Article

Not peer-reviewed version

TIMSS and PISA are not a Competition: Purpose, Difference, and Measurement Standardization by Item Response Theory (IRT) in Large-Scale International Assessments

Wan Chong Choi * and Chi In Chang

Posted Date: 28 April 2025

doi: 10.20944/preprints202504.2356.v1

Keywords: TIMSS; Trends in International Mathematics and Science Study; PISA; Programme for International Student Assessment; IRT; Item Response Theory; Large-Scale International Assessments; Purpose of Educational Assessments; Differences between TIMSS and PISA; Measurement Standardization

 doi:10.20944/preprints202504.2356.v1

Article

TIMSS and PISA Are Not a Competition: Purpose, Difference, and Measurement Standardization by Item Response Theory (IRT) in Large-Scale International Assessments

Wan Chong Choi [1,2,*] and Chi In Chang [2]

[1] Department of Computer Science, Illinois Institute of Technology, U.S.
[2] Department of Psychology, Golden Gate University, U.S.
* Correspondence: wchoi8@hawk.iit.edu

Abstract: Large-scale international assessments have long attracted attention for their comparative insights into educational systems. This paper examined the purposes, design differences, and measurement methodologies of two major assessments: TIMSS (Trends in International Mathematics and Science Study) and PISA (Programme for International Student Assessment). It demonstrated that although often perceived as competitive rankings, these studies were fundamentally intended as diagnostic tools to inform educational improvement. TIMSS was found to focus on curriculum-based, grade-level achievements, whereas PISA emphasized age-based competencies applicable to real-world contexts. The analysis further explained how Item Response Theory (IRT) enabled both assessments to generate standardized and comparable scores across different populations and assessment cycles, thereby enhancing measurement reliability. Despite rigorous scaling efforts, challenges in achieving full international comparability continued to exist, particularly in relation to age-grade mismatches, curriculum coverage differences, and cultural variations. Finally, the study argued that TIMSS and PISA should be interpreted as instruments for educational development rather than benchmarks for international competition. A thoughtful and context-sensitive use of their results is essential for driving meaningful and sustainable educational reforms.

Keywords: TIMSS; Trends in International Mathematics and Science Study; PISA; Programme for International Student Assessment; IRT; Item Response Theory; Large-Scale International Assessments; Purpose of Educational Assessments; Differences between TIMSS and PISA; Measurement Standardization

1. Introduction

1.1. Background

Large-scale international assessments have become highly visible tools for comparing educational outcomes across countries. Among these, the TIMSS and PISA surveys attract significant attention for their periodic rankings of student achievement in mathematics, science, and other domains. The results are often widely reported in the media, leading to headlines about which countries "top the tables" and which are "falling behind." Policymakers sometimes react to small shifts in rankings with concern, treating the outcomes almost as an Olympic medal count for education.

However, it is critical to recognize that TIMSS and PISA were never intended as international competitions. Rather, they were designed as research-oriented assessments to provide data on educational systems' strengths and weaknesses and to inform policy and practice improvements [1]. Superficial comparisons of country rankings can be misleading and even "dangerous" when they become emotionally charged or stripped of context [2]. This paper examines the purposes and design differences of TIMSS and PISA, explains how both use Item Response Theory (IRT) to achieve

 doi:10.20944/preprints202504.2356.v1

standardized measurement, and discusses why their results should be interpreted with caution and insight rather than as simple scorecards.

We begin by outlining what TIMSS and PISA each measure and aim to achieve. We then compare their frameworks and implementation – curriculum-based vs. competency-based focus, grade-based vs. age-based sampling, and content vs. literacy domains – to clarify how these assessments differ. Next, we describe the role of IRT in scaling scores for TIMSS and PISA, highlighting how this method improves comparability over raw scores and allows tracking of trends across assessment cycles. In addition, we address challenges in achieving perfect international comparability, such as differences in students' ages and curricula, which can influence outcomes. Finally, we discuss the interpretation of TIMSS and PISA results, cautioning against the misuse of rankings and emphasizing the proper use of these studies as tools for educational research and improvement. Through this analysis, we reinforce the message that TIMSS and PISA are not a competition, and that understanding their distinct nature and methodologies is key to using their findings constructively.

1.2. Research Questions

Based on the goals of this paper to better understand large-scale international assessments and their proper interpretation, we set out to answer the following six research questions. Each question corresponds to a major section of this study and helps structure the comparison between TIMSS and PISA, their measurement methods, and the responsible use of their results.

RQ1. What is the purpose and design of TIMSS, and how does it measure student learning based on curriculum?

RQ2. What is the purpose and design of PISA, and how does it measure student skills based on competencies?

RQ3. What are the main differences between TIMSS and PISA in goals, frameworks, and sampling methods?

RQ4. How does IRT help TIMSS and PISA produce fair and comparable scores?

RQ5. What problems make it hard to compare results internationally in TIMSS and PISA?

RQ6. How should TIMSS and PISA results be interpreted?

2. RQ1: What Is TIMSS? – Curriculum-Based Assessment of Grade-Level Achievement

TIMSS (Trends in International Mathematics and Science Study) [3] is an international assessment conducted by the International Association for the Evaluation of Educational Achievement (IEA) [4] on a regular four-year cycle (1995, 1999, 2003, … up to the present). It evaluates the mathematics and science knowledge of students in specific grade levels – traditionally 4th grade (approximately 9–10 years old) and 8th grade (13–14 years old) in each participating country. TIMSS is fundamentally curriculum-based, meaning it is designed to measure how well students have learned the school curriculum in math and science at those grade levels [2].

In fact, TIMSS uses the curriculum as an "organizing concept," considering the intended curriculum (what students are expected to learn), the implemented curriculum (what is actually taught in classrooms), and the attained curriculum (what students have learned) [5].

This focus on curriculum is reflected in TIMSS's aim: to provide data that help countries understand the effectiveness of their mathematics and science education programs, thereby highlighting areas of strength and weakness in curriculum and instruction that can guide improvements [6]. In other words, TIMSS's primary purpose is diagnostic and formative for educational systems, not to declare which country is "best".

