

Beyond Structural Indicators: Human Development, Student Self-Efficacy, and Mathematics Achievement in TIMSS 2019

João Marôco

`joao.maroco@ulusofona.pt`

Universidade Lusófona

Heidi Heidi Harju-Luukkainen

University of Jyväskylä

Juhani Rautopuro

University of Jyväskylä

Research Article

Keywords: TIMSS 2019, mathematics achievement, multilevel modelling, human development, student self-efficacy, international large-scale assessment

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11116355/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Mathematics achievement is an important educational and societal outcome, yet substantial cross-national differences persist. This study examines how student-, school/teacher-, and country-level characteristics are associated with Grade 8 mathematics achievement and whether structural educational indicators retain independent associations after accounting for student and school contexts.

Methods

Using TIMSS 2019 grade 8 data from 38 participating countries and education systems, a three-level hierarchical linear model was estimated, with students nested within schools and schools nested within countries. The analysis incorporated the five mathematics plausible values and TIMSS sampling weights and examined student characteristics, school and teacher factors, and country-level indicators, including the Human Development Index (HDI), public educational expenditure, curriculum coverage, and national or regional examinations.

Results

Mathematics confidence and home educational resources were among the strongest student-level correlates of achievement. At the country level, HDI showed the strongest association with mathematics achievement after adjustment for student- and school-level characteristics. Simple country-level relationships between structural educational indicators and TIMSS achievement were substantially attenuated when broader human-development conditions and student-, school-, and teacher-associated characteristics were modelled simultaneously. These findings indicate that simple cross-national associations between structural educational indicators and mathematics achievement are substantially conditioned by broader human-development conditions and by the composition of student and school contexts.

Conclusions

The findings should not be interpreted as evidence that educational policies or investment are ineffective. Rather, they suggest that structural indicators such as educational expenditure, curriculum coverage, and the presence of national or regional examinations cannot be interpreted independently of the wider socioeconomic and educational context in which they operate. TIMSS 2019 provides a pre-pandemic baseline for examining these relationships and offers a reference point for subsequent research using later TIMSS cycles.

Introduction

Mathematics and numeracy play a central role in contemporary societies by supporting scientific reasoning, technological development, and evidence-informed decision-making. Numeracy– defined as

the ability to interpret and apply mathematical information in everyday educational, professional, and civic contexts – is increasingly recognized as a foundational competence for individual well-being and societal functioning (Duchhardt et al., 2017; National Numeracy, 2019; OECD, 2024; Sellars, 2018).

Quantitative reasoning and mathematical modeling underpin core applications in science, technology, and public policy, enabling the analysis of complex systems and evidence-based decision-making (Sriraman, 2021). Nevertheless, functional numeracy deficits remain widespread across contemporary societies. For example, only 22% of working-age adults in the United Kingdom demonstrate functional numeracy skills (National Numeracy, 2019). Data from the Programme for the International Assessment of Adult Competencies (PIAAC) reveal that 25% of adults across participating OECD countries score at or below Level 1 in numeracy, demonstrating only basic, single-step numerical capabilities (OECD, 2024). These proportions vary substantially across nations. They reach 38% in Portugal and 56% in Chile, compared to 10%–12% in Japan and Finland. Notably, several high-income economies exhibit above-average shares of low-performing adults, including France (28%), Spain (30%), the United States (34%), and Italy (35%) (OECD, 2024). Empirical research indicates that these persistent disparities stem less from innate cognitive limitations than from negative self-perceptions, math anxiety, and culturally pervasive attitudes that normalize unfavorable views toward mathematics (Chang, 2023; Hembree, 1990; Muhaimin et al., 2024; Namkung et al., 2019; National Numeracy, 2020).

The Importance and Challenges of Secondary Mathematics Education

A solid foundation in secondary mathematics fosters cognitive development, analytical reasoning, and problem-solving skills necessary to navigate a data-driven world (Chiasson & Freiman, 2022; Golding, 2018; Harju-Luukkainen et al., 2022; Marôco, 2020, 2025). Furthermore, quality mathematics education cultivates intellectual curiosity, resilience, and academic self-efficacy, encouraging students to persevere through complex challenges (Lee & Johnston-Wilder, 2016; National Council of Teachers of Mathematics, 2014).

Despite its importance, mathematics education faces persistent challenges, including widespread student math anxiety, inadequate instructional resources, and difficulties in adapting curricula to diverse student populations (Chinn, 2009; Mierluș-Mazilu & Yilmaz, 2024). Identifying which student-, school-, teacher-, and system-level factors explain cross-national achievement differences remain a vital inquiry. However, research examining the structural correlates of large-scale assessment outcomes at the end of lower secondary education remains relatively scarce compared to the extensive literature on Grade 4 students in TIMSS or 15-year-olds in PISA (Blömeke et al., 2022; Harju-Luukkainen et al., 2022; Hernández-Torrano & Courtney, 2021; Teig & Steinmann, 2023).

TIMSS Framework, Sampling, and the Pre-Pandemic Baseline

The International Association for the Evaluation of Educational Achievement (IEA) Trends in International Mathematics and Science Study (TIMSS) has assessed global mathematics achievement since 1995 (Beaton et al., 1996; IEA, 2024). The TIMSS Grade 8 mathematics framework evaluates four content domains: Number (30%), Algebra (30%), Geometry (20%), and Data and Probability (20%), across three cognitive domains: Knowing (35%), Applying (40%), and Reasoning (25%) (Mullis & Martin, 2017). The assessment emphasizes both procedural fluency and conceptual understanding within real-world problem scenarios (Mullis & Martin, 2017).

To ensure cross-national comparability, TIMSS 2019 employed a rigorous multi-stage stratified cluster sampling design, selecting representative samples of schools and intact Grade 8 classrooms (Foy et al., 2020). In addition to student testing, contextual questionnaires were administered to students, mathematics teachers, school principals, and National Research Coordinators (Fishbein et al., 2021). To accommodate extended content coverage without burdening individual respondents, TIMSS utilized a matrix-sampling booklet design. Individual student proficiency was estimated using Item Response Theory (IRT) scaling and background conditioning models, generating five plausible values per student to accurately represent the population score distribution and account for measurement uncertainty (Martin et al., 2020; von Davier, 2020).

TIMSS 2019 is particularly appropriate for the present study because it provides a pre-pandemic baseline for examining the structural, institutional, and student-level correlates of Grade 8 mathematics achievement. Although TIMSS 2023 provides the most recent cross-national achievement data, its assessment followed the substantial disruption to schooling caused by the COVID-19 pandemic and therefore represents a different educational context from that observed in 2019. The purpose of the present study is not to estimate pandemic-related learning losses, but to examine cross-national correlates of mathematics achievement under relatively stable, pre-pandemic educational conditions. TIMSS 2019 therefore offers a theoretically and temporally coherent baseline against which subsequent changes in achievement and educational inequalities can be interpreted. This choice is not intended to imply that TIMSS 2019 is preferable for every research question; rather, it follows directly from the study's objective of identifying pre-pandemic correlates of cross-national differences in Grade 8 mathematics achievement.

The present study extends this literature in three ways. First, rather than examining student, school, teacher, and national characteristics in isolation, it evaluates their associations with mathematics achievement simultaneously within a three-level framework, allowing the contribution of country-level structural indicators to be considered after accounting for differences in student and school composition. Second, it distinguishes broader national human development from more specific educational-system characteristics, providing a more stringent test of whether indicators such as educational expenditure, curriculum coverage, and national or regional examinations retain independent associations with achievement after adjustment for contextual factors. Third, the study examines the robustness of these associations through multiple-imputation and leave-one-country-out sensitivity analyses. This approach is intended not to identify causal effects, but to clarify which observed

associations remain consistent when potentially confounding contextual and compositional factors are considered.

Study Aims and Research Hypotheses

Official TIMSS reports provide essential descriptive information on cross-national achievement and contextual relationships. However, such reporting provides limited insight into how factors operating at different levels of the educational system are jointly associated with achievement, particularly when student composition, school characteristics, and country-level context are considered simultaneously. Although previous secondary analyses have advanced this literature, comparatively fewer studies use a three-level framework that explicitly models students within schools and schools within countries while incorporating country-level contextual indicators.

To address these methodological and substantive considerations, the present study implements a three-level Hierarchical Linear Model (HLM), nesting students within schools and schools within countries. This design allows school-level associations to be estimated after adjustment for student characteristics and permits country-level indicators to be examined conditional on student and school composition. Specifically, we examine whether widely discussed structural indicators, including curriculum coverage, national or regional assessment systems, and public educational expenditure, remain associated with mathematics achievement after broader societal development, particularly the Human Development Index (HDI), and student self-efficacy are taken into account.

The study is guided by the expectation that achievement is associated with factors operating at multiple levels of the educational system. The specific research hypotheses are:

Hypothesis 1

(Student Level): Students' socioeconomic background, home educational resources, and confidence in mathematics are positively associated with mathematics achievement.

Hypothesis 2

(School Level): School-level socioeconomic composition and school academic emphasis significantly explain variations in student achievement.

Hypothesis 3

(Teacher characteristics at the school level): Teachers' professional experience, formal qualifications, and instructional practices are positively associated with student proficiency.

Hypothesis 4

(Country Level): Country-level human-development and education-system characteristics are associated with differences in Grade 8 mathematics achievement after adjustment for student- and school-level characteristics. The magnitude and statistical significance of these associations are expected to differ across indicators.

<

Data and Methodology

Data and Sample Description

The Grade 8 TIMSS 2019 assessment involved 38 participating countries and regions, representing one of the most extensive international evaluations of secondary mathematics achievement. In addition to measuring student mathematics achievement, TIMSS gathered contextual data on students, family, classroom, and school environments through standardized questionnaires administered to students, mathematics teachers, and school principals. Standardized indices were derived using partial credit Item Response Theory (IRT) models (see Yin & Fishbein, 2020, Appendix 16B). The mathematics achievement dataset is publicly accessible via the TIMSS and PIRLS International Study Center at Boston College.

Although subsequent assessment cycles, including TIMSS 2023, are available, the 2019 dataset provides the appropriate empirical context for the present research questions. It captures macro-level socioeconomic and institutional correlates of mathematics achievement before the COVID-19 disruption, avoiding the additional contextual heterogeneity introduced by pandemic-related school closures, emergency remote instruction, and differential recovery trajectories. TIMSS 2023 is therefore acknowledged as the most recent cycle, but it addresses a different temporal context; the present study deliberately establishes a pre-pandemic baseline for interpreting subsequent changes in achievement.

Data Merging and Initial Screening

Data management and preliminary screening were conducted using the `intsvy` package (v. 2.9; Caro & Biecek, 2017) within the R statistical system (v. 4.3; R Core Team, 2024). Student, school, and teacher files were merged across levels while retaining survey structure. National mathematics means and standard errors were calculated across all five plausible values (PVs) for mathematics achievement using appropriate estimation procedures to account for the complex multi-stage survey design of TIMSS 2019.

