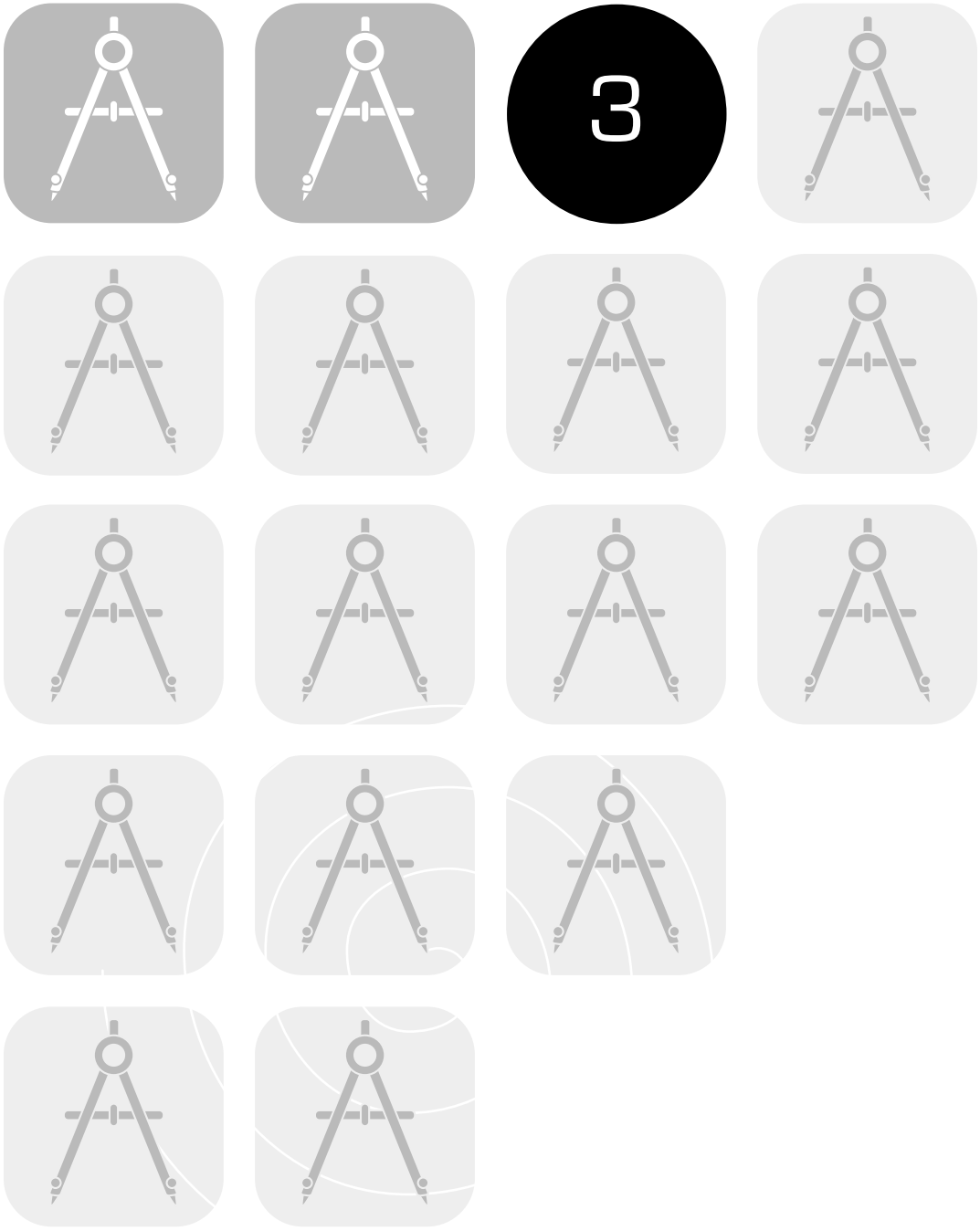
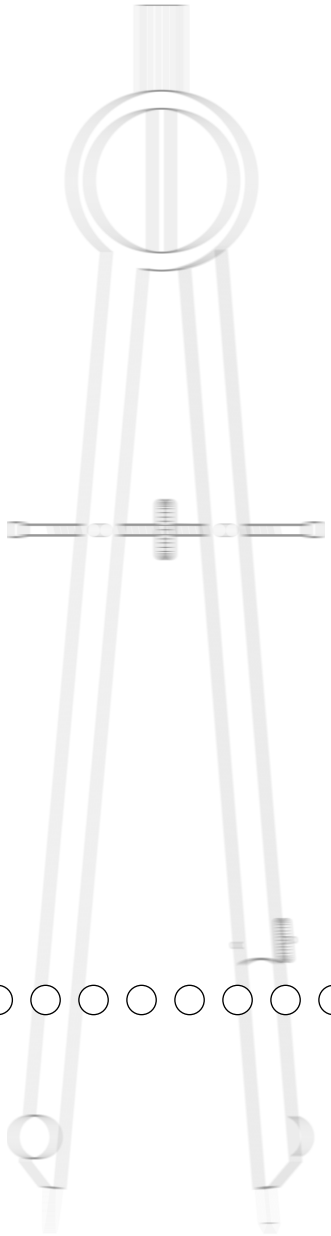