The TIMSS assessment framework is organized along two important dimensions: content domains and cognitive domains [7]. The content domains delineate the subject matter topics in mathematics and science that are assessed at each grade. For example, in mathematics, TIMSS Grade

 doi:10.20944/preprints202504.2356.v1

4 covers content areas such as Number, Geometric Shapes & Measures, and Data Display, while Grade 8 covers Number, Algebra, Geometry, and Data & Chance [8]. In science, Grade 4 content domains include Life Science, Physical Science, and Earth Science, whereas Grade 8 includes domains like Biology, Chemistry, Physics, and Earth Science [8]. These domains align closely with typical topics in school curricula.

In parallel, the cognitive domains of TIMSS describe the thinking processes and skills students must use: the TIMSS framework specifies three cognitive domains – Knowing, Applying, and Reasoning [8]. Knowing items test students' recall of facts, concepts, and procedures; Applying items require using knowledge to solve problems; and Reasoning items involve unfamiliar situations that require higher-order thinking and analysis. Every question in TIMSS is designed to assess a combination of a content topic and a cognitive skill. This two-dimensional framework ensures that TIMSS results can be interpreted in terms of what content students have learned and how well they can think through problems of varying complexity [8].

Because TIMSS is tied to curriculum, it also collects extensive background information. Participating students, their teachers, and school principals complete questionnaires about curriculum coverage, teaching practices, and resources. This context data helps link performance outcomes with curricular and instructional factors.

For instance, TIMSS publishes an international "encyclopedia" of education systems that documents each country's curriculum and policies in math and science [5]. Such information is invaluable for researchers and educators seeking to understand why students in certain systems perform better or worse on particular content areas. It underscores that TIMSS's intent is to generate insights for policy – for example, if a country's students struggle in a content domain like geometry, the TIMSS data (coupled with curriculum information) might suggest a need to strengthen that part of the national curriculum.

Another defining feature of TIMSS is that it is grade-based in its target population. TIMSS selects a grade (4th or 8th) and tests a representative sample of students in that grade in each country. This approach ensures that, across countries, the students have all had roughly the same number of years of schooling when tested (e.g., eight years of formal education by Grade 8). However, because schooling starts at different ages in different countries, the age of TIMSS test-takers can vary – one country's 4th graders may be 9 years old on average, while another's are 10 or 11 if formal schooling begins later. We will later discuss how this grade-based sampling can pose comparability challenges, but it is chosen to align with the idea of measuring attainment of the school curriculum at common educational stages.

In summary, TIMSS is a curriculum-aligned assessment targeting specific grades, structured around concrete content topics and cognitive skills taught in school. Its results are meant to help educational authorities monitor trends in achievement (hence "Trends" in its name) and to identify areas for curriculum and instructional improvement. Any comparisons of TIMSS scores across countries are most meaningful when considered alongside what is taught in each country – a high score implies students have learned their curriculum well, and a low score might indicate a gap between curriculum expectations and student learning that could be addressed.

3. RQ2: What Is PISA? – Competency-Based Assessment of 15-Year-Olds

PISA (Programme for International Student Assessment) [9] is an international assessment administered by the Organisation for Economic Co-operation and Development (OECD) [10] every three years since 2000. In contrast to TIMSS, PISA does not target a specific grade level; instead, it tests an age-defined sample: students who are 15 years old in participating countries (technically those between 15 years 3 months and 16 years 2 months at the time of testing). Fifteen-year-olds are typically nearing the end of compulsory education in many countries, so PISA aims to evaluate what students approaching school-leaving age can do with their knowledge.

 doi:10.20944/preprints202504.2356.v1

Importantly, PISA is not tied to any particular national curriculum. It is often described as competency-based or literacy-based, focusing on how well students can apply knowledge and skills in real-life contexts, rather than on whether they have mastered specific curricular content.

PISA assesses three main domains: reading literacy, mathematical literacy, and scientific literacy. Rather than traditional school subjects, these are defined broadly. For example, mathematical literacy in PISA is defined as "an individual's capacity to formulate, employ, and interpret mathematics in a variety of contexts… to describe, predict and explain phenomena" – essentially, the ability to use mathematical knowledge to solve real-world problems [2]. Similarly, scientific literacy is about using science knowledge to understand and make decisions about the natural world and science-related issues, and reading literacy involves understanding, using, and reflecting on written texts for one's goals. Each cycle of PISA typically emphasizes one of the domains (for instance, reading in 2018, mathematics in 2021, etc.), with that domain getting more assessment items, while still testing the other areas.

The PISA tasks are often set in practical or cross-curricular contexts. For example, a math problem in PISA might require students to analyze a chart from a newspaper or compute a practical quantity, rather than perform a routine textbook exercise. This approach reflects PISA's guiding question: "Are students well prepared for life after school?" The prime aim of PISA, as stated by OECD, is to determine the extent to which students near the end of compulsory schooling have acquired the knowledge and skills "that they will need in adult life" [7]. It is about measuring competencies – the ability to apply learning – rather than just recall of facts. As a result, PISA is sometimes described as assessing educational outcomes in a real-world sense, providing a check on how effectively education systems equip students with critical thinking and problem-solving skills.

Because PISA deliberately does not focus on specific curriculum content, it treats all participating students as a single cohort of 15-year-olds and does not adjust for differences in curricula. It assumes a common "literacy" framework that is intended to be internationally relevant. However, this means that PISA results speak to a somewhat different question than TIMSS. PISA's question is: how well can 15-year-olds, on average, apply reading, math, and science skills? This is framed as an indicator of how education systems as a whole prepare students for the challenges of modern society. It is not directly asking, for example, "Have Japanese or Brazilian 9th graders learned their country's algebra curriculum?" but rather, "Can Japanese or Brazilian 15-year-olds use math in various contexts?"

