CHAPTER 14

Using Scale Anchoring to Interpret TIMSS Achievement Results

Lale Khorramdel
Charlotte E. A. Aldrich
Allison Bookbinder

Introduction

TIMSS achievement results are summarized using extensions of item response theory (IRT) and reported on the TIMSS achievement mathematics and science scales (see Chapter 12). While average achievement scores for most countries cluster between 400 and 575, the TIMSS mathematics and science scales span a wide range of possible scores. Country-level average scores provide information about how fourth-grade and eighth-grade students' mathematics and science achievement compare across countries and whether achievement, on average, is improving or declining over time. However, comparing average achievement internationally is only one part of the picture. The range of mathematics and science achievement within countries, paired with information about what students at different scale levels know and can do, provide a much more nuanced picture of student ability to policymakers compared to a single average score per country.

To provide this information, it is important to understand the mathematics and science competencies associated with different regions along the TIMSS achievement scales. For example, what does it mean for a country to have an average mathematics or science achievement of 513 or 426? Are there groups of students with much higher achievement, and how are these students characterized? What are students expected to be able to do if their achievement is around 475? What further competencies do students have with achievement at around 550?

This chapter describes an approach of central importance to characterize the progression and variability of student achievement within and across countries: creating descriptions of the TIMSS International Benchmarks of Mathematics and Science Achievement. These benchmarks help contextualize TIMSS results by providing information about what students know and can do at different points along the TIMSS achievement scales. More specifically, TIMSS has identified four points along the TIMSS mathematics and science achievement scales to use as International Benchmarks of achievement: Advanced International Benchmark (625),

https://doi.org/10.6017/lse.tpisc.timss.rs7604

High International Benchmark (550), Intermediate International Benchmark (475), and Low International Benchmark (400). The cut points for these benchmarks were originally selected in TIMSS 2003 and have remained constant since that time (Gonzalez et al., 2004). However, the descriptions of the International Benchmarks, specific to each grade level (fourth and eighth grade) and subject domain (mathematics and science), are updated for each cycle as trend item blocks and new assessment blocks define what is measured over time.

The TIMSS & PIRLS International Study Center works with the TIMSS Science and Mathematics Item Review Committee (SMIRC) to conduct a scale anchoring exercise to describe student competencies at each of the benchmarks based on the international results of the current cycle. Scale anchoring is conducted separately for each grade and subject. An important feature of the scale anchoring method is that it yields descriptions of the mathematics and science competencies demonstrated by students reaching each of the International Benchmarks on the TIMSS scale, reflecting the content and cognitive areas of the TIMSS mathematics and science assessment frameworks. The following sections describe the TIMSS approach for classifying items and writing the benchmark descriptions. Then, results are presented for each step in updating the TIMSS 2019 benchmark descriptions for TIMSS 2023.