TIMSS Test Development

Robert A. Garden
Teresa A. Smith



3 TIMSS Test Development

Robert A. Garden
Teresa A. Smith

3.1 Overview

To provide as much information as possible about the nature and scope of the 1995 TIMSS achievement tests, almost two thirds of the items on the tests 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 terms of 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 richly 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 resulting mathematics and science assessments maintained the same distribution of items and testing time across content areas, performance expectations, and item formats that were specified in the original TIMSS framework[1] for the 1995 assessment.

○○○

1. The curriculum frameworks for TIMSS 1995 (Robitaille et al., 1993) resulted from an exhaustive analysis of the mathematics and science curricula of countries participating in that study. Specifications for the TIMSS tests were based on these curriculum frameworks. Mathematics and science content formed one dimension of the specifications, and performance expectations the other.

3.2 Development of Replacement Items

The major goal of test development was to produce a test that would parallel that of TIMSS 1995 in overall structure and content. The strategy used involved treating the 1995 items as a representative sample from the "pool" of all possible items within the defined test domain and selecting new items from this "pool" with the same subdomains as the released items from TIMSS 1995. In practice, each released item was evaluated to define its subdomain (mathematics or science content, performance expectation, item format, and difficulty level), and a set of potential replacement items from the same subdomain was then created. This method ensured that the final test, comprising the nonreleased and replacement items, covered the same test domain as in TIMSS 1995. The approach is described in further detail in the following sections.

3.2.1 Replacement of Item Clusters

In the 1995 TIMSS assessment, mathematics and science items were organized into 26 clusters, labeled A-Z. These clusters were rotated through eight student test booklets, with five or seven clusters in each book, according to the scheme shown in Exhibit 3.1 (Adams and Gonzalez, 1996). The same booklet design was used in TIMSS 1999. Clusters A - H, of multiple-choice items only, took about 12 minutes of testing time in both mathematics and science. Clusters I through R each took 22 minutes of testing time and contained a mixture of multiple-choice and free-response items in both mathematics and science. Clusters S through V, for mathematics, and W through Z for science, contained free-response items and took 10 minutes of testing time.

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 assessed the same basic content area and performance expectation and, as nearly as possible, matched the difficulty level of the 1995 items. The same item format was maintained for the replacement items. Thus the TIMSS 1999 tests were made to resemble closely those of TIMSS 1995 in structure and content.

Exhibit 3.1 Assignment of Item Clusters to Student Test Booklets[2] - 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		

2. Numbers in the cells indicate the position of the cluster within the booklet. For example, cluster A was the second cluster in each of the eight booklets.

3.2.2 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 item was to present students with a task similar to that addressed by the corresponding 1995 item, care was taken not to make it so similar as to favor any 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 replacement items were written in the same format as the original.[2] In the case of multiple-choice items, when feasible, each distracter was designed to depend on the same faulty reasoning, miscalculation, or misconception as in the original item.

Item-by-item matching in 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.

○○○

2. Item formats included multiple-choice, short-answer, and extended-response. Short-answer items require a numerical response, a short factual statement or sentence, or the completion of a table or sketch. Extended-response items require students to interpret text or diagrams to describe or explain procedures, processes, or mathematics and scientific concepts.

