IEA

IEA Research_
and Development_

Striving for More Inclusiveness: Understanding and Addressing Student Exclusion in TIMSS

Umut Atasever[1]
Sabine Meinck[1]
Matthias von Davier[2]
Bethany Fishbein[2]
Rolf Strietholt[1]

[1]International Association for the Evaluation of Educational Achievement
[2]Boston College, Chestnut Hill, MA, USA

DOI: https://doi.org/10/58150/rd_4_3

Summary

International large-scale assessments such as TIMSS and PIRLS depend on representative samples and inclusive coverage of the target population to generalize the sample results to the underlying population and ensure cross-national comparability over time. In recent cycles, student exclusion rates—particularly within schools—have increased, raising concerns about representativeness and potential bias in reported achievement results. Bias may arise if lower-performing students are disproportionately excluded from the assessed sample.

Using TIMSS 2023 data, we examined the performance of students classified as "excluded" who nonetheless participated in the assessment and generated achievement data. We investigate whether they demonstrated meaningful engagement with the assessment, their average performance, and whether their exclusion biased national achievement estimates. Together with published mathematics and science achievement data from "included" students, we analyze outcomes for excluded students with available data across 17 education systems at grade four and eight systems at grade eight. Analyses focus on (i) the psychometric quality of the data using indicators marking students' achievement as too low for reliable estimation, (ii) attainment of the TIMSS Low International Benchmark, (iii) differences in average achievement between included and excluded students, and (iv) the impact of inclusion on national estimates.

Results show that excluded students are more likely than included students to have lower achievement and to be flagged as too low for reliable estimation. However, such cases remain relatively rare in most systems and are often comparable to those of included students, indicating that the majority of excluded students are able to engage meaningfully with the assessment. A substantial share of excluded students reaches the Low International Benchmark. Because excluded students score systematically below included students, their exclusion introduces a modest upward bias in national averages, particularly in systems with higher exclusion rates.

Overall, this study suggests evidence that there have been differences across countries in how the studies' exclusion definitions have been interpreted and implemented to some extent. That is, many excluded students with available data could have been incorporated into achievement estimation without compromising psychometric integrity, highlighting the need to refine exclusion definitions, strengthen accommodation guidance, and improve oversight to enhance validity and inclusiveness in international assessments.

This work was supported in whole or part by the IEA Research and Development (R&D) Funds, which aim to advance and improve the science and methodology of IEA studies, ensuring that they remain at the forefront of international large-scale assessments in education. Funding is drawn from an overhead on participation fees across IEA studies. This does not constitute endorsement of the contents, which reflects the views only of the authors. IEA cannot be held responsible for any use which may be made of the information contained therein.

Title: Striving for More Inclusiveness:
Understanding and Addressing Student Exclusion in TIMSS

2026.08.20

Umut Atasever[12], Sabine Meinck[1], Matthias von Davier[3], Bethany Fishbein[3], Rolf Strietholt[1]

Corresponding Author: Umut Atasever

umut.atasever@iea-hamburg.de

umut.atasever@tu-dortmund.de

ORCID: 0009-0007-4635-4436

[1] IEA, Hamburg, Germany
[2] TU Dortmund, Dortmund, Germany
[3] Boston College, Boston, USA

Table of Contents

Abstract

International large-scale assessments such as TIMSS depend on representative samples and inclusive coverage of the target population to generalize sample results to the underlying population and ensure cross-national comparability over time. In recent cycles, student exclusion rates have increased, raising concerns about representativeness and bias if lower-performing students are disproportionately excluded.

Using TIMSS 2023 data, we examine the performance of students classified as "excluded" but who nonetheless participated in the assessment and generated achievement data. We investigate whether they demonstrated meaningful engagement with the assessment, their average performance, and whether their exclusion biased national achievement estimates across 17 education systems at grade four and eight at grade eight.

Results show that excluded students generally achieve lower scores and are more likely to be flagged as too low for reliable estimation. However, such cases remain relatively rare, and most excluded students engage meaningfully with the assessment, with many reaching the *Low International Benchmark*. Because excluded students score systematically below included students, their exclusion introduces a modest upward bias in national achievement, particularly in systems with higher exclusion rates.

Overall, this study provides evidence that many excluded students could have been included in achievement estimation without compromising psychometric integrity, highlighting the need to refine exclusion definitions, strengthen accommodation guidance, and enhance the validity and inclusiveness of international assessments.

Keywords: international large-scale assessments; TIMSS; student exclusion; sample representativeness; inclusiveness

Introduction

International large-scale educational assessments (ILSAs) such as TIMSS, PIRLS and PISA are pivotal for understanding educational achievement across countries and over time (Hernández-Torrano et al., 2021). Over more than two decades, these assessments have played a central role in informing educational policy, monitoring system performance, and shaping public and academic debates (UNESCO, 2020). Their scientific credibility, however, critically depends on the quality and representativeness of the samples drawn in participating countries (Wagemaker, 2013). Comparisons across countries, or over time, are only meaningful if the assessed samples accurately reflect the underlying target populations. A key challenge to representativeness arises from the exclusion of students from the assessment, which may be necessary in some cases, for example, if they have intellectual or functional disabilities, or are not proficient in the language of the assessment. However, when specific subgroups of the target population are excluded, they are no longer represented in the sample, potentially undermining the validity and comparability of achievement estimates (Meinck, 2020).

ILSAs are therefore intended to be as inclusive as possible, and exclusion is intended to be minimal and applied only under clearly defined circumstances. To mitigate such risks to representativeness, the IEA—consistent with other organizations conducting ILSAs, such as the OECD—has defined quality standards that limit exclusions from the target population to a maximum of five percent[4] (Siegel & Foy, 2024; OECD, 2024). While these thresholds serve as important operational guidelines, they are largely normative and provide limited empirical guidance on the magnitude of bias introduced by excluding a given proportion of the target population, because bias depends both on the extent of the exclusions and on the differences between excluded and included students (Meinck et al., 2023).

Recent cycles of ILSAs have shown a noticeable increase in student exclusion rates, driven primarily by within-school exclusions (e.g., student-level exclusions) rather than exclusions at the school level (Atasever, Jerrim & Tieck, 2023; Atasever & Meinck, 2025). This trend is particularly noteworthy in the context of expanding inclusive education policies, which aim to ensure that students with diverse learning needs are educated in regular classrooms rather than segregated settings (Weber & Meinck, 2019). While this shift reflects a positive development toward greater educational equity and inclusion, it also poses new

[4] Exclusion rates are reported as whole percentages; values are rounded to the nearest integer percentage point (e.g., 5.1% is treated as 5%).

challenges for large-scale assessments when participation guidelines do not fully align with inclusive instructional practices within schools. In many cases, students classified as excluded are nonetheless present during the assessment session and are assigned less difficult test booklets containing items of low to medium difficulty. This practice is intended to support classroom inclusiveness and reduce stigmatization during test administration. Despite participating in the assessment and producing item-level response data, the achievement data of these excluded students are currently not incorporated into international databases, weighting procedures, or psychometric analyses and scaling. As a result, a potentially informative source of data remains unused in international reporting.

This study analyzes the performance of excluded students with available data and assesses the implications of including them in achievement estimation. By doing so, the study seeks to evaluate whether there is evidence of misinterpretation of the exclusion definitions in large-scale assessments and whether current exclusion practices should therefore be reconsidered.

Theoretical Framework and Background

TIMSS Sampling Model and Exclusion Procedures

TIMSS employs a two-stage random sampling design in which schools are selected at the first stage, and intact classes within sampled schools are selected at the second stage (Siegel & Foy, 2024). This design aims to ensure representative coverage of the target population while maintaining operational feasibility across diverse national contexts.

Exclusions may occur at both the school and student levels. School-level exclusions are defined and approved prior to sampling according to criteria, specified in the sampling standards (Siegel & Foy, 2024), mostly relating to practicability and cost-efficiency. Examples of school-level exclusions include geographical inaccessibility, very small school size, radically different curricula (e.g., international or minority-language schools), and special needs schools. The school-level exclusion rate is calculated using population statistics for the target grade and reflects the proportion of students excluded before any sampling takes place. Furthermore, within sampled schools and classes, a subset of students may be classified as excluded based on internationally defined criteria specified in the TIMSS 2023

Survey Operations Procedures Unit 1 (SOP1; IEA, 2021a)[5], if they are unable to take the test. According to these guidelines, student-level exclusions are limited to three categories: (a) students with functional disabilities that prevent them from performing in the testing situation, (b) students with severe intellectual disabilities who are unable to follow even the general instructions of the assessment, and (c) non-native speakers of the test language who are unable to overcome the language barrier, typically defined as having received less than one year of instruction in the language of the test (IEA, 2021a).

Operationally, exclusion decisions are implemented within the school and documented through standardized tracking instruments, as detailed in the TIMSS School Coordinator Manual (IEA, 2021b) and Survey Operations Procedures Unit 3 (SOP3; IEA, 2021c). School coordinators, often in consultation with national study centers, teachers, and test administrators, are responsible for identifying students who meet the international exclusion criteria and recording their status in the student tracking forms (IEA, 2021c). It should be noted that manuals or forms used by school coordinators are subject to translation and adjustment to account for specific characteristics of the national education systems.

The manuals emphasize that exclusion should be applied conservatively: students should not be excluded solely because of low academic performance, behavioral issues, or general learning difficulties, and accommodation is preferred whenever feasible. In practice, this means that students who are intellectually capable of taking the assessment but face specific access barriers—such as hearing or visual impairments, dyslexia, or physical impairments—should be supported through appropriate accommodations rather than excluded (IEA, 2021b). TIMSS permits a range of accommodations, including, for example, disabling the automatic timing of the digital assessment, providing additional supervision, or arranging a separate testing room where necessary (IEA, 2021b). Additional accommodations are permitted depending on the assessment domain; for example, reading test items aloud may be permitted for mathematics or science assessments, such as TIMSS, but not for PIRLS, where reading comprehension is the construct being measured. However, the accommodations permitted within the assessment framework may not fully address all individual accessibility needs, and their implementation may vary across education systems and schools.

[5] The Survey Operation Procedures is a series of manuals accessible to national research coordinators persons that implement TIMSS in their national contexts.

In TIMSS it is possible to adapt the difficulty of the test materials to different ability levels. For this purpose, TIMSS employs a group-adaptive assessment design in which test booklets vary in difficulty, including less difficult and more difficult booklets (see Yin & Foy, 2021, for the TIMSS 2023 Assessment Design). This design aims to better match assessment difficulty to student ability and to maximize the reliability of the assessment. National study centers responsible for implementing TIMSS in their country may assign assessment booklets to students who are classified as excluded under international exclusion criteria, for example, to support classroom inclusiveness (IEA, 2021c). If excluded students do take the assessment, they are assigned less difficult assessment booklets through IEA's within-school sampling and tracking software by default (WinW3S; see Johansone, 2024, for details on survey operations procedures). In TIMSS 2023 at grades four and eight, approximately two-thirds of countries[6] assigned test booklets to at least some students classified as excluded, resulting in the availability of achievement data for these students. However, despite being scored using the same procedures as for participating students, achievement data from excluded students are not included in the international databases and do not contribute to the final weighting or psychometric scaling of achievement. Such data may be used for national purposes but are excluded from international analyses and reporting (IEA, 2021c).

After data collection, the within-school exclusion rate is computed as the weighted proportion of excluded students among all students in participating schools (Meinck, 2020). To obtain an overall estimate representative of the full target population, the within-school exclusion rate is combined with the school-level exclusion rate. The resulting overall exclusion rate represents the proportion of the target population that is not represented in the assessed sample, incorporating exclusions that occur both prior to sampling and within sampled schools (see *Methods* section).

