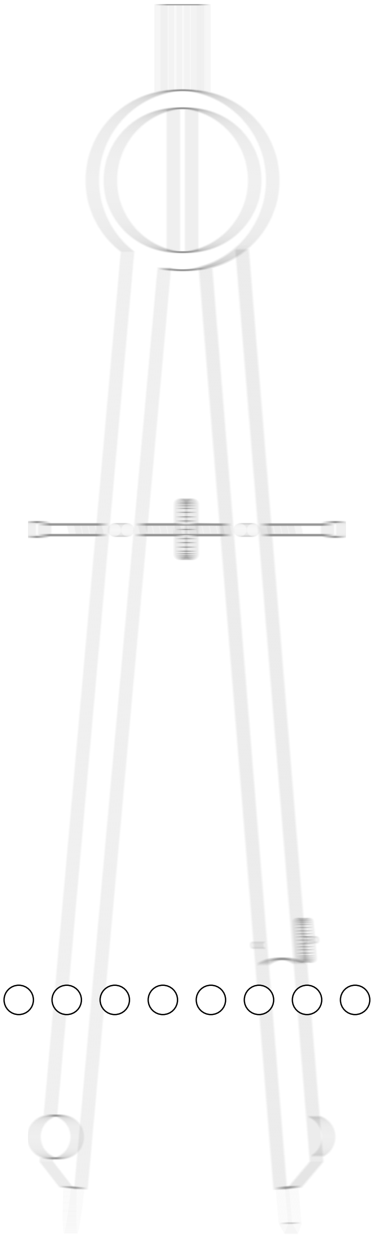
Describing TIMSS 1999 International Benchmarks of Student Achievement

Kelvin D. Gregory
Ina V. S. Mullis



14 Describing TIMSS 1999 International Benchmarks of Student Achievement[1]

Kelvin D. Gregory
Ina V. S. Mullis

14.1 Overview

To help policymakers, educators, and the public better understand student performance on the mathematics and science achievement scales, TIMSS used scale anchoring to summarize and describe student achievement at each of the international benchmarks – top 10%, upper quarter, median, and lower quarter.[2] This means that several points along a scale are selected as anchor points, and the items that students scoring at each anchor point can answer correctly (with a specified probability) are identified and grouped together. Subject-matter experts review the items that "anchor" at each point and delineate the content knowledge and conceptual understandings each item represents. The item descriptions are then summarized to yield a portrait, illustrated by example items, of what students scoring at the anchor points are likely to know and be able to do.

The theoretical underpinnings of scale anchoring and decisions related to the application of scale anchoring to the TIMSS data can be found in Kelly (1999). This chapter is derived from chapter three of Kelly's work and describes how the TIMSS 1999 International Benchmarks were developed. These benchmarks are used in TIMSS 1999 Benchmarking Reports.

Scale anchoring is a two-part process. First, the achievement data for each TIMSS scale were analyzed to identify items that students scoring at each anchor point answered correctly.

The scale-anchoring process for TIMSS 1999 capitalized on the TIMSS 1995 procedures implemented at the fourth and eighth grades. The TIMSS 1995 scale-anchoring results for mathematics are presented in Kelly, Mullis, & Martin (2000); those for science are presented in Smith, Martin, Mullis, & Kelly (2000).

[1] This chapter was mainly reproduced from Gregory & Mullis (2000) in the international technical report for TIMSS 1999 (Martin, Gregory, & Stemler, 2000).

[2] The international benchmarks - top 10%, upper quarter, median, and lower quarter - correspond to the 90th, 75th, 50th, and 25th percentiles, respectively, of the international distribution of student achievement in mathematics and science. The international benchmarks should not be confused with the TIMSS Benchmarking study, in which states and school districts compared or "benchmarked" their school systems against high-performing countries around the world.