Multicollinearity among explanatory variables was evaluated using Variance Inflation Factors (VIF) from the `car` package (v. 3.0.13; Fox & Weisberg, 2019). All calculated VIF values were between 1.11 and 1.43, falling well below the standard threshold of 5 and confirming the absence of severe multicollinearity among predictors.

Missing Data and Analytic Samples

The analytic sample comprised $n = 228,158$ students in the null model and $n = 87,739$ students in the fully specified model. Missingness in predictor variables ranged from 0.3% to 100% (see supplementary figure S1), reflecting heterogeneous item nonresponse and country-specific non-administration of some contextual variables. Planned matrix sampling of achievement items is handled through the TIMSS plausible-value methodology and is distinct from ordinary missingness in questionnaire variables. The number of observations contributing to each model is therefore reported explicitly in Table 2.

Missingness in contextual variables was addressed through the model-specific analytic samples used by the multilevel estimation procedure. Rather than treating planned matrix sampling as conventional missing data, the five TIMSS mathematics plausible values were retained to represent uncertainty in the achievement estimates. Because contextual variables differ in their patterns of country availability and nonresponse, the composition of the analytic sample can change across model specifications. Accordingly, model-specific sample sizes are reported and changes in coefficients across models are interpreted with attention to possible changes in sample composition. The analysis does not assume that contextual-variable missingness is Missing Completely at Random (MCAR), and no causal interpretation is attached to differences in estimates across samples.

Table 1

Variables included in the three-level Hierarchical Linear Model (HLM), with measurement characteristics and scale reliability

HLM Level	Unit	Variable (TIMSS code)	Description	Reliability (Cronbach's α)
Level 1	Students	Sex (BSBG01)	Student sex	Single item
		Time on maths homework (BSBM42BA)	Time devoted to mathematics homework	Single item
		Extra maths lessons (BSBM43AA)	Number of extra mathematics lessons (last 12 months)	Single item
		Home Educational Resources (BSBGHER)	Learning resources at home and parental education	0.25–0.64; M = 0.43 (SE = 0.010)
		Sense of School Belonging (BSBGSSB)	Feelings of belonging and attachment to school	0.70–0.86; M = 0.80 (SE = 0.006)
		Student Bullying (BSBGSB)	Frequency of bullying experiences	0.78–0.92; M = 0.87 (SE = 0.006)
		Students Like Learning Mathematics (BSBGSLM)	Enjoyment of and interest in mathematics	0.88–0.95; M = 0.93 (SE = 0.003)
		Instructional Clarity in Maths Lessons (BSBGICM)	Clarity of teachers' explanations and expectations	0.81–0.94; M = 0.88 (SE = 0.004)
		Disorderly Behaviour in Maths Lessons (BSBGDML)	Classroom disruptions (higher scores = less disorder)	0.73–0.93; M = 0.87 (SE = 0.008)
		Students Confident in Mathematics (BSBGSCM)	Confidence in learning and performing mathematics	0.75–0.92; M = 0.89 (SE = 0.008)
		Students Value Mathematics (BSBG SVM)	Importance attributed to mathematics	0.84–0.92; M = 0.89 (SE = 0.003)
		Self-Efficacy for Computer Use (BSBGSEC)	Confidence in using computers	0.76–0.89; M = 0.82 (SE = 0.008)
Level 2	Schools / Teachers	% Economically Disadvantaged Students (BCBG03A)	Proportion of disadvantaged students	Single item
		Instructional days per year (BCBG06A)	Total number of instructional days	Ratio variable

HLM Level	Unit	Variable (TIMSS code)	Description	Reliability (Cronbach's α)
		School library (BCBG10A)	Existence of a school library	Binary
		School Discipline Problems (BCBGDAS)	Severity of school disciplinary problems	0.81–0.98; M = 0.91 (SE = 0.008)
		School Emphasis on Academic Success (BCBGEAS)	Emphasis on academic achievement	Reliability not reported
		Years teaching (BTBG01)	Teachers' teaching experience (median per school)	Ratio variable
		Teachers' education level (BTBG04)	Highest level of formal education	Ordinal
		Teacher majored in mathematics (BTBG05F)	Mathematics as major area of study	Binary
		Class size (BTBG10)	Number of students in class	Ratio variable
		Teaching Limited by Student Needs (BTBGLSN)	Constraints due to students' needs	0.68–0.86; M = 0.78 (SE = 0.008)
		Teachers' Job Satisfaction (BTBGTJS)	Satisfaction with the teaching profession	0.78–0.95; M = 0.89 (SE = 0.006)
		Teachers majored in mathematics / math education (BTDMMME)	Major field of study	Binary
		% Students taught number topics (BTDMMNUM)	Curriculum coverage: number	Ratio variable
		% Students taught algebra topics (BTDMALG)	Curriculum coverage: algebra	Ratio variable
		% Students taught geometry topics (BTDMGEO)	Curriculum coverage: geometry	Ratio variable
		% Students taught data/probability topics (BTDMDAT)	Curriculum coverage: data/probability	Ratio variable
Level 3	Countries	Human Development Index (HDI)	National development indicator	External index
		Expenditure on education (ExpEdu)	Education spending as % of GDP	External indicator

HLM Level	Unit	Variable (TIMSS code)	Description	Reliability (Cronbach's α)
		Curriculum coverage (PercCurrCovered)	Alignment with TIMSS mathematics framework	External indicator
		National/regional exams (NatRegExam)	Presence of exams	Binary
		School maths self-evaluation (SchSelfEv)	Maths self-evaluation practices	Binary

Note: All student- and school-level variables are described at the TIMSS & PIRLS International Study Centre, Boston College, MA. Country-level variables HDI and ExpEdu were retrieved from UNDP (<https://hdr.undp.org/data-center/human-development-index#/indicies/HDI>) and the World Bank (<https://data.worldbank.org/indicator/SE.XPD.TOTL.GD.ZS>), respectively

Hierarchical Linear Modelling (HLM)

Hierarchical Linear Modelling (HLM) is specifically designed to analyze data with a nested structure, such as students clustered within schools, and schools clustered within countries, as observed in the Trends in International Mathematics and Science Study (TIMSS). Unlike single-level correlational or regression analyses common in international assessment reports, HLM accounts for the hierarchical dependence among observations and permits variance partitioning and estimation of associations at the student, school, and country levels (Raudenbush & Bryk, 2002). The three-level framework adopted in this study reflects the theoretical and empirical structure of the data, wherein student-level variables (Level 1) capture individual characteristics and learning experiences, school and teacher variables (Level 2) represent institutional and instructional contexts, and country-level variables (Level 3) reflect broader educational systems and policy environments.

This analytical approach enables the simultaneous estimation of multi-level effects, the decomposition of variance in mathematics achievement across all three levels, and the examination of how contextual factors are associated with cross-national differences beyond individual student characteristics. Consequently, HLM extends standard descriptive and bivariate TIMSS analyses by offering a theoretically grounded, multilevel analysis of variation in achievement. Substantively, this method permits a reassessment of widely assumed correlates of mathematics achievement, such as curriculum coverage, national assessments, and educational expenditure, net of student- and school-level compositional differences, allowing for stronger conclusions regarding which factors remain meaningful explanatory variables in international comparisons.

To account for this nested structure, students (i) within schools (j), and schools within countries (k), we estimated a three-level HLM with random intercepts and fixed slopes. The full model is specified as:

$$\text{Level1(Student): } Y_{ijk} = \pi_{0jk} + \sum_{p=1}^P \pi_{pjk} X_{pijk} + e_{ijk}, \quad e_{ijk} \sim N(0, \sigma^2)$$

$$\text{Level2(School): } \pi_{0jk} = \beta_{00k} + \sum_{q=1}^Q \beta_{0qk} W_{qjk} + r_{0jk}, \quad r_{0jk} \sim N(0, \tau_\pi)$$

$$\text{Level3(Country): } \beta_{00k} = \gamma_{000} + \sum_{m=1}^M \gamma_{00m} Z_{mk} + u_{00k}, \quad u_{00k} \sim N(0, \tau_\beta)$$

where Y_{ijk} represents student i 's mathematics achievement plausible value; X_{pijk} , W_{qjk} , and Z_{mk} denote vectors of student-, school/teacher-, and country-level predictors, respectively; and e_{ijk} , r_{0jk} , and u_{00k} represent random residuals at each corresponding level. Level 2 clusters were constructed by combining the country ID (IDCNTRY) and school ID (IDSCHOOL), whereas Level 3 clusters comprised countries.

To address measurement error associated with the matrix-sampling design of the Large-Scale Assessment (LSA), models were estimated using the five mathematics plausible values (PVs) through the *mixPV* function (Huang, 2024) in the WeMix package (v. 4.0; Bailey et al., 2023). Regression coefficients, standard errors, and degrees of freedom were pooled across PVs according to Rubin's (1987) combining rules. Sampling weights were incorporated to account for the unequal probabilities of selection inherent in the TIMSS multistage sampling design (Rutkowski et al., 2010). At Level 1, the student sampling weight (SENWGT) and, at Level 2, the school sampling weight (SCHWGT) were included in the multilevel estimation and standardized by their respective overall sample means. This normalization preserves the relative contribution of sampled students and schools while avoiding changes in the overall scale of the weights resulting solely from their magnitude. At Level 3, country-level weights were set to 1. This was an explicit analytical choice rather than an assumption that participating education systems have equal population sizes or equal sampling probabilities. It ensures that each participating education system contributes equally to the estimation of country-level associations. Consequently, the country-level coefficients should be interpreted as associations across the participating TIMSS education systems, rather than as student-population-weighted estimates for a global population. The resulting estimand is therefore the conditional association between country-level characteristics and mathematics achievement across participating education systems, while the Level-1 and Level-2 weights retain the within-system sampling structure. This specification is consistent with the study's focus on explaining differences between participating countries rather than estimating an association representative of the aggregate population of all TIMSS students.

Missing data on predictors and derived indices were therefore handled using the model-specific available observations in WeMix, while the five plausible values were retained for achievement estimation. This approach preserves the TIMSS survey structure and avoids treating planned matrix sampling as ordinary missingness. Because the analytic sample changes as variables with different missingness patterns are introduced, the nested models are interpreted as complementary specifications rather than as strict sequential block tests based on an identical sample.

Model performance was evaluated using Marginal R^2 (R_m^2 , variance explained by fixed effects) and Conditional R^2 (R_c^2 , total variance explained by fixed and random effects; Nakagawa & Schielzeth,

2012):

$$R_m^2 = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_r^2 + \sigma_e^2}$$

$$R_c^2 = \frac{\sigma_f^2 + \sigma_r^2}{\sigma_f^2 + \sigma_r^2 + \sigma_e^2}$$

where σ_f^2 is the variance of the fixed effects, $\sigma_r^2 (\tau_\pi + \tau_\beta)$ represents the total variance of the random effects across higher levels, and $\sigma_e^2 (\sigma^2)$ is the residual variance at the student level. Intraclass Correlation Coefficients (ICCs) were calculated to evaluate variance partitioning across levels.