Rising exclusion rates

IEA quality standards specify that the overall exclusion rate should not exceed five percent of the target population, with additional scrutiny when exclusions exceed 10 percent (Siegel & Foy, 2024). When these thresholds are exceeded, international reporting includes annotations to alert readers to potential threats to representativeness and bias in national

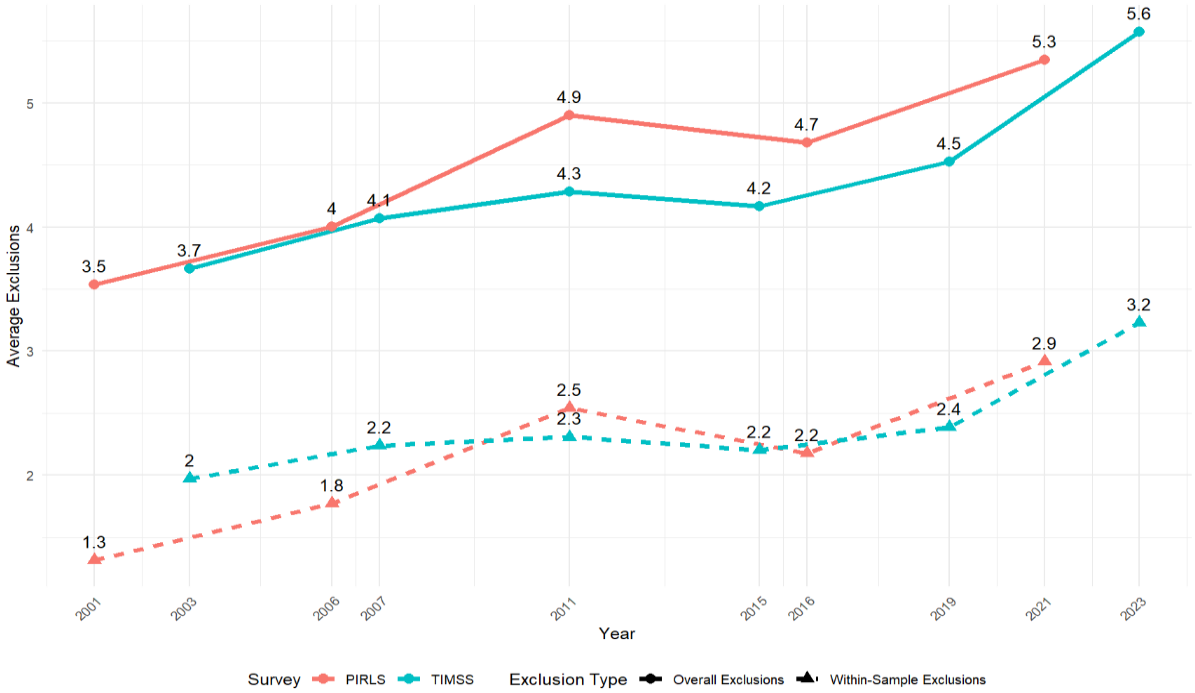
[6] When using the term "countries" we refer to educational systems and benchmark entities participating in the assessments.

achievement estimates. Beyond their informational role, the annotation thresholds encourage participating education systems to minimize exclusions by carefully applying the international exclusion criteria. Countries exceeding these thresholds receive annotations in the international reports, drawing attention to potential implications for representativeness and comparability. Still, however, a growing number of countries have exceeded this threshold in recent cycles. In their study analyzing three decades of IEA data from 1995 to 2023, Atasever et al. (2023; 2025) examine this issue, noting that while the school-level exclusion rate has remained relatively stable, within-school exclusions have significantly contributed to the rising overall exclusion rates.

Several factors may contribute to this trend. One possible explanation is the increasing inclusion of students with special educational needs (SEN) in mainstream classrooms following the implementation of the UN Convention on the Rights of Persons with Disabilities, which promotes inclusive education (United Nations, 2006). In addition, changes in national policies, increased awareness of students' support needs, and expanded guidance on accommodations may lead schools to identify more students as potentially eligible for accommodations or, where appropriate, exclusion (Weber & Meinck, 2019). Finally, Atasever et al. (2023) found consistently higher exclusion rates in primary school assessments (e.g., grade four) than in secondary school assessments (e.g., grade eight). This pattern may reflect several factors, including differences in the composition of the assessed student population, such as higher enrollment rates of students with disabilities in mainstream primary education, as well as potential differences in the implementation of exclusion criteria across grades.

In *Figure 1*, we display the average exclusion rates, both overall and within-sample, for grade four, covering all TIMSS cycles since 1995 and all PIRLS cycles since 2001.

***Figure 1*. Average overall and within-school exclusions by study (grade four)**



Note. The figure presents averages across participating countries in each cycle; changing country composition may affect comparisons over time. However, prior analyses (Atasever et al., 2023; 2025) show that the increase in exclusion rates persists after controlling for country participation, study, grade, and country fixed effects, primarily driven by increases in within-school exclusions.

The trend is evident: from 2003 to 2023, the average overall exclusion rate for TIMSS increased from 3.7 percent to 5.6 percent, a 1.9 percentage-point increase. Similarly, PIRLS saw an increase from 3.5 percent to 5.3 percent between 2001 and 2021. This increase resulted in approximately half of the participating countries exceeding the five percent exclusion limit in PIRLS 2021 and TIMSS 2023 at grade four (Almaskut et al., 2023; Siegel et al., 2024). Within-sample exclusion rates also rose, with PIRLS increasing from an average of 1.3 percent to 2.9 percent and TIMSS from 2 percent to 3.2 percent. This trend indicates that more than half of the increases in exclusion rates are due to the within-sample exclusions.

An increase in the number of students not proficient in the language of instruction is plausible in education systems that experienced substantial inflows of refugees or migrants in the years preceding data collection, and several countries have reported this as a contributing factor (Siegel et al., 2024). Other countries have reported increasing integration of students with special educational needs into mainstream schools (Weber & Meinck, 2019). Such developments may influence within-school exclusion rates by increasing the number of students requiring accommodations or whose participation requires careful consideration under the international exclusion criteria. In addition, recent research has documented increasing psychological and mental health concerns among children and adolescents (McGorry et al., 2024), although the extent to which these developments influence participation or exclusion in international large-scale assessments remains unclear. Overall, the relative contribution of these factors to the observed increase in exclusion rates has yet to be established and warrants further investigation.

Bias risks due to selective participation in ILSAs

The impact of non-response and high exclusion rates on the validity and reliability of international large-scale assessments has been widely documented, with evidence suggesting that deviations from ideal sampling conditions can introduce systematic bias in achievement estimates. Literature points to several mechanisms through which such bias may arise.

First, selective participation and non-response can lead to non-random samples of students. In the context of PISA in England, Durrant et al. (2018) and Micklewright et al. (2012) show that participation patterns are associated with achievement, likely resulting in upwardly biased estimates. Consistent with this, Jerrim (2021) reports substantial discrepancies between national examination results and PISA outcomes, indicating potential inflation of observed achievement levels. Meinck et al. (2023) compared results from a longitudinal survey in Germany, in which participation at the first measurement point (PISA 2012) was mandatory, whereas participation in the follow-up study one year later (PISA Plus) was voluntary. Analyses of non-response bias showed that students at the lower end of the achievement distribution, as well as the schools serving these students, were more likely to drop out. As a result, achievement estimates at the second measurement point were upwardly biased unless appropriate adjustments were applied.

Second, exclusion and eligibility rules may systematically reduce the representation of lower-performing students. Anders et al. (2021) highlight that in Canada, relatively high exclusion rates combined with low response rates may distort the achieved sample. Similarly, the National Audit Office (2021) reports that exclusion practices in Sweden did not fully meet OECD criteria, raising concerns about comparability. Spaull (2018) further argues that eligibility rules in ILSAs may structurally overestimate achievement, particularly when exclusion is correlated with lower performance, with similar concerns noted for Vietnam (Education Datalab, 2017).

Third, evidence from validation studies using external data suggests that these mechanisms can materially affect estimates. Borger et al. (2024), comparing Swedish PISA 2018 results with registry data, find substantial overestimation of achievement levels, reinforcing concerns that sampling and exclusion processes can have non-trivial effects on reported outcomes.

Overall, these studies indicate that non-response, selective participation, and exclusion practices can jointly bias ILSA results, with implications for cross-national comparability and interpretation of achievement differences.

Current Study

The purpose of this study is to examine the implications of student exclusion practices in TIMSS 2023 by analyzing data from students who were formally classified as excluded but nevertheless completed the assessment. Specifically, the study seeks to assess whether these students, overall or in part, demonstrate the capacity to participate meaningfully in the assessment, where meaningful participation is defined as the ability to generate psychometrically reliable achievement data within the TIMSS scaling framework, and, hence, should not have been excluded according to the study's definitions. We evaluate the feasibility of including these students in achievement estimation without compromising measurement quality (reliability) and quantify the extent to which their inclusion affects national achievement results.

Whereas nonresponse has been widely studied as a source of potential bias in ILSAs, considerably less attention has been devoted to exclusion practices, particularly in cross-national perspective. Existing studies often rely on single-country analyses or external administrative data available in only a limited number of contexts. This study contributes to the literature by providing systematic, multi-country empirical evidence based directly on the observed performance of excluded students.

Guided by this aim, the study tests the following research questions.

RQ1. Are excluded students who completed the assessment psychometrically estimable within the TIMSS scaling framework, and able to meaningfully engage with the test, as evidenced by the prevalence of students flagged as too low for reliable estimation?

RQ2. What is the achievement level of these excluded students in the TIMSS assessment, as indicated by (a) the percentage of excluded students reaching the *TIMSS Low International Benchmark (400)*, and (b) their average achievement in comparison with included students?

RQ3. How does the inclusion of these excluded students affect national-level achievement estimates in TIMSS, including the relative performance of included and excluded students and the resulting overall achievement levels?

Methods

Data

This study uses TIMSS 2023 data at grades four and eight. Achievement and background data for students classified as participating (i.e., included) are publicly available through the international TIMSS databases (Fishbein et al., 2025[7]). In contrast, data from students classified as excluded are not included in the international public-use files. Although these students may generate response data that are scored using the same procedures as for participating students, their data is excluded from international analyses, reporting, weighting, and psychometric scaling. These data arise because, in practice, excluded students are often present during the assessment session and complete the test alongside their classmates, as whole-class administration is used to avoid stigmatization and maintain standard classroom procedures. Such data are retained internally by the IEA for operational and quality-control purposes and are shared with national study centers during the preparation of the international databases for use in country-level analyses.

As outlined earlier, national education systems retain the discretion to determine whether students who are flagged for exclusion are ultimately asked to take part in the assessment. This decision may not only vary across countries but could also differ between schools within the same country. Consequently, the sample of excluded students included in this analysis represents a non-systematic, potentially unrepresentative subset of the broader group. It is therefore possible that additional students with similar characteristics remain unaccounted for, suggesting that the true extent of the issue may not be fully covered by the present study.

[7] https://timss2023.org/data/

Accordingly, analyses were restricted to countries[8] with at least 50 excluded students with available data to ensure reliable subgroup analyses. Countries with fewer excluded students, as well as countries that did not assign booklets to excluded students, were omitted from the analysis[9]. Overall, the analytic sample comprises 17 countries at grade four, and eight countries at grade eight (*Table 1*)[10]. Across these countries, the number of excluded students with available data ranged from 50 in Armenia to 1,337 in Canada, with an average of approximately 193 excluded students per country across grade four and grade eight.

