*Corresponding Author's Email: mariajose.ramirez001@gmail.com
Proceedings of the International Conference on Modern Research in Education, Teaching and Learning, Vol. 5, Issue. 1, 2026, pp. 1-17
DOI: https://doi.org/10.33422/icmetl.v5i1.2081

ISSN: 2783-7734 online

Global Evidence of Wide Variability in Mathematics Achievement within Classrooms:
The First Global Analysis of Within-Classroom Achievement Variability Using TIMSS 2023

María José Ramírez

EduMetrica LLC, USA

Abstract

Teachers teach students in classrooms; therefore, knowing how diverse their classrooms are in terms of student achievement is critical for educational policy and practice. This study provides the first global analysis of mathematics achievement variability within intact grade 4 classrooms using data from the Trends in International Mathematics and Science Study (TIMSS) 2023. Analysis of variance (ANOVA) served to determine the proportion of total mathematics score variance that lies (i) within classrooms, (ii) between schools, and (iii) between classrooms within the same schools. Descriptive statistics were used to further quantify the variability of mathematics scores existing within the classrooms. Results showed that in 60 of the 63 participating countries and education systems, within-classroom variance exceeded variation between classrooms and schools. On average, 70% of the mathematics score variance was within classrooms, considerably more than the 30% found between classrooms and schools. The spread of achievement in an average classroom covered three full standard deviations of the national achievement distributions. This was equivalent to almost five years of primary school. Most classrooms contained students from every TIMSS achievement level. On average across all entities, 10% of the students were at the Advanced level, 24% High, 29% Intermediate, 21% Low, and 16% at the Below Low level. This study challenges the perception that the main differences in student achievement are between schools instead of within classrooms. It also challenges the vision that in a high quality and egalitarian school system, all students have similar achievement levels. Classroom heterogeneity is a global characteristic of primary education. The findings suggest that policies aimed at improving learning should place greater emphasis on supporting teachers to address within-classroom heterogeneity. For instance, by introducing more flexible curricula and differentiated instruction, countries may be in a better position to address the learning needs of all their students.

Keywords: within-classroom achievement variability, classroom heterogeneity, mathematics achievement, TIMSS 2023, differentiated instruction

1. Introduction

International and national large-scale assessments have encouraged researchers and policymakers to compare differences in student achievement between schools and countries. This, in turn, has led to the wrong perception that most of the achievement differences are concentrated at these levels.

Little attention has been given to the achievement differences that exist within the classrooms. Yet teachers teach students within classrooms. One of the main challenges that teachers face every day is how to deal with students that are at different stages of learning. In classrooms where all students have more or less the same achievement, similar learning experiences may be appropriate to boost learning. On the contrary, in classrooms where students have a wide range of knowledge and skills, a more differentiated approach would be more consistent with learning theories and effective instruction.

Despite the critical importance of how homogeneous or heterogeneous the classrooms are in terms of student achievement, this topic has been largely absent from the global discussion in education. Because of this research gap, critical questions are left unanswered: How homogeneous or heterogeneous are the classrooms from different countries in terms of student achievement? How is student achievement distributed within and between classrooms and schools in the countries? What is the range of student achievement that a teacher typically encounters when teaching to a whole classroom in different countries?

This study shows that, all over the world, it is within the classrooms—not between the schools—that most of the student achievement variation is found. This is the achievement variation that teachers must address to support student learning. Nevertheless, the public debate and educational policy have traditionally devoted greater attention to differences between schools than to achievement variability within classrooms.

This paper fills a gap in the literature by examining the mathematics achievement variability existing within the classrooms of different countries. For the first time, in-depth analyses of the within-classroom mathematics achievement variance are presented at a global scale using data from the TIMSS 2023 international assessment.

Although variance decomposition has long been used to study differences between schools, surprisingly little attention has been devoted to interpreting the within-classroom component itself. This paper fills that gap by examining classroom achievement heterogeneity at a global scale.

The findings provide strong evidence that achievement diversity in the classroom is the norm, not the exception. The study highlights the implications for teaching and learning in classrooms with students that are at diverse learning levels. Drawing on learning theories that converge on the notion that learning is always built on previous learning, and that instruction should be sensitive to the actual learning level of the students, this paper argues for pedagogical approaches that take into account the diversity of achievement levels in the classroom.

2. Literature Review

2.1. Achievement Variability

International large-scale assessments have provided evidence of the strong differences in achievement that exist between and within countries, but have fallen short to report on the achievement variability that exist within the classrooms.

In the case of the international assessment PISA, it is not possible to compute the within-classroom variance given that no intact classrooms are selected. Instead, 15-year-old students are sampled from different classrooms and grades. Nevertheless, PISA 2022 reports that, across all 81 countries and economies participating in this study, 67% of the mathematics achievement variance was within schools, while the remaining 33% was between schools (OECD, 2023). Interestingly, the data shows a positive correlation between within-country achievement variation and mean achievement in mathematics.

Other international assessments do not to report on within-classroom variability, despite the fact of sampling intact classrooms within schools. TIMSS 2023 result report (von Davier, Kennedy et al., 2024) highlights that there are important differences in the average achievement of the 63 countries and education systems that participated in this study at grade 4, and that there is even more variability within each jurisdiction. However, no further analyses are provided regarding the achievement variability within schools and classrooms.

Pedersen et al. (2023) went one step beyond, analyzing the within-classroom variability existing within schools in the United States using TIMSS 2019 data. The authors reported that while most of the mathematics achievement variance (68%) was within the classrooms at grade 4, at grade 8 the figure was only 37%. This drop in classroom variability as students progress in schools is probably the consequence of school tracking, a more common practice in upper grades this country (NCES, 2021).

The 2019 Estudio Regional Comparativo y Explicativo (ERCE) refers to differences in achievement between and within the 16 Latin American countries participating in the study (UNESCO, 2023). It reports that 45% of the achievement variance of 3rd grade students was between schools. Given that only one classroom was sampled per school, this should be interpreted as variance that is between classrooms from different schools. It is implicit that the other 55% corresponds to within-classroom achievement variance.

Using national large-scale assessments, very few studies on the achievement variability within the classrooms seems to exist. But the few ones tell a very compelling story of wide classroom heterogeneity. Data from Chile's census of national learning assessment SIMCE, shows that around 70% of the mathematics achievement variance was within 4th grade classrooms (Fontaine & Urzúa, 2018; Ramírez, 2007). A similar finding was reported in the state of Guanajuato in Mexico, where data from the census assessment RIMA shows that 72% of the Spanish score variance was found within the 6th grade classrooms (Ramírez, 2023; Secretaría de Educación de Guanajuato, 2022).

No other studies using national or international large-scale assessments that quantify the achievement variability within the classrooms were found. This research gap may originate in the tendency to focus on independent variables (e.g., attendance, poverty) to explain differences in performance between students and schools.