Standardized beta coefficients (β) were used to describe the magnitude of conditional associations ($p < .05$). Following Cohen (1988) and Nieminen (2022), effects were categorized as trivial ($\beta < 0.10$), small ($0.10 \leq \beta < 0.20$), medium ($0.20 \leq \beta < 0.30$), or large ($\beta \geq 0.30$) ($\beta \geq 0.30$).

Sensitivity Analysis to Missing Data and Leave-one-country-out

A sensitivity analysis was conducted to assess the robustness of the TIMSS 2019 mathematics achievement model to non-structural missing data. Teacher-level variables were aggregated to the school level using the median across teachers within each school. Missing non-structural student-, school-, and country-level covariates were handled using multiple imputation by chained equations (MICE; van Buuren & Groothuis-Oudshoorn, 2011), with two-level predictive mean matching (2l.pmm for student-level variables and 2lonly.pmm for school-level variables). Ten imputed datasets were generated using 10 iterations per imputation. Country-level expenditure on education was imputed separately using predictive mean matching and subsequently merged with the imputed datasets.

For each imputed dataset, mathematics achievement was modelled using the five TIMSS plausible values in a weighted three-level linear mixed model, with students nested within schools and schools nested within countries. Sampling weights were incorporated using TIMSS student-, school-, and country-level weights. Estimates across plausible values were obtained using mixPV and WeMix (Baumert et al., 2020) and estimates across the 10 imputed datasets were pooled using Rubin's rules (Rubin, 1987). Complete-case estimates were used as the reference for evaluating differences in parameter estimates and standard errors.

To assess whether the estimates depended on any single country, we conducted a leave-one-country-out (LOCO) sensitivity analysis, following the logic of influence diagnostics for multilevel models, in which the model is re-estimated after omitting each higher-level unit in turn (Nieuwenhuis et al., 2012; Van der Meer et al., 2010). Because refitting the full weighted model with plausible values for each of the 38 countries was computationally prohibitive, we used a screening model: a two-level random-intercept linear model (schools nested in countries) fitted with the lme4 package in R (Bates et al., 2015), with the

mean of the five mathematics plausible values as the outcome and the same fixed effects as the main model, on completed dataset from the pooled multiple imputations. The model was refitted 38 times, each time excluding all students from one country. For each fixed effect we computed the change in the estimate relative to the full-sample estimate, expressed in standard errors of the full-sample model (a DFBETAS-type measure; Belsley et al., 1980), and compared it with the size-adjusted cutoff of $2/\sqrt{G}$, where G is the number of countries ($G = 38$, cutoff = 0.32), used as a guideline and not as a formal test. We also counted the exclusions that changed the sign of a coefficient or its significance at the 5% level ($|t| > 1.96$). We paid particular attention to the country-level predictors (HDI, public expenditure on education) and to the teacher-reported variables (class size, teaching limited by student needs, algebra topics taught).

Results

Global Distribution of Mathematics Achievement

As illustrated in Fig. 2, Grade 8 mathematics achievement exhibits substantial global variation, based on replicated means across the five plausible values. African and Middle Eastern nations demonstrated the lowest achievement, falling below the 400-point threshold (e.g., Morocco: $M = 388.2$, $SE = 2.28$; South Africa: $M = 389.6$, $SE = 2.30$; Saudi Arabia: $M = 393.8$, $SE = 2.37$). Most participating European, North American, and Oceanian systems clustered near the international TIMSS benchmark of 500 points (e.g., New Zealand: $M = 481.8$, $SE = 3.39$; Italy: $M = 497.5$, $SE = 2.73$; Portugal: $M = 500.0$, $SE = 3.18$; Norway: $M = 501.5$, $SE = 2.55$; Finland: $M = 508.1$, $SE = 2.57$; United States: $M = 515.4$, $SE = 4.78$; Australia: $M = 517.6$, $SE = 3.74$). East Asian systems occupied the highest tier, averaging near 600 points (Japan: $M = 595.3$, $SE = 2.80$; Republic of Korea: $M = 607.1$, $SE = 2.88$; Taiwan: $M = 612.1$, $SE = 2.70$; Singapore: $M = 615.0$, $SE = 3.98$). Overall, a gap exceeding 200 TIMSS points separates the lowest- and highest-performing regions worldwide.

Macro-Level Correlates and System-Level Disconnects

Cross-national differences in mean mathematics achievement correlated strongly with the United Nations Development Programme (UNDP) Human Development Index ($r = .72$, $p < .001$), indicating that higher national development aligns with elevated student performance. Conversely, total educational expenditure as a percentage of GDP displayed no linear association with achievement ($r = -.04$, $p = .860$), suggesting that higher spending levels do not automatically yield superior performance (Fig. 3).

A linear model predicting achievement from HDI revealed distinct regional deviations (Fig. 3A). High-performing East Asian systems (e.g., Japan, Singapore, Republic of Korea) substantially outperformed predictions based on their HDI values. In contrast, several Middle Eastern countries (e.g., Saudi Arabia, Kuwait, Oman) achieved scores lower than expected given their development metrics. When examining expenditure (Fig. 3B), East Asian systems with below-average spending achieved top scores, whereas African nations such as Morocco and South Africa allocated approximately 5% of GDP to education yet

scored below 400 points. Nordic nations (Norway, Sweden, Finland) reported the highest spending rates but clustered around the scale mean.

Figure 3.

TIMSS 2019 mathematics mean scores at grade 8 by country as a function of UNDP's Human Development Index (A) and Total Expenditure in Education (as % of GDP) (B). R is the Pearson correlation coefficient ($R = 0.72$; $p < 0.001$ for Maths Country mean score vs. Human Development Index; and $R = -0.04$, $p = 0.86$ for Maths Country mean score vs. Expenditure). Gray bands are the 95% confidence intervals for the regression lines.

Intended curriculum coverage and accountability infrastructure showed similarly weak associations with system outcomes (Fig. 4). East Asian systems covering approximately 90% of the TIMSS framework outperformed systems reporting 100% coverage, whereas Middle Eastern systems with over 95% coverage performed below average. Neither national examinations nor standardized evaluation mechanisms systematically tracked performance differences. These descriptive patterns do not by themselves establish policy effects; in the multilevel models, the examined structural indicators did not show positive independent associations with mathematics achievement after adjustment for other contextual factors.

Figure 4.

Relationship between the National Curriculum match with the TIMSS grade 8 mathematics assessed curriculum (% of items in the TIMSS test covered by national curricula) and Maths Mean Scores (A) and the existence of national or regional exams and the countries means (B). R is the Pearson correlation coefficient (A) or the Biserial Correlation (B). The boxes in panel B are the boxplots of the data in each group.

Multilevel Variance Partitioning and Conditional Associations

The unconditional baseline model revealed that 37.7% of total variance in mathematics achievement was attributable to differences between countries ($ICC_{\text{Country}} = 0.377$), while 31.1% was attributable to differences between schools within countries ($ICC_{\text{School}} = 0.311$), leaving 31.2% attributable to differences among students within schools (Table 2).

Adding student-level predictors reduced residual variance from the null model by 32.8%, with fixed effects accounting for 32.0% of total variance ($R_m^2 = .320$) and the full model explaining 63.3% ($R_c^2 = .633$; Table 2). Student confidence in mathematics emerged as a prominent correlate ($\beta = 0.304$, $SE = 0.015$, $p < .001$), alongside home resources for learning ($\beta = 0.170$, $SE = 0.015$, $p < .001$). Extra mathematics lessons ($\beta = 0.058$, $SE = 0.028$, $p = .040$) and computer self-efficacy ($\beta = 0.029$, $SE = 0.008$, $p = .002$) were statistically significant but trivial in magnitude ($\beta < 0.10$). These results partially support Hypothesis 1.

The pronounced effect of student confidence ($\beta = 0.299$ in the final model) aligns with Social Cognitive Theory (Bandura, 1997), wherein self-efficacy operates as a proximal correlate that governs task engagement, cognitive persistence, and problem-solving success. While structural policy mechanisms act as distal background conditions, academic confidence is closely associated with learning behaviors.

Table 2

Fixed Effects Standardized regression coefficients (β) with SE between parentheses for the student level variables, school and teacher level, and country level, and Random Effects. *** - $p \leq .001$; ** - $0.001 < p \leq .01$; * - $0.01 < p \leq 0.05$, + $0.05 < p \leq 0.10$

Variables	Basal	Student Level	School Level	Country Level	Fully adjusted
Fixed Effects β (SE)					
(Intercept)	0.035 (0.085)	0.245 (0.095)*	0.096 (0.086)	0.026 (0.041)	0.174 (0.075)*
Fixed Effects					
BSBG01		-0.003 (0.008)			
BSBGHER		0.170 (0.015)***			0.155 (0.013)***
BSBM42BA		-0.009 (0.010)			
BSBM43AA		0.058 (0.028)*			
BSBGSSB		-0.017 (0.011)			
BSBGSB		-0.010 (0.008)			
BSBGSLM		0.008 (0.012)			
BSBGICM		-0.011 (0.009)			
BSBGDML		-0.001 (0.012)			
BSBGSCM		0.304 (0.015)***			0.299 (0.013)***
BSBG SVM		0.001 (0.013)			
BSBGSEC		0.029 (0.008)***			0.036 (0.007)***
BCBG03A			-0.140 (0.023)***		-0.106 (0.029)***
BCBG06A			0.005 (0.012)		

Variables	Basal	Student Level	School Level	Country Level	Fully adjusted
BCBG13BA			0.005 (0.019)		
BCBG10A			0.038 (0.013)**		
BCBGDAS			0.062 (0.013)***		0.071 (0.016)***
BCBGEAS			0.140 (0.019)***		0.088 (0.022)***
BCBGMRS			0.034 (0.022)		
BTBG01			0.019 (0.011)+		
BTBG04			0.024 (0.015)		
BTBG05F			0.013 (0.009)		
BTBG10			0.126 (0.035)***		0.114 (0.035)**
BTBGLSN			0.157 (0.025)***		0.135 (0.027)***
BTBGTJS			-0.023 (0.010)*		-0.036 (0.013)**
BTDMME			-0.021 (0.013)		
BTDMNUM			0.016 (0.007)*		0.015 (0.010)
BTDMALG			0.080 (0.036)*		0.082 (0.046)+
BTDMGEO			0.006 (0.013)		
BTDMDAT			0.019 (0.009)*		0.025 (0.008)**
HDI				0.479 (0.058)***	0.368 (0.131)**
ExpEdu				-0.172 (0.050)***	-0.136 (0.075)+