***Table 1*. Student counts by country and grade before and after including students with data flagged as excluded**

	Student counts and inclusion		
	Before		After
Country	**Participating (Included)**	**Excluded with data**	**Participating (Included + Excluded)**
Panel A. Grade 4			
Armenia	4855	50	4905
Australia	5497	59	5556
Belgium (French)	4875	172	5047
Brazil	22152	544	22696
Canada	13724	1337	15061
Canada (Ontario)	4266	471	4737
Canada (Quebec)	3881	277	4158
Chinese Taipei	5321	94	5415
Czech Republic	6822	96	6918
Denmark	4953	213	5166
Hungary	5241	58	5299
Italy	4453	175	4628
Oman	8095	55	8150
Poland	4666	166	4832
Türkiye	4541	149	4690
United Arab Emirates	34872	90	34962
United Arab Emirates (Dubai)	9301	82	9383
Panel B. Grade 8			
Australia	8538	87	8625
Bahrain	5168	54	5222
Brazil	22823	409	23232
Chinese Taipei	5550	61	5611
Cyprus	4210	62	4272
Czech Republic	8110	102	8212
Italy	4591	105	4696
Türkiye	4925	103	5028

Note. Counts are unweighted.

[8] In this paper, "countries" is used as a shorthand term for all participating education systems in TIMSS, including both sovereign states and subnational or benchmarking participants that participate as separate entities in the assessment framework.
[9] While some countries use the option of including excluded students in the assessment session, there is a significant portion of students for whom no data is available.
[10] Note that the samples of Ontario and Quebec were also part of the sample for Canada, and Dubai was a part of the United Arab Emirates sample.

Denmark's grade-four data was also used, which provides a particularly noteworthy example in TIMSS 2023. Denmark initially marked 213 students as excluded who took the assessment. Following a country-initiated review conducted in consultation with the TIMSS international study center and the TIMSS sampling team, these students were subsequently included in the final international dataset. As a result, Denmark's overall exclusion rate was reduced by approximately four percentage points, from about nine percent to 5 percent. In the present study, we merged assessment data for excluded students with the TIMSS 2023 grade four and grade eight datasets, as well as the full Danish dataset containing both included and formerly excluded students.

Re-weighting Procedure

Including students who were previously excluded requires reweighting and recomputing adjustments for nonresponse. In TIMSS, exclusion rates are defined as the combination of school-level exclusions and within-sample exclusions. The school-level exclusion rate (r_1) is calculated prior to sampling using population statistics provided with the school frame and reflects the proportion of students excluded before any sample selection:

$$r_1 = \left(\frac{\text{Total number of excluded schools' students}}{\text{Total number of students}}\right) \times 100 \tag{1}$$

Within-sample exclusions are estimated after data collection using weighted student data. Let denumw (weighted denominator) denote the sum of final student weights for all students in participating schools with participation status codes 0 (excluded, no data available), 3 (not excluded, participating), and 5 (excluded, data available), and let numw (weighted numerator) denote the sum of final student weights for excluded students with participation status codes 0 and 5[11]. The within-sample exclusion rate (r_2) is defined as: $r_2 = 100 \times \frac{\text{numw}}{\text{denumw}}$. This rate is adjusted for school-level exclusions as:

$$r_{2\_adj} = \left(1 - \frac{r_1}{100}\right) \times r_2 \tag{2}$$

[11] Student participation status codes: - 0: Student excluded, did not participate - 1: Student left school - 2: Student absent - 3: Student participated - 4: Parental permission denied - 5: Student excluded, but data available

The overall exclusion rate (r_3) is then given by:

$$r_3 = r_1 + r_{2\_adj} \tag{3}$$

Under standard TIMSS procedures, only students classified as participating (status code 3) are treated as part of the assessed sample. Students classified as excluded but with available data (status code 5) are not included in the estimation of participating student totals or in the calculation of student non-participation adjustments. For the purposes of this study, students with participation status code 5 were recoded as participating students (status code 3). School-level and class-level weights were left unchanged. Student weights were updated by recalculating the student nonparticipation adjustment $A_{st1}^{i,j}$ for the *j*th classroom in the *i*th school, which is calculated as:

$$A_{st}^{i,j} = \frac{s_{rs}^{i,j} + s_{nr}^{i,j}}{s_{rs}^{i,j}} \tag{4}$$

where $s_{rs}^{i,j}$ is the number of participating students[12] in the *j*th class of the *i*th school, and $s_{nr}^{i,j}$ is the number of students sampled in this class who were selected but did not participate in the assessment (Siegel & Foy, 2024 p. 3.21.).

In the TIMSS sampling framework, the final student weight is defined as the product of the school weight, class weight, and student weight (See Siegel & Foy, 2024 for further details and other weighting components). By recoding students with participation status code 5 as participating students, the number of participating students increased while the set of eligible students denumw remained unchanged. As a result, the student non-participation adjustment factor decreased, resulting in a downward adjustment to final student weights. Further, excluded students with available data now contribute directly to the total weighted count of participating students, thereby expanding the represented target population in affected countries.

Achievement Estimation

Following the same scaling implementation procedures used for TIMSS 2023 (Fishbein et al., 2024b), plausible values (PVs) of student achievement in mathematics and

[12] Participating students {3} carry the weight of non-responding students {2,4}.

science were imputed for all students in the dataset through a psychometric analysis that uses latent regression population models and multiple imputation. The procedure relied on the published TIMSS 2023 item parameters and scale transformation constants to produce PVs on the TIMSS reporting metric, using data from the excluded students and applying the reestimated sampling weights.

For each country, latent regression population models were estimated, and plausible values were imputed using MGROUP (Sheehan, 1985). The program used students' responses to the achievement items, the published item parameters from TIMSS 2023, and new conditioning variables created from the study dataset, including the formerly excluded students. PVs were imputed simultaneously for the overall mathematics and science scales using a two-dimensional model. The resulting PVs were converted to the TIMSS reporting metric using the TIMSS 2023 published scale transformation constants.

Analysis

We applied the same statistical procedures that are used in the TIMSS international reports, including weighted descriptive statistics, benchmark analyses, and comparisons of average achievement. This ensures that the results are directly comparable to officially reported TIMSS outcomes. Following the reweighting and scaling procedures described above, excluded students with available data were treated as participating students and included in the analytic datasets.

To address the research questions, we analyzed achievement outcomes for included students (i.e., students classified as participating in the official TIMSS 2023 datasets), excluded students with available data, and the combined sample including both groups. In Denmark, where all students appear in the international database, an additional indicator was constructed to identify students who had previously been classified as excluded.

For the descriptive results, we examined exclusion rates before and after treating excluded students with available achievement data as participants, in order to assess the potential implications for overall exclusion levels and compliance with international quality standards[13]. We then analyzed student sample sizes and the prevalence of testing

---

[13] We also examined the composition of excluded students with available data by gender and grade. These descriptive statistics provide contextual information on whether excluded students differ systematically from included students in observable characteristics (see *Appendix*

Appendix A. **1**).

accommodations, distinguishing between included students and excluded students with available data. Descriptive statistics on accommodations—such as reported testing supports and digital timing adjustments—were used to assess whether students receive additional support that may have enabled their participation in the assessment.

Next, we examined whether excluded students differ from included students in the proportions flagged as having achievement too low for reliable estimation, to assess whether excluded students provide sufficient information for psychometric scaling. This approach also allows for assessing whether excluded students demonstrate achievement levels consistent with meaningful participation in the assessment. TIMSS considers students to have achievement too low for estimation if their performance on the assessment was no better than that expected from random guessing on the multiple-choice items. Within the TIMSS scaling framework, the proportion of students flagged as too low for estimation is monitored carefully, as values exceeding 15 percent raise concerns regarding the stability of achievement estimates, and values above 25 percent are considered critical and may threaten the psychometric validity of country-level results (von Davier et al., 2024). Countries exceeding these thresholds are annotated in international reporting to alert readers to potential reliability concerns. Students flagged as having achievement too low for reliable estimation nevertheless receive plausible values and are retained in international reporting and in all subsequent analyses in this study.

Student achievement was then evaluated using the TIMSS International Benchmarks, with a focus on the *Low International Benchmark (400)*. TIMSS benchmarks provide an interpretable link between student performance and the TIMSS achievement scales and are widely used in international reporting and secondary analyses. The *Low International Benchmark* is a core reference point in TIMSS reporting and is commonly interpreted as indicating basic proficiency with foundational knowledge and skills necessary to engage with the curriculum (see von Davier et al., 2024). Moreover, the TIMSS benchmarks are internationally recognized and have been acknowledged by UNESCO as contributing to the monitoring of progress toward Sustainable Development Goal (SDG) 4.1.1b (UNESCO, 2020). Benchmark analyses were conducted by estimating percentages of students reaching the *Low International Benchmark* separately for included students and excluded students with available data.

In addition, we examined differences in average achievement between included and excluded students and evaluated the impact of the excluded students' data on national achievement estimates. Weighted mean achievement was estimated separately for included students, excluded students with available data, and the combined sample. Differences between groups and changes in overall national means were evaluated using standard errors that account for the complex TIMSS sampling design (Fishbein, 2024a). Differences in point estimates were tested using two-sided tests at the $\alpha = 0.05$ level. A dependent-samples formula was applied, as the included, excluded, and overall groups are not independent (see Fishbein, 2024a for the dependent-samples approach). In addition to point differences, standardized mean differences (Cohen's d) are reported to provide a magnitude-based interpretation of the observed gaps[14]. Cohen's d values of approximately 0.20, 0.50, and 0.80 are commonly interpreted as small, medium, and large effects, respectively (Cohen, 1988).

All analyses were conducted separately by grade (grade four and grade eight) and by subject (mathematics and science), using the IEA IDB Analyzer implemented in R (IEA, 2025).

Results

Descriptive Results

Exclusion Rates Before and After Including Excluded Students with Data

We first examine how within-school and overall exclusion rates change when students who were flagged as excluded but nevertheless produced achievement data are treated as participants. *Table 2* presents school-level and within-school exclusion rates before and after this inclusion, highlighting countries that exceed the five-percent exclusion threshold specified by TIMSS quality standards. The effect is especially pronounced in grade four: 11 countries exceeded the international exclusion threshold of five percent under standard TIMSS definitions. When excluded students with available data are treated as participants, this number is reduced to two.

[14] Cohen's d was computed as the difference between included and excluded group means divided by the pooled standard deviation, $d = (M_{inc} - M_{exc})/s_p$, where $s_p = \sqrt{\frac{(n_{inc}-1)SD^2_{inc}+(n_{exc}-1)SD^2_{exc}}{n_{inc}+n_{exc}-2}}$.

Table 1. **Exclusion rates by country and grade before and after including excluded students with data**

	Exclusion rates				
		Before		After	
Country	**School-level (%)**	**Within-school (%)**	**Overall (%)**	**Within-school (%)**	**Overall (%)**
Panel A. Grade 4					
Armenia	0.9	5.3	**6.1**	4.1	5
Australia	1.9	1.9	3.8	0.8	2.7
Belgium (French)	5.3	3.6	**8.9**	0.4	**5.7**
Brazil	3.9	2.6	**6.5**	0.5	4.4
Canada	3.2	11.1	**14.3**	2.0	5.2
Canada (Ontario)	2.0	11.2	**13.2**	1.1	3
Canada (Quebec)	2.7	6.1	**8.8**	0.3	3
Chinese Taipei	0.3	2.3	2.6	0.6	0.9
Czech Republic	2.8	4.5	**7.3**	3.1	**6**
Denmark	2.3	7	**9.3**	3.0	5.3
Hungary	2.6	2.5	5.1	1.0	3.6
Italy	0.9	6.1	**7**	2.5	3.4
Oman	1.0	1.4	2.4	0.6	1.6
Poland	1.9	4.6	**6.5**	1.3	3.3
Türkiye	1.2	20.7	**21.9**	17.8	**19**
United Arab Emirates	0.6	1.1	1.7	0.9	1.5
United Arab Emirates (Dubai)	1.6	2	3.6	1.2	2.8
Panel B. Grade 8					
Australia	2.0	2.4	4.4	1.4	3.5
Bahrain	0.2	0.7	0.9	0.0	0.2
Brazil	2.4	1.9	4.3	0.3	2.7
Chinese Taipei	0.3	2	2.3	1.1	1.4
Cyprus	0.9	1.7	2.6	0.5	1.4
Czech Republic	3.0	4	**7.1**	2.9	**5.9**
Italy	0.8	4.4	5.2	2.2	3
Türkiye	1.0	19	**20**	17.2	**18.2**

Note. School-level exclusions are defined prior to sampling. Within-school exclusions refer to students excluded within sampled classes or students in excluded classes. Overall exclusions combine school-level and within-school exclusions. Exclusions after inclusion treat excluded students with available data as participants. Highlighted in **bold** cells flag countries with overall exclusion rates exceeding the TIMSS 5% threshold.