While the direct comparisons of the results described above may have limitations due to differences in sampling and test design, among others, the picture that emerges is that students from different countries learn in schools and classrooms that are very diverse in term of their mathematics achievement. It is within the classrooms, and not between the schools or the countries, were most of the achievement variability is found.

The decomposition of achievement variance has a long tradition in educational effectiveness research. Multilevel models and variance partition coefficients have frequently been used to estimate the proportion of achievement differences attributable to students, classrooms, and schools (Raudenbush & Bryk, 2002; Teddlie & Reynolds, 2000). These studies have contributed substantially to understanding school effects. However, they have generally treated the within-classroom component as residual variance rather than as an object of substantive interest. Consequently, relatively little attention has been devoted to characterizing the magnitude of achievement heterogeneity that teachers encounter within intact classrooms.

2.2. Learning Theories

Evidence regarding classroom heterogeneity raises an important pedagogical question: if classrooms are highly heterogeneous, what guidance do learning theories provide for instruction?

Learning theories converge on the notion that learning is always built on the foundations of previous knowledge and skills. It is organized in a hierarchical way, with new learning constructed on already existing cognitive structures (Piaget, 1947).

This constructivist vision of learning finds its roots in Vygotsky (1978), who heralded the notion that learning must be challenging but attainable, within the Zone of Proximal Development of the students. ZPD refers to tasks that children cannot do all by themselves, but can do with the support of a more knowledgeable mediator (e.g. teacher, parent, peer).

More recent research from neurosciences states that learning occurs on previously established neural paths, repurposing their existing organization. Learning needs to be anchored within a neural space that precedes it. Learning stimulates the brain, and this stimulation is experienced as curiosity and motivation, key factors for fostering continuous learning. Children lose curiosity when they lack mental stimulation. This happens in schools when instruction is not targeted at the right level, because it is too easy or too hard. Therefore, the importance of adapting instruction to the learning level of students (Dehaene, 2021).

2.3. Curriculum and Instruction

If learning theories converge on the idea that learning is built on the foundations of previous learning, what are the implications for curriculum and instruction?

The specialized literature indicates that good teaching requires more curriculum and instructional flexibility to address the diverse learning needs of students (CEPPE, 2013; CAST, 2024; Gordon, 2024; OECD, 2024; Tomlinson et al., 2003). That is, official curricula that recognizes that students in the same classrooms have very different competencies; flexibility for teachers to decide when to introduce new learning objectives and for which students in her classroom; flexibility to use different activities with different levels of complexity, so that they are challenging but attainable for different groups of students in her classroom; flexibility to work in small groups with students who have similar learning levels; and flexibility to assess her students using different learning criteria.

Pedagogical practices cannot follow a one-size-fits-all curriculum or "teaching to the mean" approach. These are not effective teaching practices for heterogeneous classrooms (Bondie et al., 2019). This is so because students at different achievement levels do not benefit equally from the same learning experiences (Pedersen et al., 2023; Tomlinson et al., 2003).

The need for adaptive or differentiated instruction depends, at least in part, on the degree of heterogeneity within classrooms. If students have relatively similar achievement levels, whole-class instruction may adequately meet their learning needs. Conversely, when classrooms encompass a wide range of achievement levels, teachers are more likely to need instructional approaches that respond to students' differing levels of readiness. Consequently, documenting the magnitude of within-classroom achievement variability is not merely a descriptive exercise; it provides essential evidence for understanding the educational relevance of differentiated and adaptive teaching practices.

Consistent with this perspective, evidence syntheses identify adaptive teaching, formative assessment, and targeted instruction among the instructional approaches associated with improved learning outcomes (Black & Wiliam, 1998; Hattie, 2023; Slavin, 1987).

Similarly, Teaching at the Right Level demonstrates the benefits of organizing instruction around students' current learning levels rather than solely by age or grade (J-PAL, 2022). Tailored or differentiated instruction seems to be one of the most cost-effective approaches to maximize learning (J-PAL, 2019).

These findings suggest that when classrooms contain students with widely differing levels of achievement, instructional flexibility may be particularly important for maximizing learning opportunities for all students.

3. Methods

This paper used data from the Trends in International Mathematics and Science Study (TIMSS) 2023 at grade 4 to compute mathematics achievement variability within and between classrooms and schools. TIMSS is a pioneer international assessment program that has been administered for over 30 years. Countries use its results to monitor and support student achievement, and to compare their performance to other countries (von Davier, Kennedy et al., 2024).

TIMSS 2023 administered mathematics tests to nationally representative samples of students within each country or education system, encompassing near 400,000 students in 63 jurisdictions. Within each education system, a two-stage stratified random sample design was used, with a sample of schools drawn as a first stage and one or more intact grade 4 classrooms selected from each of the sampled schools as a second stage. Probability-proportional-to-size (PPS) sampling was used so that larger schools had a higher probability of being selected. Each country and education system sampled at least 150 schools and around 6,000 students (ranging from around 2,000 in Albania to almost 35,000 in United Arab Emirates); with one or more intact 4th grade classrooms sampled per school. The number of classrooms sampled per school varied between 1-4 in most of the jurisdictions. Once the schools were selected, intact classrooms were sampled rather than students from across the grade level because instructional experiences are typically organized on a classroom basis (Siegel & Foy, 2024).

TIMSS was designed to draw optimal samples of students, but these may not necessary be optimal samples of classrooms or schools within countries and education systems. Nevertheless, these samples provide reasonable estimates of the distribution of achievement variability within the classrooms of each education system, which is the focus of this paper.

The TIMSS 2023 database was the main source of information for this study. All the analyses were performed according to the guidelines provided in the TIMSS 2024 Technical Report (von Davier, Fishbein, et al. 2024) and User Guide (Fishbein et al., 2025).

The International Association for the Evaluation of Educational Achievement (IEA) IDB Analyzer (IEA, 2025) was used to compute descriptive statistics, whereas IBM SPSS Statistics 23 was used to aggregate data and to perform analysis of variance (ANOVA); this technique was chosen given that it is much simpler and intuitive than multilevel models. The overall mathematics score variance was analyzed at three levels: (i) within classrooms, (ii) between schools, and (iii) between classrooms within the same schools. However, the third level was only computed in 32 out of the 63 countries and education systems that sampled more than 1.5 intact classrooms on average per school.

The mathematics score variance was decomposed independently for each plausible value (ASMMAT01–ASMMAT05) within each participating country or education system. Country-level estimates were obtained by averaging the percentages of within-classroom variance across the five plausible values. All analyses were weighted using the overall student sampling weight (TOTWGT). Standard errors and 95% confidence intervals were estimated using the TIMSS jackknife repeated replication method with the 150 replicate weights (JKZONE and JKREP), and were combined across the five plausible values.