Variables	Basal	Student Level	School Level	Country Level	Fully adjusted
PercCurrCovered				-0.045 (0.045)	
NatRegExams				0.006 (0.050)	
SchSelfEvaluation				0.005 (0.029)	
Random Effects Var(SE)					
cntschid. (Intercept)	0.217 (0.021)***	0.126 (0.016)***	0.139 (0.013)***	0.217 (0.023)***	0.104 (0.012)***
IDCNTRY. (Intercept)	0.263 (0.059)***	0.131 (0.055)*	0.214 (0.043)***	0.069 (0.021)***	0.096 (0.040)*
Residual	0.448 (0.020)***	0.301 (0.025)***	0.442 (0.023)***	0.444 (0.022)***	0.310 (0.022)***
Statistics					
ICC _{School}	0.311	0.181	0.199	0.311	0.149
ICC _{Country}	0.377	0.188	0.307	0.099	0.138
R ² _m	0.000	0.320	0.226	0.206	0.549
R ² _c	0.517	0.633	0.570	0.517	0.726
Nobs	228158	76553	160160	199775	87739
N.pv	5	5	5	5	5
AIC	470850.4936	141996.3583	341525.523	408589.9883	172155.789
BIC	470891.8448	142144.2901	341745.1695	408681.8329	172324.6672

Note: Variables' selection was based on the step-up models and that the final model's Nobs (87,739, ~ 38% of the full sample) reflects joint completeness across all selected predictors

Evaluating school and teacher factors reduced between-school variance by 36% ($R_m^2 = .226$, $R_c^2 = .570$). Schools with larger proportions of economically disadvantaged students achieved lower scores ($\beta = -0.140$, $SE = 0.023$, $p < .001$), whereas those prioritizing academic success scored higher ($\beta = 0.140$, $SE = 0.019$, $p < .001$). Disciplinary climate and library availability yielded statistically significant but trivial effects. Among teacher-associated measures, reduced instructional limitations ($\beta = 0.157$, $SE = 0.025$, $p < .001$) and larger class sizes ($\beta = 0.126$, $SE = 0.035$, $p < .001$) were positively associated with higher

performance, the latter mirroring structural class size dynamics observed in East Asian contexts (Hanushek & Woessmann, 2011). These findings partially support Hypotheses 2 and 3.

At the country level, fixed effects explained 20.6% of total variance ($R_m^2 = .206$, $R_c^2 = .517$), accounting for 74% of the between-country variance. HDI showed a strong positive association ($\beta = 0.479$, $SE = 0.058$, $p < .001$), while public educational expenditure maintained a negative conditional association ($\beta = -0.172$, $SE = 0.050$, $p < .001$). Curriculum coverage and assessment infrastructure remained non-significant. These results partially support H4.

In the final fully specified model, HDI showed the largest conditional country-level association among the country-level indicators included in the model. ($\beta = 0.368$, $SE = 0.131$, $p < .01$), while student confidence was among the strongest student-level correlates ($\beta = 0.299$, $SE = 0.013$, $p < .001$). Home resources ($\beta = 0.155$, $SE = 0.013$, $p < .001$), instructional limitations ($\beta = 0.135$, $SE = 0.027$, $p < .001$), and class size ($\beta = 0.114$, $SE = 0.035$, $p < .01$) remained associated with achievement, while educational expenditure retained its slight negative conditional association ($\beta = -0.136$, $SE = 0.075$, $p = .075$). The combined fixed effects explained 54.9% of total variance ($R_m^2 = .549$), with the full conditional model accounting for 72.6% ($R_c^2 = .726$). Overall, these findings reveal that the examined policy indicators were not independently associated with achievement in the fully adjusted model, regardless of societal development and student self-efficacy.

Sensitivity Analysis to Missing Data in TIMSS and Country outliers

A sensitivity analysis was conducted to examine the robustness of the multilevel model of TIMSS 2019 mathematics to missing data. Complete-case estimates ($n = 87,739$) were compared with estimates obtained from multiple imputations of non-structural missing data ($n = 228,158$). The overall pattern of results was largely consistent across the two approaches, although the magnitude of several associations changed. Results are summarized in Fig. 5.

The strongest changes were observed for the teacher-level predictors. For example, the association between BTBG10 and mathematics achievement decreased from $\beta = 0.114$ to $\beta = 0.023$ (– 79.8%), whereas the association for BTDMALG decreased from $\beta = 0.082$ to $\beta = 0.013$ (– 84.1%). The coefficients for BSBGHER, BSBGSCM, BSBGSEC, BCBG03A, BCBGDAS, and HDI remained in the same direction but generally decreased in magnitude. In contrast, BCBGEAS increased from $\beta = 0.088$ to $\beta = 0.114$ (+ 29.5%). Several predictors that were statistically significant in the complete-case analysis were no longer significant after imputation, including BTBG10, BTBGTJS, BTDMALG, and BTDMDAT. Conversely, BTDMNUM was significant in the imputed analysis but not in the complete-case analysis.

The contextual effects were comparatively stable. The HDI coefficient changed from $\beta = 0.368$ to $\beta = 0.347$ (– 5.8%), while the coefficient for ExpEdu changed from $\beta = -0.136$ to $\beta = -0.085$ (+ 37.5%), although neither estimate for ExpEdu was statistically significant. The estimated school- and country-

level variance components increased under multiple imputation by 14.4% and 44.8%, respectively, and the residual variance increased by 12.3%

The leave-one-country-out (LOCO) analysis indicates that the estimated coefficients are generally stable across country exclusions, although a few countries exert appreciable influence on specific estimates. Using the conventional DFBETAS threshold of $2/\sqrt{38} = 0.32$, South Africa (ZAF) shows substantial influence on the Class size coefficient (DFBETAS = - 3.87), while the United Arab Emirates (ARE) has a marked influence on Teaching limited by student needs (2.18) and Algebra coverage (- 3.40). Smaller but above-threshold influences are observed for Singapore on Education Expenditure (- 0.56) and HDI (0.51), and for Hong Kong on HDI (0.35). Importantly, the LOCO coefficient plots show that, despite these influential observations, the overall direction and magnitude of most associations remain broadly consistent across country exclusions.

Discussion

The present study examined student-, school/teacher-, and country-level correlates of Grade 8 mathematics achievement in TIMSS 2019 using a three-level hierarchical model. The central aim was to determine whether structural educational indicators, including curriculum coverage, national or regional examinations, and public educational expenditure, retained independent associations with mathematics achievement after accounting for student characteristics, school context, and broader country-level development. Overall, the results indicate that these associations are highly conditional on the broader context in which educational systems operate. Student confidence in mathematics and home educational resources were among the strongest student-level correlates, whereas the Human Development Index (HDI) was the strongest observed country-level correlate in the fully adjusted model. The results therefore support a multilevel interpretation of cross-national achievement differences rather than a simple correspondence between individual structural indicators and national performance.

Student-Level Correlates: Hypothesis 1

Hypothesis 1

proposed that socioeconomic background, home educational resources, and confidence in mathematics would be positively associated with achievement. The findings provide partial support for this hypothesis. Student confidence in mathematics was one of the strongest student-level correlates in the final model ($\beta = 0.299$, $SE = 0.013$, $p < .001$), and home educational resources were also positively associated with achievement ($\beta = 0.155$, $SE = 0.013$, $p < .001$). These findings are consistent with previous research showing that both students' academic beliefs and the resources available in the home are systematically related to educational achievement (Eklöf, 2022; Lee & Stankov, 2018; Lee & Shute, 2010; Marôco, 2021).

The association between mathematics confidence and achievement is also consistent with Social Cognitive Theory (Bandura, 1997), which proposes that self-efficacy beliefs are related to students'

engagement, persistence, and regulation of learning. However, the cross-sectional design does not permit determination of the direction of this relationship. Higher confidence may be associated with greater engagement and achievement, but previous achievement may also contribute to students' confidence. The present findings therefore identify confidence as an important correlate of mathematics achievement rather than demonstrating a causal effect of self-efficacy.

Home educational resources showed a similarly consistent positive association. This result is compatible with evidence linking socioeconomic and educational resources in the home to differences in learning opportunities and achievement (Broer et al., 2019; Perry & McConney, 2010; Wang et al., 2023). Importantly, these associations remained evident after accounting for school- and country-level characteristics, indicating that student and family context remains relevant even within a multilevel model that incorporates broader educational-system conditions.

School and Teacher Factors: Hypotheses 2 and 3

Hypothesis 2

proposed that school socioeconomic composition and academic emphasis would be associated with mathematics achievement. The results provide support for this hypothesis. Schools with a larger proportion of economically disadvantaged students had lower mathematics achievement ($\beta = -0.140$, $SE = 0.023$, $p < .001$), whereas school academic emphasis was positively associated with achievement ($\beta = 0.140$, $SE = 0.019$, $p < .001$). These findings are consistent with previous research showing that school composition and the academic environment are associated with student achievement beyond individual student characteristics (Broer et al., 2019; Perry & McConney, 2010).

Hypothesis 3

proposed positive associations between teacher characteristics, formal qualifications, and instructional practices and achievement. The evidence provides only partial support for this hypothesis. Reduced instructional limitations was positively associated with achievement in the complete-case model, but several teacher-associated coefficients changed substantially after multiple imputation. In particular, the association for class size was reduced from $\beta = 0.114$ to $\beta = 0.023$, and the association for algebra coverage decreased from $\beta = 0.082$ to $\beta = 0.013$. Several teacher-associated variables that were statistically significant in the complete-case analysis were no longer significant after imputation. Thus, the evidence for teacher-associated factors is considerably less stable than that for the main student- and country-level correlates.

This sensitivity is important because teacher characteristics were aggregated at the school level. The resulting measures represent school-level profiles of the teachers included in the data rather than individual teacher effects. Missing teacher information can therefore alter both the aggregated measure and its between-school variability. The results should consequently not be interpreted as demonstrating that particular teacher characteristics or instructional practices have no relationship with achievement.

Rather, they indicate that their estimated associations are sensitive to the treatment of missing data and to the way teacher information is aggregated.

Country-Level Correlates: Hypothesis 4

Hypothesis 4

proposed that human-development and education-system characteristics would be associated with mathematics achievement after adjustment for student- and school-level characteristics. The evidence provides partial support for this hypothesis. HDI was the strongest observed country-level correlate in the fully specified model ($\beta = 0.368$, $SE = 0.131$, $p < .01$), whereas curriculum coverage and the presence of national or regional examinations did not show statistically significant independent associations with achievement. Public educational expenditure also did not show a robust association: its coefficient was negative in the complete-case model ($\beta = -0.136$, $p = .075$) and became smaller and remained non-significant after multiple imputation ($\beta = -0.085$). The findings therefore do not provide evidence of a reliable independent association between expenditure as a percentage of GDP and mathematics achievement in these data.

