes are small—e.g. there is only a 0.06 percentage point decrease in student response rates per each 1.0 percentage point increase in student exclusion rate—and are not statistically significant at conventional thresholds. Thus, overall, there is no clear evidence of any trade-off between exclusion rates and response rates at the country level; those jurisdictions that exclude more students do not seem to gain any benefit in terms of maximising their survey response rate.

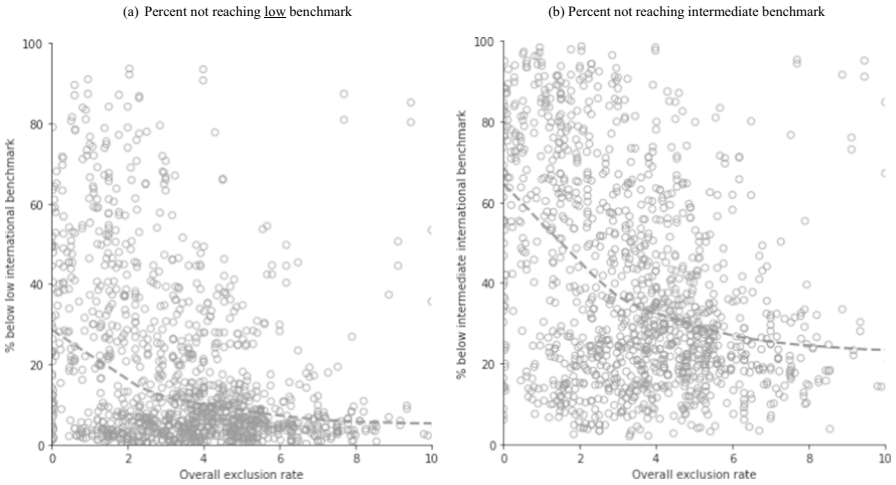

Fig. 4 The relationship between overall exclusion rates and percent of pupils not reaching international benchmarks. Notes: Dashed grey line generated by locally weighted scatterplot smoothing (LOWESS). Graphs based upon 1299 observations

3.6 Research question 6. Is there an association between exclusion rates and the proportion of students who fail to reach (a) the low international proficiency benchmark and (b) the intermediate international proficiency benchmark?

Figure 4 displays the unconditional association between exclusion rates (horizontal axis) and the percentage of pupils failing to meet the low (Fig. 4a) and intermediate (Fig. 4b) benchmarks. There is a clear negative relationship whereby higher levels of exclusion are associated with higher levels of achievement in the tests (i.e. fewer pupils failing to reach the benchmarks). The link is particularly strong in Fig. 4b, with some indication that the relationship may be non-linear; the gradient to the fitted LOWESS line is sharpest when the exclusion rate increases from around 0 to 3% and then becomes falter thereafter.

Table 5 supplements these provisional findings via regression modelling. Estimates from model M1 confirm that the raw, conditional association is sizable and statistically significant; each percentage point increase in the exclusion rate is associated with around a two-percentage-point reduction in children failing to achieve the low and intermediate benchmarks. The inclusion of country-fixed effects in model 2 leads to a substantial reduction in the size of these estimates, although they remain moderate in size and statistically significant. In particular, we now estimate that each percentage point increase in the exclusion rate is associated with around a 0.35 percentage point decline in students failing to reach the intermediate proficiency threshold. Note also the extremely high model R^2 (~0.85), meaning that we are able to detect this association even though the vast majority of the variation in the data has been explained. The addition of further controls in models 3 to 5 leads

Table 5 OLS estimates of the association between exclusion rates and the percentage of students failing to reach key international benchmarks