As a robustness check, unconditional two-level random-intercept models were estimated separately for each participating country or education system in STATA 15, with students nested within classrooms. Models were fitted independently for each of the five TIMSS mathematics plausible values using the student sampling weight (TOTWGT). For each model, the intraclass correlation coefficient (ICC) was estimated to quantify the proportion of total achievement variance attributable to differences between classrooms. The percentage of variance occurring within classrooms was then calculated as the complement of the ICC. Country-level estimates were obtained by averaging the ICCs across the five plausible values, and the corresponding within-classroom variance percentages were derived from these averaged ICCs.

To provide a familiar way to understand achievement variability within the classroom, this study reports classroom heterogeneity in equivalent years of schooling. In each education system, this was done by: (i) computing the raw range of scores within an average classroom; (ii) computing the standardized range of scores by dividing the raw range by the standard deviation of the score distribution, and (iii) dividing the standardized range by a measure of the distance existing between two adjacent school grades.

The measure of the distance between adjacent grades was taken from TIMSS 1995, which used nationally representative samples of 3rd and 4th grade students (Mullis et al., 1997). TIMSS 1995 is the only large-scale international assessment that assessed both Grades 3 and 4 on a common scale and using nationally representative samples of students. Later cycles target grades 4 and 8, so it is not possible to compute the distance between adjacent grades. Therefore, the 1995 estimate provides the best available benchmark.

Considering that in 1995 the distance between these two adjacent grades was equivalent to 0.64 standard deviations on average across all participant countries, it is reasonable to assume that a similar distance existed in 2023. This approach to compute equivalent years of schooling is similar to the one used by PISA[1] (OECD, 2023).

The relationship between within-classroom achievement variability and national mathematics achievement was examined using the Pearson product-moment correlation coefficient. Countries and education systems were the units of analysis, with each of the 63 participating entities contributing equally to the analysis. The two variables entering the correlation—the national mean mathematics achievement and the percentage of within-classroom variance—were country-level estimates previously computed following the TIMSS 2023 methodology, using the five plausible values, student sampling weights, and jackknife repeated replication procedures. Confidence intervals for the correlation coefficient were computed using Fisher's z transformation, the standard method for statistical inference on Pearson correlations because it provides an approximately normal sampling distribution from which confidence intervals can be derived.

[1] PISA (OECD, 2023) uses a much liberal estimate of 0.20 standard deviations as a reference for the distance between two adjacent school grades.

4. Results

This section presents evidence of wide achievement variability in the classrooms from four different angles: (1) The distribution of mathematics achievement variance that is (i) within classrooms, (ii) between schools, and (iii) between classrooms within the same schools; (2) The range of mathematics achievement in the classrooms; (3) The diversity of mathematics achievement levels in the classrooms; and (4) The country classroom profiles. Additional analyses show (5) the relationship existing between classroom achievement variance and overall mathematics performance of the countries.

4.1. Distribution of Mathematics Achievement Variance within and Between Classrooms

The distribution of mathematics achievement variance within and between classrooms was computed for all 63 countries and education systems participating in TIMSS 2023 at grade 4. The within-classroom variance accounted for the largest proportion of mathematics achievement variance in 60 out of the 63 cases (Figure 1).

Across the 63 participating countries and education systems, an average of 69.53% of the total mathematics score variance occurred within classrooms (SE = 0.30 percentage points, 95% CI [68.94%, 70.11%]). This is a substantially higher proportion than the 30.47% that was found between classrooms and schools.

The countries with the highest proportion of within-classroom variance were Slovenia (88%), Chinese Taipei and Japan (87% each); with the two Asian countries also being among the top performing TIMSS countries. On the other side, the countries with the lowest proportion of within-classroom variance were the United Arab Emirates (45%) and South Africa (47%); the latter being a bottom performing country in TIMSS. See Annex A.

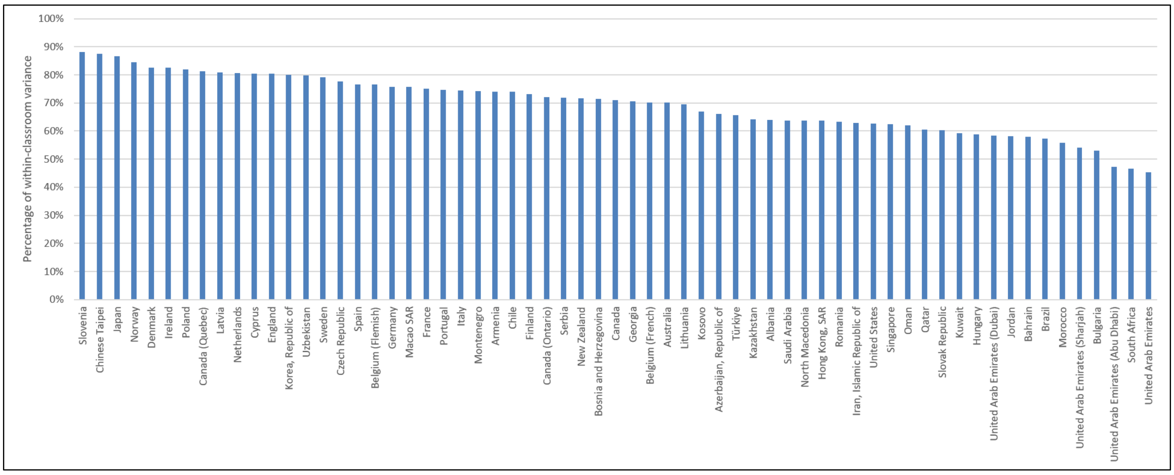

Figure 1. Percentage of total mathematics score variance that is within the classrooms

Note. Results based on the 63 countries and education systems that participated in TIMSS 2023 at grade 4. Results computed using all five plausible values of the overall mathematics score and weighted by the overall student sampling weight.

As a robustness check, unconditional two-level multilevel models were estimated separately for each participating country. On average, 73.97% of the total variance in mathematics achievement occurred within classrooms, with values ranging from 48.03% in South Africa and 92.79% in Slovenia. The models produced results that were consistent with the ANOVA-based variance decomposition, confirming that the majority of achievement variation lies among students within the same classroom rather than between classrooms.

In 32 out of the 63 education systems participating in TIMSS 2023 at grade 4, it was possible to compute the amount of variance between classrooms from the same schools. These jurisdictions sampled more than 1.5 classrooms on average per school. This allowed to break down the achievement variance that was (i) within classrooms, (ii) between schools, and (iii) between classrooms within the same schools. On average across these 32 jurisdictions, 70.06% of the mathematics score variance was within the classrooms; 25.13% was between the schools and; the remaining 4.81% was between the classrooms within the same schools (Figure 2).