The sampling method (age-based sampling) introduces some diversity in the schooling background of the tested students. In one country, most 15-year-olds might be in Grade 10, whereas in another they might predominantly be in Grade 9 (if school starts later or if grade repetition is common). PISA requires that students be 15 and have at least some schooling (typically it excludes those in lower than 7th grade even if age-eligible) [7]. Still, the age–grade mismatch can be significant: in some countries a notable proportion of 15-year-olds may still be in lower secondary grades or have left school, whereas in others 15-year-olds are in higher grades. OECD chose an age-based design with the argument that comparing by age provides a common reference for preparedness for future life, since age is a consistent marker across countries (unlike grade, which can represent different schooling durations). Proponents argue this places countries on "more equal footing" for judging student readiness. Yet, as we will discuss, it also means that PISA results can be influenced by differences in average years of schooling that 15-year-olds have received in each country [7].

PISA's methodology includes extensive background questionnaires as well, asking students about their home, attitudes, and learning experiences, and school principals about school contexts. These data allow OECD to analyze factors associated with performance (for example, socioeconomic status, or amount of reading for pleasure). PISA reports often highlight policy-relevant findings such as the impact of socioeconomics on scores or students' attitudes toward learning. The OECD also explicitly positions PISA as a tool for policy development, claiming that PISA "provides insights into the factors that influence the development of skills at home and at school and how these factors interact and what the implications are for policy".

 doi:10.20944/preprints202504.2356.v1

In summary, PISA is a competency-oriented, age-based assessment that evaluates broad literacy in reading, math, and science among 15-year-olds. Its purpose is to measure how well education systems worldwide prepare students for real-life challenges and continued learning, rather than to test mastery of each country's school syllabus. PISA's outcome is a profile of students' abilities to apply knowledge, intended to inform policy on a broad level (such as identifying if students can think critically and solve novel problems) rather than on specific curriculum adjustments. As with TIMSS, however, PISA's results are often oversimplified into country rankings, which can obscure the subtleties of what is being measured.

4. RQ3: Key Differences Between TIMSS and PISA

Although both TIMSS and PISA provide valuable information on student achievement in an international context, they differ significantly in purpose, content focus, and population. These differences mean that the two assessments are not directly interchangeable and often should not be directly compared as if they were measuring the exact same thing. Here we outline several key distinctions.

4.1. Curriculum-Based (TIMSS) vs. Competency-Based (PISA)

TIMSS is explicitly curriculum-driven, rooted in the specific content that students are expected to learn in school by 4th or 8th grade [2]. In contrast, PISA is not curriculum-driven; it is constructed to be as curriculum-independent as possible, emphasizing cross-cutting skills and competencies. This fundamental difference means TIMSS questions might look more like typical school exam questions (e.g., arithmetic operations, science facts taught in class), whereas PISA questions are often set in real-life scenarios and may require reasoning that goes beyond any single year's syllabus. For example, a TIMSS science item might ask about a concept like photosynthesis exactly as taught in the curriculum, while a PISA science item might present a hypothetical experiment or everyday problem (like interpreting a nutrition label or an environmental issue) requiring scientific reasoning.

4.2. Grade-Based Sampling (TIMSS) vs. Age-Based Sampling (PISA)

TIMSS tests students in specific grades (4 and 8), aiming to directly measure what students in those grades have learned in school. PISA tests a specific age group (15-year-olds), regardless of grade, to measure the cumulative outcome of education by that age. The practical consequence is that TIMSS and PISA target overlapping but not identical student populations. In many countries, 8th graders (TIMSS) are 13-14 years old, while PISA's 15-year-olds might mostly be 9th or 10th graders. Some countries participating in both studies have noted that their TIMSS and PISA samples differ – one might include slightly younger or older students than the other.

As a result, a nation's performance can differ between TIMSS and PISA partly because of this sampling difference alone. Research has shown that performance differences at the country level can be partly attributed to the sampling definitions – age-based vs. grade-based – of PISA and TIMSS. Neither approach is inherently superior; each addresses fairness in a different way (years of schooling vs. age cohort, as discussed in Section 6).

4.3. Subjects/Content (TIMSS) vs. Literacy Domains (PISA)

TIMSS is organized by traditional subject content areas (mathematics and science, each subdivided into topics). PISA is organized by broader literacy domains (mathematical literacy, scientific literacy, etc.). For instance, TIMSS mathematics might have separate scores or sub-scores for content areas like algebra or geometry at eighth grade. PISA's mathematical literacy, however, does not explicitly report by school-taught branches of math; instead, it might report proficiency levels or subscales related to processes or contexts.

In fact, the balance of content in the tests differs: TIMSS covers a wide range of curriculum topics in math and science, whereas PISA's test items in a domain may not cover all school topics evenly

 doi:10.20944/preprints202504.2356.v1

(for example, one analysis found PISA had relatively few algebra and geometry items compared to number and data, since PISA's focus is applied math, whereas TIMSS had a more balanced content coverage). These differences in content balance can advantage or disadvantage countries depending on their curriculum. A country with a strong algebra curriculum might excel in TIMSS algebra items, but if PISA has minimal algebra content, that strength won't translate as much to PISA scores.

4.4. Assessment Cycle and Scope

TIMSS assesses math and science at two education stages (primary and lower secondary) every four years. PISA assesses older students in multiple domains (including reading) every three years. PISA thus includes reading literacy as a major domain, which TIMSS does not address (TIMSS focuses only on math and science).

Additionally, PISA periodically includes optional innovative domains (such as collaborative problem solving or financial literacy) and places emphasis on trends in those literacies. TIMSS, on the other hand, has a companion study PIRLS for reading at 4th grade, but within TIMSS itself, reading is not tested. The inclusion of reading in PISA means that PISA offers a broader view of general education outcomes (reading, math, science) at age 15, whereas TIMSS provides a more detailed look at math and science achievement at specific grades.

4.5. Framework and Data Collection Differences

As noted, TIMSS includes an extensive curriculum questionnaire and collects detailed data about teaching practices related to the subjects tested [5]. PISA's background questionnaires often focus on socio-economic factors, learning environment, and attitudes (for example, students' enjoyment of reading or confidence in math). Each reflects the study's focus: TIMSS wants to correlate achievement with curriculum and instruction variables; PISA is interested in broader socioeconomic and attitudinal correlates of competence.

Another difference is in test design: both TIMSS and PISA use multiple test booklets with subsets of items given to different students (matrix sampling), but PISA's design, especially with a major domain, might have a longer, more integrated set of tasks for that domain (including open-ended tasks that TIMSS, being more traditional, might have fewer of).