The HDI finding should nevertheless be interpreted cautiously. HDI is a composite measure incorporating dimensions of health, education, and income, and therefore represents a broad indicator of national development rather than a specific educational policy or mechanism (Klugman et al., 2011). Its relatively large coefficient indicates that broader country-level development was strongly associated with mathematics achievement after adjustment for the variables included in the model. It should not be interpreted as evidence that increasing HDI itself would cause higher achievement. Moreover, the analysis includes only 38 country-level units, which limits the precision with which country-level relationships can be characterized and increases the potential influence of individual education systems.

The absence of statistically significant independent associations for curriculum coverage and national or regional examinations should similarly not be interpreted as evidence that these institutional arrangements are ineffective. Their aggregate presence or coverage may not adequately capture implementation quality, instructional practice, enforcement, selectivity, or the conditions under which policies operate. More generally, the results indicate that simple cross-national associations between structural indicators and achievement can change substantially once broader national development and student and school composition are considered simultaneously. This is consistent with the broader literature cautioning against interpreting international large-scale assessment data as direct tests of policy effectiveness (Feniger, 2020; Keitel & Kilpatrick, 1999). Taken together, these findings suggest that the apparent strength of structural indicators depends strongly on conditioning on broader human-development and compositional factors.

Educational Expenditure and Class Size

The expenditure result illustrates the importance of distinguishing conditional statistical associations from policy effects. Public educational expenditure expressed as a percentage of GDP is a relative

indicator and may capture different institutional circumstances across countries. Once HDI and other country-level characteristics were included, the expenditure coefficient was neither statistically significant nor stable across missing-data specifications. The result therefore does not support an interpretation in terms of a negative effect of educational expenditure. Rather, it suggests that expenditure as a percentage of GDP alone is insufficient to characterize the resources available to students or how effectively those resources are translated into learning opportunities.

Class size requires particular caution. In the complete-case model, larger class size was positively associated with mathematics achievement ($\beta = 0.114$, $SE = 0.035$, $p < .01$). However, this association was substantially attenuated after multiple imputation ($\beta = 0.023$), and the LOCO analysis identified South Africa as highly influential for the class-size coefficient (DFBETAS = -3.87). The positive class-size association should therefore be interpreted as sensitive to influential country-level observations and not as evidence that larger classes improve student achievement. The association may instead reflect differences between education systems in which class size co-occurs with other institutional, socioeconomic, or instructional characteristics that are not fully captured by the model.

Robustness to Missing Data and Influential Countries

The sensitivity analyses provide an important qualification to the interpretation of the findings. The country-level HDI association was relatively stable between the complete-case and multiply imputed analyses, changing by approximately -5.8% , whereas several teacher-associated estimates changed substantially. This pattern suggests that the principal country-level finding and the strongest student-level associations are comparatively robust, while conclusions concerning several teacher-associated variables should be treated more cautiously.

The LOCO analysis provides a complementary assessment of cross-national influence. Most coefficients retained broadly similar directions across country exclusions, but specific estimates were sensitive to individual education systems. South Africa exerted substantial influence on the class-size coefficient, while the United Arab Emirates influenced estimates for teaching limitations and algebra coverage. These findings reinforce the importance of treating country-level coefficients in international large-scale assessments as associations across a relatively small number of heterogeneous education systems rather than as universally generalizable effects.

Implications and Limitations

Taken together, the findings suggest that cross-national differences in mathematics achievement cannot be adequately characterized by examining structural educational indicators in isolation. Student confidence, home educational resources, school socioeconomic composition, and academic emphasis were associated with achievement within the multilevel framework, while the strongest country-level association was observed for the broader HDI measure. At the same time, several structural indicators showed limited or non-robust independent associations after adjustment.

These findings should not be interpreted as implying that structural policies, educational expenditure, curriculum design, or assessment systems are unimportant. Rather, the results indicate that their observed associations with achievement depend on the broader socioeconomic and educational contexts in which they are implemented. Longitudinal, quasi-experimental, and policy-evaluation studies are needed to determine whether particular reforms produce changes in achievement and under which contextual conditions such effects emerge.

Several limitations constrain interpretation. First, the cross-sectional design of a single TIMSS cycle precludes causal inference and does not permit evaluation of the long-term effects of specific educational reforms. Second, the analysis includes approximately 38 country-level units, limiting statistical power for country-level inference and making some coefficients potentially sensitive to influential education systems. Third, substantial missingness was present in some background variables, and the sensitivity analysis showed that several teacher-associated estimates were affected materially by the treatment of missing data. Fourth, TIMSS is a low-stakes assessment for students, and test-taking effort may differ from that observed in high-stakes examinations (OECD, 2019). Finally, the country-level indicators are necessarily broad measures and cannot capture the full complexity of policy implementation, instructional practice, institutional arrangements, or cultural context.

Concluding remarks

Using TIMSS 2019 Grade 8 mathematics data from 38 participating countries and education systems, this study examined how student-, school-, teacher-, and country-level characteristics were associated with mathematics achievement within a multilevel framework. Student confidence in mathematics and home educational resources showed the strongest and most consistent associations at the student level. At the country level, HDI remained positively associated with achievement in the fully adjusted model, whereas associations involving educational expenditure and other structural indicators were substantially attenuated after accounting for student, school, teacher, and broader country-level characteristics.

These findings suggest that the apparent strength of structural indicators depends strongly on the level of analysis and on conditioning on broader human-development and compositional factors. They therefore caution against interpreting bivariate or partially adjusted associations between national educational characteristics and achievement as evidence that system-level characteristics are independently associated with higher performance. At the same time, the results do not establish causal relationships and should not be interpreted as demonstrating that structural indicators are unimportant for educational outcomes.

The study is also subject to the limitations of cross-sectional large-scale assessment data, including the observational nature of the analyses, the measurement of contextual constructs, and sensitivity of some estimates to missing-data treatment and influential countries or education systems. Future research using longitudinal or repeated large-scale assessment data, together with more detailed measures of

educational policy and implementation, could help determine whether the associations identified here are replicated across assessment cycles and how they vary across national contexts.

Declarations

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

This study involved secondary analysis of publicly available, de-identified TIMSS 2019 data and involved no direct interaction with human participants. Consequently, no additional institutional ethics approval or participant consent was required for this secondary analysis.

Declarations of Generative AI and AI-assisted technologies in the writing process

During the preparation of this manuscript, the authors used ChatGPT (OpenAI) to assist with language editing, improving clarity and readability. All analyses, methodological decisions, interpretations, and conclusions were made and verified by the authors. The authors reviewed and edited all AI-assisted output and take full responsibility for the accuracy, integrity, and originality of the final manuscript.

Funding

The first author was the recipient of the XXX Visiting Fellow Grant 2024. High-volume data analysis was performed with the support of INCD funded by FCT and FEDER under the project 01/SAICT/2016 n° 022153 and UID/04011/2025 (<https://doi.org/10.54499/UID/04011/2025>)

Author Contribution

JM conceived the study, conducted the analyses, and drafted the manuscript. HHL contributed to the study design and interpretation of the results. JR contributed to the interpretation of the results and manuscript revision. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Data availability

and code availability

The data used in this study are publicly available from the International Association for the Evaluation of Educational Achievement (IEA) TIMSS 2019 Grade 8 International Database. The data and accompanying documentation can be accessed through the IEA Data Repository at https://doi.org/10.58150/IEA_TIMSS_2019_G8. The TIMSS 2019 International Database and User Guide are also available from the TIMSS & PIRLS International Study Center at Boston College. The R code used for the analyses is available from the corresponding author upon reasonable request.

References

1. Bailey, P., Webb, B., Kelley, C., Nguyen, T., & Huo, H. (2023). WeMix: Weighted mixed-effects models using multilevel pseudo maximum likelihood estimation (Version 4.0.0) [R package]. CRAN. <https://CRAN.R-project.org/package=WeMix>
2. Bakan Kalaycioglu, D., Reynolds, K. A., Fishbein, B., Gonzalez, E., & von Davier, M. (2026). *Intended, implemented, and attained mathematics and science curricula at the eighth grade: Findings from TIMSS 2023*. TIMSS & PIRLS International Study Center. <https://doi.org/10.6017/lse.tpisc.timss.wf9486>. Boston College.
3. Banerjee, P. A., & Lamb, S. (2016). A systematic review of factors linked to poor academic performance of disadvantaged students in science and maths in schools. *Cogent Education*, 3(1). <https://doi.org/10.1080/2331186X.2016.1178401>. Article 1178401.
4. Bandura, A. (1997). *Self-efficacy: The exercise of control*. W H Freeman/Times Books/ Henry Holt & Co.
5. Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
6. Van der Meer, T., Grotenhuis, T., M., & Pelzer, B. (2010). Influential cases in multilevel modeling: A methodological comment. *American Sociological Review*, 75(1), 173–178. <https://doi.org/10.1177/0003122409359166>
7. Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1996). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study*. Center for the Study of Testing, Evaluation, and Educational Policy.
8. Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. Wiley.

9. Blömeke, S., Nilsen, T., Olsen, R. V., & Gustafsson, J. E. (2022). Conceptual and methodological accomplishments of ILSAs, remaining criticism and limitations. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International handbook of comparative large-scale studies in education* (pp. 415–439). Springer. https://doi.org/10.1007/978-3-030-38298-8_22-1
10. Broer, M., Bai, Y., Fonseca, F. Socioeconomic Inequality and Educational Outcomes: Evidence from Twenty Years of TIMSS., & Springer (2019). <https://doi.org/10.1007/978-3-030-11991-1>
11. Caro, D. H., & Biecek, P. (2017). intsvy: An R package for analysing international large-scale assessment data. *Journal of Statistical Software*, 81(7), 1–44. <https://doi.org/10.18637/jss.v081.i07>
12. Chang, I. (2023). Early numeracy and literacy skills and their influences on fourth-grade mathematics achievement: A moderated mediation model. *Large-Scale Assessments in Education*, 11(1). <https://doi.org/10.1186/s40536-023-00168-6>. Article 15.
13. Chiasson, M., & Freiman, V. (2022). Rethinking the 21st-century school: New citizens' skills for the digital era and their interaction with mathematics teaching and learning. In C. Michelsen, A. Beckmann, V. Freiman, U. T. Jankvist, & A. Savard (Eds.), *Mathematics and its connections to the arts and sciences (MACAS)* (Vol. 19, pp. 155–174). Springer. https://doi.org/10.1007/978-3-031-10518-0_9
14. Chinn, S. (2009). Mathematics anxiety in secondary students in England. *Dyslexia*, 15(1), 61–68. <https://doi.org/10.1002/dys.381>
15. Clarke, D. (2003). International comparative research in mathematics education. In A. J. Bishop, M. A. Clements, C. Keitel, J. Kilpatrick, & F. K. S. Leung (Eds.), *Second international handbook of mathematics education* (pp. 143–184). Kluwer Academic. https://doi.org/10.1007/978-94-010-0273-8_6
16. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
17. Conger, A. J. (1974). A revised definition for suppressor variables: A guide to their identification and interpretation. *Educational and Psychological Measurement*, 34(1), 35–46. <https://doi.org/10.1177/001316447403400105>
18. Duchhardt, C., Jordan, A. K., & Ehmke, T. (2017). Adults' use of mathematics and its influence on mathematical competence. *International Journal of Science and Mathematics Education*, 15(1), 155–174. <https://doi.org/10.1007/s10763-015-9678-y>
19. Eklöf, H. (2022). Student motivation and self-beliefs. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International handbook of comparative large-scale studies in education* (pp. 825–845). Springer. https://doi.org/10.1007/978-3-030-88178-8_44
20. Feniger, Y. (2020). Evidence-based decision making or a tunnel vision effect? TIMSS, problem definition and policy change in Israeli mathematics education. *Critical Studies in Education*, 61(3), 363–379. <https://doi.org/10.1080/17508487.2018.1448877>
21. Fishbein, B., Foy, P., & Yin, L. (2021). TIMSS 2019 international database. *TIMSS 2019 user guide for the international database* (2nd ed.). TIMSS & PIRLS International Study Center.