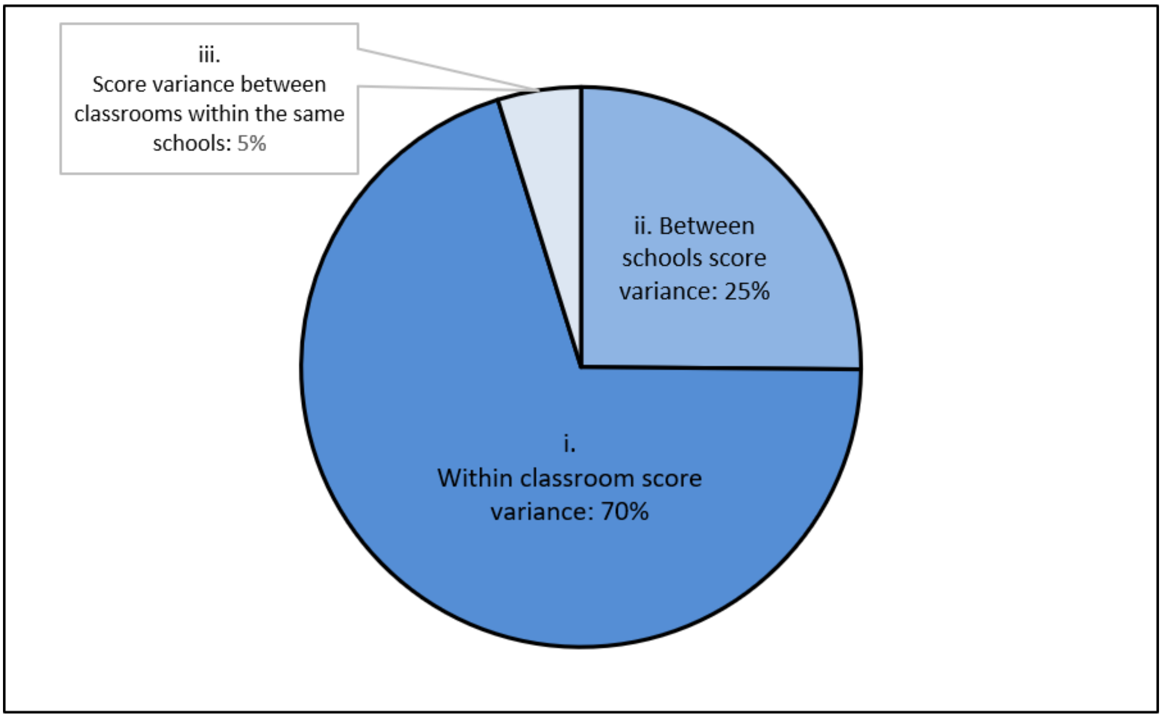

Figure 2. Decomposition of mathematics score variance that is (i) within classrooms, (ii) between schools, and (iii) between classrooms within the same schools.

Note. The figure shows the overall average result of 32 countries and education systems that sampled more than 1.5 classrooms on average per school in TIMSS 2023 at grade 4. Results computed using all five plausible values of the overall mathematics score and weighted by the overall student sampling weight.

These results, based on 32 jurisdictions, almost exactly replicated the variance decomposition across all 63 education systems participating in TIMSS 2023 at grade 4. With some caution, it is possible to assume that the same proportion of variance between classrooms within the same schools (4.81%) would be found across all 63 participants. Globally, the variance break down would then be 69.53% within the classrooms, 25.66% (30.47-4.81) between schools, and 4.81% between classrooms within the same schools.

4.2. Range of Mathematics Achievement in the Classrooms

On average, the range of mathematics scores across all 4th grade classrooms from the 63 education systems participating in TIMSS 2023, was equal to 266 score points. Given that the average standard deviation across all education systems was equal to 85 score points, the average range of scores within the classrooms was equivalent to 3.14 (266/85) standard deviations in the mathematics scale.

This result was translated into equivalent years of schooling. Assuming that the distance between two adjacent grades was equal to 0.64 standard deviations (Mullis et al., 1997), the average range of scores across all countries was equivalent to 4.91 (3.14/0.64) years of primary school. The range varied from a minimum of 4.17 years of schooling in the United Arab Emirates to a maximum of 5.99 in Japan.

Sensitivity analyses assuming that the distance between adjacent grades was equal to 0.55 and 0.75 standard deviations yielded classroom ranges between approximately 4.21 and 5.74 equivalent years of schooling.

Similar results were obtained when the analyses were restricted to the 14 countries that participated in both TIMSS 1995 and TIMSS 2023. For these countries, the average range of scores within classrooms was 3.17 standard deviations. Dividing this value by 0.64 yielded an average within-classroom achievement range equivalent to 4.95 years of primary schooling.

Together, these findings suggest that the conclusion that classrooms encompass several years of learning is robust to reasonable alternative assumptions and to differences in country participation. The estimate of equivalent years of schooling should be interpreted as a heuristic intended to facilitate understanding of the magnitude of within-classroom achievement variability, rather than as a direct measure of learning progression.

4.3. Diversity of Mathematics Achievement Levels in the Classrooms

Another measure of within-classroom heterogeneity is the proportion of students in different achievement levels found in an average classroom. On average across all 63 education systems under analysis, an average 4th grade classroom contained students from every TIMSS achievement level: 10% Advanced, 24% High, 29% Intermediate, 21% Low, and 16% Below Low (Figure 3).

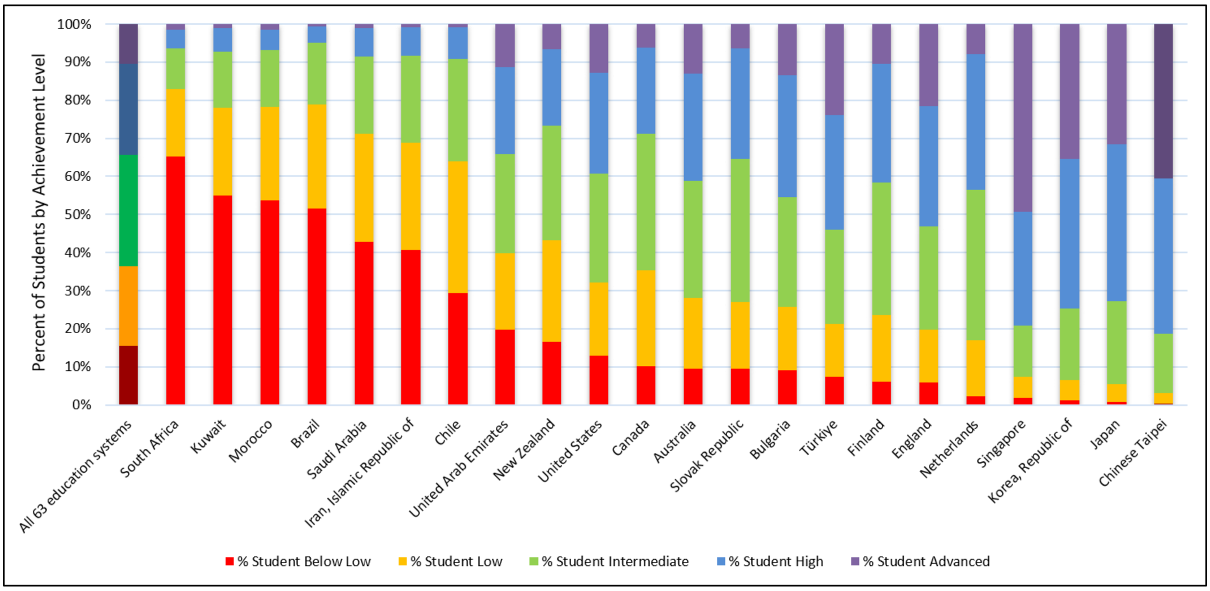

Figure 3. Percent of students by mathematics achievement level within an average classroom