14.2 Scale Anchoring Data Analysis

In conducting the data analysis for the scale anchoring, TIMSS used a five-step procedure that involved:

- Selecting anchor points and forming groups of examinees at each anchor point
- Calculating the proportion of students at each anchor point answering the items correctly
- Determining the anchor items for the lowest anchor point for each subject
- Determining the anchor items for the remaining anchor points

14.2.1 Anchor Points

An important feature of the scale-anchoring method is that it yields descriptions of the knowledge and skills of students reaching certain performance levels on a scale, and that these descriptions reflect demonstrably different accomplishments from point to point. The process entails the delineation of sets of items that students at each anchor point are very likely to answer correctly and that discriminate between performance levels. Criteria are applied to identify the items that are answered correctly by most of the students at an anchor point, and by fewer students at the next lower point.

TIMSS 1999, like TIMSS 1995, based the scale-anchoring descriptions on the international benchmarks, the 25th, 50th, 75th and 90th percentiles. These percentiles were labelled the lower quarter, median, upper quarter, and top 10% international benchmarks, respectively. The international percentiles were computed using the combined data from the countries that participated. Exhibit 14.1 shows the scale scores representing the international benchmarks for mathematics and science, respectively.

Exhibit 14.1 TIMSS 1999 International Benchmarks for Eighth Grade*—Mathematics and Science

	25th Percentile	50th Percentile	75th Percentile	90th Percentile
Mathematics	396	479	555	616
Science	410	488	558	616

*Eighth grade in most countries.

The performance data analysis was based on students scoring in a range around each anchor point. These ranges are designed to allow an adequate sample in each group, yet be small enough so each anchor point is still distinguishable from the next. Following the procedures used for TIMSS 1995, a range of plus and minus five scale points was used. The ranges around the international percentiles and the number of observations within each range are shown in Exhibit 14.2.

Exhibit 14.2 Range around Each Anchor Point and Number of Observations within Ranges

	25th Percentile	50th Percentile	75th Percentile	90th Percentile
Mathematics				
Range	391-401	474-484	550-560	611-621
Observations	3540	5690	5531	3703
Science				
Range	405-415	483-493	553-563	611-621
Observations	3632	6090	5806	3426

14.3 Anchoring Criteria

In scale anchoring, the anchor items for each point are intended to be those that differentiate between adjacent anchor points. To meet this goal, the criteria for identifying the items must take into consideration performance at more than one anchor point. Therefore, in addition to a criterion for the percentage of students at a particular anchor point correctly answering an item, it is necessary to use a criterion for the percentage of students scoring at the next lower anchor point who correctly answer an item. Once again, following the procedures used for TIMSS 1995, the criterion of 65% was used for the anchor point, since students would be likely (about two-thirds of the time) to answer the item correctly. The criterion of fewer than 50% was used for the next lower point, because with this response probability, students were more likely to have answered the item incorrectly than correctly.

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

For the 25th percentile, an item anchored if

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

Because the 25th percentile is the lowest point, items were not identified in terms of performance at a lower point.

For the 50th percentile, an item anchored if

- At least 65% of students scoring in the range answered the item correctly and
- Fewer than 50% of students at the 25th percentile answered the item correctly

For the 75th percentile, an item anchored if

- At least 65% of students scoring in the range answered the item correctly and
- Fewer than 50% of students at the 50th percentile answered the item correctly

For the 90th percentile, an item anchored if

- At least 65% of students scoring in the range answered the item correctly and
- Fewer than 50% of students at the 75th percentile answered the item correctly

To supplement the pool of anchor items, items that met a slightly less stringent set of criteria were also identified. The criteria to identify items that "almost anchored" were the following:

For the 25th percentile, an item almost anchored if

- At least 60% of students scoring in the range answered the item correctly

Because the 25th percentile is the lowest point, items were not identified in terms of performance at a lower point.

For the 50th percentile, an item almost anchored if

- At least 60% of students scoring in the range answered the item correctly and
- Fewer than 50% of students at the 25th percentile answered the item correctly

For the 75th percentile, an item almost anchored if

- At least 60% of students scoring in the range answered the item correctly and
- Fewer than 50% of students at the 50th percentile answered the item correctly

For the 90th percentile, an item almost anchored if

- At least 60% of students scoring in the range answered the item correctly and
- Fewer than 50% of students at the 75th percentile answered the item correctly

Items answered correctly by at least 60% to 65% of the students regardless of the performance of students at the next lower point were identified to further supplement the item pool. Items that anchored, almost anchored, and met the 60% to 65% criterion were placed into three mutually exclusive categories. Each of these items helped to inform the descriptions of student achievement at the anchor levels.

