Chapter 13

The TIMSS 2007 International Benchmarks of Student Achievement in Mathematics and Science

Ina V.S. Mullis, Ebru Erberber, and Corinna Preuschoff

13.1 Overview

It is important for users of the TIMSS achievement results to understand what the scores on the TIMSS mathematics and science achievement scales mean. That is, what does it mean to have a scale score of 513 or 426? To describe student performance at various points along the TIMSS mathematics and science achievement scales, TIMSS 2007 used scale anchoring to summarize and describe student achievement at four points on the mathematics and science scales—Advanced International Benchmark (625), High International Benchmark (550), Intermediate International Benchmark (475), and Low International Benchmark (400). For the description of performance at the international benchmarks please see *TIMSS 2007 International Mathematics Report* (Mullis, Martin, & Foy, 2008) and *TIMSS 2007 International Science Report* (Martin, Mullis, & Foy, 2008).

This chapter describes the scale anchoring procedures that were applied to describe student performance at these benchmarks. Information about the TIMSS 2007 achievement scales and details about the methods used for scaling were presented in Chapter 11. In brief, scale anchoring involves selecting benchmarks (scale points) on the TIMSS achievement scales to be described in terms of student performance and then identifying items that students scoring at the anchor points (the international benchmarks) can answer correctly. The items, so identified, are grouped by content domain within benchmarks for review by mathematics and science experts. For TIMSS 2007, the Science and Mathematics Item Review Committee conducted the review. The committee members examined the content of

each item and determined the kind of mathematics or science knowledge and/or skill demonstrated by students answering the item correctly. They then summarized the detailed list of item competencies in a brief description of achievement at each international benchmark. This procedure resulted in a content-referenced interpretation of the achievement results that can be considered in light of the TIMSS 2007 mathematics and science frameworks. The item-by-item descriptions developed as part of the scale anchoring procedures are provided in Appendix F.

13.2 History of Identifying the International Benchmarks[1]

Identifying the scale points to serve as international benchmarks initially was a challenge for TIMSS in the context of measuring trends. For the TIMSS 1995 and 1999 assessments, the scales were anchored using percentiles. That is, the scale anchoring analysis was conducted using the Top 10 percent (90^{th} percentile), the Top Quarter (75^{th} percentile), the Top Half (50^{th} percentile), and the Bottom Quarter (25^{th} percentile). However, with different participating countries in each TIMSS cycle and different achievement for countries participating in previous cycles, the percentile points had changed between 1995 and 1999.

In planning for reporting the results of TIMSS 2003, it was clear that TIMSS needed a set of points to serve as benchmarks, that would not change in the future, that would look sensible, and that were similar to the points used in 1999. After much consideration, a set of four points with equal intervals on the mathematics and science achievement scales was identified to be used as the international benchmarks, namely 400, 475, 550, and 625. These points were selected to be as close as possible to the percentile points anchored in 1999 at the eighth grade (i.e., Top 10 percent was 616 for mathematics and science, Top Quarter was 555 for mathematics and 558 for science, Top Half was 479 for mathematics and 488 for science, and Bottom Quarter was 396 for mathematics and 410 for science). The newly defined benchmark scale points were used as the basis for the scale anchoring descriptions in TIMSS 2003 and again in TIMSS 2007.

1 The description of the scale anchoring procedure was adapted from Kelly (1999), and Gregory and Mullis (2000).

13.3 Identifying the Students Achieving at the International Benchmarks

The first step in the scale-anchoring procedure was to identify those students scoring at the international benchmarks. Following the procedure used in previous IEA studies, students scoring within plus and minus 5 scale score points of each benchmark were identified for the benchmark analysis. The score ranges around each international benchmark and the number of students scoring in each range at the fourth and eighth grades for mathematics are shown in Exhibit 13.1 and for science in Exhibit 13.2. The range of plus and minus 5 points around a benchmark is intended to provide an adequate sample in each group, yet be small enough so that performance at each benchmark anchor point is still distinguishable from the next. The data analysis for the scale anchoring was based on these students scoring at each benchmark range.

Exhibit 13.1 Range Around Each International Benchmark and Number of Students Within Each Range – Mathematics

	Low Benchmark	Intermediate Benchmark	High Benchmark	Advanced Benchmark
Range of Scale Scores	**395–405**	**470–480**	**545–555**	**620–630**
Fourth Grade	3151	5243	5732	2755
Eighth Grade	6969	7649	5639	2335

Exhibit 13.2 Range Around Each International Benchmark and Number of Students Within Each Range – Science

	Low Benchmark	Intermediate Benchmark	High Benchmark	Advanced Benchmark
Range of Scale Scores	**395–405**	**470–480**	**545–555**	**620–630**
Fourth Grade	2950	5091	6321	2981
Eighth Grade	6393	8366	6749	2767

13.4 The Scale Anchoring Criteria

Having identified the number of students scoring at each benchmark anchor point, the next step was determining which particular items anchored at each of the anchor points. An important feature of the scale anchoring method is that it yields descriptions of the performance demonstrated by students reaching each of the benchmarks on the TIMSS mathematics and science achievement scales, and that these descriptions reflect demonstrably

different accomplishments by students reaching each successively higher benchmark. The process entails the delineation of sets of items that students at each international benchmark are very likely to answer correctly and that discriminate between one benchmark and the next. Criteria were applied to identify the items that were answered correctly by most of the students at a particular benchmark, but by fewer students at the next lower benchmark.