	Model 1		Model 2		Model 3		Model 4		Model 5	
	Beta	SE	Beta	SE	Beta	SE	Beta	SE	Beta	SE
Failed to reach low benchmark										
Change in % not meeting benchmark per each % increase in overall exclusion rate	−1.73**	0.79	−0.37**	0.18	−0.23	0.16	−0.25	0.16	−0.25	0.16
Observations	1299		1299		1299		1299		1299	
R-squared	0.068		0.863		0.882		0.882		0.882	
Controls										
Country fixed effects	–		Yes		Yes		Yes		Yes	
Study fixed effects	–		–		Yes		Yes		Yes	
Year	–		–		Yes		Yes		Yes	
Primary or secondary study	–		–		–		Yes		Yes	
Overall response rate	–		–		–		–		Yes	
Failed to reach intermediate benchmark										
Change in % not meeting benchmark per each % increase in overall exclusion rate	−2.10**		0.96		−0.58**		0.22		−0.37**	
Observations	1299		1299		1299		1299		1299	
R-squared	0.067		0.832		0.886		0.887		0.887	
Controls										
Country fixed effects	–		Yes		Yes		Yes		Yes	
Study fixed effects	–		–		Yes		Yes		Yes	
Year	–		–		Yes		Yes		Yes	
Primary or secondary study	–		–		–		Yes		Yes	
Overall response rate	–		–		–		–		Yes	

Asterisks (* and **) indicate statistical significance at the 10% and 5% levels respectively. Estimates refer to the change in the percentage of students that fail to achieve the low/intermediate benchmark for a 1% increase in the overall exclusion rate. Standard errors have been clustered by country

to a further slight reduction in the size of the estimates, with statistical significance only achieved at the 10% level for the intermediate benchmark. Thus, our overall interpretation of the results is that they provide some moderate evidence of a relatively modest association between exclusion rates and proficiency levels.

4 Conclusions

International large-scale assessments have become a key resource for many education systems across the globe. Two of their central aims are to facilitate comparisons of educational achievement across countries and to investigate changes over time. To facilitate such goals, obtaining unbiased, representative samples is vital. Although many focus upon the issue of response rates when considering potential bias in a country's sample, other factors may have an important impact as well. One such factor—which has received comparatively little attention—is exclusion rates; countries removing certain schools and students so that they cannot appear in the final sample. This paper has presented a new summary of exclusion rates from international studies conducted by the IEA over the last 20 years. The analysis includes how such exclusion rates compare across countries and over time and an investigation of potential differences across subjects and school grades, while also considering how exclusions are linked to response rates and low achievement rates at the country level. In doing so, we have presented the most comprehensive investigation of this issue to date.

We have found there to be significant differences in exclusion rates across countries. Although this may be anticipated, given the heterogeneity of the countries that participate in ILSAs, it is nevertheless important given its potential to impact upon the results. For instance, if country A excludes no students but country B excludes 10% or more, and this clearly raises questions about whether their results should be compared given that they represent different fractions of their original student population. This is a particularly pertinent issue given that many countries now struggle to consistently stay below the maximum 5% threshold.

Such exclusions tend to be higher within studies focusing on primary school students than secondary school students, though with no clear difference by subject (reading versus science/mathematics). This is important as it highlights how particular ILSAs focusing on primary pupils may be at greater risk of bias from this issue than ILSAs focusing on secondary pupils. Although we can only speculate as to why this may be the case, it may be related to schools/teachers' perceptions around test accessibility, with younger students being more likely to be excluded on such grounds. Either way, this result points towards a need for further research using administrative records to understand the characteristics of who gets excluded from ILSAs and the potential impact this has on the results—particularly those focused on primary school pupils.

We also find evidence of a modest increase in exclusion rates over the last 20 years, driven by quite steep increases in around a dozen counties. We find no evidence of an inverse relationship between exclusion rates and response rates at the country level. On the other hand, there is a weak negative relationship between the

percentage of countries excluded from the sample and the percentage of students failing to reach the international intermediate achievement benchmark.