- <https://timssandpirls.bc.edu/timss2019/international-database/>
22. Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage.
 23. Foy, P., Fishbein, B., von Davier, M., & Yin, L. (2020). Implementing the TIMSS 2019 scaling methodology. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report* (p. 12.1–12.146). TIMSS & PIRLS International Study Center, Boston College. <https://timssandpirls.bc.edu/timss2019/methods/chapter-12.html>
 24. Freudenthal, H. (1991). *Revisiting mathematics education: China lectures*. Kluwer Academic.
 25. Golding, J. (2018). Mathematics education in the spotlight: Its purpose and some implications. *London Review of Education*, 16(3), 460–473. <https://doi.org/10.18546/LRE.16.3.08>
 26. Hanushek, E. A., & Woessmann, L. (2011). The economics of international differences in educational achievement. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (Vol. 3, pp. 89–200). Elsevier. <https://doi.org/10.1016/B978-0-444-53429-3.00002-8>
 27. Harju-Luukkainen, H., Nissinen, K., & Vettenranta, J. (2022). Early skills as predictors for later educational outcome in mathematics and science in Finland and Sweden: A further analysis on TIMSS 2015 data. In H. Harju-Luukkainen, J. Kangas, & S. Garvis (Eds.), *Finnish early childhood education and care: A multi-theoretical perspective on research and practice* (pp. 181–192). Springer. https://doi.org/10.1007/978-3-030-95512-0_12
 28. Hembree, R. (1990). The nature, effects, and relief of mathematics anxiety. *Journal for Research in Mathematics Education*, 21(1), 33–46. <https://doi.org/10.2307/749455>
 29. Hernández-Torrano, D., & Courtney, M. (2021). Modern international large-scale assessment in education: An integrative review and mapping of the literature. *Large-Scale Assessments in Education*, 9(1). <https://doi.org/10.1186/s40536-021-00109-1>. Article 17.
 30. Huang, F. L. (2024). Using plausible values when fitting multilevel models with large-scale assessment data using R. *Large-Scale Assessments in Education*, 12(1). <https://doi.org/10.1186/s40536-024-00192-0>. Article 7.
 31. International Association for the Evaluation of Educational Achievement (2024). Trends in International Mathematics and Science Study (TIMSS). <https://www.iea.nl/studies/iea/timss>
 32. Jablonka, E. (2003). Mathematical literacy. In A. J. Bishop, M. A. Clements, C. Keitel, J. Kilpatrick, & F. K. S. Leung (Eds.), *Second international handbook of mathematics education* (pp. 75–102). Kluwer Academic. https://doi.org/10.1007/978-94-010-0273-8_4
 33. Kadijevich, D. M., Stephens, M., Solares-Rojas, A., & Guberman, R. (2023). Impacts of TIMSS and PISA on mathematics curriculum reforms. In Y. Shimizu & R. Vithal (Eds.), *Mathematics curriculum reforms around the world* (pp. 359–374). Springer. https://doi.org/10.1007/978-3-031-33880-9_20
 34. Keitel, C., & Kilpatrick, J. (1999). The rationality and irrationality of international comparative studies. In G. Kaiser, E. Luna, & I. Huntley (Eds.), *International comparisons in mathematics education* (pp. 241–256). Falmer.
 35. Klugman, J., Rodríguez, F., & Choi, H. J. (2011). The HDI 2010: New controversies, old critiques. *Journal of Economic Inequality*, 9(2), 249–288. <https://doi.org/10.1007/s10888-011-9178-z>

36. Lee, C., & Johnston-Wilder, S. (2016). Learning mathematics: An affective focus. In S. Johnston-Wilder, C. Lee, & D. Pimm (Eds.), *Learning to teach mathematics in secondary school: A companion to school experience* (4th ed., pp. 52–69). Routledge.
37. Lee, J., & Shute, V. J. (2010). Personal and social-contextual factors in K–12 academic performance: An integrative perspective on student learning. *Educational Psychologist*, 45(3), 185–202. <https://doi.org/10.1080/00461520.2010.493471>
38. Lee, J., & Stankov, L. (2018). Non-cognitive predictors of academic achievement: Evidence from TIMSS and PISA. *Learning and Individual Differences*, 65, 50–64. <https://doi.org/10.1016/j.lindif.2018.05.009>
39. Leung, F. K. S., Graf, K. D., & Lopez-Real, F. J. (Eds.). (2006). Mathematics education in different cultural traditions: A comparative study of East Asia and the West. Springer. <https://doi.org/10.1007/0-387-29822-6>
40. Marôco, J. (2020). International large-scale assessments: Trends and effects on the Portuguese public education system. In H. Harju-Luukkainen, N. McElvany, & J. Stang (Eds.), *Monitoring student achievement in the 21st century: European policy perspectives* (pp. 249–265). Springer. https://doi.org/10.1007/978-3-030-38969-7_17
41. Marôco, J. (2021). What makes a good reader? Worldwide insights from PIRLS 2016. *Reading and Writing*, 34(1), 231–272. <https://doi.org/10.1007/s11145-020-10068-8>.
42. Marôco, J. (2025). Portugal: COVID-19 learning losses through the lenses of ILSA and local low stake assessments. In N. Crato & H. A. Patrinos (Eds.), *Improving national education systems after COVID-19* (pp. 175–196). Springer. https://doi.org/10.1007/978-3-031-69284-0_10
43. Martin, M. O., von Davier, M., & Mullis, I. V. S. (Eds.). (2020). Methods and procedures: TIMSS 2019 technical report. TIMSS & PIRLS International Study Center, Boston College. <https://timssandpirls.bc.edu/timss2019/methods>
44. Mierluș-Mazilu, I., & Yilmaz, F. (2024). Teaching mathematics in STEM education. In V. Gayoso Martínez, F. Yilmaz, A. Queiruga-Dios, D. M. Rasteiro, J. Martín-Vaquero, & I. Mierluș-Mazilu (Eds.), *Mathematical methods for engineering applications: ICMASE 2023* (pp. 143–161). Springer. https://doi.org/10.1007/978-3-031-52565-0_11
45. Muhaimin, L. H., Sholikhakh, R. A., Yulianti, S., Ardani, A., Hendriyanto, A., & Sahara, S. (2024). Unlocking the secrets of students' mathematical literacy to solve mathematical problems: A systematic literature review. *Eurasia Journal of Mathematics Science and Technology Education*, 20(4). <https://doi.org/10.29333/ejmste/14404>. Article em2424.
46. Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). *TIMSS 2019 assessment frameworks*. TIMSS & PIRLS International Study Center. <http://timssandpirls.bc.edu/timss2019/frameworks/>
47. Namkung, J. M., Peng, P., & Lin, X. (2019). The relation between mathematics anxiety and mathematics performance among school-aged students: A meta-analysis. *Review of Educational Research*, 89(3), 459–496. <https://doi.org/10.3102/0034654319843494>

48. National Council of Teachers of Mathematics (2014). Principles to actions: Ensuring mathematical success for all.
49. National Numeracy (2019). Building a numerate nation: Confidence, belief, and skills. <https://www.nationalnumeracy.org.uk/research-and-resources/building-numerate-nation-confidence-belief-and-skills>
50. National Numeracy (2020). Attitudes towards maths. <https://www.nationalnumeracy.org.uk/research-and-resources/attitudes-towards-maths>
51. Nieminen, P. (2022). Application of standardised regression coefficient in meta-analysis. *BioMedInformatics*, 2(3), 434–458. <https://doi.org/10.3390/biomedinformatics2030028>
52. Nieuwenhuis, R., te Grotenhuis, M., & Pelzer, B. (2012). influence.ME: Tools for detecting influential data in mixed effects models. *The R Journal*, 4(2), 38–47. <https://doi.org/10.32614/RJ-2012-011>
53. OECD (2019). PISA 2018 results (I): What students know and can do. OECD Publishing. <https://doi.org/10.1787/5f07c754-en>
54. OECD (2024). Do adults have the skills they need to thrive in a changing world? Survey of Adult Skills 2023. OECD Publishing. <https://doi.org/10.1787/b263dc5d-en>
55. Perry, L. B., & McConney, A. (2010). Does the SES of the school matter? An examination of socioeconomic status and student achievement using PISA 2003. *Teachers College Record*, 112(4), 1137–1162. <https://doi.org/10.1177/016146811011200401>
56. Peters, A. (2024). Using the TIMSS curriculum model to develop a framework for coherence and its role in developing mathematical connections. *Research in Mathematics Education*, 26(2), 300–324. <https://doi.org/10.1080/14794802.2023.2252562>
57. R Core Team (2024). R: A language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
58. Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage.
59. Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, 39(2), 142–151. <https://doi.org/10.3102/0013189X10363170>
60. Sellars, M. (2018). Mathematics and numeracy in a global society. In M. Sellars (Ed.), *Numeracy in authentic contexts: Making meaning in the multidisciplinary classroom* (pp. 1–18). Springer. https://doi.org/10.1007/978-981-10-5736-6_1
61. Sriraman, B. (2021). Mathematics in society: An introduction. In B. Sriraman (Ed.), *Handbook of the mathematics of the arts and sciences* (pp. 141–157). Springer. https://doi.org/10.1007/978-3-319-57072-3_80
62. Sullivan, P. (2018). The aspirations of the Australian curriculum prompting reform of school mathematics. In Y. Shimizu & R. Vithal (Eds.), *School mathematics curriculum reforms: Challenges, changes and opportunities. Proceedings of the twenty-fourth ICMI Study conference* (pp. 85–92). International Commission on Mathematical Instruction.