4.6. Outcome Reporting

TIMSS reports scores typically on scales for math and science (separately for Grade 4 and 8) with an international scale average of 500 (often based on a baseline year) [11]. PISA reports scores for reading, math, and science literacy with the OECD average set around 500 in a baseline year (2000 for reading, 2003 for math, 2006 for science). While both use a similar scale metric (500 point mean, 100 point standard deviation convention), their interpretation differs.

For example, a 500 in TIMSS Grade 8 math and a 500 in PISA math literacy are not equivalent in content – one reflects mastery of Grade 8 math curriculum, the other reflects proficiency in applying math at age 15. Additionally, PISA emphasizes proficiency levels (Level 1 to 6) that describe what tasks students at various score ranges can do, framing results in terms of real-world skill levels. TIMSS reports sometimes highlight the percentage of students reaching international benchmarks (e.g., low, intermediate, high, and advanced benchmark), which are analogous to proficiency levels but tied to the curriculum content difficulty.

4.7. Summary

In light of these differences, it becomes clear that TIMSS and PISA serve complementary, not competing, roles. TIMSS is often seen as a measure of how effectively a country's curriculum is being learned by students, while PISA is seen as a measure of how well students can apply knowledge and skills in novel situations. It is not surprising, then, that some countries perform differently on the two assessments. For instance, analysts have observed that certain Western countries (e.g., some English-

 doi:10.20944/preprints202504.2356.v1

speaking or European nations) tend to rank relatively higher in PISA than in TIMSS, whereas some East Asian or Eastern European countries have exceptionally high TIMSS scores and also high (but sometimes comparatively slightly lower) PISA scores. These patterns might reflect differences in curriculum focus, instructional styles, or cultural factors affecting learning. For example, East Asian education systems often have curricula that emphasize procedural mastery in math and rigorous content coverage by Grade 8, yielding top-tier TIMSS performance; the same systems also do well on PISA, but the margin by which they lead others might be less, possibly because PISA's problem contexts can involve more reading or applied creativity. On the other hand, some countries with less emphasis on early formal academics but strong problem-solving culture might appear relatively stronger in PISA. All these nuances reinforce that one assessment is not "better" than the other – they are different lenses on student achievement.

5. RQ4: Measurement and IRT: How Scores Are Standardized

Both TIMSS and PISA rely on advanced psychometric techniques to measure student achievement in a fair and comparable way. The cornerstone of their scoring methodology is IRT, a family of statistical models that transform raw test results into scale scores that can be meaningfully compared across different sets of test questions and groups of students. In this section, we explain how IRT is used in these assessments and why it is superior to simple raw scores for large-scale comparisons.

5.1. Why Not Just Use Raw Scores?

In a traditional classroom test, a student's score might simply be the number or percentage of questions answered correctly. However, using raw scores to compare performance has limitations, especially in an international survey.

For one, not every student in TIMSS or PISA takes the exact same set of questions – both programs use a matrix sampling design where the entire pool of test items is distributed across several test booklets and each student completes only one booklet. This design enables broad content coverage without overburdening individual students, but it means no single student's raw score captures the whole assessment.

Moreover, even if all students took the same questions, raw scores are influenced by the difficulty of those specific items. A country's students might score higher in percentage terms on an easier test than on a harder test, even if their true ability remained the same. If we gave one group of students a set of very easy questions and another group very hard questions, a raw score comparison would be meaningless – the first group would obviously have more correct answers on average due to easier items, not necessarily higher ability. In large-scale assessments, where test forms and item difficulties can vary, using raw scores "as is" would yield unfair and potentially misleading comparisons. As noted in analysis of TIMSS, a raw-score based test is susceptible to manipulation by choosing easier or harder items, and differences in test difficulty could be misconstrued as differences in student ability.

5.2. Item Response Theory (IRT) Basics

Item Response Theory (IRT) is a family of statistical models used to analyze test data by considering the probability that a student answers a given question correctly as a function of two key factors: the student's latent ability and the item's properties, primarily its difficulty. Unlike traditional raw score methods, which simply count the number of correct responses, IRT acknowledges that different items vary in difficulty and diagnostic value.

In IRT, both students and test items are positioned on the same scale. A student with higher ability has a greater likelihood of correctly answering more difficult items. Importantly, IRT allows simultaneous estimation of student abilities and item difficulties based on observed response patterns across many students and items. This process ensures that ability estimates are not biased

 doi:10.20944/preprints202504.2356.v1

by which specific items were administered, making scores more comparable even when different students answer different sets of items.

Large-scale assessments like TIMSS and PISA use matrix sampling designs, where no student answers every test item. IRT supports this structure by linking items across booklets and calibrating them to a common scale. It also supports more complex models that include item discrimination (how well an item differentiates between students of different ability levels) and guessing behavior, depending on the assessment's design needs.

By applying IRT, TIMSS and PISA can report ability scores that are independent of the specific test forms used, facilitating fair comparisons across different groups, countries, and over time. The method ensures that observed differences in performance reflect real differences in proficiency rather than variations in test difficulty or composition. Thus, IRT provides the psychometric foundation that makes trend measurement and international comparison scientifically valid and reliable.

5.3. How IRT Enables Comparability

Once item difficulties are known, we can compare two students who took completely different questions. For example, if Student A answered many difficult items correctly and Student B answered only easy items correctly, IRT might assign Student A a higher ability estimate even if B's raw score was numerically higher. The model recognizes that Student B's correct answers did not demonstrate as high a proficiency because the questions were easier. In this way, scores become independent of the particular mix of easy or hard items a student saw [11].

IRT also provides a framework for linking scores across years. Both TIMSS and PISA include some common items from previous cycles or use other linking techniques to maintain the same scale over time. By anchoring item parameters across cycles, they ensure that a score of, say, 500 in 2019 means the same level of proficiency as a 500 in 2015.