Note. Results based on a selection of 22 countries participating in TIMSS 2023 at grade 4. Results computed using all five plausible values of the overall mathematics score and weighted by the overall student sampling weight.

These results indicate that in an average classroom, there are students with qualitatively different mathematics knowledge and skills. For instance, students in the Advanced level can select and relate information to implement appropriate operations to solve problems; whereas students in the Low level only demonstrate basic mathematical understanding of operations, geometry, measurement and data. In the meantime, students in the Below Low level do not even have the knowledge and skills of the Low level. This means that in practically every country and education system, an average classroom also includes students that do not even have a basic mathematical understanding (see what students at each achievement level know and can do in Annex B).

Annex C presents publicly released TIMSS 2023 mathematics items that students from different achievement levels were likely to answer correctly. There is one item from the Advanced, High and Intermediate achievement levels[2]. The items measure different cognitive domains (applying or reasoning) within the same content domain of Numbers. This allows for a better visualization of the range of skills of students in the same classrooms, and of the challenges of teaching students with a wide range of achievement within the classrooms.

4.4. Country Classroom Profiles

Country classroom profiles provide visualizations of the distribution of achievement variance within and between the classrooms. Figure 4 presents the profiles of four countries, including two with very high and low within-classroom variability (Japan and South Africa, respectively), plus two others that are more average in terms of their within-classroom variability (Chile and France).

In Japan, one of the top performing TIMSS countries, up to 87% of the country score variance was within the classrooms (among the most of any participating entity). In this country, all classrooms have relatively similar score range (in the figure, represented by the length of the vertical lines) and median performance (represented by the relatively horizontal band of lines). Therefore, it does not matter too much which classroom (or school) a student attends. Moreover, in practically every classroom even the least proficient students scored above the 400-point reference line (separating the Low from the Below Low achievement level).

On the other extreme there is South Africa, one of the bottom performing TIMSS countries, where only 47% of the country variance was within the classrooms (one of the least of any education system under analysis). While in the figure the vertical lines representing the score range within each classroom are relatively long, they represent a smaller share of the overall country achievement variance. In this African country, it makes a big difference which classroom (and school) a student attends. Moreover, only the most proficient students within each classroom scored above the 400-point reference line.

Somewhat in between were Chile and France, with 74% and 75% of achievement variance within the classrooms, respectively. Both these countries were closer to the average achievement variance found within the classrooms across all 63 education systems (70%). In the case of France, only the bottom of the vertical lines fell below the 400-point reference line, whereas in the case of Chile, nearly half did, indicating important differences in the average performance of these countries.

[2] No items from the Low achievement level were published in the TIMSS 2023 result report. Since the Below Low level correspond to a default level for students who do not reach the Low level, there are no test items associated to this level.