63. Teig, N., & Steinmann, I. (2023). Leveraging large-scale assessments for effective and equitable school practices: The case of the Nordic countries. *Large-Scale Assessments in Education*, 11(1). <https://doi.org/10.1186/s40536-023-00171-1>. Article 21.
64. Van den Heuvel-Panhuizen, M., & Drijvers, P. (2014). Realistic mathematics education. In S. Lerman (Ed.), *Encyclopedia of mathematics education* (pp. 521–525). Springer. https://doi.org/10.1007/978-94-007-4978-8_170
65. Ventistas, G., Ventista, O. M., & Tsani, P. (2024). The impact of realistic mathematics education on secondary school students' problem-solving skills: A comparative evaluation study. *Research in Mathematics Education*, 1–25. <https://doi.org/10.1080/14794802.2024.2335122>
66. von Davier, M. (2020). TIMSS 2019 scaling methodology: Item response theory, population models, and linking across modes. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report* (p. 11.1–11.25). TIMSS & PIRLS International Study Center, Boston College. <https://timssandpirls.bc.edu/timss2019/methods/chapter-11.html>
67. 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*. TIMSS & PIRLS International Study Center. <https://doi.org/10.6017/lse.tpisc.timss.rs6460>
68. Wang, X. S., Perry, L. B., Malpique, A., & Ide, T. (2023). Factors predicting mathematics achievement in PISA: A systematic review. *Large-Scale Assessments in Education*, 11(1). <https://doi.org/10.1186/s40536-023-00174-8>. Article 35.
69. Wößmann, L., & West, M. R. (2006). Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS. *European Economic Review*, 50(3), 695–736. <https://doi.org/10.1016/j.euroecorev.2004.11.005>
70. Yin, L., & Fishbein, B. (2020). Creating and interpreting the TIMSS 2019 context questionnaire scales. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report* (p. 11.1–11.25). TIMSS & PIRLS International Study Center, Boston College. <https://timssandpirls.bc.edu/timss2019/methods>
71. Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley. <https://doi.org/10.1002/9780470316696>
72. van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67. <https://doi.org/10.18637/jss.v045.i03>

Figures

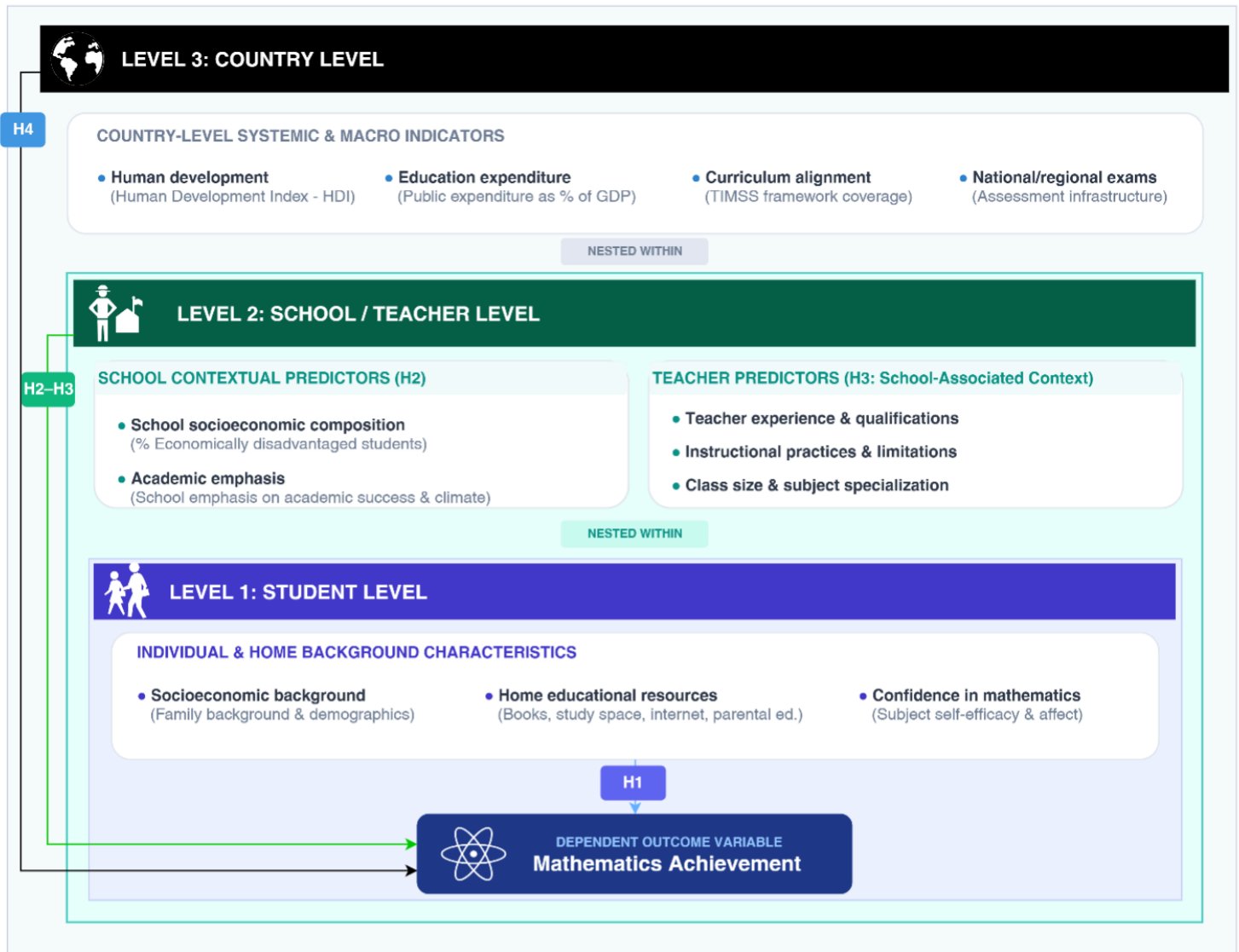


Figure 1

Conceptual organization of the three-level Hierarchical Linear Model (HLM) for Grade 8 TIMSS mathematics achievement

Note: In this three-level Hierarchical Linear Model (Students → Schools → Countries), teacher characteristics are modeled as school-associated contextual predictors at Level 2. Labeling Level 2 as "School / Teacher Level" communicates their role as contextual predictors while accurately reflecting the statistical architecture, avoiding the implication of an unestimated fourth teacher-associated random effect.

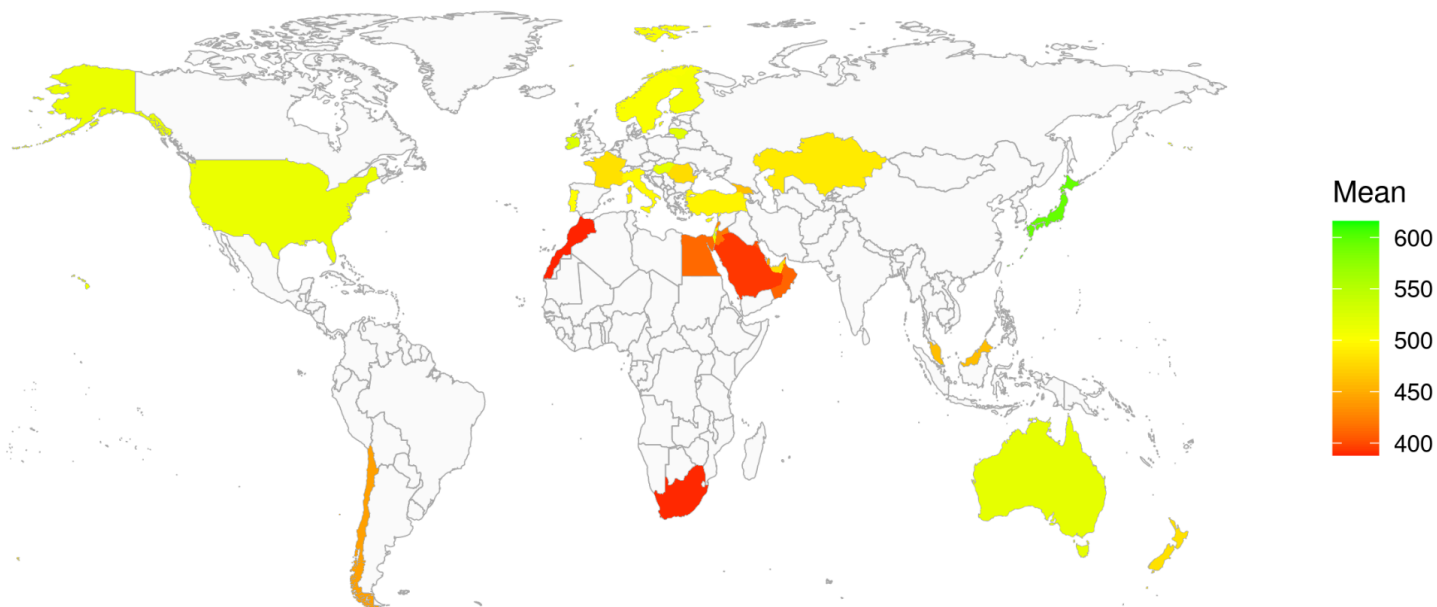


Figure 2

Mean mathematics achievement from countries who participated in the TIMSS 2019 grade 8 edition. The TIMSS scale has a mean of 500 points with a standard deviation of 100 points.