For instance, TIMSS initially set its scale in 1995 such that the international average was 500 with a standard deviation of 100; subsequent rounds (1999, 2003, etc.) were linked to this metric so that scores remain comparable [11]. Concretely, after each new TIMSS assessment, data from the new test and the previous one (for countries that participated in both) are analyzed together to recalibrate item difficulties and then split apart on the same scale. This way, trends can be measured – an increase from 500 to 520 for a country between cycles suggests an improvement that is not due to an easier test but a real gain in ability as per the common scale.

PISA does similarly: it defines scale scores with an OECD mean of 500 (and standard deviation ~100) in the first cycle for each domain, and then keeps that scale for trend reporting. For example, PISA mathematics was scaled so that the OECD average in 2003 was 500; in later years, changes in a country's math score reflect changes relative to that baseline (accounting for the fact that the test content changes but is linked via common items and IRT calibration). Without IRT, establishing such comparability over time would be extremely hard – a rise or drop could simply result from a tougher or easier set of questions. With IRT, the difficulty of each item is quantified, and by extension, the difficulty of each assessment is controlled for when interpreting scores.

In summary, IRT provides a stable ruler for measurement. It allows TIMSS and PISA to report scores that are more than just percentages correct – instead, they are estimates of student proficiency on a latent scale. This is why both assessments report scores around a midpoint of 500 rather than raw percent correct. That 500 is not a percentage; it is a location on the proficiency scale. The use of IRT means that if one test is slightly more difficult than another, or if one student happened to get harder items, the scoring adjusts for that. It also means that scores can be compared meaningfully across countries and years, as they have been placed on a common scale through a rigorous modeling process [11]. IRT also produces measures of uncertainty (standard errors) for scores and enables the generation of plausible values (multiple imputed proficiency scores for each student) which are used to correctly aggregate data and run analyses without bias [11]. These advanced techniques underpin the reliability of reports that, for example, "Country X improved by 20 points since the last cycle" or "students in Country Y outperform the international mean by 50 points."

 doi:10.20944/preprints202504.2356.v1

5.4. Contrast with Raw Score Interpretation

To illustrate the benefit, consider a scenario without IRT: Say TIMSS 2019 and TIMSS 2023 had entirely different sets of math questions. If 2019's test was slightly easier overall than 2023's, a raw score average could go down in 2023 even if students knew more, simply because the test got harder. Thanks to IRT linking, we know the difficulty of 2019 items and 2023 items on one scale, so TIMSS can adjust and report scores such that 2019's 500 and 2023's 500 are equivalent in proficiency terms. As the TIMSS documentation explains, after IRT calibration, "TIMSS 2019's 600 points is the same as TIMSS 2023's 600 points, because item difficulty parameters are the same." [11] This property is invaluable for trend analysis.

In PISA's case, IRT also allows fair comparison between countries despite each country administering the test in its own language. While language differences can affect item difficulty (e.g., a question might be slightly harder to comprehend in one language due to translation nuances), the IRT scaling can help adjust for that by examining how items function in different countries. Items that show differential difficulty by country can be investigated, and overall country scores are derived from the pattern of performance across many items, reducing the influence of any single item's quirk.

In conclusion, the use of IRT is a key aspect of measurement standardization in TIMSS and PISA. It represents a modern approach to testing that moves beyond raw scores to a more sophisticated, model-based estimate of ability. This approach ensures that the comparisons drawn from these studies – whether comparisons over time, between countries, or between subgroups – rest on a fair and scientifically-grounded basis. It is one of the reasons we can trust that when a report says "students in Country A scored higher than in Country B," it is not merely because Country A's students answered more questions (they might have taken harder questions, in fact), but because on balance they demonstrated higher underlying proficiency.

6. RQ5: International Comparability Challenges

Even with careful test design and IRT-based scaling, comparing educational performance across dozens of countries is a complex task. TIMSS and PISA each face inherent challenges in achieving perfect apples-to-apples comparisons. It is important for consumers of these results – especially media and policymakers – to be aware of these challenges, which include differences in student age vs. grade, variations in curricula, and cultural/contextual factors that can influence performance. Recognizing these issues helps prevent misinterpretation of the data.

6.1. Age–Grade Mismatch (PISA's Age-Based vs TIMSS's Grade-Based Populations)

One comparability issue arises from the different target populations. Because PISA tests 15-year-olds, and TIMSS tests a particular grade, the cohorts are not aligned by schooling. In some education systems, most 15-year-olds are in Grade 10; in others, many are in Grade 9. Some 15-year-olds may even be in vocational tracks or have left school (though PISA tries to test those still enrolled up to grade 7 minimum). Meanwhile, TIMSS's grade-based approach means a fixed schooling year, but with different ages.

Why does this matter? The number of years of formal schooling a student has completed by the time of the test can affect performance. A 15-year-old in Grade 10 has had one more year of instruction than a 15-year-old in Grade 9. Conversely, a 8th grader who is 15 (maybe due to late school start or repeating a year) might perform differently than an 8th grader who is 13. A detailed OECD analysis noted that neither purely age-based nor purely grade-based sampling provides a perfectly uniform basis for comparison – they are just different compromises. Proponents of PISA's age-based approach argue that aligning by age is fairer to compare outcomes of schooling as of a certain life stage, because countries vary in the typical age of finishing compulsory education. On the other hand, proponents of TIMSS's grade-based approach argue that aligning by years of schooling (grade) is more fair

 doi:10.20944/preprints202504.2356.v1

educationally, since it compares students with similar instructional exposure and allows us to compare curricula and school factors more directly.

Studies have shown that these definitions can indeed shift results. For example, if PISA were grade-based (say testing all 9th graders) some countries might score higher or lower than in the age-based design. Conversely, if TIMSS were age-based (testing all 14-year-olds for an "8th grade" comparison), rank orders might change. One empirical finding is that in countries where students start school later or have higher rates of grade repetition, an age-based sample (PISA) tends to include more students with one year less of schooling, which can lower the average performance relative to a grade-based sample. For instance, consider a country where children start school at age 7; by age 15, they've had 8 years of schooling (just finished Grade 8). In PISA they'll be compared with age-mates from a country where kids started at 5 and are in Grade 10 (10 years of schooling by age 15). The latter have two more years of education, which likely gives them an advantage in PISA's tested skills. TIMSS would compare Grade 8 in both countries – in that case, the country that starts later would be testing 15-year-olds against the other's 13-year-olds, which could advantage the former (older students might score higher due to maturity, even with the same schooling years). Clearly, both scenarios have trade-offs.