At grade four, the impact of including excluded students with data is particularly pronounced in systems with high within-school exclusion rates. In Canada, the overall exclusion rate would decrease from 14.3 percent to 5.2 percent, a reduction of 9.1 percentage points, corresponding to the inclusion of 1,337 previously excluded students. For the remaining 5.2 students no test data was available. An even larger reduction is observed in Ontario, where the overall exclusion rate drops by 10.2 percentage points (from 13.2 percent to 3.0 percent). Substantial reductions are also observed in Quebec (–5.5 percentage points), Denmark (–4.0 percentage points), Italy (–3.6 percentage points), Belgium (French; –3.2 percentage points), and Poland (–3.2 percentage points). In contrast, countries with relatively low within-school exclusion rates—such as Oman, the United Arab Emirates, and Chinese Taipei—show smaller absolute reductions in overall exclusions, typically below one percentage point. This pattern suggests that the potential to reduce overall exclusion rates lies mainly in addressing within-school exclusions rather than school-level exclusions.

At grade eight, overall exclusion rates are generally lower, and the impact of inclusion is correspondingly more modest. Nevertheless, meaningful reductions are observed in systems with higher within-school exclusions, including Italy (–2.2 percentage points), Brazil (–1.6 percentage points), and Türkiye (–1.8 percentage points). In systems where excluded students with data constitute less than one percent of the population, changes in overall exclusion rates are minimal.

Testing Accommodations Among Included and Excluded Students

As part of a descriptive check of assessment implementation, we examined the use of testing accommodations among included and excluded students. *Table 3* presents sample sizes and weighted percentages of students receiving testing accommodations, disaggregated by exclusion status, grade, and country.

***Table 3.* Sample sizes with testing accommodations by country and grade for included and excluded students**

	Students with Testing Accommodations* (Tracking)				Students with Digital Timer Disabled**			
	Included		Excluded		Included		Excluded	
Country	**Student counts**	**%**	**Student counts**	**%**	**Student counts**	**%**	**Student counts**	**%**
				Panel A. Grade 4				
Armenia	0	0	0	0	0	0	0	0
Australia	53	1	12	21	52	1	12	25
Belgium (French)	14	0	6	6	0	0	0	0
Brazil	45	0	102	20	48	0	113	22
Canada	130	1	336	24	76	1	148	13
Canada (Ontario)	44	1	140	31	14	0	76	17
Canada (Quebec)	19	0	51	17	40	1	38	16
Chinese Taipei	0	0	12	14	0	0	7	8
Czech Republic	220	3	11	12	188	3	8	10
Denmark	30	1	149	67	8	0	17	6
Hungary	6	0	6	11	0	0	0	0
Italy	168	3	43	23	168	4	44	26
Oman	6	0	4	8	7	0	0	0
Poland	0	0	0	0	0	0	0	0
Türkiye	0	0	0	0	0	0	0	0
United Arab Emirates	12	0	2	4	51	0	7	16
United Arab Emirates (Dubai)	2	0	0	0	41	0	5	13
				Panel B. Grade 8				
Australia	56	1	9	6	56	1	9	7
Bahrain	5	0	4	5	2	0	1	3
Brazil	33	0	70	13	33	0	70	13
Chinese Taipei	0	0	0	0	1	0	0	0
Cyprus	3	0	0	0	1	0	0	0
Czech Republic	171	2	5	5	159	2	9	6
Italy	203	4	28	21	240	5	20	17
Türkiye	0	0	0	0	0	0	0	0

Note. * Indicator for testing accommodations provided through tracking forms (reported by individuals within country).
** Indicator for timing accommodation provided through digital assessment system. Students with this accommodation had no time limits during the assessment.
Counts are unweighted. Percentages are weighted estimates.

Across both grade four and grade eight, excluded students with available data are consistently more likely than included students to receive testing accommodations. This pattern is evident in accommodations reported through tracking forms as well as the accommodation in timing implemented via the digital assessment system (i.e., digital timer disabled). As shown in *Table 3,* Denmark provides a particularly striking example: at grade four, 67 percent of excluded students received testing accommodations. Similar patterns—though less pronounced—are observed in countries such as Canada (including Ontario and Quebec), Brazil, Italy, and Australia, where excluded students frequently receive either general testing accommodations or timing accommodations through the digital assessment system.

At the same time, *Table 3* highlights substantial cross-national variation in accommodation practices. In several systems, the use of accommodations for excluded (and often also for included) students is limited or absent. For example, Armenia, Poland, Türkiye, and Chinese Taipei report little to no use of testing accommodations or timing adjustments for excluded students at either grade four or grade eight. These differences suggest that the likelihood of excluded students participating in the assessment could depend, besides other factors, on national policies and practices regarding accommodations. There could be policy-related or operational reasons for not providing testing accommodations. For example, countries or schools may not have policies for identifying and providing accommodations. Operationally, under-resourced schools may not be able to provide extended testing time or additional space. Nevertheless, this finding raises further questions on whether testing accommodations are consistently understood and implemented.

Overall, the findings suggest that accommodations may support participation among students at risk of exclusion, but their availability and use are far from uniform across education systems. Moreover, the observed variation highlights that currently available accommodations do not necessarily meet the full range of students' accessibility needs. Together, these findings point to considerable scope for strengthening and harmonizing accommodation practices while further improving accessibility in international large-scale assessments.

Percentages of Students with Mathematics and Science Achievement Too Low for Estimation

To examine excluded students' available data for its measurement properties, and whether they can engage meaningfully with the test, we investigated how many students were flagged as having achievement too low for reliable estimation. We summarized the results presented in *Table 4,* which reports weighted percentages of students that were flagged accordingly in mathematics and science, at grades four and eight, and contrast the findings against the respective percentages of included students.

Table 4. **Percentages of students with mathematics and science achievement too low for estimation by exclusion status and overall, grade, and subject**

Country	Math			Science		
	Included	Excluded	Overall	Included	Excluded	Overall
	% (SE)	% (SE)	% (SE)	% (SE)	% (SE)	% (SE)
Panel A. Grade 4						
Armenia	2 (0.3)	0 (0)	2 (0.3)	4 (0.4)	6 (5.7)	4 (0.4)
Australia	2 (0.2)	12 (5.5)	2 (0.3)	1 (0.2)	5 (4.3)	1 (0.2)
Belgium (French)	2 (0.2)	6 (2.0)	2 (0.2)	2 (0.3)	10 (3.1)	2 (0.3)
Brazil	**16 (0.8)**	**18 (2.8)**	**16 (0.8)**	9 (0.4)	**16 (2.8)**	9 (0.4)
Canada	2 (0.2)	8 (1.3)	2 (0.2)	1 (0.1)	5 (0.7)	1 (0.2)
Canada (Ontario)	2 (0.3)	9 (1.8)	3 (0.3)	1 (0.2)	5 (1.1)	2 (0.2)
Canada (Quebec)	1 (0.2)	3 (1.2)	1 (0.2)	1 (0.2)	4 (1.3)	1 (0.2)
Chinese Taipei	0 (0.0)	11 (3.7)	0 (0.1)	0 (0.1)	11 (3.4)	1 (0.1)
Czech Republic	1 (0.2)	5 (3.1)	1 (0.2)	1 (0.1)	4 (2.9)	1 (0.1)
Denmark	1 (0.1)	6 (2.2)	1 (0.2)	1 (0.2)	3 (1.1)	1 (0.2)
Hungary	2 (0.3)	7 (4.0)	2 (0.3)	2 (0.4)	**22 (9.0)**	2 (0.4)
Italy	2 (0.2)	12 (2.3)	2 (0.3)	1 (0.2)	10 (2.5)	2 (0.3)
Oman	10 (0.6)	**15 (5.1)**	10 (0.6)	7 (0.5)	14 (4.3)	8 (0.5)
Poland	1 (0.1)	7 (2.3)	1 (0.1)	0 (0.1)	3 (1.2)	0 (0.1)
Türkiye	1 (0.2)	9 (2.4)	1 (0.2)	1 (0.1)	4 (1.4)	1 (0.1)
United Arab Emirates	4 (0.1)	6 (3.2)	4 (0.1)	5 (0.2)	4 (2.6)	5 (0.2)
United Arab Emirates (Dubai)	1 (0.1)	4 (2.3)	1 (0.1)	1 (0.1)	2 (1.2)	1 (0.1)
Panel B. Grade 8						
Australia	4 (0.4)	6 (3.3)	4 (0.4)	2 (0.3)	4 (1.0)	2 (0.3)
Bahrain	10 (0.5)	**17 (5.8)**	10 (0.5)	5 (0.4)	11 (4.9)	5 (0.4)
Brazil	**18 (0.5)**	**27 (3.0)**	**18 (0.5)**	6 (0.3)	13 (2.3)	6 (0.3)
Chinese Taipei	1 (0.2)	**18 (6.7)**	1 (0.2)	1 (0.1)	11 (4.7)	1 (0.1)
Cyprus	5 (0.3)	10 (4.6)	5 (0.3)	3 (0.3)	**22 (6.1)**	4 (0.3)
Czech Republic	2 (0.2)	4 (1.8)	2 (0.2)	0 (0.1)	2 (1.3)	0 (0.1)
Italy	3 (0.4)	14 (3.5)	3 (0.4)	1 (0.2)	8 (2.8)	1 (0.2)
Türkiye	5 (0.4)	**15 (3.7)**	5 (0.4)	1 (0.3)	8 (3.0)	1 (0.3)

Note. Results are shown for students classified as participating (Included), excluded but have data available (Excluded), and Overall. Percentages are weighted percentage estimates; standard errors are shown in parentheses. Estimates refer to percentage of students that do not have data for reliable estimation in mathematics and science achievement. Students were considered to have achievement too low for estimation if their performance on the assessment was no better than could be achieved by simply guessing on the multiple-choice assessment items. However, such students were assigned scale scores (plausible values) by the achievement scaling procedure, despite concerns about their reliability. Results are highlighted in **bold** if they exceed 15% for reservations with reliable estimation.

Across all participating systems, students classified as excluded are more likely than included students to be flagged as too low for estimation in both mathematics and science, and in both grades. This pattern is expected, as excluded students often include those with intellectual or functional disabilities, language-related barriers, or other characteristics that may limit their ability to engage fully with the assessment. Nevertheless, the overall prevalence of too-low-for-estimation flags remains modest across countries, subjects, and grades. An important consideration when interpreting these results is that standard errors are generally larger for excluded students than for included students. This reflects the smaller number of excluded students and, in particular, the relatively small number flagged as too low for estimation.

At grade four, in several systems—such as Australia, Italy, and Canada (including Ontario and Quebec)—the proportion among excluded students is several times higher than among included students. However, absolute levels remain low in most cases, typically in the low single digits. No country exceeds the 25 percent threshold at grade four in either subject. Only a small number of systems approach or exceed the 15 percent threshold, most notably Brazil and Oman. In these systems, relatively high rates are also observed among included students, suggesting that these patterns reflect broader system-level performance challenges rather than exclusion status alone.