In scale anchoring, the anchor items for each point are intended to be those that differentiate between adjacent anchor points (e.g., between the Advanced and the High International Benchmarks). To meet this goal, the criteria for identifying the items must take into consideration performance at more than one benchmark. Therefore, in addition to a criterion for the percentage of students at a particular benchmark correctly answering an item, it also was necessary to use a criterion for the percentage of students scoring at the next lower benchmark who correctly answer an item. For multiple-choice items, the criterion of 65 percent was used for the benchmark, since students would be likely (about two thirds of the time) to answer the item correctly. The 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. A somewhat less strict criterion was used for constructed-response items, because students have much less possibility of guessing. For constructed-response items, the criterion of 50 percent was used for the benchmark without any discrimination criterion for the next lower benchmark.

The criteria used to identify multiple-choice items that "anchored" are outlined below:

For the Low International Benchmark (400), a multiple-choice item anchored 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).

For the Intermediate International Benchmark (475), a multiple-choice item anchored if

- At least 65 percent of students scoring in the range answered the item correctly and

- Less than 50 percent of students at the Low International Benchmark answered the item correctly.

For the High International Benchmark (550), a multiple-choice item anchored 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.

For the Advanced International Benchmark (625), a multiple-choice item anchored 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 items in the anchoring process and provide information about content domains and cognitive processes that might not have had many items anchor exactly, items that met a slightly less stringent set of criteria were also identified. The criteria to identify multiple-choice items that "almost anchored" were the following:

For the Low International Benchmark (400), a multiple-choice item almost anchored if

- At least 60 percent of students scoring in the range answered the item correctly (because this was the lowest benchmark no further criteria were used).

For the Intermediate International Benchmark (475), a multiple-choice item almost anchored if

- At least 60 percent of students scoring in the range answered the item correctly and
- Less than 50 percent of students at the Low International Benchmark answered the item correctly.

For the High International Benchmark (550), a multiple-choice item almost anchored if

- At least 60 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.

For the Advanced International Benchmark (625), a multiple-choice item almost anchored if

- At least 60 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 be completely inclusive for all items, items that met only the criterion that at least 60 percent of the students answered correctly (regardless of the performance of students at the next lower point) were also identified. The three 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 60 percent of students at the advanced benchmark answered the item correctly.

Different criteria were used to identify constructed-response items that "anchored." A constructed-response item anchored at one of the international benchmarks if at least 50 percent of students at that benchmark answer 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.

13.5 Identifying the Anchor Items at Each International Benchmark

For the students scoring in the range around each international benchmark, the percentage of those students that answered each item correctly was computed. To compute these percentages, students in each country were weighted to contribute proportional to the size of the student population in a country. Most of the TIMSS 2007 items were scored 1-point for a correct answer and 0 points for other answers. For these items, the percentage of students at each benchmark who answered each item correctly was computed. For relatively few constructed-response items scored for partial or full credit, percentages were computed for the students receiving full credit.

The criteria described above were applied to identify the items that anchored, almost anchored, and met only the 60 to 65 percent criteria. For mathematics at the fourth grade 118 items anchored, 19 almost anchored,

and 40 met the 60 to 65 percent criteria. At the eighth grade, 151 mathematics items anchored, 27 almost anchored, and 36 met the 60 to 65 percent criteria. For science 111 items anchored, 16 almost anchored, and 43 met the 60 to 65 percent criteria at the fourth grade. At the eighth grade 152 science items anchored, 16 almost anchored, and 42 met the 60 to 65 percent criteria, respectively.

Broadening the anchor criteria on each benchmark to include items meeting the less stringent criteria, enabled the Science and Mathematics Item Review Committee to use all of the items included in the TIMSS 2007 assessment to characterize performance at each benchmark. Even though these items did not meet the 65 percent anchoring criteria, they were still items that students scoring at the benchmarks had a high degree of probability of answering correctly.

Exhibit 13.3 presents the number of mathematics items by content domain that anchored at each international benchmark at the fourth grade. Exhibit 13.4 presents the corresponding information for the eighth grade. Exhibit 13.5 and Exhibit 13.6 present the number of science items by content domain at each international benchmark at fourth and the eighth grades, respectively.