Research by Wu [2] quantified that one additional year of schooling can increase average performance by a significant margin – on the order of 20 to 40 score points in these assessments. This suggests we must be cautious: some of the gap between certain countries in PISA might simply be because in one country the tested 15-year-olds have had a year more of schooling. PISA somewhat mitigates this by its design (since it only includes students who are at least in 7th grade, avoiding completely unschooled 15-year-olds), but differences remain. In TIMSS, differences in age can also play a role: a country where 4th graders are mostly 10 years old might do better than one where 4th graders are 9, just due to additional cognitive development with age. Analysts have debated which sampling method yields more "comparable" results. In practice, both PISA and TIMSS results are usually analyzed in the context of the respective target definitions, and it's understood that they are somewhat different comparisons. The key takeaway is that when interpreting scores, one should account for the educational exposure of the students tested. If a country performs below another in PISA, but one realizes that its students had significantly less schooling by that age, that factor should be considered before jumping to conclusions about instructional quality.

6.2. Curriculum Coverage and Content Alignment

Another comparability issue is the alignment (or lack thereof) of test content with what students have actually been taught. This is primarily a concern for TIMSS, given its curriculum-based nature, but it also affects PISA in a different way. In TIMSS, test developers strive to include content that is common to many countries' curricula. However, not all countries introduce topics at the same grade. For example, if probability (part of "Data and Chance" domain) is taught by Grade 8 in Country A but only introduced in Grade 9 in Country B, then B's 8th graders will likely struggle on those TIMSS items simply because they haven't learned that material yet. This would depress Country B's score in that content area, reflecting a curriculum opportunity-to-learn gap rather than an innate ability gap. TIMSS data analysis often includes looking at curriculum questionnaires to see if low performance correlates with topics that were not covered in the national curriculum by that grade. Thus, differences in national curricula can impair direct comparability on TIMSS – one must check to what extent the test content was taught. IEA tries to minimize this problem by getting input from countries during test development (to avoid including an item that is completely foreign to a majority of participants), but some variation is inevitable. In fact, TIMSS explicitly reports on the intended curriculum vs. implemented curriculum [5], acknowledging that part of the story behind scores is whether students had the chance to learn the material.

PISA, by contrast, deliberately does not align to curricula – which ironically brings a different comparability challenge: the relevance of tasks across cultures. PISA's real-world scenarios might be more familiar to students in some settings than others. For instance, a question involving reading a

 doi:10.20944/preprints202504.2356.v1

train timetable might be easier for students who use public transit frequently. While not a curriculum issue per se, it's a contextual factor that could bias results. Moreover, even though PISA doesn't follow school curricula, it cannot escape the fact that what students have learned in school will affect their ability to answer PISA questions. If a country's curriculum heavily emphasizes advanced algebra and geometry and less on data interpretation, their students might find PISA's data-oriented questions relatively easy and the few algebra questions trivial, resulting in high performance. If another country's schools focus more on pure theory and less on real-world applications, students might be less practiced in the style of problem PISA presents, possibly lowering performance despite having learned the content in abstract terms. A study comparing PISA and TIMSS content found exactly this: TIMSS covers a wider range of traditional curriculum content, whereas PISA's content balance is different – with fewer algebra/geometry items and more number/data items – which means countries strong in algebra (often an area of strength for some Asian countries) might not get as much benefit on PISA. Conversely, countries that focus on data literacy might shine more on PISA. Indeed, analyses concluded that differences in content balance between PISA and TIMSS can lead to differences in country rankings. This is a comparability issue when people naively compare PISA and TIMSS outcomes – it may seem like one test is "harder" or a country is "worse," but in reality the tests emphasize different content.

6.3. Language and Cultural Context

Both assessments must be translated into many languages and administered in varied cultural contexts. Every effort is made to maintain equivalence, but certain items might still favor some cultures. For example, word problems requiring a lot of reading could disadvantage students who are less proficient in the test's language of instruction (especially in PISA, which often has lengthy scenarios to read). It's documented that reading load can affect performance – in PISA, items often have more text to read than in TIMSS, so students who are weaker readers may struggle, which could disproportionately affect some countries. In fact, one analysis showed that differences in national reading proficiency could partly explain differences between PISA and TIMSS math results (countries with strong reading literacy tended to do relatively better on PISA math, presumably because they could navigate the word problems more easily).

Cultural attitudes towards tests can also differ. TIMSS and PISA are low-stakes for students (no individual consequences), but in some places students might still be motivated to try hard (perhaps due to national pride or encouragement) while in others they might be less engaged. These subtle factors can introduce noise in the comparisons. Additionally, curricula may differ not just in content but in skills emphasis – some systems train students in test-taking and problem-solving more than others.

6.4. Socio-Economic and School Structure Factors

Countries vary in terms of which students are in school at the target age/grade. For example, dropout rates by age 15 differ – PISA tries to test a representative sample of 15-year-olds, but in countries where many 15-year-olds have left school or are in non-academic tracks not covered by the assessment, there's a selection effect. Similarly for TIMSS Grade 8: in some countries nearly all 14-year-olds are in school at grade 8, whereas in others there might be attrition or tracking that changes the pool. These factors can influence average scores and complicate direct comparison. OECD and IEA usually publish information on enrollment rates and exclusion rates (students with disabilities or language issues who weren't tested) to contextualize this.

In summary, while TIMSS and PISA make enormous efforts to standardize administration and scoring, one must remember that each country's context is unique. Ages, grades, curricula, languages, and student backgrounds differ. Therefore, international rankings should not be taken at face value without context. A country's position in TIMSS or PISA is a starting point for questions, not a final judgment. For instance, if Country X scores below Country Y, it should prompt investigation: Does Country X have a different curriculum sequence? Did its students have fewer years of schooling by

 doi:10.20944/preprints202504.2356.v1

age 15? Are there cultural factors at play? Responsible use of the data involves digging into such questions rather than assuming one system is categorically "better" based solely on scores.