Clearer exclusion-specific differences emerge in some contexts, such as Hungary in science, where excluded students are substantially more likely than included students to be flagged as too low for estimation. By contrast, systems such as Armenia, Denmark, Poland, and the United Arab Emirates show very low rates for both included and excluded students, indicating that even excluded students typically provide sufficient item-level information for reliable estimation.

At grade eight, overall rates of too-low-for-estimation flags are generally lower than at grade four. In many systems, differences between included and excluded students are small. Nevertheless, several notable exceptions stand out, particularly in mathematics. In Brazil, Bahrain, Chinese Taipei, and Türkiye, excluded students are substantially more likely than included students to be flagged as too low for estimation, with Brazil exceeding the 25 percent threshold among excluded students. In science, high rates among excluded students are observed in Cyprus, although overall levels remain lower than in mathematics. In some cases, included students also show non-negligible rates, again pointing to system-level influences beyond exclusion status alone.

Overall, the findings indicate that including formerly excluded students with available data does not pose a substantial threat to psychometric integrity at the country level, as reflected in the overall estimates, and that only few students are not able to engage meaningfully with the test. When excluded students with data are included, only one country (Brazil) exceeds the 15 percent monitoring threshold at both grades in the overall results, and Brazil already exceeded the threshold based on included students, so therefore is not attributable to the inclusion of excluded students.

On the contrary, the vast majority of excluded students provide sufficient item-level information to support reliable achievement estimation within the TIMSS scaling framework. This finding is noteworthy, given that excluded students are often classified as such due to disabilities or language barriers. The results, therefore, provide strong evidence for the feasibility of incorporating excluded students with available data into achievement estimation, provided that standard quality-control procedures—such as monitoring too-low-for-estimation flags—are maintained.

Percentages of Students reaching the Mathematics and Science Low International Benchmark

To assess achievement levels of excluded students with available data, we examined first the percentages of students reaching the TIMSS International Benchmarks of Achievement in mathematics and science. Specifically, we focused on the *Low International Benchmark (400)*.

Table 5 reports the weighted percentages of students reaching the *Low International Benchmark (400)* in mathematics and science, disaggregated by inclusion status, grade, and subject. The results provide insight into whether excluded students with available data demonstrate sufficient achievement levels and how their performance compares with that of included students. It should be noted, however, that percentage estimates for excluded students are based on comparatively small sample sizes in some systems and are therefore subject to greater uncertainty.

***Table 5.* Percentages of students reaching the low international benchmark by exclusion status, grade and subject**

Country	Math		Science	
	Included	**Excluded**	**Included**	**Excluded**
Panel A. Grade 4				
Armenia	93 (0.8)	90 (5.2)	83 (1.4)	74 (10.8)
Australia	90 (0.7)	56 (11.7)	95 (0.6)	65 (13.6)
Belgium (French)	88 (1.0)	72 (4.7)	84 (1.2)	59 (5.1)
Brazil	49 (1.5)	46 (6.1)	60 (1.3)	51 (6.0)
Canada	89 (0.6)	65 (3.0)	93 (0.5)	75 (2.4)
Canada (Ontario)	89 (1.1)	60 (4.9)	94 (0.8)	73 (3.8)
Canada (Quebec)	93 (0.8)	73 (4.1)	92 (0.8)	71 (4.5)
Chinese Taipei	100 (0.1)	61 (6.9)	98 (0.2)	38 (7.9)
Czech Republic	95 (0.5)	86 (5.3)	95 (0.5)	70 (6.1)
Denmark	96 (0.5)	89 (3.0)	94 (0.5)	87 (2.7)
Hungary	89 (1.1)	55 (10.4)	91 (1.0)	58 (11.0)
Italy	91 (0.9)	67 (5.6)	93 (0.7)	68 (6.2)
Oman	57 (1.7)	41 (7.7)	62 (1.5)	43 (8.0)
Poland	96 (0.5)	70 (4.8)	97 (0.4)	66 (4.7)
Türkiye	92 (0.8)	51 (4.9)	96 (0.6)	62 (4.7)
United Arab Emirates	80 (0.5)	78 (5.8)	78 (0.5)	79 (6.7)
United Arab Emirates (Dubai)	94 (0.4)	84 (5.0)	94 (0.4)	83 (6.0)
Panel B. Grade 8				
Australia	87 (1.3)	59 (9.4)	89 (1.0)	64 (7.2)
Bahrain	58 (1.3)	29 (6.7)	67 (1.1)	30 (6.8)
Brazil	38 (1.1)	19 (3.2)	58 (1.3)	31 (5.4)
Chinese Taipei	97 (0.4)	41 (10.7)	96 (0.4)	39 (8.2)
Cyprus	84 (0.8)	43 (6.8)	75 (0.9)	27 (6.7)
Czech Republic	92 (0.7)	74 (5.2)	95 (0.6)	58 (6.4)
Italy	87 (1.1)	52 (6.6)	88 (1.1)	46 (6.1)
Türkiye	81 (1.4)	42 (6.5)	90 (0.9)	60 (8.1)

Note. Results are shown for students classified as participating (Included) and excluded but with available achievement data (Excluded). Percentages are weighted percentage estimates; standard errors are shown in parentheses. Estimates refer to the percentage of students reaching a specified achievement threshold on the TIMSS scale, namely the *Low International Benchmark (400)*.

Overall, the results indicate that many excluded students reach the *Low International Benchmark*, although at consistently lower rates than included students. The magnitude of differences between included and excluded students varies substantially across systems. In some contexts, excluded students attain the *Low Benchmark* at levels close to those of included students, while in others sizeable performance gaps suggest a stronger association between exclusion status and lower achievement.

At grade four, three broad patterns emerge. First, several systems show only small differences between included and excluded students, suggesting that exclusion status is weakly associated with benchmark attainment. This pattern is evident in Armenia, Denmark, the United Arab Emirates (including Dubai), Brazil, Oman, and, for mathematics, the Czech Republic. Although overall attainment levels differ across these systems, the gaps between included and excluded students are generally small.

Second, a larger group of systems shows moderate differences between included and excluded students. Belgium (French), Canada (including Ontario and Quebec), Italy, and Poland fall into this category, with gaps typically ranging from approximately 15 to 25 percentage points.

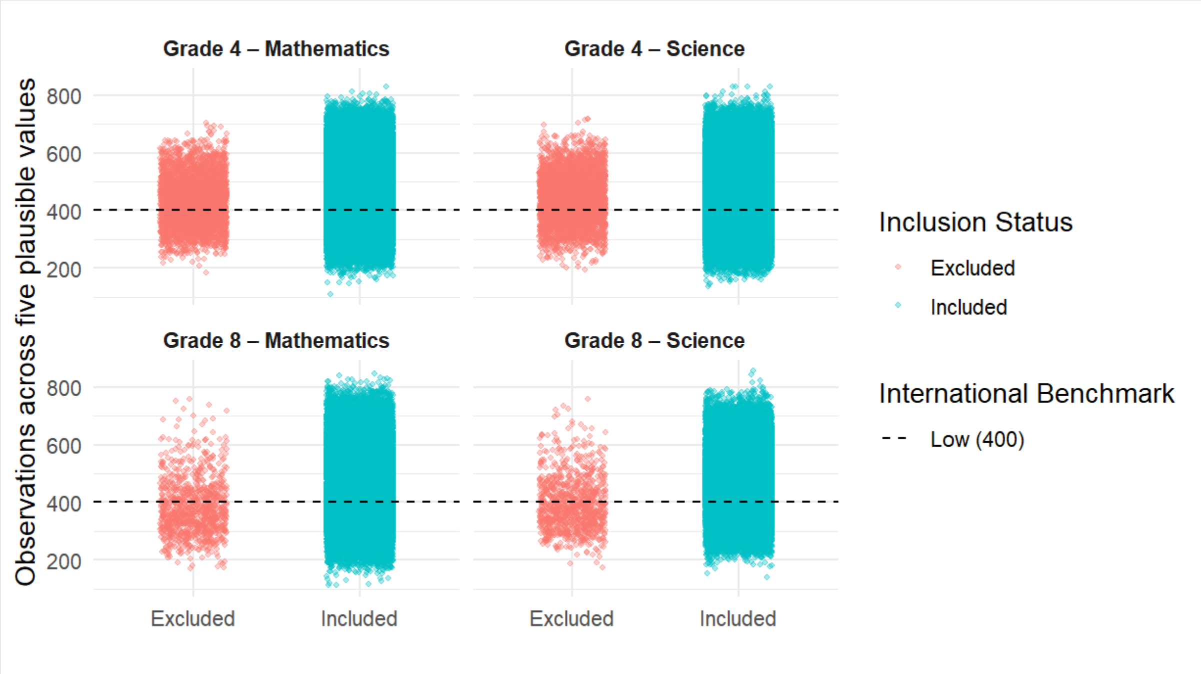
Third, Australia, Hungary, Türkiye, and Chinese Taipei show the largest disparities between included and excluded students, with gaps often exceeding 25 or 30 percentage points. In these systems, exclusion status appears more strongly associated with lower achievement. Nevertheless, a substantial share of excluded students still reaches the *Low International Benchmark*, indicating that many were able to engage meaningfully with the assessment.

At grade eight, attainment among excluded students is generally lower than at grade four, particularly in mathematics, and large differences between included and excluded students are more common. In Brazil, Bahrain, and Türkiye, fewer than half of excluded students reach the *Low International Benchmark* in mathematics. In Australia, Chinese Taipei, Cyprus, Italy, and the Czech Republic, gaps relative to included students frequently exceed 25 percentage points. Overall, grade eight shows more consistent and pronounced disparities associated with exclusion status.

Differences in Average Achievement Between Included and Excluded Students

Figure 2 displays the distributions of achievement scores (i.e., five plausible values) in mathematics and science for included and excluded students at grade four and eight. The dashed horizontal reference line indicates the *Low International Benchmark (400)*.

***Figure 2.* Mathematics and science achievement distributions by inclusion status, grade, and subject**



Across subjects and grades, included students exhibit higher average achievement and greater dispersion than excluded students. Their distributions extend further into the upper tail, whereas excluded students show more compact distributions with fewer high-achievement outliers. Despite this separation, most excluded students cluster well above the lower end of the scale, and, as described in the previous section, a substantial share exceeds the *Low International Benchmark*. This evidence supports the subsequent analyses examining average achievement differences and the impact of including excluded students on national achievement estimates.

Table 6 presents average mathematics achievement by exclusion status and grade, reporting average achievement scores and differences between included and excluded students, alongside the resulting overall estimates after inclusion.