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* to expand the item pool and permit the results in these two content areas to be reported separately for TIMSS 1999 (see section 3.5 for a discussion of the final TIMSS 1999 science test).

3.2.3 Scoring Guides for Free-Response Items

The TIMSS 1999 item replacement task focused heavily on developing free-response items, questions where 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 3.2 presents the number of free-response and multiple-choice questions included in the field test.

Exhibit 3.2 Number of Free-Response and Multiple-Choice Items in the TIMSS 1999 Field Test

	Free-Response	Multiple-Choice	Total
Mathematics	38	108	146
Science	53	78	131
Total	91	186	277

In TIMSS 1995 and TIMSS 1999 both short-answer and extended-response items were scored using two-digit codes with rubrics specific to each item (Lie, Taylor, and Harmon, 1996). The first digit designates the correctness level of the response. The second digit, combined with the first, represents a diagnostic code used to identify specific types of approaches, 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 3.3.

Exhibit 3.3 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 7 and 8 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 9) at the appropriate score level.

3.2.4 Item Review

Once drafted, the proposed replacement items and scoring guides were reviewed by the subject-matter coordinators, the mathematics and science consultants, International Study Center staff, the Subject Matter Item Replacement Committee (SMIRC), and the National Research Coordinators (NRCs). The items were evaluated individually by the mathematics and science coordinators, consultants, and International Study Center staff to check that the item addressed its intended objective. Any technical deficiencies found were rectified. In addition, some possible sources of bias due to cultural, national, or gender differences were eliminated. Three item development and review meetings of the item writers and International Study Center staff were held during October and November, 1997.

3.2.5 Subject Matter Item Replacement Committee

An international committee of mathematics and science experts was formed to scrutinize the initial pool of items and make suggestions for revisions, select items from the item pool for the field test, review the item statistics from the field test, and select final test items for the main survey. The Subject Matter Item Replacement Committee (SMIRC) consisted of prominent mathematics and science educators nominated by participating countries, and thus represented a variety of nations and cultures.[3] The committee was responsible for ensuring that items were mathematically and scientifically accurate, and could be readily translated into the many languages and cultural contexts of the study. The committee contributed greatly to the quality of the item pool and played a critical role in identifying and modifying or deleting items that had the potential for cultural or national bias.

At its first meeting in November 1997, the committee met to review, revise, and select the items for the field test. Committee members were asked to consider whether each item was a reasonable replacement for the original item in terms of the content measured, and whether the answer key or scoring guide for the item was appropriate. A high-quality item needed to be unambiguous in meaning, with appropriate reading demands, clear graphics, and a defensible key or scoring guide. For free-response

○○○

3. See Appendix A for a list of the members of the Subject Matter Item Review Committee.

items, a good scoring guide needed to capture major student responses with a clear distinction between score points. The committee review resulted in a number of improvements to both the items and scoring guides.

Selecting items for the field test also demanded the expertise the committee brought to the task. Although it would have been desirable, the time available in the field test precluded piloting two candidate replacement items for every TIMSS 1995 released item. It was, therefore, necessary to distinguish between proposed items that were almost certain to be effective replacements ("preferred" items), and less certain replacements ("alternate" items). For every item released in 1995, one preferred replacement item was selected to be field-tested. In addition, for about 40% of the released items, a second alternate item was field-tested in case the preferred replacement did not perform well. The judgment of the committee was important in identifying items most likely to be effective replacements and those for which alternates should also be field-tested.

3.3 Field Test

A total of 277 potential replacement items was selected for the field test, including 190 preferred replacements and 87 alternates. These items were organized into five booklets and administered to approximately 200 students in each of 31 countries.[4] The following sections describe the item analyses of results from the field test and the process used to select items for the main survey based on these results.

3.3.1 Field-Test Item Analyses

International item analysis of results from the field test was used to inform the review and selection of mathematics and science items for the main survey. Item statistics were computed to determine the difficulty of each item, how well items discriminated between high- and low-performing students, the reliability of the scoring of free-response items, and whether there were any biases towards or against any particular country, or in favor of boys or girls. These statistics also included the distributions of responses across multiple-choice response options or across the diagnostic response codes for the free-response items. The results of these analyses were summarized in a series of data almanacs that were used to review the field test results.