14.4 Computing the Item Percent Correct at Each Level

The percentage of students scoring in the range around each anchor point and who answered a given item correctly was computed. To that end, students were weighted to contribute proportionally to the size of the student population in a country. Most of the TIMSS 1999 items were scored dichotomously. For these items, the percentage of students at each anchor point who answered each item correctly was computed. Some of the open-ended items, however, are scored on a partial-credit basis (one or two points); these were transformed into a series of dichotomously scored items, as follows. Consider an item that was scored zero, one, or two. Two variables were created:

$v_1 = 1$ if the student received a one or two, and

$v_1 = 0$ otherwise, and

$v_2 = 1$ if the student received a two and

$v_2 = 0$ otherwise.

The percentage of students receiving a 1 on v_1 and of those receiving a 1 on v_2 was computed. This yielded the percentage of students receiving at least one point and a percentage of students receiving full credit. For mathematics, the descriptions used only the percentages of students receiving full credit on such items, whereas science sometimes also took the results for partial credit into consideration.

14.5 Identifying Anchor Items

For the TIMSS 1999 mathematics and science scales, the criteria described above were applied to identify the items that anchored, almost anchored, and met only the 60% to 65% criterion. Exhibits 14.3 and 14.4 present the number of these items at each anchor point. Altogether, six mathematics items met the anchoring criteria at the 25th percentile, 36 did so for the 50th percentile, 73 for the 75th percentile, and 43 for the 90th percentile. Eleven items were too difficult for the 90th percentile. In science, 15 items met one of the criteria for anchoring at the 25th percentile, 33 for the 50th percentile, 39 for the 75th percentile, and 41 for the 90th percentile. Twenty-eight items were too difficult to anchor at the 90th percentile.

Including items meeting the less stringent anchoring criteria substantially increased the number of items that could be used to characterize performance at each anchor point, beyond what would have been available if only the items that met the 65%/50% criteria were included. Despite not meeting the 65%/50% criteria, these were still items that students scoring at the anchor points had a high probability of answering correctly.

Exhibit 14.3 Number of Items Anchoring at Each Anchor Level—Eighth Grade Mathematics

	Anchored	Almost Anchored	Met 60-65% Criterion	Total
25th Percentile	4	2	0	6
50th Percentile	16	7	13	36
75th Percentile	34	14	25	73
90th Percentile	17	4	22	43
Too difficult for 90th				11
Total	71	27	60	158

Exhibit 14.4 Number of Items Anchoring at Each Anchor Level—Eighth Grade Science

	Anchored	Almost Anchored	Met 60-65% Criterion	Total
25th Percentile	10	5	0	15
50th Percentile	6	3	24	33
75th Percentile	5	8	26	39
90th Percentile	7	9	25	41
Too difficult for 90th				28
Total	29	25	75	156

14.6 Expert Review of Anchor Items by Subject and Content Areas

The purpose of scale anchoring was to describe the mathematics and science that students know and can do at the four international benchmarks. In preparation for review by the subject-matter experts, the items were organized in binders grouped by anchor point and within anchor point by content area. One binder was prepared for each subject area, with each binder having four sections, corresponding to the four anchor levels. Within each section, 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% criteria. The following information was included for each item: its TIMSS 1999 content area and performance expectation categories; its answer key; percent correct at each anchor point; overall international percent correct by grade; and item difficulty. For open-ended items, the scoring guides were included.

When going through each section of a binder, the panelists examined the items grouped by content area to determine what students at an anchor point knew and could do in each content area. Exhibits 14.5 and 14.6 present, for each scale, the number of items per content area that met one of the anchoring criteria discussed above, at each international percentile, and the number of items that were too difficult for the 90th percentile.