Classifying the Items

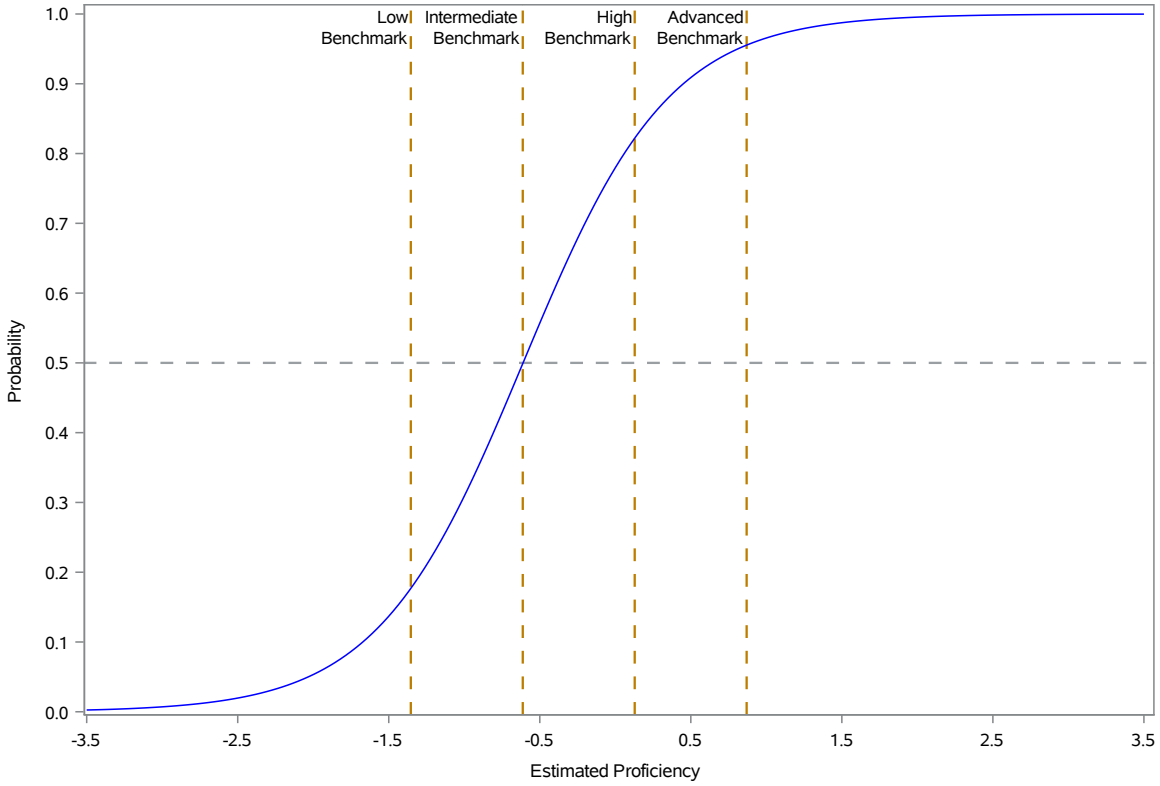
The first stage in the TIMSS scale anchoring process involves identifying items that anchor at each International Benchmark based on the IRT parameters estimated as part of achievement scaling for the current cycle. As described in Chapter 11, the probability of a correct response to each item can be determined for a respondent with a certain ability, given the item's characteristics. In IRT, these item-specific characteristics are referred to as item parameters. The IRT model provides an item-level probability model—an item function—describing how the probability of a correct response relates to the student ability and the item parameters. For each item, the probability of a correct response is computed at each International Benchmark level, given the item's parameters. The item parameters that serve as the basis for these probabilities were estimated using all available data from the achievement scaling and give theoretical probabilities based on the international distribution of student proficiency.

As a first step, ability values in the IRT logit metric are calculated for each International Benchmark. Using the ability values on the IRT logit metric enables analysis on the same metric as the item parameters. This is accomplished using the linear transformation of proficiency scores for trend measurement described in Chapter 11 and provided in Chapter 12 for TIMSS 2023.

The second step involves computing the item-level probabilities of a correct response at each benchmark. For multiple-choice items and constructed-response items worth one point, it is straightforward to compute the probabilities using the two- and three-parameter IRT models (see Chapter 11). For constructed-response items scored for partial and full credit, up to two points, probabilities are computed using the generalized partial credit model (GPCM).

To illustrate the procedure, an example item characteristic function for a 1-point item is shown in Exhibit 14.1. The vertical reference lines indicate the location of each International Benchmark in the logit metric, and the values on the vertical axis correspond to the model-based probability of answering the item correctly at each benchmark.

Exhibit 14.1: Example Item Characteristic Function for Scale Anchoring Analysis



As can be seen, the probability of a correct response increases from the lower to the higher International Benchmarks. Students with achievement that meets or exceeds the Advanced Benchmark have the highest probability, while students at lower benchmarks are less likely to give a correct response.

