Using Scale Anchoring to Interpret the TIMSS and PIRLS 2011 Achievement Scales

Ina Mullis

Introduction

As described in **Scaling the TIMSS and PIRLS 2011 Achievement Data**, the TIMSS and PIRLS achievement results are summarized using item response theory (IRT) scaling and reported on 0 to 1,000 achievement scales, with most achievement scores ranging from 300 to 700. Although the focus typically is on countries' average scores, the country-by-country distributions of achievement scores provide users of the data with information about how achievement compares among countries and whether scores are improving or declining over time.

To provide important information for policy and curriculum reform, however, it is important to understand the mathematics, science, and reading competencies associated with the range of scores on the achievement scales. For example, in terms of levels of student understanding, what does it mean for a country to have average achievement of 513 or 426, and how are these scores different?

The TIMSS and PIRLS International Benchmarks provide information about what students know and can do at different points along the achievement scales. More specifically, TIMSS and PIRLS have identified four points along the achievement scales to use as international benchmarks of achievement—Advanced International Benchmark (625), High International Benchmark (550), Intermediate International Benchmark (475), and Low International Benchmark (400). With each successive assessment, TIMSS and PIRLS work with the expert international committees [Science and Mathematics Item Review Committee (SMIRC) for TIMSS and the Reading Development Group (RDG) for PIRLS] to conduct a scale anchoring analysis to describe student competencies at the benchmarks.

This chapter describes the scale anchoring procedures that were applied to describe student performance at the international benchmarks for PIRLS 2011

and for TIMSS 2011.The analysis was conducted separately for PIRLS reading at the fourth grade and for mathematics and for science at the fourth and eighth grades. In brief, scale anchoring involved identifying items that students scoring at the international benchmarks answered correctly, and then having experts examine the content of each item to determine the kind of knowledge, skill, or reasoning demonstrated by students who responded correctly to the item. The experts then summarized the detailed list of item competencies in a brief description of achievement at each international benchmark. Thus, the scale anchoring procedure yielded a content-referenced interpretation of the achievement results that can be considered in light of the TIMSS and PIRLS frameworks for assessing mathematics, science, and reading.

Classifying the Items

As the first step, students scoring within 5 scale-score points of each benchmark (i.e., the benchmark plus or minus 5) were identified for the benchmark analysis. The range of 10 points provided an adequate sample of students scoring at the benchmark, yet was small enough so that performance at one international benchmark was still distinguishable from the next. The score ranges around each international benchmark and the number of students scoring in each range are shown in Exhibit 1.

Exhibit 1: Range Around Each International Benchmark and Number of Students Within Each Range

	Low (400)	Intermediate (475)	High (550)	Advanced (625)
Range of Scale Scores	*395-405*	*470-480*	*545-555*	*620-630*
TIMSS Grade 4 Mathematics	5179	9077	9921	4520
TIMSS Grade 4 Science	4660	8907	10458	4636
TIMSS Grade 8 Mathematics	6992	8446	6735	3028
TIMSS Grade 8 Science	6305	8697	8033	3550
PIRLS	3999	8503	12259	5872

The second step involved computing the percentage of those students scoring in the range around each international benchmark that answered each item correctly. To compute these percentages, students in each country were weighted proportionally to the size of the student population in the country. For multiple choice items and constructed response items worth 1 point, it was a straightforward matter of computing the percentage of students at each benchmark who answered each item correctly. For constructed response items

scored for partial and full credit, percentages were computed for students receiving partial credit as well as for the students receiving full credit. This was particularly important in the PIRLS 2011 scale anchoring.

Third, the criteria described below were applied to identify the items that anchored at each benchmark. An important feature of the scale anchoring method is that it yields descriptions of the performance demonstrated by students reaching each of the international benchmarks on the scales, and that the descriptions reflect demonstrably different accomplishments by students reaching each successively higher benchmark. Because the process entails the delineation of sets of items that students at each international benchmark are likely to answer correctly and that discriminate between one benchmark and the next, the criteria for identifying the items that anchor considers performance at more than one benchmark.

For multiple choice items, 65 percent was used as the criterion for anchoring at each benchmark being analyzed, since students would be likely (about two thirds of the time) to answer the item correctly. A criterion of less than 50 percent was used for the next lower benchmark, because with this response probability, students were more likely to have answered the item incorrectly than correctly. The criteria for each benchmark are outlined below.

- A multiple choice item anchored at the Low International Benchmark (400) if at least 65 percent of students scoring in the range answered the item correctly. Because this was the lowest benchmark described, there were no further criteria.
- A multiple choice item anchored at the Intermediate International Benchmark (475) if at least 65 percent of students scoring in the range answered the item correctly, and less than 50 percent of students at the Intermediate International Benchmark answered the item correctly.
- A multiple choice item anchored at the High International Benchmark (550) if at least 65 percent of students scoring in the range answered the item correctly, and less than 50 percent of students at the Intermediate International Benchmark answered the item correctly.
- A multiple choice item anchored at the Advanced International Benchmark (625) if at least 65 percent of students scoring in the range answered the item correctly, and less than 50 percent of students at the High International Benchmark answered the item correctly.