○○○

4. See Chapter 6 of this report for more information about the field test.

International Study Center Review: Field-test item statistics were reviewed in several phases. By June 19, 1998, preliminary field-test results for 12 countries were analyzed as a trial run. The International Study Center staff reviewed the preliminary field-test data results for each field-test item in both mathematics and science. A second preliminary analysis for 20 countries was completed for review, July 1-2, 1998. The results were further 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 also 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 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.

3.3.2 Selection of Items for the Main Survey

The results from the field test indicated that the pool of replacement items was of high quality. Of the 277 field test-items, 202 were selected for the main survey.[5] Some 80% of the mathematics items selected were used in the main survey without change, and only minor revisions were made in the others. Similarly, 75% of the science field-test items selected were essentially unchanged in the main survey. Revisions made included improving the clarity and print quality of graphics and drawings, clarifying item stems, and revising distracters that were selected by very low percentages of students.

3.3.3 Revising the Scoring Guides

The TIMSS International Study Center used information collected in the field test to make a number of revisions to the scoring guides. Although analyses of the reliability of the free-response scoring in the field test showed substantial agreement between scorers in each country, they also identified some scoring guides that needed revision and areas where improvements were desirable. Revisions to the scoring guides included:

- deleting categories with very few responses
- adding categories with very frequent responses as reported by the NRCs
- clarifying or sometimes combining less reliable categories and
- including additional international examples of student responses supplied by NRCs to illustrate the various diagnostic codes.

Particular attention was given to the number of score points awarded to each item or part of an item, and to ways of improving scoring reliability. Consistent with the approach used in TIMSS 1995, some free-response items were awarded 1 point, others 2 points, and some had more than one part, each worth 1 or 2 points. In general, 1 point was allocated for short-answer items (essentially scored correct or incorrect) that required students to provide a brief response to a question. In mathematics, these questions usually called for a numerical result. In science, the 1-point items usually required a short explanation or factual

○○○

5. Nearly all items selected for the main survey had international mean discrimination indices above 0.3.

description in one or two sentences. In both subjects, 2-point items were those judged to demand more than a numerical response or a short written response. In mathematics, students were asked to show their work or explain their methods, and these responses were taken into account in scoring their correctness. In science, the 2-point items required a fuller explanation demonstrating knowledge of science concepts. The distinction between the 1- and 2-point items was sometimes hazy in science, and for some 2-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 2 points, as compared with those meriting only 1 point. The generalized scoring guides for mathematics are presented in Exhibit 3.4 and those for science in Exhibit 3.5.

Exhibit 3.4 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 • May contain an incorrect solution but applies a mathematically appropriate process
0 Points: A zero-point response is completely incorrect, 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.

Exhibit 3.5 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 were responsive to their suggestions. A few last suggestions were made before the scoring guides were prepared for use in training in the Southern Hemisphere countries in Wellington, New Zealand, in October 1998. During this first scoring training session, a few additional revisions were made. These were incorporated into the final version of the TIMSS 1999 scoring guides used during the scoring training for the Northern Hemisphere countries in February 1999.

3.4 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 scor-

ers, 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 the within-country reliability studies.

Training procedures for the scoring of free-response items in TIMSS 1999 were based on the same "train-the-trainers" approach that had produced highly reliable scores in TIMSS 1995 (see Mullis and 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 training sessions, the general TIMSS 1999 scoring approach was reviewed. Participants then were trained on a subset of the mathematics and science free-response items that were selected to represent 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 general procedures were followed for each item:

- Participants read the item and its scoring guide
- Trainers discussed the rationale and methodology 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 opportunity to 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 the following materials: scoring guides, manuals, and packets of example and practice papers for each of the items covered in the training. The training teams emphasized the need for the NRCs to prepare comparable training materials for training in their own country, including all of the free-response items rather than only the sample of items 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.

For the field test, scoring training was conducted for 10 mathematics items and 12 science items. At the Berlin NRC meeting, NRCs and/or their scoring coordinators participated in a two-day training session for scoring these items. Using a round-robin scheme, half of the NRCs were trained first on mathematics items and then on science items, while the other half were trained first on science items and then on mathematics items. The training was provided by the subject-area coordinators and consultants with support from International Study Center staff. During the field-test training sessions, the NRCs made many good suggestions for improving the scoring guides in both content and clarity. The revisions were made before the final field-test scoring guides were assembled into the final manual and distributed to the countries participating in the TIMSS 1999 field test.