The third step includes applying criteria that consider performance at adjacent benchmarks, described below, to identify items that students at each International Benchmark are likely to answer correctly and that discriminate between one benchmark and the next. These criteria help ensure that the performance descriptions at each benchmark reflect demonstrably different accomplishments by students reaching each successively higher benchmark.

For multiple-choice items, a probability of 0.65 (i.e., 65% expected correct answers) is used as the criterion for anchoring at each benchmark because students would likely (about two-thirds of the time) answer the item correctly, given the possibility of guessing the correct answer. In addition, a criterion requiring less than 50% expected correct answers on the next lower benchmark is applied such that there is a clear distinction that students at lower benchmarks

are more likely to have answered the item incorrectly than correctly. A somewhat less strict criterion was used for the constructed-response items because students had much less scope for guessing. For constructed-response items, a probability of 0.50 of answering correctly (50% expected correct answers) was used without any discrimination criterion for the next lower benchmark. It should be noted that the different score points of a two-point item can anchor at different benchmarks, typically at a higher benchmark for full credit (2 of 2 points) and at a lower benchmark for partial credit (1 of 2 points). Still, sometimes both can anchor at the same benchmark.

To illustrate the described criteria, consider a multiple-choice item as an example. The criteria for each benchmark are as follows:

- Multiple-choice items anchor at the Low International Benchmark (400) if students at the ability level corresponding to 400 scale score points have at least 65% expected correct answers. Because this is the lowest benchmark described, there were no further criteria.
- Multiple-choice items anchor at the Intermediate International Benchmark (475) if students at the ability level corresponding to 475 scale score points have at least 65% expected correct answers and if the students at the Low International Benchmark have less than 50% expected correct answers.
- Multiple-choice items anchor at the High International Benchmark (550) if students at the ability level corresponding to 550 scale score points have at least 65% expected correct answers and if the students at the Intermediate Benchmark have less than 50% expected correct answers.
- Multiple-choice items anchor at the Advanced International Benchmark (625) if students at the ability level corresponding to 625 scale score points have at least 65% expected correct answers and if the students at the High International Benchmark have less than 50% expected correct answers.

The classification of items that “almost anchor” is used for multiple-choice items that meet slightly less stringent criteria for the IRT-based percent correct estimates. The criteria to identify multiple-choice items that almost anchor are between 60% and 65% expected correct answers and less than 50% correct at the next lowest benchmark. To be completely inclusive for all items, items that met only the 60–65% criterion (regardless of the probability at the next lower benchmark) are also identified. The categories of items are mutually exclusive and ensure that all the items are available to inform the descriptions of student achievement at the anchor levels. Using the additional less-stringent criteria to maximize the inclusion of multiple-choice items in the anchoring process helps provide information about content domains and cognitive processes that might not otherwise be represented at the anchor levels.

A multiple-choice item is considered “too difficult” to anchor if students at the Advanced Benchmark have less than 60% expected correct responses. A constructed-response item is considered “too difficult” to anchor with less than 50% of students expected to answer correctly at the Advanced Benchmark.

Writing the International Benchmark Descriptions

Based on the analysis and resulting classification, the items are organized according to the benchmark they anchor and presented to the experts of the SMIRC for review. Each item is summarized to describe the kind of knowledge, skill, or cognitive process demonstrated by students who responded correctly to the item. The item descriptions from the previous cycle are used for trend items carried over from the previous cycle. The TIMSS & PIRLS International Study Center drafts the item descriptions for items newly developed for the current cycle in preparation for the review. The SMIRC members review the item descriptions to confirm that each description captures the content knowledge and cognitive process needed to respond correctly to the item and propose any necessary revisions. The confirmed item descriptions are then used to update the TIMSS International Benchmark descriptions for mathematics and science.