In mathematics, each of the five reporting categories had the most items anchoring at the 75th percentile. Fractions and number sense, data representation, analysis and probability, and algebra had at least one item anchoring at the 25th percentile, while the geometry and measurement categories did not. The science items for earth science, life science, physics and chemistry were reasonably spread out across the anchoring categories. The categories of environmental and resource issues, and scientific inquiry and the nature of science had no items that anchored at the 25th percentile, but it should be remembered that they contained the fewest items.

Exhibit 14.5 Number of Items Anchoring at Each Anchor Level, by Content Area—Eighth Grade Mathematics

	25th Percentile	50th Percentile	75th Percentile	90th Percentile	Too Difficult for 90th Percentile	Total
Fractions and Number Sense	3	14	27	14	4	62
Measurement	0	3	9	12	2	26
Data Representation Analysis, and Probability	2	8	10	1	1	22
Geometry	0	4	10	7	0	21
Algebra	1	7	17	9	4	38
Total	6	36	73	43	11	169

Exhibit 14.6 Number of Items Anchoring at Each Anchor Level, by Content Area—Eighth Grade Science

	25th Percentile	50th Percentile	75th Percentile	90th Percentile	Too Difficult for 90th Percentile	Total
Earth Science	3	5	6	6	3	23
Life Science	8	9	11	10	4	42
Physics	5	12	7	7	8	39
Chemistry	2	2	7	7	4	22
Environmental and Resource Issues	0	4	5	2	3	14
Scientific Inquiry and the Nature of Science	0	1	5	1	6	13
Total	18	33	41	33	28	153

14.7 The Anchoring Expert Panels

Two panels of experts in mathematics and science were assembled to examine the items and draft descriptions of performance at the anchor levels. The mathematics anchor panel had 11 members, and the science anchor panel seven, listed in Exhibits 14.7 and 14.8, respectively. The members had extensive experience in their subject areas and a thorough knowledge of the TIMSS curriculum frameworks and achievement tests.

Exhibit 14.7 Mathematics Scale Anchoring Panel Members

Lillie Albert Boston College United States	Anica Aleksova Pedagosiki Zawod na Makedonija Republic of Macedonia
Kiril Bankov University of Sofia Bulgaria	Jau-D Chen Taiwan Normal University Taiwan
John Dossey Consultant United States	Barbara Japelj Educational Research Institute Slovenia
Mary Lindquist National Council of Teachers of Mathematics United States	David Robitaille University of British Columbia Canada
Graham Ruddock National Foundation for Education Research England	Hanako Senuma National Institute for Educational Research Japan
Pauline Vos University of Twente Netherlands	

Exhibit 14.8 Science Scale Anchoring Panel Members

Audrey Champagne State University of New York United States	Galina Kovalyova Center for Evaluating the Quality of Education Russian Federation
Jan Lokan Australian Council for Educational Research Australia	Jana Paleckova Institute for Information on Education Czech Republic
Senta Raizen National Center for Improving Science Education United States	Vivien Talisayon Institute of Science and Mathematics Education Development University of the Philippines
Hong Kim Tan Ministry of Education Research and Evaluation Singapore	

14.8 Development of Anchor Level Descriptions

The TIMSS International Study Center convened the two expert panels for a three-day meeting, May 7 to 10, 2000, at Martha's Vineyard, Massachusetts. The panelists' were assigned 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% criterion, draft a description of the knowledge, understandings, and skills demonstrated by students at each anchor point; and (3) select example items to support and illustrate the anchor point descriptions. These drafts were then edited and revised as necessary, and the panelists reviewed and approved the item descriptions, anchor point descriptions, and example items for use in the TIMSS 1999 International Reports.

References

Gregory, K., & 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* (pp. 267-278). Chestnut Hill, MA: Boston College.

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

Kelly, D.L., Mullis, I.V.S., & Martin, M.O. (2000). *Profiles of student achievement in mathematics at the TIMSS international benchmarks: U.S. performance and standards in an international context.* Chestnut Hill, MA: Boston College.

Smith, T.A., Martin, M.O., Mullis, I.V.S., & Kelly, D.L. (2000). *Profiles of student achievement in science at the TIMSS international benchmarks: U.S. performance and standards in an international context,* Chestnut Hill, MA: Boston College.