7. RQ6: "Not a Competition": Interpreting and Using Results Wisely

Both TIMSS and PISA were conceived as research tools to improve education, not as contests. The organizations behind these studies emphasize that the assessments are designed to allow countries to learn from each other and to monitor internal progress. Unfortunately, the narrative of competition often overtakes the discourse once results are released. It is common to see news reports focusing on which country ranked first or how a nation "slipped in the rankings," sometimes with alarmist overtones. Policymakers, under public pressure, may feel compelled to react swiftly to such rankings. In this section, we stress the importance of interpreting TIMSS and PISA results in the right spirit and caution against common misuses.

7.1. Educational Improvement, Not Medals

The proper goal of TIMSS and PISA is to identify strengths and weaknesses in students' knowledge and skills so that informed decisions can be made to enhance teaching and learning. For example, if TIMSS reveals that students in a country struggle with a particular science topic, that insight can spur curriculum developers to address that gap. If PISA shows that students lack higher-order problem-solving skills, perhaps teaching methods can be adjusted. In short, these studies are tools for diagnosis and inspiration for reform. Many countries have indeed used international assessment results to drive reforms (e.g., adding more emphasis on reasoning skills after PISA, or focusing on foundational knowledge after TIMSS) [12]. The international aspect allows for benchmarking against a variety of school systems: a country can observe what top performers are doing and consider whether those practices are adaptable. The intent is collaborative improvement – very much in line with the idea of a global laboratory of education policies.

7.2. Media and Misinterpretation

Despite the intentions, media portrayals often reduce results to a simplistic ranking of winners and losers. Headlines might proclaim "Country X falls to 20th place in math" or "Country Y leaps ahead." Such framing can be misleading. First, as discussed, differences in rank may not be statistically significant or educationally meaningful. A country ranked 10th and another ranked 18th might have only a few points difference, well within the margin of error. Treating those ranks as meaningful ordering can be an overinterpretation. Second, the league table mentality ignores the confidence intervals and overlapping performance ranges. It also obscures the fact that these tests cover limited domains – a country's lower rank in PISA science, for instance, doesn't mean its entire education system is inferior; it might be doing well in aspects that aren't captured by that metric (arts, civic education, etc., or even aspects of science not well measured by PISA's approach).

Some researchers have criticized the almost sports-like fervor that surrounds PISA releases. Yong Zhao, for example, argues that PISA has "successfully created an illusion of education quality" that seduces countries into viewing its scores as the ultimate measure of educational success [13]. Zhao points out that nations participate "out of the belief that this triennial test accurately measures the quality of their education systems, the effectiveness of their teachers, the ability of their students, and the future prosperity of their society" [13]. This is a sweeping set of conclusions to draw from test scores, and Zhao warns that it's an illusion – in reality, test scores are one indicator, not a definitive oracle of future economic or societal outcomes. In a similar vein, Carnoy et al. [1] note that international test rankings have come to dominate public perception of educational quality, but that these data are "notoriously limited" in diagnosing the causes of differences [1]. They caution that policy prescriptions drawn directly from simplistic interpretations of rankings are often misguided [1]. In the U.S., for instance, mediocre PISA rankings have been cited as evidence of an education crisis threatening economic competitiveness, leading to calls for drastic reforms. Carnoy and

 doi:10.20944/preprints202504.2356.v1

colleagues challenge that narrative, indicating it's an oversimplification to tie national economic fate to PISA results without deeper analysis of context [1].

7.3. The Danger of Overreaction

When a country treats TIMSS or PISA as a competition, there is a risk of knee-jerk policy changes that may not truly address underlying issues. For example, if a country drops in rank, leaders might push for adopting whatever curriculum the top-ranked country uses, or increase testing and drills in hopes of boosting scores. These responses might overlook local context and other consequences. Moreover, chasing a higher ranking might lead to narrowing the curriculum (teaching only what is tested) or teaching to the test, which can undermine real learning. It is important to remember that improving test scores is not the ultimate goal – improving education is. A rise in scores that comes from superficial test-prep measures does not necessarily mean students are better prepared for life and work; it might just mean they are better prepared for that particular test format.

7.4. Responsible Use of Data

A more productive approach is to use TIMSS and PISA results as a starting point for inquiry. For instance, if a nation's mathematics score is stagnant, delve into the data: Are there particular content areas or question types where students struggle? What do the questionnaires suggest (perhaps students have low confidence or certain ineffective instructional practices are common)? How does this compare to countries with similar challenges? By analyzing the rich data beyond the headline score – including gender differences, regional differences, trends over time, etc. – policymakers can make targeted decisions. International assessments also encourage setting realistic benchmarks. Rather than fixating on being "number 1," a country might focus on achieving a certain absolute score target or closing a gap. For example, if analysis shows that eighth graders in Country A are on average a year behind those in Country B in a subject, the country can aim to close that gap by specific interventions, and later verify through another TIMSS cycle if progress occurred.

It's also crucial to communicate results to the public in a nuanced way. Educators and officials have the responsibility to explain what the scores mean – and what they do not mean. Emphasizing the confidence intervals, the idea of statistical ties, and the specific domains tested can help temper the tendency to overinterpret small ranking differences. Some countries even avoid public "rankings" in press releases and instead highlight their results in terms of achievement bands or long-term trends. This can shift the conversation from "Did we beat country X?" to "Did we improve and are we educating our children better than before?".

7.5. Learning from Each Other, Not Blaming

Perhaps the most positive use of TIMSS and PISA is as platforms for international learning. The fact that these assessments include background information means one can investigate, for example, how the highest-performing countries teach math – do they emphasize basics or problem-solving? How much homework do their students get? What are their teacher qualification levels? While this doesn't prove causation, it provides hypotheses: if top performers all share a certain practice, it might be worth considering. Conversely, if a low-performing country has a practice not seen elsewhere, it might reconsider that practice. The key is not to shame countries with lower scores, but to help them find actionable insights. In many cases, countries have formed partnerships or consulted international experts based on assessment outcomes, fostering a collaborative rather than competitive spirit.

