2 TIMSS Test Development[1]

Robert A. Garden
Teresa A. Smith

2.1 Overview

The mathematics and science achievement tests used to measure achievement in the TIMSS benchmarking project were identical to those used by the United States as part of the international data collection for TIMSS 1999. This chapter describes how the 1999 tests followed the same design and used many of the same items as the TIMSS 1995 assessment, how the items released into the public domain following the 1995 assessment were replaced for 1999, and the procedure for field testing the replacement items and reviewing the results.

To provide as much information as possible about the nature and scope of the 1995 TIMSS achievement tests, almost two-thirds of the test items were released to the public. The remaining one-third were kept secure as a basis for accurately measuring trends in student achievement from 1995 to 1999. Releasing most of the 1995 items enabled more meaningful reports, both national and international, to be published and also provided information for secondary research. But it also meant that students in the TIMSS 1999 samples may have been exposed to these items, which necessitated the development of new mathematics and science items for TIMSS 1999.

The challenge for TIMSS 1999 was to develop tests containing replacement items that were similar in subject matter content and expectations for student performance to those released in 1995, to be used alongside the secure items from 1995. This would provide a reliable and informative assessment of student achievement in mathematics and science in 1999, comparable in scope and coverage to the 1995 assessment, while also providing a valid measure of the changes in achievement since 1995.

This chapter describes the TIMSS 1999 test development, including the development and construction of the replacement items, the item review process, field testing and item analysis, selection of the final item set, scoring guide development, and the resulting main survey test design. The new mathematics and science

Items in clusters A-H were kept secure for future use in trend studies, and the remaining 18 clusters (I-Z) were released to the public. The secure clusters A-H were used in TIMSS 1999 exactly as in TIMSS 1995. The 103 mathematics and 87 science items released in 1995 were replaced with similar items. Replacement items retained the same format, assessed the same basic content area and performance expectation and, as nearly as possible matched the difficulty level of the 1995 items.

2.2.2 Construction of Replacement Items

An initial pool of over 300 science and mathematics items, with scoring guides, was developed as potential replacement items, with most TIMSS 1995 released items having at least two possible replacements. Item development took place from July to November 1997. Replacement items and scoring guides for science were developed by Teresa Smith and Christine O'Sullivan, science coordinator and science consultant, respectively, and by the National Foundation for Educational Research in England and Wales. Robert Garden and Chancey Jones, mathematics coordinator and mathematics consultant, respectively, developed the mathematics items and scoring guides.

While each mathematics replacement item was to present students with a task similar to that in the corresponding 1995 item, care was taken not to make it so similar as to favor students who had encountered the original item. Replacement items were designed not only to satisfy the original content and performance expectation requirements but, wherever possible, to cue students to similar reasoning or preferred methods of solution, and were written in the same format as the original.

Exhibit 2.1 Assignment of Item Clusters to Student Test Booklets * — TIMSS 1995 and 1999

Cluster Type	Cluster Label	Booklet 1	2	3	4	5	6	7	8
Core Cluster (12 minutes) (Mathematics and Science Items - Multiple-Choice)	A	2	2	2	2	2	2	2	2
Focus Clusters (12 minutes) (Mathematics and Science Items - Multiple-Choice)	B	1				5		3	1
	C	3	1				5		
	D		3	1				5	
	E	5		3	1				
	F		5		3	1			
	G			5		3	1		
	H				5		3	1	
Breadth Clusters (22 minutes) (Mathematics and Science Items - Multiple-Choice and Free-Response)	I	6							
	J		6						
	K			6					
	L				6				
	M					6			
	N						6		
	O							6	
	P								6
	Q								3
	R								5
Mathematics Free-Response Clusters (10 minutes)	S	4							
	T	7		4					
	U			7		4			
	V					7		4	
Science Free-Response Clusters (10 minutes)	W		4					7	
	X		7		4				
	Y				7		4		
	Z						7		

*

Item-by-item matching of the science items was more difficult because of more specific topic area knowledge, which affected both the nature and difficulty of the item. While general skills can be assessed with a number of very similar items, specific topic area knowledge is more difficult to replicate in different contexts. In writing science replacement items, the main goal was to cover the same general content area knowledge that was defined in the TIMSS 1995 framework. For many of the original science items, quite similar replacement items could be generated. For others, while the same general science content area was maintained, the specific topic area, performance expectation, and difficulty of the 1999 item may have been altered somewhat.

In addition to the replacements for released items from TIMSS 1995, several new science items were written in the areas of *Environmental and Resource Issues* and *Scientific Inquiry and the Nature of Science*. This was done to expand the item pool and permit the results in these two content areas to be reported separately for TIMSS 1999 (see section 2.5 for a discussion of the final TIMSS 1999 science test).

2.2.3 Scoring Guides for Free-Response Items

The TIMSS 1999 item replacement efforts focused heavily on developing free-response items, questions to which students were asked to construct their own answers. Because creating such questions and scoring guides that work well in an international context is quite difficult, many more free-response items and scoring guides were developed and included in the field test than were required for the main survey. Exhibit 2.2 presents the number of free-response and multiple-choice questions included in the field test.