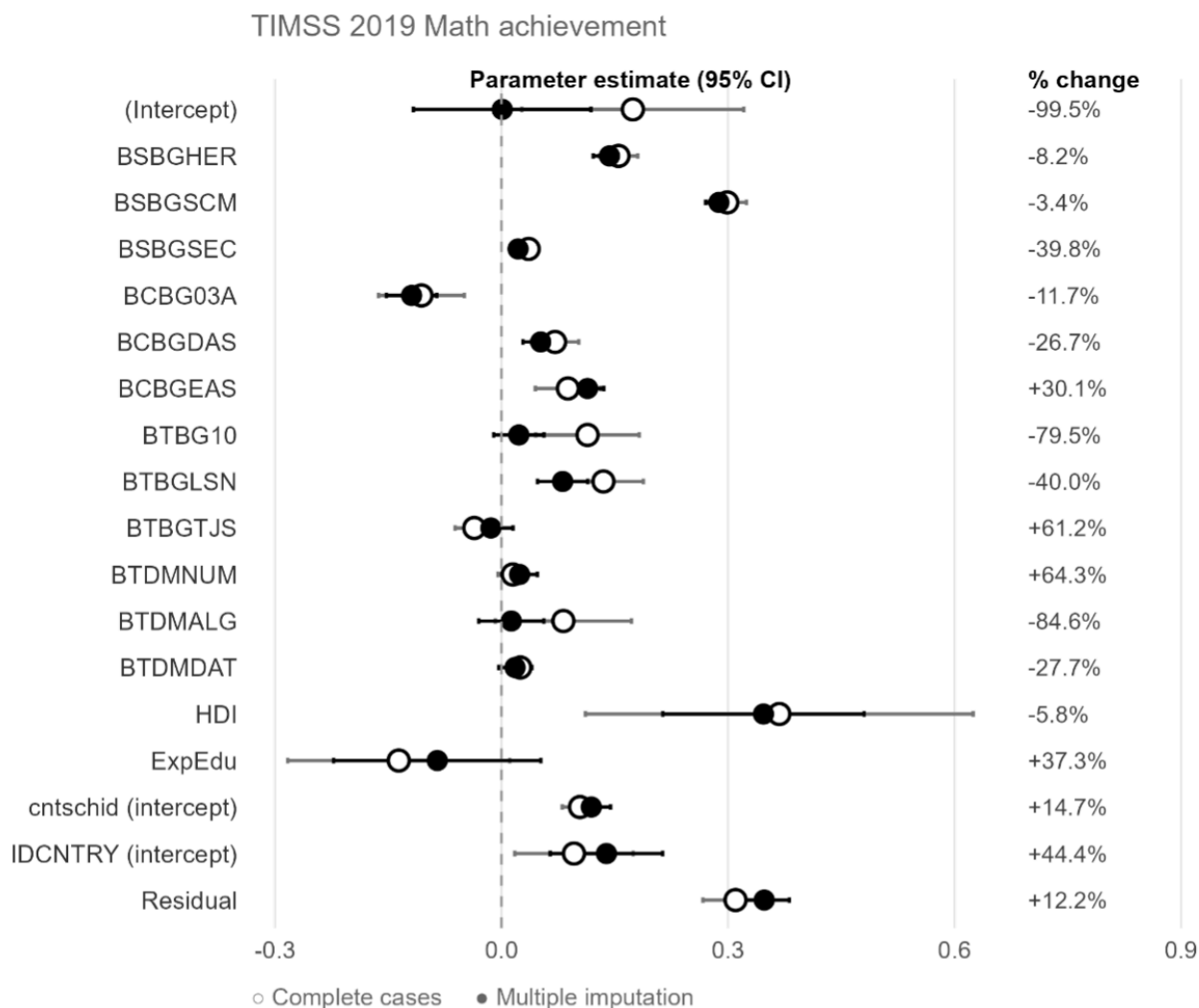


Figure 3

TIMSS 2019 mathematics mean scores at grade 8 by country as a function of UNDP's Human Development Index (A) and Total Expenditure in Education (as % of GDP) (B). R is the Pearson correlation coefficient ($R = 0.72$; $p < 0.001$ for Maths Country mean score vs. Human Development Index; and $R = -0.04$, $p = 0.86$ for Maths Country mean score vs. Expenditure). Gray bands are the 95% confidence intervals for the regression lines.

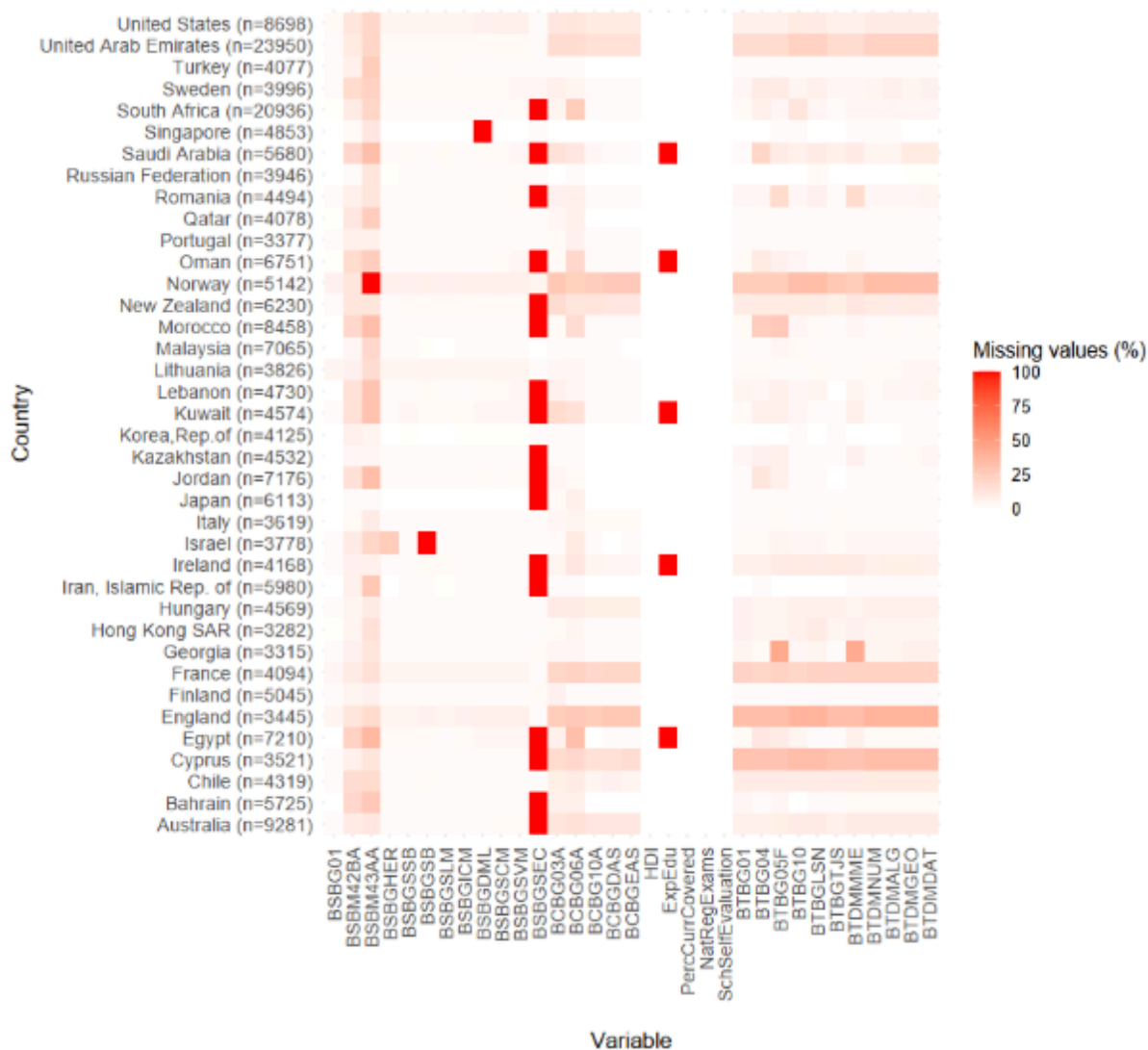


Figure 4

Relationship between the National Curriculum match with the TIMSS grade 8 mathematics assessed curriculum (% of items in the TIMSS test covered by national curricula) and Maths Mean Scores (A) and the existence of national or regional exams and the countries means (B). R is the Pearson correlation coefficient (A) or the Biserial Correlation (B). The boxes in panel B are the boxplots of the data in each group.

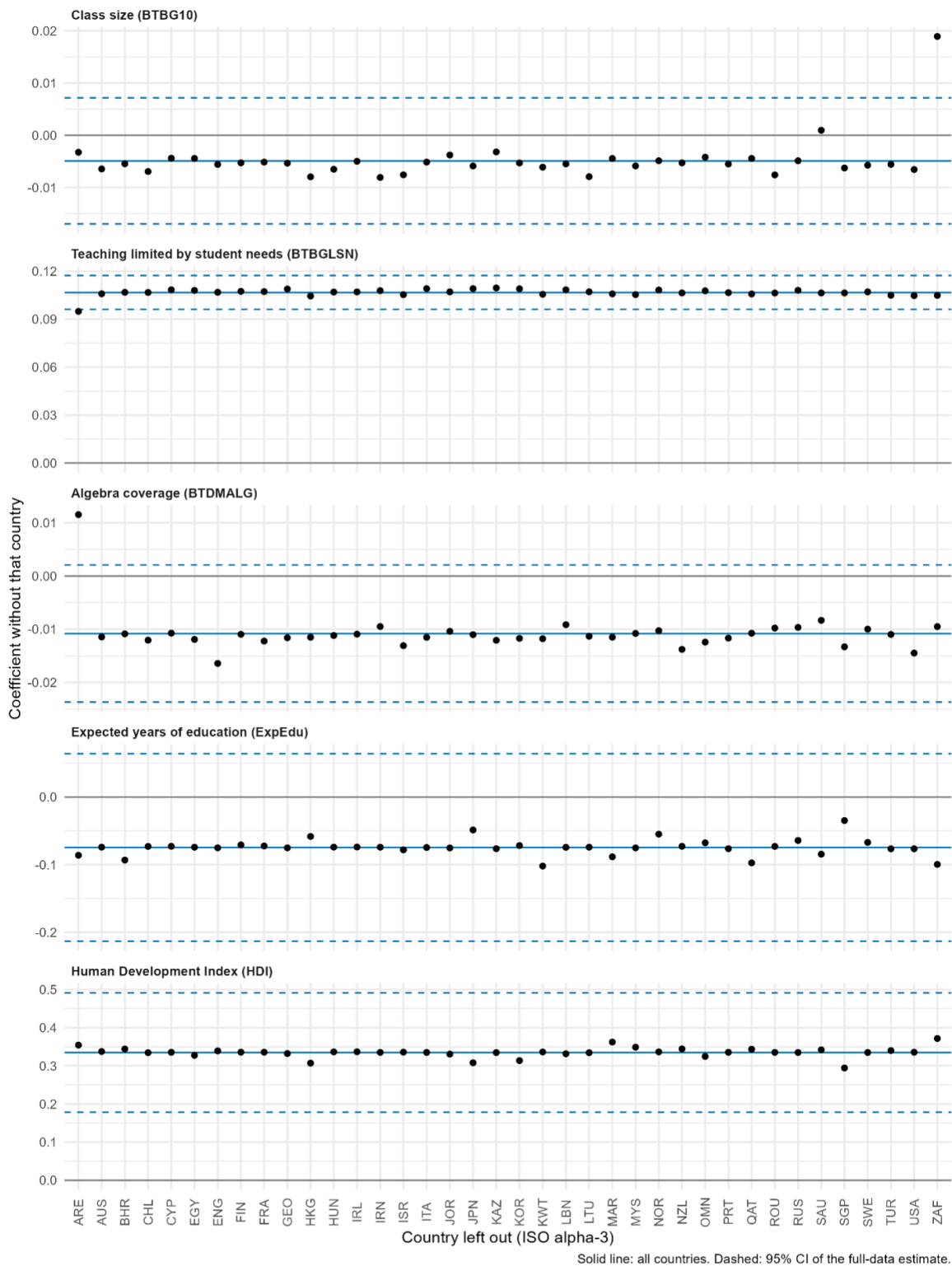


Figure 5

Complete-Case and Multiple-Imputation Estimates and 95% Confidence Intervals for the Multilevel Model of TIMSS 2019 Mathematics Achievement. Open circle represent estimated for the complete cases (n= 87,739); closed circles represent the estimates for the mice imputed data set (n = 228,158), m = 10 imputations, each fitted with 5 plausible values; estimates pooled with Rubin's rules.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementarymaterial.docx](#)