Exhibit 13.3 Number of Items Anchoring at Each International Benchmark by Content Domain – Fourth Grade Mathematics*

	Low (400)	Intermediate (475)	High (550)	Advanced (625)	Too Difficult to Anchor	Total
Number	6	15	36	30	4	91
Geometric Shapes and Measures	5	13	20	18	4	60
Data Display	3	11	9	3	–	26
Total	**14**	**39**	**65**	**51**	**8**	**177**

* Following the item review, 2 items were deleted out of 179 items in the mathematics fourth grade test, resulting in 177 items (see Chapter 10 for more details on the item review process).

Exhibit 13.4 Number of Items Anchoring at Each International Benchmark by Content Domain – Eighth Grade Mathematics*

	Low (400)	Intermediate (475)	High (550)	Advanced (625)	Too Difficult to Anchor	Total
Number	5	17	24	14	3	63
Algebra	1	7	29	26	1	64
Geometry	–	7	22	17	1	47
Data and Chance	3	9	19	8	1	40
Total	**9**	**40**	**94**	**65**	**6**	**214**

* Following the item review, 1 item was deleted out of 215 items in the mathematics eighth grade test, resulting in 214 items (see Chapter 10 for more details on the item review process).

Exhibit 13.5 Number of Items Anchoring at Each International Benchmark by Content Domain – Fourth Grade Science*

	Low (400)	Intermediate (475)	High (550)	Advanced (625)	Too Difficult to Anchor	Total
Life Science	7	17	15	20	12	71
Physical Science	7	9	28	15	5	64
Earth Science	1	6	11	13	4	35
Total	**15**	**32**	**54**	**48**	**21**	**170**

* Following the item review, 3 items were deleted out of 174 items in the science fourth grade test, resulting in 171 items. Also, 1 two-part item was combined to form a single item, further reducing the number of items to 170 (see Chapter 10 for more details on the item review process).

Exhibit 13.6 Number of Items Anchoring at Each International Benchmark by Content Domain – Eighth Grade Science*

	Low (400)	Intermediate (475)	High (550)	Advanced (625)	Too Difficult to Anchor	Total
Biology	2	11	26	25	11	75
Chemistry	3	4	11	16	7	41
Physics	2	2	14	24	12	54
Earth Science	-	6	17	12	5	40
Total	**7**	**23**	**68**	**77**	**35**	**210**

* Following the item review, 4 items were deleted out of 214 items in the science eighth grade test, resulting in 210 items (see Chapter 10 for more details on the item review process).

13.6 Experts Review Anchor Items by International Benchmark and Content Domains to Develop the Descriptions of Achievement

Having identified the items that anchored at each of the international benchmarks, the next step was to have the items reviewed by the TIMSS 2007 Science and Mathematics Item Review Committee to develop descriptions

of student performance. In preparation for the review by the members of the TIMSS 2007 Science and Mathematics Item Review Committee, the mathematics and science items, respectively, were organized in binders grouped by international benchmark and within benchmark, the items were sorted by content area and then by the anchoring criteria they met - items that anchored, followed by items that almost anchored, followed by items that met only the 60 to 65 percent criteria. The following information was included for each item: content area, topic area, cognitive domain, maximum points, answer key, release status, percent correct at each benchmark, and overall international percent correct. For constructed-response items, the scoring guides were included.

The TIMSS & PIRLS International Study Center staff convened the TIMSS 2007 Science and Mathematics Item Review Committee for a four-day meeting in Kaohsiung, Taiwan. The work involved in completing the scale anchoring for the international benchmarks consisted of three tasks: (1) work through each item in each binder and arrive at a short description of the knowledge, understanding, and/or skills demonstrated by students answering the item correctly; (2) based on the items that anchored, almost anchored, and met only the 60 to 65 percent criterion, develop a description (in detailed and summary form) of the level of mathematics or science proficiency demonstrated by students at each of the four international benchmarks to publish in the TIMSS 2007 international reports; and (3) select example items that supported and illustrated the benchmark descriptions to publish together with the descriptions.

References

Gregory, K. D. & Mullis, I. V. S. (2000). Describing international benchmarks of student achievement. In M. O. Martin, K. D. Gregory, & S. E. Stemler (Eds.), *TIMSS 1999 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Kelly, D. L. (1999). *Interpreting the Third International Mathematics and Science Study (TIMSS) achievement scales using scale anchoring*. Unpublished doctoral dissertation, Boston College, Chestnut Hill, MA.

Martin, M.O., Mullis, I.V.S., & Foy, P. (with Olson, J.F., Erberber, E., Preuschoff, C., & Galia, J.). (2008). *TIMSS 2007 international science report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I.V.S., Martin, M.O., & Foy, P. (with Olson, J.F., Preuschoff, C., Erberber, E., Arora, A., & Galia, J.). (2008). *TIMSS 2007 international mathematics report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.