These findings should be considered in light of the limitations of this research. First, we have focused upon ILSAs conducted by the IEA that attempt to measure the skills of primary and secondary students. ILSAs led by the OECD and/or those examining other populations (e.g. teachers and adults) have not been included. Future work may thus seek to extend our analysis to investigate the evidence on exclusions for other populations. Second, all our analyses have been conducted at the country level. Ideally, future work will further probe some of our research questions at the individual level, possibly by linking ILSA data to administrative records (similar in spirit to the work of Micklewright et al., 2012). Such work should include further exploration of the trade-offs between exclusions and non-response and how excluding students may introduce bias into estimates of the percentage of the population who are classified as low achievers. Third, relatedly, all of the analyses presented in this paper refer to correlations only and do not attempt to capture cause and effect.

Our findings nevertheless have important implications. Given the relatively high and rising level of exclusions in some countries, it is important that exclusion rates do not increase any further and—ideally—start to decline. Not only does minimising exclusion rates help ensure the face validity of ILSAs, but it also minimises risks of bias in the estimation and comparison of population parameters. Although we have found there to only be a modest association between exclusion rates and achievement at the country level, further individual-level analysis using administrative microdata is still needed to assess whether high exclusion levels may lead to some bias in ILSA results. As this is unlikely to be possible—at least on a widescale international basis—any time soon, continuing to minimise exclusion rates should continue to be an important element of large-scale international studies for the foreseeable future. This is particularly important for those conducted in primary schools—where such exclusions tend to be highest—and in countries that have seen recent rises.

One option could be for the international consortia that conduct such studies to introduce a five-star rating system, providing an easy-to-communicate overview of the quality of each country's sample. Exclusion rates—along with school and student response rates—would be one key component of this. Currently, if a country exceeds the stated 5% maximum exclusion rate, a footnote is added to the results, the meaning of which may not be easily understood by non-specialist audiences. Such a five-star rating system of sample quality would help such non-specialists to better appreciate the limitations of the data available. In particular, such a system would likely offer three important advantages. First, as argued by Jerrim (2021) and Anders et al. (2021), it is not exclusions alone that matter for the comparability of ILSAs across countries and over time, but how this combines with other factors such as school and student response rates. A five-star rating system would potentially communicate the combined effect of potential sample deficiencies rather than getting "stuck in the weeds" of various different technical points. Second, such a system would encourage countries to focus on the overall quality of the sample (i.e. ensuring the overall representativeness of the data) rather than meeting certain "targets" (e.g. exclusion rate, student response rate and school response rate) in isolation. Finally, although ILSAs provide various footnotes where exclusion rates

are above the target (or there are other potential issues with the sample), it is not clear how widely they are understood or even taken notice of. It is likely that a five-star rating system of data quality—as other organisations such as the Education Endowment Foundation use—would be a much more effective communication tool.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11092-023-09416-3.

Data availability The data used in this paper are available in the public domain as part of the technical reports for each of the respective studies.

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

Anders, J., Has, S., Jerrim, J., et al. (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

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.

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 05/05/2022 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/.

Jerrim, J. (2021). PISA 2018 in England, Northern Ireland, Scotland and Wales: Is the data really representative of all four corners of the UK? *The Review of Education, 9*, e3270. https://doi.org/10.1002/rev3.3270

Martin, M, von Davier, M, & Mullis, I. (2020). TIMSS 2019 technical report.

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, 175*, 915–938.

Pereira, M. (2011). An analysis of Portuguese students' performance in the OECD programme for international student assessment (PISA). Accessed 05/05/2022 from https://www.bportugal.pt/sites/default/files/anexos/papers/ab201111_e.pdf

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

Swedish National Audit Office.(2021). The 2018 PISA survey – ensuring reliable student participation. Accessed 30/06/2022 from https://www.riksrevisionen.se/en/audit-reports/audit-reports/2021/the-2018-pisa-survey---ensuring-reliable-student-participation.html . Swedish language version available from https://www.riksrevisionen.se/rapporter/granskningsrapporter/2021/pisa-undersokningen-2018---arbetet-med-att-sakerstalla-ett-tillforlitligt-elevdeltagande.html

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