Exhibit 2.2 Number of Free-Response and Multiple-Choice Items in the TIMSS 1999 Field Test

In TIMSS 1995 and TIMSS 1999 both short-answer and extended-response items were scored using two-digit codes with rubrics spe-

combined with the first, represents a diagnostic code used to identify specific approaches or strategies, or common errors and misconceptions. The general scoring scheme used for a two-point and a one-point item in TIMSS 1995 is shown in Exhibit 2.3.

Exhibit 2.3 TIMSS Two-Digit Scoring Scheme for Free-Response Items

Two-Point Item Codes	
Code	**Definition**
20	fully-correct response; answer category/method #1
21	fully-correct response; answer category/method #2
22	fully-correct response; answer category/method #3
29	fully-correct response; some other method used
10	partially-correct response; answer category/method #1
11	partially-correct response; answer category/method #2
12	partially-correct response; answer category/method #3
19	partially-correct response; some other method used
70	incorrect response; common misconception/error #1
71	incorrect response; common misconception/error #2
76	incorrect response; information in stem repeated
79	incorrect response; some other error made
90	crossed out/erased, illegible, or impossible to interpret
99	Blank

One-Point Item Codes	
Code	**Definition**
10	correct response; answer category/method #1
11	correct response; answer category/method #2
12	correct response; answer category/method #3
19	correct response; some other method used
70	incorrect response; common misconception/error #1
71	incorrect response; common misconception/error #2
76	incorrect response; information in stem repeated
79	incorrect response; some other error made
90	crossed out/erased, illegible, or impossible to interpret
99	Blank

In TIMSS 1999, the same scoring scheme was retained with minor modifications. The use of code 76 for responses that merely repeated information in the stem of the item was discontinued for TIMSS 1999. Code 90 was also deleted, and responses in this category were coded as 79. For both surveys, the second-digit codes of seven and eight were reserved for nationally-defined diagnostic codes used by the national centers to monitor the occurrence of certain common response types in individual countries that were not already captured with the internationally-

defined diagnostic codes. In processing the data for the international database, these country-specific codes were recoded to the "other" response category (second digit nine) at the appropriate score level.

2.2.4 Item Review

Once drafted, the proposed replacement items and scoring guides were reviewed by the subject-matter coordinators, the mathemati

International Study Center Review

Field-test item statistics were reviewed in several phases. By June 19, 1998, preliminary field-test results for 12 countries had been analyzed as a trial run. The International Study Center staff reviewed these data for each item in both mathematics and science. A second preliminary analysis for 20 countries was completed July 1-2, 1998. The results were again reviewed by International Study Center staff on July 6-8, 1998. These reviews identified specific problems in items and item translations. In a few instances, the translated versions of the field test were compared with the international version and found to diverge. Discrepancies included changes in the meaning of the question, altered graphics, and changed order of response options. These issues were taken into account when the field-test data were reviewed and test questions for the main survey selected. In addition, the comment sheets that NRCs were asked to submit, reporting field-test items and scoring guides found to be problematic in their country, were reviewed. Such feedback clarified problems with specific items and with the use of the free-response scoring guides. These comments, problems, and suggestions were organized into a database and used during each phase of item review.

Subject Matter Item Replacement Committee Review

International Study Center staff met with the committee July 15-17, 1998, in London, England, to review the results of the field test and to identify the best replacement items for the main TIMSS 1999 survey. Item statistics for 21 countries were available at that time. Materials containing TIMSS 1995 released items, TIMSS 1999 field-test items, field-test scoring guides, field-test item analysis results, and suggestions from NRCs were compiled for the review. The committee reviewed the field-test item analysis results, suggested some item and scoring guide revisions, and proposed items for the main survey.

NRC Review

At the third NRC Meeting in Boston, Massachusetts, in August 1998, NRCs reviewed the items selected by the SMIRC for the main survey, the scoring guides, and the data almanacs from the field test. Data from 29 countries were available. NRCs accepted the main survey items subject to agreed-upon editing and modifications incorporated by the International Study Center.

mathematics, students were asked to show their work or explain their methods, and these explications were taken into account in scoring. In science, the two-point items required a fuller explanation demonstrating knowledge of science concepts. The distinction between the one- and two-point items was sometimes hazy in science, and for some two-point field-test items, the field-test data suggested little discrimination between the two score points.

Generalized scoring guides were developed for TIMSS 1999 to clarify the types of responses that would merit two points vs. only one point. The generalized scoring guides for mathematics are presented in Exhibit 2.4 and those for science in Exhibit 2.5.