Table 6. **Average mathematics achievement by exclusion status and overall, by grade.**

Country	**Included**	**Excluded**	**Difference in Achievement**	**Cohen's *d***	**Overall**	
	Average Achievement	**Average Achievement**	**Included - Excluded**	**Included - Excluded**	**Average Achievement**	**Difference (Overall-Included)**
Panel A. Grade 4						
Armenia	512 (2.9)	495 (12.1)	17 (11.9)	0.2	512 (2.9)	0 (2.6)
Australia	525 (2.6)	421 (19.9)	**104 (19.8)**	**1.1**	524 (2.6)	1 (2.6)
Belgium (French)	490 (2.5)	441 (6.7)	**49 (6.9)**	**0.6**	488 (2.5)	2 (2.3)
Brazil	400 (3.3)	393 (11.3)	7 (12.2)	0.1	400 (3.3)	0 (3.1)
Canada	503 (2.1)	434 (4.4)	**69 (4.2)**	**0.8**	497 (2.1)	6 (2.0)
Canada (Ontario)	502 (3.5)	423 (6.8)	**79 (6.5)**	**1.0**	494 (3.6)	**8 (3.5)**
Canada (Quebec)	514 (2.9)	448 (6.8)	**66 (6.7)**	**0.9**	510 (2.9)	**4 (2.6)**
Chinese Taipei	607 (1.6)	433 (10.7)	**174 (10.4)**	**2.6**	604 (1.7)	3 (1.6)
Czech Republic	530 (2.3)	480 (11.2)	**50 (10.6)**	**0.7**	530 (2.3)	1 (2.0)
Denmark	526 (2.1)	482 (4.9)	**44 (5.3)**	**0.6**	524 (2.1)	2 (1.9)
Hungary	520 (3.4)	434 (19.4)	**87 (18.5)**	**0.9**	519 (3.5)	1 (3.3)
Italy	513 (2.7)	445 (9.9)	**68 (9.3)**	**0.8**	510 (2.8)	2 (2.7)
Oman	421 (4.1)	393 (15.6)	28 (15.0)	0.3	421 (4.1)	0 (3.9)
Poland	546 (2.0)	458 (10.2)	**88 (10.7)**	**1.1**	543 (2.0)	3 (2.0)
Türkiye	552 (4.1)	417 (12.1)	**136 (11.9)**	**1.4**	548 (4.2)	4 (4.2)
United Arab Emirates	498 (1.2)	478 (13.7)	20 (13.6)	0.2	498 (1.2)	0 (1.0)
United Arab Emirates (Dubai)	557 (1.4)	492 (14.7)	**65 (14.6)**	**0.7**	556 (1.5)	1 (1.3)
Panel B. Grade 8						
Australia	509 (3.5)	433 (15.7)	**76 (16.0)**	**0.8**	508 (3.5)	1 (3.3)
Bahrain	427 (2.4)	359 (15.7)	**67 (15.4)**	**0.7**	426 (2.4)	0 (2.1)
Brazil	378 (2.8)	335 (7.2)	**43 (7.0)**	**0.5**	377 (2.8)	1 (2.6)
Chinese Taipei	602 (3.0)	400 (19.4)	**202 (19.3)**	**2.0**	600 (3.1)	2 (2.8)
Cyprus	494 (2.7)	393 (10.8)	**101 (11.0)**	**1.1**	493 (2.7)	1 (2.6)
Czech Republic	518 (2.3)	449 (10.3)	**69 (10.3)**	**0.8**	517 (2.3)	1 (2.2)
Italy	500 (3.1)	409 (10.0)	**92 (10.3)**	**1.1**	498 (3.0)	2 (3.0)
Türkiye	508 (4.4)	414 (17.9)	**94 (18.3)**	**0.8**	507 (4.3)	2 (4.2)

Note. Results are shown for students classified as participating (Included), excluded but with available achievement data (Excluded), and Overall. Estimates are weighted mean achievement scores; standard errors are shown in parentheses. Differences in achievement are tested at the $\alpha = 0.05$ level (two-sided), assuming approximate normality of the sampling distribution. Statistically significant differences are shown in **bold**. Standardized mean differences (Cohen's *d*) are also reported; effect sizes with $|d| \geq 0.50$ are shown in **bold**.

At grade four, substantial differences in mathematics achievement between included and excluded students are observed in the majority of participating systems. The largest gap is observed in Chinese Taipei (174 points; $d = 2.6$), followed by Türkiye (136 points; $d = 1.4$) and Australia (104 points; $d = 1.1$). Several additional systems show gaps between roughly 80 and 90 points (e.g., Poland, Hungary, and Ontario), corresponding to large standardized effects ($d \geq 0.8$). In these contexts, excluded students perform close to or more than one standard deviation below included students.

More moderate differences, typically between 40 and 70 points ($d \approx 0.5 - 0.8$), are observed in systems such as Canada, Quebec, Italy, Belgium (French), the Czech Republic, Denmark, and Dubai. In contrast, small gaps below 30 points ($d \leq 0.3$) are found in Armenia, Brazil, Oman, and the United Arab Emirates, suggesting a weaker association between exclusion status and achievement in these systems. Thus, while excluded students consistently score lower than included students, the magnitude of the gap varies considerably across countries.

Despite in some cases large point differences, the impact on overall national mean achievement at grade four is generally modest. In most systems, including excluded students changes the national means by fewer than five points. This reflects the relatively small proportion of excluded students in the weighted sample. Expectedly, the magnitude of potential bias depends jointly on the size of the achievement gap (in points or d) and the share of excluded students. Ontario and Quebec illustrate this relationship: because excluded students with data represent a comparatively larger share of the population (approximately 10 percent and 5.5 percent, respectively), their inclusion reduces national means by about eight and four points respectively.

At grade eight, differences in achievement between included and excluded students follow a similar pattern to grade four. Chinese Taipei shows the largest gap (202 points; $d = 2.0$), followed by Cyprus (101 points, $d = 1.1$) and Türkiye (94 points, $d = 0.8$), with several other systems exhibiting differences above 75 points ($d \geq 0.8$). These gaps correspond to medium-to-large standardized effects, confirming meaningful separation between the achievement distributions of included and excluded students. Nevertheless, changes in overall national means at grade eight remain small—typically one to two points—because excluded students with available data constitute only a limited share of the assessed sample.

Patterns in *science achievement* closely mirror those observed for mathematics, which is why we, in the interest of space, refrain from discussing the results in detail. Excluded students consistently performed below included students across both grades, with statistically significant differences in most participating systems (see Appendix A. 2). Standardized effect sizes further support these findings. As in mathematics, the magnitude of these gaps varies considerably across countries, while their impact on national averages remains limited.

Overall, the findings indicate that while exclusion is often associated with lower achievement, including excluded students with available data does not substantially alter national means, but does improve representativeness by incorporating lower-performing students who would otherwise be omitted.

We additionally examined differences in the standard deviations of the achievement estimates (see *Appendix A. 3* and *Appendix A. 4)*. The standard deviations were generally similar for included and excluded students, as well as for the composite estimates. In other words, a more inclusive assessment would be expected to have little to no effect on the observed variability in achievement.

Discussion and Conclusions

This study examined whether students classified as excluded in TIMSS 2023, but who nevertheless took the assessment and produced achievement data, are capable of meaningful participation and whether their exclusion biases national achievement estimates. By reweighting these students and analyzing achievement outcomes in mathematics and science at grades four and eight, the study provides new empirical evidence on the consequences of current exclusion practices for representativeness, bias, and measurement quality in large-scale international assessments.

Overall, the results provide consistent empirical answers to the research questions.

Addressing the first research question, only a small subset of the investigated excluded students was flagged as performing too poorly at the item level to support reliable achievement estimation. From a measurement perspective, concerns regarding psychometric reliability appear limited. Although excluded students are more likely to be flagged as too low for estimation than included students, exceedance of established monitoring thresholds is rare and confined to a small number of low-performing systems. In most countries, including excluded students with available data does not compromise the stability or validity of country-level achievement estimates. The evidence therefore suggests that many students initially classified as excluded who nevertheless participated in the assessment generated sufficient information to be psychometrically estimable and demonstrated meaningful engagement with the test. However, these findings apply only to the subgroup of excluded students for whom achievement data were available and should not be generalized to all students classified as excluded. The observed cross-national variation may reflect differences in how exclusion criteria are interpreted and implemented, differences in assessment accessibility and accommodation practices, or a combination of these factors.

Responding to the second research question, an important proportion of students who were classified as excluded but produced achievement data demonstrated sufficient performance levels. Across grades and subjects, many of these students reached the *Low International Benchmark* (400 points), indicating that their performance aligns sufficiently well with the test and that exclusion criteria may at times be applied too broadly. However, they may also reflect situations in which exclusion decisions are based on considerations beyond students' cognitive ability, such as the accessibility of the assessment, the adequacy of available accommodations, or concerns about students' well-being and their ability to participate under appropriate conditions.

Addressing the third research question, excluded students had lower average achievement than included students in most countries, with substantial achievement gaps observed particularly in mathematics. These point differences correspond to medium-to-large standardized effects in many systems, often approaching or exceeding one standard deviation. Therefore, exclusion practices can introduce systematic upward bias in national achievement estimates. While the magnitude of this bias is modest in systems with very small excluded populations, it becomes non-trivial in countries where excluded students constitute a larger share of the population. A negligible degree of upward bias is expected when exclusions are applied strictly according to their intended purpose and only up to the threshold of five percent—that is, when students are excluded because they are genuinely unable to meaningfully engage with the assessment. However, the empirical results show that a substantial share of excluded students demonstrate achievement well above minimal proficiency thresholds, indicating that the observed bias cannot be fully attributed to appropriate exclusion alone.

Taken together, these findings suggest that incorporating excluded students with data improves representativeness while maintaining the reliability of achievement estimates in TIMSS.

From a broader assessment perspective, these findings are directly relevant to the long-standing objective of IEA studies: maximizing inclusiveness to maintain representative coverage of the target population. This study shows that classifying excluded students with available data as participants can substantially reduce overall exclusion rates in several countries and, in many cases, bring national exclusion rates below the commonly referenced five percent threshold. The results indicate that a meaningful share of current exclusions does not faithfully reflect students' inability to engage with the assessment. More likely, students are being incorrectly excluded due to (i) a misunderstanding of the exclusion guidelines; (ii) variation in national or local policies for identifying and accommodating students with special needs; or (iii) operational limitations regarding test administration and data use. While the inclusion of excluded students in data should not be viewed as a standalone solution, it represents one concrete, immediately actionable measure to improve inclusiveness and representativeness.

In this study, we were able to investigate only data from excluded students who were allowed to participate in the test session, a decision at the discretion of national centers or school staff. It is reasonable to assume that a part of the students excluded but without data (i.e., also excluded from the test session) would be able to engage meaningfully with the test. Hence, it is likely that our study underestimates the scope of the problem, and that even more students currently excluded should be, in fact, included.

Moreover, there is little reason to assume that the findings of this study are relevant only to TIMSS. Increasing exclusion rates have also been observed across other ILSAs and may plausibly be attributable to mechanisms similar to those identified in the TIMSS context. Hence, the recommendations provided at the end of this paper should be seen in the broader ILSA context.

Limitations

Several limitations should be considered when interpreting the findings. First, students included in this study represent only the subset of students initially classified as excluded who nevertheless participated in the assessment and generated achievement data.

Because countries differ in whether they administer test booklets to students flagged for exclusion, the study cannot draw conclusions about students who remained excluded and did not participate. Consequently, the findings should not be generalized to the broader population of excluded students, and cross-country comparisons should be interpreted with caution.

Second, the concept of meaningful participation in this study is defined from a psychometric perspective, referring to students' ability to generate reliable achievement estimates within the TIMSS scaling framework. This should not be interpreted as evidence that the assessment conditions were fully appropriate, accessible, or beneficial for every student. Some students may have participated despite encountering physical, cognitive, linguistic, or emotional barriers that limited their opportunity to demonstrate their full proficiency. Likewise, lower achievement may partly reflect limitations in assessment accessibility rather than students' underlying proficiency. Although TIMSS provides testing accommodations, accommodations during administration cannot fully substitute for assessment designs that are accessible from the outset.

Third, data gaps exist in some countries, as background questionnaire information was not available for all excluded students with achievement data, limiting the ability to fully characterize this subgroup. Fourth, the relatively small number of excluded students in several education systems reduces the precision of subgroup estimates. Where excluded students constitute only a small share of the population, estimates are subject to greater sampling variability, particularly for percentage-based indicators. Fifth, the analysis is based on a single assessment cycle, making it unclear whether the observed patterns reflect persistent structural differences or cycle-specific variation. Finally, this study focuses primarily on achievement outcomes. The implications of exclusion practices for analyses of background and contextual variables, particularly where questionnaire data are unavailable, remain an important topic for future research.

Recommendations

Overall, the findings indicate that incorporating excluded students with available data improves representativeness while maintaining psychometric integrity. Exclusion is clearly associated with lower average performance, but the majority of excluded students demonstrate sufficient competence to engage meaningfully with the assessment. These results support reconsidering current exclusion practices in TIMSS and similar international studies, warranting a set of practical recommendations.