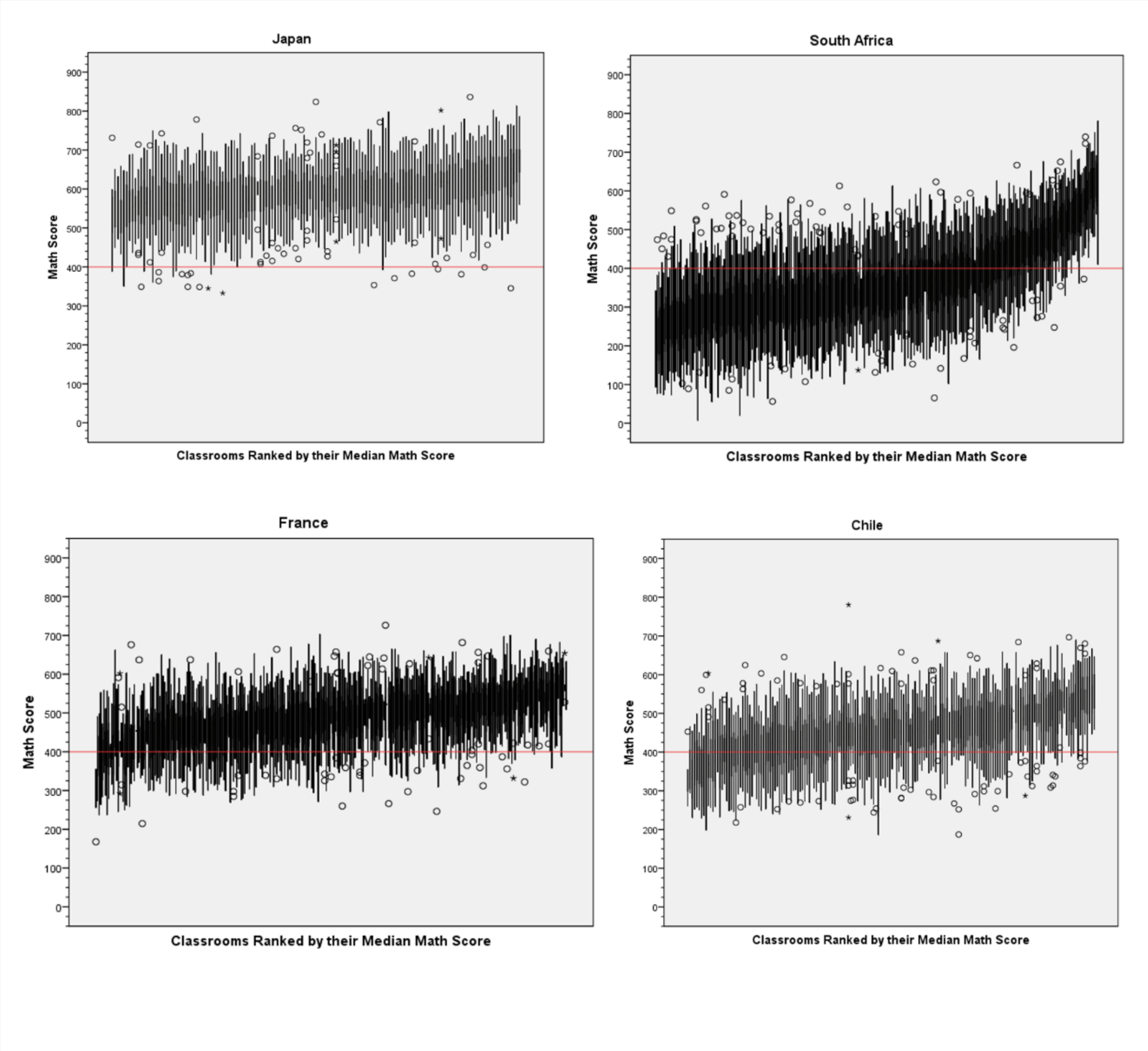

Figure 4. Country classroom profiles for four selected countries

Note. Results based on data from TIMSS 2023 at grade 4. Each vertical line represents a classroom. The length of the lines depicts the range of mathematics scores within each classroom. The circles and asterisks are outliers that depart from the classroom score distribution. The classrooms were sorted by their median mathematics score. The 400-point reference line signals the limit between the Low and Below Low mathematics achievement level.

4.5. Relationship between Within-classroom Variability and Mathematics Achievement

A key question addressed by this study is if there is any relationship between within-classroom achievement variability and the overall mathematics performance of the countries. Exploratory analyses showed that within-classroom variability was positively correlated with national mathematics achievement ($r = .40$, 95% CI [.17, .59], $p = .001$), indicating a moderate association between greater classroom heterogeneity and higher average country performance (Figure 5). This means that in countries where classrooms cover a wider range of scores relative to their national achievement distribution, average national mathematics achievement tends to be higher. Because this analysis is correlational and cross-sectional, no causal interpretation should be inferred.

It is worth noting the cases of top performing Japan and Chinese Taipei (both among the countries with the highest within-classroom ranges) and, on the other extreme, bottom performing Brazil, Morocco and South Africa (all three among the countries with the lowest within-classroom ranges).

A notable exception is Singapore, a top performing country with relatively low within-classroom achievement variability. This may be the consequence of some level of tracking or streaming that starts occurring in primary education in some schools.

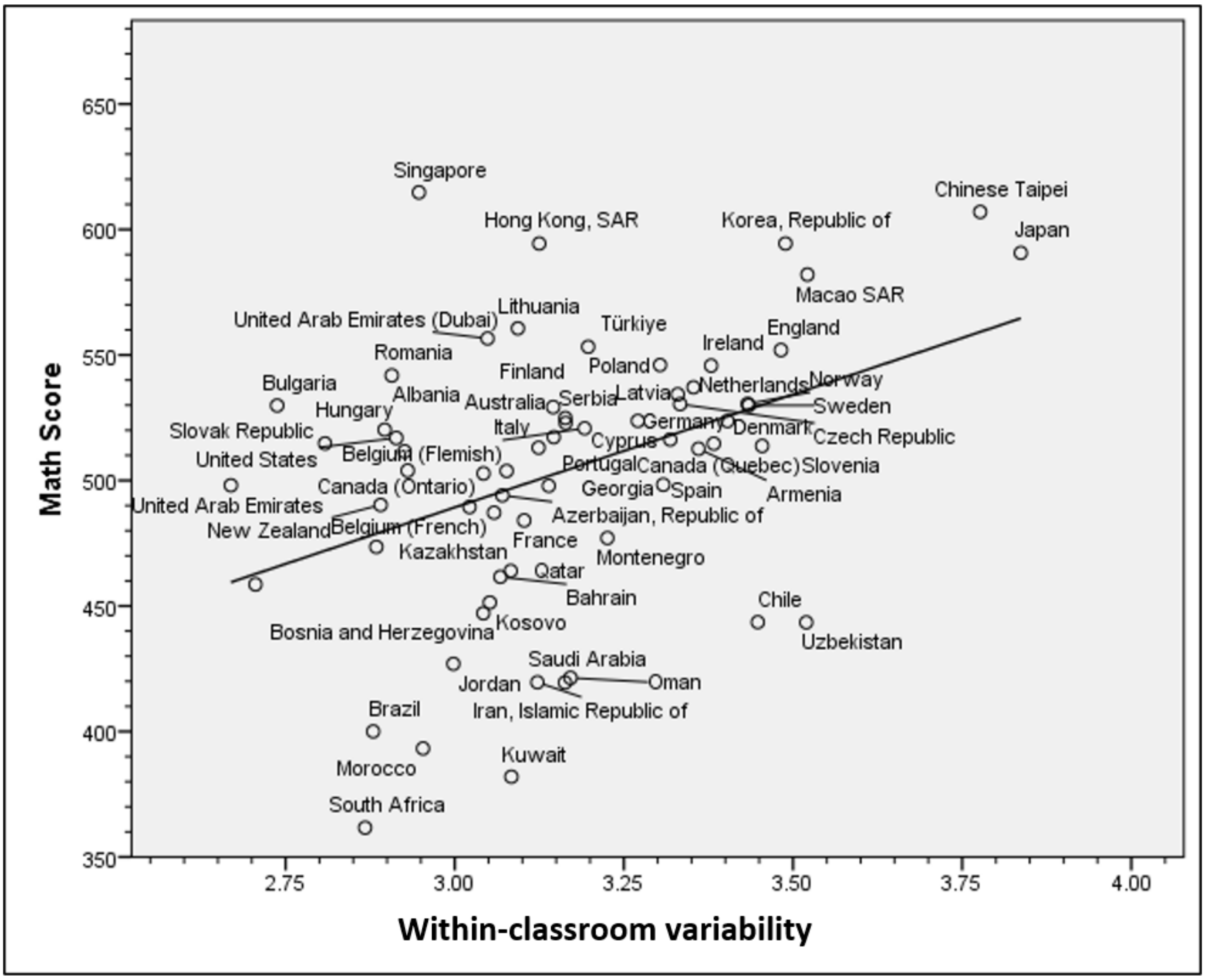

Figure 5. Relationship between within-classroom variability and overall mathematics performance

Note. Results based on the 63 countries and education systems that participated in TIMSS 2023 at grade 4. Within-classroom variability was measured in standard deviation units of the national achievement distribution.

5. Discussion

The findings of this study suggest that classroom heterogeneity is a defining characteristic of primary education worldwide. The results consistently show that most achievement variation occurs within rather than between classrooms and schools, that the range of scores within an average classroom is equivalent to several years of schooling, and that classrooms encompass students from very different achievement levels.

These findings contribute to the broader literature on variance decomposition in education. Previous studies have primarily used variance partitioning to quantify the relative contribution of schools and classrooms to overall achievement differences or to estimate school effects. In contrast, the present study interprets the within-classroom component from a pedagogical perspective, highlighting that the largest share of achievement variation represents the diversity of learning levels that teachers must address in their daily practice.