Exhibit 2.4 TIMSS 1999 Mathematics Generalized Scoring Guide

Score Points for Extended-Response Items
2 Points: A two-point response is complete and correct. The response demonstrates a thorough understanding of the mathematical concepts and/or procedures embodied in the task. • Indicates that the student has completed the task, showing mathematically sound procedures • Contains clear, complete explanations and/or adequate work when required
1 Point: A one-point response is only partially correct. The response demonstrates only a partial understanding of the mathematical concepts and/or procedures embodied in the task. • Addresses some elements of the task correctly but may be incomplete or contain some procedural or conceptual flaws • May contain a correct solution with incorrect, unrelated, or no work and/or explanation when required •

Exhibit 2.5 TIMSS 1999 Science Generalized Scoring Guide

Score Points for Extended-Response Items
2 Points: A two-point response is complete and correct. The response demonstrates a thorough understanding of the science concepts and/or procedures embodied in the task. • Indicates that the student has completed all aspects of the task, showing the correct application of scientific concepts and/or procedures • Contains clear, complete explanations and/or adequate work when required
1 Point: A one-point response is only partially correct. The response demonstrates only a partial understanding of the scientific concepts and/or procedures embodied in the task. • Addresses some elements of the task correctly but may be incomplete or contain some procedural or conceptual flaws • May contain a correct answer but with an incomplete explanation • May contain an incorrect answer but with an explanation indicating a correct understanding of some of the scientific concepts
0 Points: A zero-point response is seriously inaccurate or inadequate, irrelevant, or incoherent.
Score Points for Short-Answer Items
1 Point: A one-point response is correct. The response indicates that the student has completed the task correctly.
0 Points: A zero-point response is completely incorrect, irrelevant, or incoherent.

The revised scoring guides were thoroughly reviewed by the Subject Matter Item Review Committee at its second meeting in London, July 1998, and further refinements were made. They were then reviewed by NRCs at their third meeting in Boston, August 1998. In general, NRCs agreed that the revisions reflected their comments. A few last suggestions were made before the scoring guides were issued for use in training in the Southern Hemisphere countries in Wellington, New Zealand, in October 1998. During this first training session, a few additional revisions were made. These were incorporated in the final TIMSS 1999 scoring guides used during scoring training for the Northern Hemisphere countries in February 1999.

2.4 Training Country Representatives for Free-Response Scoring

At both the first (Amsterdam) and second (Berlin) meetings of the NRCs, the International Study Center provided training in TIMSS procedures for free-response scoring. During plenary sessions, all of the NRCs were introduced to the TIMSS scoring approach. They learned about the significance of the first and second digits in the TIMSS codes – that the first digit is a correctness score, and that the second digit, when combined with the first, provides diagnostic information about the type of response. Other topics covered included the importance of maintaining high reliability in scoring, the necessary qualifications of the scorers, the process for training scorers in each country, and the

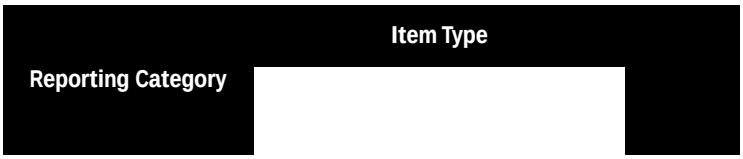
scope of work involved for the entire free-response scoring effort. NRCs who had participated in TIMSS 1995 shared information about the time required to score the free-response items. NRCs were also trained in the procedures for actual free-response scoring and for the within-country reliability studies.

Training procedures used the same "train-the-trainers" approach that had produced highly reliable scores in TIMSS 1995 (see Mullis & Smith, 1996). Personnel who were to be responsible for training scorers in each country participated in training sessions for the field test and for the main survey. In these training sessions, the TIMSS 1999 scoring approach was reviewed. Participants then were trained on a subset of the mathematics and science free-response items representing a range of situations that would be encountered in the scoring and included many of the items with the most complicated scoring guides. The following procedures were followed for each item:

- Participants read the item and its scoring guide
- Trainers discussed the rationale and method of the scoring guide
- Trainers presented and discussed a set of prescored example student responses illustrating the diagnostic codes and the rationale used to score the responses
- Participants scored a set of 10-30 practice student responses
- Trainers led group discussion of the scores given to the practice responses, with the aim of having all participants reach a common understanding

The purpose of the training sessions was to present a model for use in each country and to provide practice with the most difficult items. For example, NRCs learned how to select example responses and create training practice sets. They also learned the process for training. At the international training sessions, the participants received scoring guides, manuals, and packets of example and practice papers for each of item covered in the training. The training teams emphasized the need for the NRCs to prepare comparable materials for training in their own country, including all of the free-response items rather than only the sample included in the international training sessions. In addition, it was pointed out that for more difficult items and scoring guides, as many as 50 example and practice responses might be needed to help scorers reach a high degree of reliability.

Exhibit 2.7 Number of TIMSS 1999 Test Items and Score Points by Type and Reporting Category—Science



The TIMSS 1999 final test items were organized into the 26 main survey item clusters (A-Z) and assigned to eight different test booklets using the rotated test design of the original TIMSS study. Assignment to item clusters generally followed the original design, with most of the replacement items being assigned to the same cluster as the released 1995 items they were replacing. In TIMSS 1999, the final test contained four more mathematics items and eight more science items than the 1995 test. These extra 12 items were incorporated in the item clusters so that each booklet included one or two of them. Experience with TIMSS 1995 indicated that students would still have ample time to complete the test.

Exhibits 2.8 and 2.9 present the distribution of items in each content area across the eight test booklets for mathematics and science, respectively.

Exhibit 2.10 Maximum Number of TIMSS 1999 Score Points in Each Booklet by