Refining Exclusion Definitions

One key implication of the findings is the need to clarify and refine exclusion definitions. The substantial variation across countries in both exclusion rates and the achievement levels of excluded students suggests that current definitions may be interpreted too broadly, or inconsistently, or misinterpreted altogether. In several systems, excluded students achieved well above internationally recognized benchmarks, indicating that exclusion decisions were not always driven by an inability to meaningfully engage with the test content. More precise operational definitions—supported by concrete examples and decision rules—could help ensure that exclusion is applied only in cases where participation would genuinely compromise measurement validity.

Strengthening Guidance on Accommodations

Closely related to the refinement of exclusion definitions is the need to strengthen guidance on accommodations. The results indicate that many excluded students perform at levels comparable to those of low-performing included students, suggesting that exclusion may, in some cases, substitute for accommodations that would otherwise enable regular participation. Evidence from the special accommodations indicators further supports this interpretation, as excluded students are substantially more likely to require special accommodations. At the same time, the results suggest that such accommodations are not always used in practice, even when they are available and permissible. However, this may be related to policy or operational circumstances within and across countries.

In this context, the transition to computer-based assessment offers additional opportunities to support inclusion, for example, through accommodations such as disabling time limits or adjusting presentation features (e.g., screen magnifiers and adjustable display settings). Providing clearer and more detailed guidance on allowable accommodations—particularly for students with special educational needs or limited language proficiency—could therefore help reduce unnecessary exclusions while preserving cross-country comparability. Such guidance should emphasize inclusion wherever feasible and clarify when accommodations are preferable to exclusion.

Future cycles of TIMSS may therefore benefit not only from clearer and more detailed guidance on allowable accommodations—particularly for students with special educational needs or limited language proficiency—but also from greater attention to accessibility during assessment development. Incorporating inclusive assessment design principles from the outset could reduce unnecessary exclusions, improve consistency in implementation across education systems, and enable a broader range of students to demonstrate their knowledge under appropriate assessment conditions. Such developments would complement existing accommodation practices while preserving cross-country comparability and further strengthening the inclusiveness and validity of international large-scale assessments.

Revising Test Assignment Defaults

This study also highlights the importance of test assignment practices as a mechanism for promoting inclusive participation. In several countries, excluded students were allowed to participate in the assessment and were assigned less difficult test booklets, thereby generating usable achievement data. These cases demonstrate that participation can be allowed without compromising test administration or psychometric quality. Rather than assigning booklets only after students are formally marked as excluded, students with special educational needs could be predefined within the sampling and tracking systems to receive a set of less demanding booklets as part of a structured special accommodation strategy. Such an approach treats participation using an easier booklet as an accommodation rather than an exception, while maintaining transparency about students' exclusion status.

Enhancing Oversight of Exclusion Practices

Given the observed cross-national variability, stronger oversight of within-school exclusion practices is warranted. National centers could be encouraged to monitor exclusion patterns more systematically, focusing not only on overall exclusion rates but also on the characteristics and performance of excluded students. Enhanced documentation and review procedures —such as a rigorous process for translating international exclusion guidelines to national contexts, along with dedicated training and support for school coordinators— would help ensure closer alignment with international definitions. These measures could reduce the risk of exclusion decisions being driven by administrative convenience rather than substantive need.

Considering Test Design Adjustments

This study points to the potential value of test design adjustments to support participation among students at risk of exclusion. The relatively low prevalence of students flagged as too low for reliable estimation suggests that many excluded students can respond meaningfully when the assessment difficulty is appropriately aligned with their proficiency. Providing easier item blocks or tailored test forms appears to be a viable strategy for enabling participation without compromising measurement quality. This interpretation is consistent with the observation that a substantial share of excluded students reaches the *Low International Benchmark,* suggesting that difficulty—not engagement—is often the primary barrier to reliable estimation. These results are particularly relevant in light of the redesigned assessment framework planned for TIMSS 2027.

Building on the group-adaptive design introduced in TIMSS 2023, the TIMSS 2027 Assessment Design again incorporates easy, medium, and difficult item blocks arranged into more and less demanding booklets, but with a new performance-based routing mechanism after the first half of the assessment (von Davier et al., 2025). If students cannot answer a few items correctly in the first half, they are assigned very easy content in the second half to ensure some information about their performance can be provided. This design reduces the likelihood that students will be flagged as too low for estimation solely due to misalignment between test difficulty and ability.

Implications for Validity and Representativeness

Overall, the results indicate that current exclusion practices may inadvertently reduce the representativeness of national samples and introduce upward bias in reported achievement, particularly in systems with higher proportions of excluded students. At the same time, the limited number of students whose data are unsuitable for reliable estimation suggests that concerns about psychometric integrity should not preclude broader inclusion. By refining exclusion criteria, strengthening accommodation guidance, revising operational defaults, enhancing oversight, and considering targeted test design adjustments, international assessments such as TIMSS can better balance measurement quality with inclusiveness.

References

Almaskut, A., LaRoche, S., & Foy, P. (2023). Sample implementation in PIRLS 2021. In M. von Davier, I. V. S. Mullis, B. Fishbein, & P. Foy (Eds.), *Methods and Procedures: PIRLS 2021 Technical Report* (pp. 8.1-8.154). Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.tr2103.kb2743

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(2), 229–249. https://doi.org/10.1007/s11092-020-09329

Atasever, U., Jerrim, J., & Tieck, S. (2023). Exclusion rates from international large-scale assessments: An analysis of 20 years of IEA data. Educational Assessment, Evaluation and Accountability. https://doi.org/10.1007/s11092-023-09416-3

Atasever, U., Meinck, S. (2025). Trends in Exclusion Rates in IEA studies over Three Decades. Paper presentation at the 11th IEA's International Research Conference (IRC), June 25-27, Italy, Rome.

Borger, L., Johansson, S., & Strietholt, R. (2024). How representative is the Swedish PISA sample? A comparison of PISA and register data. Educational Assessment, Evaluation and Accountability. https://doi.org/10.1007/s11092-024-09438-5

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Hillsdale, NJ: L. Erlbaum Associates.

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 (Statistics in Society), 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://fteducationdatalab.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/

Fishbein, B., Siegel, P., Atasever, U., Cooney, D., & Gonzalez, G. (2024a). Standard error estimation in TIMSS. In M. von Davier, B. Fishbein, & A. Kennedy (Eds.), *TIMSS 2023*

Technical Report (Methods and Procedures) (pp. 13.1-13.10). Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.timss.rs6950

Fishbein, B., Yin, L., & Bezirhan, U. (2024b). TIMSS 2023 achievement scaling implementation. In M. von Davier, B. Fishbein, & A. Kennedy (Eds.), *TIMSS 2023 Technical Report (Methods and Procedures)* (pp. 12.1-12.23). Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.timss.rs4586

Fishbein, B., Taneva, M., & Kowolik, K. (2025). *TIMSS 2023 User Guide for International Database*. Boston College, TIMSS & PIRLS International Study Center. https://timss2023.org/data

IEA. (2021a). *TIMSS 2023 Survey Operations Procedures: Unit 1 – Sampling Schools and Obtaining their Cooperation [unpublished internal document].* IEA.

IEA. (2021b). *TIMSS School coordinator manual [unpublished internal document].* IEA

IEA. (2021c). *TIMSS 2023 Survey Operations Procedures: Unit 3 – Contacting Schools and Sampling Classes [unpublished internal document]*. IEA.

IEA. (2025). Help Manual for the IEA IDB Analyzer (Version 5.0). (Available

from www.iea.nl)

Hernández-Torrano, D., & Courtney, M. G. R. (2021). Modern international large-scale assessment in education: An integrative review and mapping of the literature. Large-scale Assessments in Education, 9(1), 17. https://doi.org/10.1186/s40536-021-00109-1

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

Johansone, I. (2024). TIMSS survey operations procedures. In M. von Davier, B. Fishbein, & A. Kennedy (Eds.), TIMSS 2023 Technical Report (Methods and Procedures) (pp. 4.1-4.19). Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.timss.rs2399

Kowolik, K., & Taneva, M. (2024). Creating the TIMSS international databases. In M. von Davier, B. Fishbein, & A. Kennedy (Eds.), *TIMSS 2023 Technical Report (Methods and*

Procedures) (pp. 8.1-8.13). Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.timss.rs4924

McGorry, P. D., et al. (2024). The Lancet Psychiatry Commission on youth mental health. The Lancet Psychiatry, 11(9). https://doi.org/10.1016/S2215-0366(24)00134-7

Meinck, S. (2020). Sampling, weighting, and variance estimation. In Reliability and Validity of International Large-Scale Assessment (pp. 113–129).

Meinck, S., Heine, H.-J., Mang, J., & Nagy, G. (2023). Bias risks in ILSA related to non-participation: evidence from a longitudinal large-scale survey in Germany (PISA Plus). Educational Assessment, Evaluation and Accountability https://doi.org/10.1007/s11092-023-09422-5

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 (Statistics in Society), 175(4), 915–938.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2019). PIRLS 2021 Assessment Frameworks. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/pirls2021/frameworks/

Mullis, I.V.S., Martin, M.O., & von Davier, M. (Eds.). (2021). TIMSS 2023 Assessment Frameworks. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2023

Organization for Economic Co-operation and Development [OECD] (2024). PISA 2022 Technical Report. PISA OECD Publishing, Paris, https://doi.org/10.1787/01820d6d-en.

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

Sheehan, K. M. (1985). M-Group: Estimation of group effects in multivariate models [computer software, Version 3.2]. Princeton, NJ: Educational Testing Service.

Siegel, P., & Foy, P. (2024). TIMSS sample design. In M. von Davier, B. Fishbein, & A. Kennedy (Eds.), *TIMSS 2023 Technical Report (Methods and Procedures)* (pp. 3.1-3.30). Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.timss.rs7952

Siegel, P., Cooney, D., Ridenhour, J., & Atasever, U. (2024). TIMSS 2023 sampling implementation. In M. von Davier, B. Fishbein, & A. Kennedy (Eds.), *TIMSS 2023 Technical Report (Methods and Procedures)* (pp. 9.1-9.124). Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.timss.rs7763

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(3), 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

UNESCO. (2020). Measuring global education goals: How TIMSS helps; Monitoring progress towards Sustainable Development Goal 4 using TIMSS (Trends in International Mathematics and Science Study). https://unesdoc.unesco.org/ark:/48223/pf0000375119

United Nations. (2006). Convention on the Rights of Persons with Disabilities. https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-persons-disabilities

von Davier, M., Kennedy, A., Reynolds, K., Fishbein, B., Khorramdel, L., Aldrich, C., Bookbinder, A., Bezirhan, U., & Yin, L. (2024). TIMSS 2023 International Results in Mathematics and Science. Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.timss.rs6460

von Davier, M., Khorramdel, L., Reynolds, K., Aldrich, C. E. A., Bookbinder, A., Kennedy, A., Gonzalez, E. (2025). LaNA 2023 Linking Study Results. Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.bk4721

von Davier, M. & Kennedy, A. M. (2025). *TIMSS 2027 Assessment Frameworks*. Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.timss.vp1245

Wagemaker, H. (2013). International large-scale assessments: From research to policy. In L. Rutkowsky, M. von Davier, & D. Rutkowsky (Eds.), Handbook of International Large-Scale

Assessment: Background, Technical Issues, and Methods of Data Analysis (pp. 11–36). CRC Press.

Weber, S., Meinck, S. (2019). Students with special education needs: suggestions to improve IEA study designs to accommodate related research. Paper presentation at the 63rd Annual Conference of the Comparative and International Education Society (CIES), April 14-19, USA, San Francisco.