A commonly held belief among policy makers is that the main differences in academic achievement are between schools. This paper challenges this belief, providing global evidence that most of the differences in academic achievement are within classrooms and schools. The implication is that teachers face far greater variation in mathematics achievement within their own classrooms than exists between classrooms or schools.

Unfortunately, the achievement variability that exists within schools and classrooms is usually hidden behind the average mean scores of the schools. Teachers do not teach to average students; they teach to real students who have wide range of achievement variability. Therefore, recognizing this reality should become a central consideration in educational research, policy and practice. Pedagogical practices cannot follow a one-size-fits-all curriculum or "teaching to the mean" approach. These are not effective teaching practices for heterogeneous classrooms (Bondie et al., 2019).

Comparisons of student achievement suggest that differences between schools are less pronounced than is often believed in at least some countries. Data show that in fact there is a lot of overlap in achievement range between different schools (and classrooms). To support learning, what happens in the classrooms should receive much more attention than the debate about the quality of different schools (e.g., public, private).

This study does not deny that important differences in achievement exist between schools in the same country. What this study highlights is that these differences are relatively small compared to the wide achievement variance that exists within the classrooms.

The vision of an egalitarian school system where all students have similar achievement is not supported by this research. Even in the most egalitarian countries, where all classrooms and schools have similar performance, students in the same classrooms have very diverse achievement levels. Look at Japan, a high quality and egalitarian country where all schools and classrooms have very similar achievement levels. Nevertheless, Japanese classrooms are among the most heterogenous in the world. Equality seems to be a more relevant concept when discussing achievement differences between classrooms and schools, but not within the same classrooms.

The findings reported in this paper show that diversity in the classroom may be a good thing. In education systems where classrooms are more diverse in the sense that they share a wider range of the national achievement distribution, the overall country achievement tends to be higher. Because these findings are very exploratory, caution should be used in interpreting them.

A basic principle that emerges from learning theories and teaching practices is that for learning to occur, this must be anchored in students' previous learning. The findings suggest that policies supporting teachers to respond to classroom heterogeneity deserve greater consideration. Countries can follow complementary approaches to implement such policies:

- *Curriculum flexibility.* This requires official documents to recognize that students within the same grades may have very different learning needs (CEPPE, 2013; CAST, 2024; Gordon, 2024; OECD, 2024; Tomlinson et al., 2003). This contrasts with curricula that are too rigid and overloaded.
- *Instructional differentiation.* Teachers use professional judgement to adapt learning objectives and teaching practices to their classrooms and students. Teachers can use an evidence-informed teaching approach, adaptive teaching and learning, formative and/or

adaptive assessments to elicit, interpret, and use data on actual learning levels of the students to make decisions about next steps in instruction (Andersson & Palm, 2017). Learning activities can be tailored to the different learning needs of students. This can be done, for instance, by organizing students into small groups with similar achievement levels for specific activities.

- *Tailored pedagogical resources*. Countries can offer learning guides, activity sheets or digital learning applications aligned with official curriculum, but scaffolded at different stages of learning, to be used by teachers and students from different grades. This approach is at the core of the success of the worldwide recognized New School Model (Chi et al., 2017).
- *Tutoring* refers to a specific type of remedial education that supplements regular instruction. It can be provided one on one or in small group to students that are lagging in their learning. Tutoring programs have consistent and positive effects on student learning (J-PAL, 2020; World Bank et al., 2022).
- *Grouping students according to their current learning level*. A more radical approach is not to group students in grades and classrooms according to their age but rather to do so based on their actual learning level. This approach has been successfully supported by programs such as Teaching at the Right Level (J-PAL, 2022).
- *Teacher training.* Changes in teacher practices would require training that highlight achievement differences in the classrooms, and support teachers to address this diversity.

The findings reported in this paper seem to be very stable in time and robust to variations in country participation. Similar analyses performed by the author using TIMSS 2019 data in 64 education systems (of which 55 also participated in TIMSS 2023) show essentially the same results (Ramírez, 2025). For instance, the within-classroom variance only changed from 69% in 2019 to 70% in 2023. The implications of these findings are relevant to countries with different levels of economic development and from different geographic regions.

In interpreting the findings of this study, it is important to bear in mind some limitations. First, the findings of this study are probably more applicable in the context of early primary school. Achievement variance within classrooms may drop in upper grades as a consequence of academic tracking. Second, the TIMSS sampling design was optimized to estimate achievement at the student rather than at the school or classroom level. Therefore, the representativeness of results computed at the school and classroom levels may be weaker compared to the ones computed at the student level. Third, by selecting schools with probability proportional to size some bias may have been introduced. The findings may differ if smaller (and less urban) schools had been sampled. Fourth, there may be a floor effect that limits the achievement variability of lower performing countries. Therefore, the positive correlation between within-classroom variability and overall country performance may be overestimated.

Future research should investigate how classroom heterogeneity changes across grades, subject areas, and countries, and examine which instructional approaches are most effective in highly heterogeneous classrooms.

6. Conclusions

This study provides the first global analysis of mathematics achievement variability within grade 4 classrooms using TIMSS 2023 data from 63 countries and education systems. The findings consistently show that classroom heterogeneity is a defining characteristic of primary education worldwide. In 60 of the 63 participating education systems, most of the mathematics achievement variance was found within classrooms rather than between classrooms and schools. On average, approximately 70% of the total variance occurred within classrooms, indicating that students sharing the same classroom often differ substantially in their mathematics knowledge and skills.

The findings also demonstrate that the achievement range within an average classroom is remarkably wide. Across participating education systems, classrooms typically spanned more than three standard deviations of the national mathematics achievement distribution, which is equivalent to five full years of schooling. An average classroom included students representing every TIMSS international benchmark level, from Below Low to Advanced. These results suggest that teachers routinely teach students with markedly different levels of prior knowledge and learning needs, even within education systems that are considered equitable.

The study contributes to the literature by shifting attention from differences between countries and schools to differences within classrooms. International assessments have traditionally been used to compare the performance of education systems and schools, yet comparatively little attention has been paid to the achievement diversity that teachers encounter in their daily practice. By documenting the magnitude and consistency of within-classroom variability across a large number of education systems, this study provides a new perspective on how student achievement is distributed and where much of the challenge of improving learning resides.

The findings have important implications for educational policy and practice. If most achievement variation occurs within classrooms, then efforts to improve learning should complement system-level reforms with strategies that help teachers respond effectively to diverse learning needs within their own classrooms. Approaches such as curriculum flexibility, differentiated instruction, formative assessment, targeted tutoring, and adaptive learning environments represent possible ways of supporting teaching in highly heterogeneous classrooms. Future research should examine how classroom heterogeneity evolves across grades and subjects areas, how it differs across educational contexts, and which instructional approaches are most effective in helping teachers respond to this diversity.