In conclusion of this section, TIMSS and PISA should be seen not as high-stakes scoreboards, but as diagnostic tools and guides. The excitement of seeing "who is on top" should be replaced by an emphasis on understanding why certain results occurred and how to apply lessons learned. Both studies are large scientific endeavors in comparative education research. Treating them like a simple contest misses their rich value and can lead to counterproductive responses. Ultimately, the real

 doi:10.20944/preprints202504.2356.v1

victory is not scoring highest, but improving one's own education system to better serve students – and in that endeavor, TIMSS and PISA are invaluable resources.

8. Conclusions

TIMSS and PISA are two pillars of international educational assessment, each with distinct goals and designs, yet both contributing to our understanding of student learning around the world. This paper has highlighted that TIMSS is a curriculum-based, grade-centered assessment focusing on what students learn in school (and how they learn it), while PISA is a competency-based, age-centered assessment focusing on how students can apply their knowledge in real-world contexts. These differences are purposeful, stemming from different questions: TIMSS asks, "How well are students learning the school curriculum in math and science?" and seeks to help improve that curriculum, whereas PISA asks, "How well can students use their learning to solve problems as they near adulthood?" and seeks to inform broader policy and lifelong learning goals.

Despite often being mentioned in the same breath, TIMSS and PISA should not be treated as rival measures of national achievement in a global competition. Their results are best interpreted within their own frameworks and purposes. Both assessments leverage Item Response Theory (IRT) to ensure that their scoring is fair and comparable – a sophisticated approach that allows these studies to report trends and differences with greater validity than simple test scores would. Through IRT scaling, each assessment creates a stable metric (with an average around 500) that enables meaningful comparisons over time and across diverse populations [11]. This technical backbone is part of what makes TIMSS and PISA influential; it gives confidence that changes in scores reflect real changes in student proficiency, not artifacts of test difficulty or sample variation.

We also discussed how international comparability is a nuanced issue. Factors like whether students are compared by age or grade, differences in curricula, language and cultural context, and varying student backgrounds all color the interpretation of results. A savvy interpretation requires looking beyond the rank or score – understanding who was tested and what was tested is crucial. For example, a gap in PISA might signal a gap in years of schooling or emphasis on certain skills, rather than an absolute shortfall in ability. Awareness of these factors guards against the misperception that a single number can fully encapsulate an education system's quality.

Finally, we underscored that TIMSS and PISA are tools for improvement, not trophies. The true value of these assessments lies in the wealth of data and analysis they offer to educators and policymakers. When used properly, they can guide evidence-based reforms: aligning curriculum with proven effective practices, addressing weaknesses in problem-solving or content knowledge, and learning from other countries' successes and failures. However, when their results are reduced to simplistic rankings and sensational headlines, the nuance is lost and there is a risk of misusing the data – from unfounded policy panic to neglecting important but unmeasured aspects of education.

As we move forward, the conversation around international assessments is gradually evolving. There is increasing recognition in the educational research community that collaboration and context matter more than competition. Initiatives are underway to ensure results are reported with appropriate caveats and to help stakeholders focus on trends and deeper insights rather than just league tables. Both IEA (for TIMSS) and OECD (for PISA) have been providing more tools (such as interactive data portals and detailed framework documents) to aid in proper interpretation.

In conclusion, TIMSS and PISA, each in their way, shine a light on student learning across the globe. They are complements rather than competitors: TIMSS informs us about success in delivering curriculum at key grade levels, and PISA informs us about the broader competences students have developed by age 15. By understanding their differences and the sophisticated measurement methods they employ, one can better appreciate the insights they offer. Stakeholders should embrace the rich information these studies provide while resisting the urge to oversimplify the story into who came in first. Ultimately, the goal is to use TIMSS and PISA results to support all countries in providing better education – so that every student, no matter where they are, can benefit from the collective knowledge gained through these large-scale assessments.

 doi:10.20944/preprints202504.2356.v1

References

1. M. Carnoy, E. Garcia, and T. Khavenson, 'Bringing it back home: Why state comparisons are more useful than international comparisons for improving US education policy', 2015.
2. M. Wu, 'Comparing the similarities and differences of PISA 2003 and TIMSS', *OECD*, 2010.
3. International Association for the Evaluation of Educational Achievement (IEA), 'TIMSS - Trends in International Mathematics and Science Study'. 2024. [Online]. Available: https://www.iea.nl/studies/iea/timss
4. IEA, 'International Association for the Evaluation of Educational Achievement (IEA)'. [Online]. Available: https://www.iea.nl/
5. I. V. Mullis and M. O. Martin, *TIMSS 2019 Assessment Frameworks*. ERIC, 2017.
6. M. O. Martin, M. Von Davier, and I. V. Mullis, 'Methods and procedures: TIMSS 2019 Technical Report', *International Association for the Evaluation of Educational Achievement*, 2020.
7. D. Hutchison and I. Schagen, 'Comparisons between PISA and TIMSS: Are we the man with two watches', *Lessons learned: What international assessments tell us about math achievement*, pp. 227–261, 2007.
8. National Center for Education Statistics (NCES), 'The TIMSS Mathematics and Science Assessments – Content and Cognitive Domains'. 2017. [Online]. Available: https://nces.ed.gov/timss/timss15_assessments.asp
9. Organisation for Economic Co-operation and Development (OECD), 'Programme for International Student Assessment (PISA)'. [Online]. Available: https://www.oecd.org/en/about/programmes/pisa.html
10. OECD, 'Organisation for Economic Co-operation and Development (OECD)'. [Online]. Available: https://www.oecd.org/
11. National Center for Education Statistics (NCES), 'Weighting, Scaling, and Plausible Values (TIMSS 2015 Technical Notes)'. 2017. [Online]. Available: https://nces.ed.gov/timss/timss15technotes_weighting.asp
12. D. M. Kadijevich, M. Stephens, A. Solares-Rojas, and R. Guberman, 'Impacts of TIMSS and PISA on mathematics curriculum reforms', in *Mathematics Curriculum Reforms Around the World: The 24th ICMI Study*, Springer International Publishing Cham, 2023, pp. 359–374.
13. Y. Zhao, 'Yong Zhao: The PISA Illusion', *Assessment*, 2019.