Yin, L. & Foy, P. (2021). TIMSS 2023 Assessment Design. In I.V.S. Mullis, M.O. Martin, & M. von Davier (Eds.), *TIMSS 2023 Assessment Frameworks*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2023

Appendix

***Appendix A. 1.* Percentages of students reaching the basic international benchmark by exclusion status, grade, and subject**

Country	Girls		Boys	
	n	% (SE)	n	% (SE)
		Panel A. Grade 4		
Armenia	16	32 (7.8)	34	68 (7.8)
Australia	15	31 (8.0)	44	69 (8.0)
Belgium (French)	80	48 (3.9)	92	52 (3.9)
Brazil	159	37 (4.4)	385	63 (4.4)
Canada	534	40 (1.8)	801	60 (1.8)
Canada (Ontario)	196	42 (2.3)	275	58 (2.3)
Canada (Quebec)	117	42 (3.2)	160	58 (3.2)
Chinese Taipei	31	33 (5.7)	63	67 (5.7)
Czech Republic	45	49 (6.4)	51	51 (6.4)
Denmark	102	48 (3.0)	111	52 (3.0)
Hungary	24	35 (6.3)	34	65 (6.3)
Italy	58	33 (3.8)	117	67 (3.8)
Oman	20	34 (6.5)	35	66 (6.5)
Poland	59	35 (4.6)	107	65 (4.6)
Türkiye	67	40 (5.0)	82	60 (5.0)
United Arab Emirates	37	41 (7.3)	53	59 (7.3)
United Arab Emirates (Dubai)	33	38 (7.6)	49	62 (7.6)
		Panel B. Grade 8		
Australia	29	47 (6.8)	57	53 (6.8)
Bahrain	28	47 (6.7)	26	53 (6.7)
Brazil	134	33 (3.2)	274	67 (3.2)
Chinese Taipei	22	35 (7.7)	39	65 (7.7)
Cyprus	18	27 (5.0)	42	73 (5.0)
Czech Republic	46	45 (5.1)	56	55 (5.1)
Italy	46	42 (6.0)	59	58 (6.0)
Türkiye	37	38 (5.5)	66	62 (5.5)

Note. Results are shown for students classified as excluded but have data available. Percentages are weighted percentage estimates; standard errors are shown in parentheses. *n* refers to unweighted counts. Gender information was collected from school administration records.

Appendix A. 2. **Average science achievement by exclusion status and overall by grade**

Country	Included	Excluded	Difference in Achievement	Cohen's *d*	Overall	
	Average Achievement	Average Achievement	Included - Excluded	Included - Excluded	Average Achievement	Difference (Overall-Included)
Panel A. Grade 4						
Armenia	458 (2.7)	446 (11.8)	11 (11.5)	0.2	458 (2.7)	0 (2.3)
Australia	549 (2.3)	451 (21.8)	**99 (21.9)**	**1.2**	548 (2.4)	1 (2.1)
Belgium (French)	480 (2.8)	418 (10.3)	**62 (9.6)**	**0.8**	478 (2.9)	2 (2.2)
Brazil	425 (3.4)	405 (11.6)	20 (12.6)	0.2	424 (3.4)	0 (3.1)
Canada	521 (1.9)	460 (4.9)	**60 (4.7)**	**0.8**	515 (2.0)	**6 (1.9)**
Canada (Ontario)	526 (3.2)	455 (7.6)	**71 (7.0)**	**0.9**	519 (3.3)	**7 (3.2)**
Canada (Quebec)	506 (2.7)	447 (6.6)	**59 (6.9)**	**0.8**	503 (2.6)	3 (2.4)
Chinese Taipei	572 (1.8)	389 (11.6)	**183 (11.5)**	**2.5**	569 (1.8)	3 (1.7)
Czech Republic	525 (2.3)	440 (10.3)	**85 (9.8)**	**1.2**	524 (2.3)	1 (2.0)
Denmark	524 (2.6)	486 (6.0)	**37 (6.1)**	**0.5**	522 (2.6)	2 (2.3)
Hungary	523 (3.3)	433 (20.1)	**91 (19.7)**	**1.1**	522 (3.4)	1 (3.2)
Italy	509 (2.5)	446 (9.3)	**63 (8.8)**	**0.9**	507 (2.6)	2 (2.3)
Oman	433 (4.1)	389 (13.8)	**44 (13.6)**	0.4	432 (4.1)	0 (3.9)
Poland	549 (2.0)	435 (8.4)	**115 (8.6)**	**1.6**	546 (2.1)	4 (2.0)
Türkiye	570 (3.5)	448 (11.5)	**122 (11.5)**	**1.4**	567 (3.6)	4 (3.5)
United Arab Emirates	495 (1.8)	480 (14.6)	15 (14.6)	0.1	495 (1.8)	0 (1.2)
United Arab Emirates (Dubai)	562 (1.7)	493 (13.5)	**69 (13.5)**	**0.7**	562 (1.7)	1 (1.4)
Panel B. Grade 8						
Australia	520 (3.2)	454 (19.7)	**66 (20.0)**	**0.7**	519 (3.1)	1 (3.0)
Bahrain	452 (2.5)	355 (15.9)	**96 (15.5)**	**0.9**	451 (2.6)	1 (2.2)
Brazil	420 (2.5)	366 (7.7)	**54 (7.3)**	**0.6**	419 (2.5)	1 (2.3)
Chinese Taipei	572 (2.6)	394 (17.3)	**178 (17.6)**	**2.0**	570 (2.6)	2 (2.3)
Cyprus	464 (2.8)	355 (13.2)	**110 (13.5)**	**1.2**	463 (2.7)	1 (2.7)
Czech Republic	527 (2.1)	428 (8.7)	**99 (8.5)**	**1.3**	526 (2.1)	1 (2.0)
Italy	500 (3.2)	399 (11.5)	**101 (11.8)**	**1.2**	498 (3.1)	2 (3.1)
Türkiye	529 (3.5)	443 (16.7)	**86 (16.8)**	**0.9**	528 (3.5)	2 (3.5)

Note. Results are shown for students classified as participating (Included), excluded but with available achievement data (Excluded), and Overall. Estimates are weighted mean achievement scores; standard errors are shown in parentheses. Differences in achievement are tested at the $\alpha = 0.05$ level (two-sided), assuming approximate normality of the sampling distribution. Statistically significant differences are shown in **bold**. Standardized mean differences (Cohen's *d*) are also reported; effect sizes with $|d| \geq 0.5$ are shown in **bold**.

Appendix A. 3. **Standard deviation of mathematics achievement by exclusion status and overall, by grade**

Country	Included	Excluded	Overall
	Standard Deviation of Achievement	**Standard Deviation of Achievement**	**Standard Deviation of Achievement**
	Panel A. Grade 4		
Armenia	70 (1.2)	68 (7.4)	70 (1.2)
Australia	92 (1.6)	91 (11.9)	92 (1.6)
Belgium (French)	75 (1.2)	72 (5.6)	76 (1.2)
Brazil	90 (2.2)	87 (5.6)	90 (2.2)
Canada	82 (1.1)	81 (2.1)	84 (1.1)
Canada (Ontario)	82 (1.8)	80 (3.5)	86 (1.8)
Canada (Quebec)	75 (1.8)	76 (3.8)	77 (1.7)
Chinese Taipei	67 (1.1)	79 (6.6)	71 (1.1)
Czech Republic	74 (1.2)	74 (8.6)	74 (1.3)
Denmark	71 (1.1)	66 (4.1)	72 (1.1)
Hungary	92 (2.4)	105 (9.4)	92 (2.5)
Italy	80 (1.3)	89 (5.5)	81 (1.3)
Oman	99 (2.4)	104 (12.4)	99 (2.4)
Poland	77 (1.2)	91 (7.2)	79 (1.2)
Türkiye	100 (1.9)	109 (7.5)	103 (1.9)
United Arab Emirates	107 (0.9)	95 (7.8)	107 (0.9)
United Arab Emirates (Dubai)	91 (1.2)	89 (7.5)	91 (1.2)
	Panel B. Grade 8		
Australia	93 (2.3)	93 (10.7)	93 (2.4)
Bahrain	99 (1.6)	84 (12.7)	99 (1.6)
Brazil	89 (2.7)	80 (6.7)	89 (2.7)
Chinese Taipei	99 (1.8)	107 (18.8)	101 (1.8)
Cyprus	92 (2.0)	78 (8.5)	93 (2.0)
Czech Republic	84 (1.4)	80 (6.3)	84 (1.4)
Italy	86 (1.7)	77 (6.4)	86 (1.6)
Türkiye	115 (2.4)	130 (12.5)	116 (2.4)

Note. Results are shown for students classified as participating (Included), excluded but with available achievement data (Excluded), and Overall. Estimates are standard deviation of achievement scores; standard errors are shown in parentheses.

Appendix A. 4. **Standard deviation of science achievement by exclusion status and overall, by grade**

Country	**Included**	**Excluded**	**Overall**
	Standard Deviation of Achievement	**Standard Deviation of Achievement**	**Standard Deviation of Achievement**
	Panel A. Grade 4		
Armenia	61 (1.2)	67 (9.9)	61 (1.2)
Australia	81 (1.4)	94 (11.7)	82 (1.4)
Belgium (French)	78 (1.5)	82 (7.8)	79 (1.5)
Brazil	95 (2.0)	98 (5.2)	95 (2.0)
Canada	78 (1.0)	84 (2.7)	80 (1.1)
Canada (Ontario)	78 (1.7)	83 (4.2)	81 (1.7)
Canada (Quebec)	73 (1.1)	79 (4.5)	74 (1.1)
Chinese Taipei	72 (0.9)	79 (7.2)	76 (1.2)
Czech Republic	72 (1.1)	72 (7.9)	73 (1.1)
Denmark	75 (1.1)	70 (4.2)	75 (1.2)
Hungary	86 (2.6)	97 (9.4)	87 (2.6)
Italy	71 (1.2)	81 (4.4)	73 (1.2)
Oman	103 (2.5)	107 (11.2)	103 (2.5)
Poland	72 (1.3)	80 (5.4)	76 (1.4)
Türkiye	88 (1.8)	102 (6.5)	91 (1.7)
United Arab Emirates	119 (0.8)	103 (10.1)	119 (0.8)
United Arab Emirates (Dubai)	95 (1.3)	97 (10.1)	95 (1.3)
	Panel B. Grade 8		
Australia	94 (2.0)	107 (12.2)	95 (2.0)
Bahrain	106 (1.5)	88 (8.5)	107 (1.5)
Brazil	88 (2.0)	83 (5.0)	89 (2.0)
Chinese Taipei	91 (1.4)	99 (19.0)	93 (1.4)
Cyprus	92 (2.0)	77 (7.9)	93 (2.0)
Czech Republic	77 (1.4)	75 (5.5)	78 (1.4)
Italy	83 (1.7)	83 (8.3)	85 (1.7)
Türkiye	98 (1.8)	115 (8.9)	99 (1.8)

Note. Results are shown for students classified as participating (Included), excluded but with available achievement data (Excluded), and Overall. Estimates are standard deviation of achievement scores; standard errors are shown in parentheses.

Declarations

Acknowledgements

We thank the International Association for the Evaluation of Educational Achievement (IEA) for funding this project through the IEA's fourth round of Research and Development Funds (R&D).

Funding

This work was supported by the International Association for the Evaluation of Educational Achievement (IEA) under the fourth round of the Research and Development Fund (Grant title: Striving for More Inclusiveness: Understanding and Addressing Student Exclusions in TIMSS, 2025; EUR 55,000).

Author Contributions

Umut Atasever: Conceptualization, Funding acquisition, Methodology, Software, Data curation, Formal analysis, Investigation, Visualization, Writing – original draft**; Sabine Meinck:** Conceptualization, Funding acquisition, Methodology, Supervision, Resources, Writing – review & editing**; Matthias von Davier:** Methodology, Writing – review & editing; **Bethany Fishbein:** Methodology, Writing – review & editing**; Rolf Strietholt:** Supervision, Resources, Writing – review & editing.