Ultimately, teachers do not teach average students, average classrooms, or average schools. They teach highly diverse groups of learners. Recognizing and responding to this diversity should become a central consideration in educational research, policy, and classroom practice.

References

Abdul Latif Jameel Poverty Action Lab (J-PAL). (2019). *Tailoring instruction to students' learning levels to increase learning*. J-PAL Policy Insights. https://doi.org/10.31485/pi.2522.2019

Abdul Latif Jameel Poverty Action Lab (J-PAL). (2020). *The transformative potential of tutoring for PreK–12 learning outcomes: Lessons from randomized evaluations.* https://www.povertyactionlab.org/publication/transformative-potential-tutoring-pre-k-12-learning-outcomes-lessons-randomized

Abdul Latif Jameel Poverty Action Lab (J-PAL). (2022). Teaching at the Right Level to Improve Learning: Evidence to Policy Case Study.

Andersson, C., & Palm, T. (2017). The impact of formative assessment on student achievement: A study of the effects of changes to classroom practice after a comprehensive professional development programme. *Learning and Instruction, 49*, 92–102. https://doi.org/10.1016/j.learninstruc.2016.12.006

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Bondie, R. S., Dahnke, C., & Zusho, A. (2019). How does changing "one-size-fits-all" to differentiated instruction affect teaching? *Review of Research in Education, 43*(1), 336–362. https://doi.org/10.3102/0091732X18821130

Dehaene, S. (2021). *How we learn: Why brains learn better than any machine... for now*. Penguin Press.

CAST. (2024). *CAST Universal Design for Learning Guidelines version 3.0*. https://udlguidelines.cast.org

CEPPE. (2013). *Learning standards, teaching standards and standards for school principals: A comparative study* (OECD Education Working Papers No. 99). OECD Publishing. https://doi.org/10.1787/5k3tsjqtp90v-en

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

Fontaine, A., & Urzúa, S. (2018). *Educación con patines*. Ediciones El Mercurio.

Gordon, D. (Ed.). (2024). *Universal design for learning: Principles, framework, and practice* (Updated ed.). CAST Professional Publishing.

Hattie, J. (2023). *Visible learning: The sequel*. Routledge.

International Association for the Evaluation of Educational Achievement. (2025). *Help manual for the IEA IDB Analyzer* (Version 5.0). https://www.iea.nl

Mullis, I. V. S., Martin, M. O., Beaton, A. E., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1997). *Mathematics achievement in the primary school years: IEA's Third International Mathematics and Science Study (TIMSS).* Boston College.

National Center for Education Statistics. (2021). *Pre-COVID ability grouping in U.S. public school classrooms* (NCES 2021-139). U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2021139

Organisation for Economic Co-operation and Development (OECD). (2023). *PISA 2022 results (volume I): The state of learning and equity in education*. OECD Publishing. https://doi.org/10.1787/53f23881-en.

Organisation for Economic Co-operation and Development (OECD). (2024). *Curriculum flexibility and autonomy: Promoting a thriving learning environment.* OECD Publishing. https://doi.org/10.1787/eccbbac2-en.

Pedersen, B., Makel, M. C., Rambo-Hernandez, K. E., Peters, S. J., & Plucker, J. (2023). Most mathematics classrooms contain wide-ranging achievement levels. Gifted Child Quarterly, 67(3), 220-234. https://doi.org/10.1177/00169862231166074

Piaget, J. (1947). *The psychology of intelligence.* Routledge.

Ramírez, M. J. (2007). Diferencias dentro de las salas de clases: Distribución del rendimiento en matemáticas [Differences within classrooms: Distribution of mathematics achievement within classrooms]. *Estudios Públicos, 106*, 5–22. https://www.cepchile.cl/cep/site/artic/20160304/asocfile/20160304093218/rev106_ramirez.pdf

Ramírez, M. J. (2023). *Pérdida y recuperación de aprendizajes en contexto de pandemia en Guanajuato: Resultados de la evaluación RIMA 2020, 2021 y 2023 y reflexiones para mejorar* [Learning loss and recovery in the context of the pandemic in Guanajuato: Results of the RIMA 2020, 2021, and 2023 assessments and reflections for improvement] [Conference presentation]. Segundo Foro Internacional *Repensar la Educación: Construir el Mejor Sistema Educativo para Guanajuato*, Guanajuato, Mexico.

Ramírez, M. J. (2025). *Global evidence of wide variability in mathematics achievement within classrooms using TIMSS 2019 data* [Conference presentation]. 11th IEA International Research Conference (IRC 2025), Rome, Italy.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage Publications.

Siegel, P., & Foy, P. (2024). TIMSS sample design. In M. von Davier, B. Fishbein, & A. Kennedy (Eds.), *TIMSS 2023 technical report: Methods and procedures*. TIMSS & PIRLS International Study Center, Boston College. https://timss2023.org/methods

Secretaría de Educación de Guanajuato. (2022). *RIMA: Recopilación de información para la mejora de los aprendizajes. Informe de resultados de la evaluación de matemáticas y español RIMA 2020 y 2021* [RIMA: Information collection for improving learning. Results report of the RIMA 2020 and 2021 mathematics and Spanish assessments].

Slavin, R. E. (1987). Ability grouping and student achievement in elementary schools: A best-evidence synthesis. *Review of Educational Research, 57*(3), 293–336. https://doi.org/10.3102/00346543057003293

Teddlie, C., & Reynolds, D. (Eds.). (2000). *The international handbook of school effectiveness research*. Falmer Press.

Tomlinson, C. A., Brighton, C., Hertberg, H., Callahan, C. M., Moon, T. R., Brimijoin, K., Conover, L. A., & Reynolds, T. (2003). Differentiating instruction in response to student readiness, interest, and learning profile in academically diverse classrooms: A review of the literature. *Journal for the Education of the Gifted, 27*(2–3), 119–145. https://doi.org/10.1177/016235320302700203

UNESCO. (2023). *Fundamental learnings in Latin America and the Caribbean: Student learning achievement assessment. Regional Comparative and Explanatory Study (ERCE 2019): Executive summary.* UNESCO Regional Office for Latin America and the Caribbean.

von Davier, M., Kennedy, A., Reynolds, K. A., 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, Boston College. https://doi.org/10.6017/lse.tpisc.timss.rs6460

von Davier, M., Fishbein, B., & Kennedy, A. (Eds.). (2024). *TIMSS 2023 technical report: Methods and procedures.* TIMSS & PIRLS International Study Center, Boston College. https://timss2023.org/methods

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Harvard University Press.

World Bank, UNICEF, & UNESCO. (2022). *Two years after: Saving a generation*. World Bank. https://openknowledge.worldbank.org/handle/10986/37586