Each TIMSS cycle uses the International Benchmark descriptions written from the previous cycle as a basis for the exercise. Using the item descriptions, the experts synthesize across the items to draft the competencies displayed at each benchmark and to create distinct descriptions of student achievement at each International Benchmark. The summaries describe students' achievement separately for mathematics and science at each grade and provide a short summary across the content domains, yielding a content-referenced interpretation of the achievement results.

TIMSS 2023 Scale Anchoring

The TIMSS 2023 scale anchoring analysis was conducted using the item parameters estimated as part of the TIMSS 2023 achievement scaling implementation (see Chapter 12). Exhibit 14.2 presents the number of TIMSS 2023 items anchoring at each International Benchmark.

Exhibit 14.2: Number of Items Anchoring and Almost Anchoring at Each TIMSS 2023 International Benchmark

Grade / Subject	Low (400)	Intermediate (475)	High (550)	Advanced (625)	Above Advanced	Total Items
Grade 4 Mathematics	20	37	65	62	8	192
Grade 4 Science	18	48	59	40	20	185
Grade 8 Mathematics	10	35	75	70	17	207
Grade 8 Science	23	47	65	58	37	230

The TIMSS 2023 scale anchoring results were compared to the TIMSS 2019 scale anchoring results to ensure there was no unusual change in the levels at which those items administered in both assessment cycles were benchmarked, such as changing from "Low" in one cycle to "Advanced" in the next. Across all trend items on average, there was 81% agreement between the TIMSS 2019 and TIMSS 2023 anchor-level classifications for fourth-grade mathematics, 81% agreement for fourth-grade science, 78% agreement for eighth-grade mathematics, and 76% agreement for eighth-grade science. These results confirmed that the benchmark descriptions from TIMSS 2019 could be used as a starting point to write the TIMSS 2023 benchmark descriptions.

Scale anchoring was conducted with the TIMSS 2023 SMIRC in June 2024.[1] In preparation for the SMIRC meeting, staff at the TIMSS & PIRLS International Study Center created detailed documentation for each item that included a draft item description, the framework classification, and the answer key or scoring guide, along with the scale anchoring analysis results and international average percent correct.

The item-by-item documentation was grouped by International Benchmark, then by content domain, by topic area, and finally by anchoring criteria: items that anchored, followed by items that almost anchored, and then by items that met only the 60–65% criterion. Item descriptions provided a short summary of the student competencies demonstrated by a correct (or partially correct) response to each item.

At the scale anchoring meeting, the SMIRC members performed three major tasks: 1) reviewed each item to finalize the description of the student competencies demonstrated by a correct (or partially correct) response; 2) summarized the proficiency demonstrated by students reaching each International Benchmark for publication in the TIMSS 2023 International Report; and 3) selected example items for publication in the TIMSS 2023 International Report that illustrated the types of items answered correctly by students at each of the four benchmarks.

Item descriptions used as the basis for the full International Benchmark descriptions can be found in Appendices 14A, 14B, 14C, and 14D, for each grade-subject combination. Summaries of student performance at the International Benchmarks in mathematics and science at the fourth and eighth grade with example items from the TIMSS 2023 assessment are presented in the achievement results section of the *TIMSS 2023 International Results in Mathematics and Science* report.

Reference

Gonzalez, E. J., Galia, J., Arora, A., Erberber, E., & Diaconu, D. (2004). Reporting student achievement in mathematics and science. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 Technical Report* (pp. 274–207). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
https://timss.bc.edu/timss2003i/technicalD.html

1 The role of the SMIRC and a list of TIMSS 2023 SMIRC members are detailed in Chapter 1.

- Appendix 14A: TIMSS 2023 Scale Anchoring Item Descriptions – Grade 4 Mathematics
- Appendix 14B: TIMSS 2023 Scale Anchoring Item Descriptions – Grade 4 Science
- Appendix 14C: TIMSS 2023 Scale Anchoring Item Descriptions – Grade 8 Mathematics
- Appendix 14D: TIMSS 2023 Scale Anchoring Item Descriptions – Grade 8 Science