To include all of the multiple choice items in the anchoring process and provide information about content domains and cognitive processes that might

not otherwise have had many anchor items, the concept of items that "almost anchored" was introduced. These were items that met slightly less stringent criteria for being answered correctly. The criteria to identify multiple choice items that "almost anchored" were that at least 55 percent of students scoring in the range answered the item correctly and less than 50 percent of students at the next lowest benchmark answered the item correctly. To be completely inclusive for all items, items that met only the criterion that at least 55 percent of the students answered correctly (regardless of the performance of students at the next lower point) were also identified. The categories of items were mutually exclusive, and ensured that all of the items were available to inform the descriptions of student achievement at the anchor levels. A multiple choice item was considered to be "too difficult" to anchor if less than 55 percent of students at the advanced benchmark answered the item correctly.

A somewhat less strict criterion also was used for all the constructed response items, because students had much less scope for guessing. For constructed response items, the criterion of 50 percent was used for the benchmark without any discrimination criterion for the next lower benchmark. A constructed response item anchored at one of the international benchmarks if at least 50 percent of students at that benchmark answered the item correctly. A constructed response item was considered to be "too difficult" to anchor if less than 50 percent of students at the advanced benchmark answered the item correctly.

Exhibit 2 presents the number of TIMSS 2011 mathematics and science items, and PIRLS 2011 items that anchored at each international benchmark. A description of the items can be found at **Item Descriptions Developed During the TIMSS 2011 Benchmarking** and **Item Descriptions Developed During the PIRLS 2011 Benchmarking**. It should be noted that for PIRLS an item can anchor at two benchmarks, at higher benchmark for full credit, and a lower benchmark or partia credit.

Exhibit 2: Number of Items Anchoring and Almost Anchoring at Each International Benchmark

	Low (400)	Intermediate (475)	High (550)	Advanced (625)	Above Advanced	Total
TIMSS Grade 4 Number	6	13	30	34	5	88
TIMSS Grade 4 Geometric Shapes and Measures	4	10	25	18	4	61
TIMSS Grade 4 Data Display	6	7	11	2	0	26
TIMSS Grade 4 Mathematics	**16**	**30**	**66**	**54**	**9**	**175**
TIMSS Grade 4 Life Science	6	12	21	22	13	74
TIMSS Grade 4 Physical Science	4	12	19	20	6	61
TIMSS Grade 4 Earth Science	0	4	12	12	5	33
TIMSS Grade 4 Science Total	**10**	**28**	**52**	**54**	**24**	**168**
TIMSS Grade 8 Number	3	9	30	18	1	61
TIMSS Grade 8 Algebra	1	6	29	27	6	69
TIMSS Grade 8 Geometry	0	3	20	17	2	42
TIMSS Grade 8 Data and Chance	3	13	15	8	4	43
TIMSS Grade 8 Mathematics	**7**	**31**	**94**	**70**	**13**	**215**
TIMSS Grade 8 Biology	2	13	21	31	12	79
TIMSS Grade 8 Chemistry	3	5	15	17	4	44
TIMSS Grade 8 Physics	1	3	15	22	13	54
TIMSS Grade 8 Earth Science	2	8	10	16	3	39
TIMSS Grade 8 Science Total	**8**	**29**	**61**	**86**	**32**	**216**
PIRLS Literary	2	28	43	14	3	90
PIRLS Informational	3	13	34	27	7	84
PIRLS	**5**	**41**	**77**	**41**	**10**	**174**

In preparation for review by SMIRC (Science and Mathematics Item Review Committee), the mathematics and science items were organized in binders by grade, grouped by international benchmark, and within each benchmark the items were sorted by content area. In preparation for review by the RDG (Reading Development Group), the PIRLS reading items were sorted by reading purpose (literary and informational), then grouped by the international benchmark, since the PIRLS anchoring is conducted separately for the two reading purposes. For both TIMSS and PIRLS 2011, the final categorization was by the anchoring criteria the items met: items that anchored, followed by items that almost anchored, followed by items that met only the

55 to 65 percent criteria. Also, for both TIMSS and PIRLS, the following information was included for each item: framework classification, answer key or scoring guide, release status, percent correct at each benchmark, and overall international percent correct.

The scale anchoring was conducted in the spring of 2012—TIMSS 2011 at a four-day meeting in Helsinki and PIRLS 2011 at a three-day meeting in Stockholm. At the scale anchoring meetings, the expert committees 1) worked through each item to arrive at a short description of the student competencies demonstrated by responding correctly (or responding with a partially correct response), 2) summarized the proficiency demonstrated by students reaching each international benchmark for publication in reports, and 3) selected example items that supported and illustrated the benchmark descriptions to publish together with the descriptions.