The experience gained from the field test was also used to inform the design of the free-response scoring training sessions for the main survey. After the field-test training, both the training staff and NRCs indicated that additional training time would be desirable, particularly for the science items. Therefore, the two-day training format used in the field test (one day for mathematics and one day for science) was expanded to a three-day session, allotting one day for mathematics and two days for science. This additional session permitted training on a total of 26 free-response items, 7 in mathematics and 19 in science. These 26 items represented nearly all of those identified in the field test as being most problematic to score. Feedback from NRCs and review of the field-test scoring reliability results were essential in identifying the items to use in the training. In addition, an international set of student papers from the field test was collected from NRCs for use in the training, giving a broader range of experiences with the types of responses and student language encountered.

Two scoring training sessions were conducted for the main survey. The first was held in October, 1998, for scoring trainers for countries (mainly Southern Hemisphere countries) where the TIMSS 1999 tests would be administered near the end of 1998. The second was held in February, 1999, for countries where the

tests would be administered around April, 1999. In contrast to the field test, all NRCs and scoring coordinators participated as a single group. Scoring guides used for the main survey sessions reflected refinements made in light of field test data and comments from national research coordinators.

3.5 Main Survey Test Design

The item development, review, and field test process achieved the desired goal of replacing the TIMSS 1995 items released to the public with new items that had similar characteristics. For both mathematics and science, coverage by content area reporting category in TIMSS 1999 was very similar to that in TIMSS 1995. TIMSS 1999 was modified in some respects, however, in order to improve the stability of trend comparisons. In mathematics, TIMSS 1995 had six reporting categories, including *Proportionality*, with only 11 items classified in this content area. For TIMSS 1999 reporting, these items were allocated to other content categories, mainly *Fractions and Number Sense*. In TIMSS 1995, there were five science reporting categories. *Environmental Issues and the Nature of Science* was included as a combined reporting category, with 14 items. For TIMSS 1999, an additional 11 items were developed, permitting the reporting of achievement results separately for the content areas of *Environmental and Resource Issues* and *Scientific Inquiry and the Nature of Science*.

Exhibits 3.6 and 3.7 show the number of items by item type and the associated maximum number of score points for each of the content-based reporting categories for the TIMSS 1999 test. Since some of the free-response items were evaluated for partial credit and were awarded a maximum of two points, the number of score points exceeds the number of items.

Exhibit 3.6 Number of Mathematics Test Items and Score Points by Type and Reporting Category - TIMSS 1999

Reporting Category	Item Type			Number of Items	Score Points
	Multiple-Choice	Short-Answer	Extended-Response		
Fractions and Number Sense	47	11	3	61	62
Measurement	15	4	5	24	26
Data Representation, Analysis and Probability	19	1	1	21	22
Geometry	20	1	-	21	21
Algebra	24	4	7	35	38
Total	125	21	16	162	169

Exhibit 3.7 Number of Science Test Items and Score Points by Type and Reporting Category- TIMSS 1999

Reporting Category	Item Type			Number of Items	Score Points
	Multiple-Choice	Short-Answer	Extended-Response		
Earth Science	17	4	1	22	23
Life Science	28	7	5	40	42
Physics	28	11	-	39	39
Chemistry	15	2	3	20	22
Environmental and Resource Issues	7	2	4	13	14
Scientific Inquiry and the Nature of Science	9	2	1	12	13
Total	104	28	14	146	153

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 used for the original TIMSS study. Assignment to item clusters generally followed the original TIMSS design, with most of the 1999 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 into 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 3.8 and 3.9 present the distribution of items in each content area across the eight test booklets for mathematics and science, respectively.

Exhibit 3.8 Number of Mathematics Items in Each Booklet by Subject Matter Content Category - TIMSS 1999

Content Category	Booklet							
	1	2	3	4	5	6	7	8
Fractions and Number Sense	16	12	15	12	15	12	14	18
Measurement	9	5	9	4	7	4	3	4
Data Representation, Analysis, and Probability	5	4	4	6	7	6	7	5
Geometry	5	6	6	3	6	4	5	5
Algebra	10	6	8	9	8	7	10	9
Total	45	33	42	34	43	33	39	41

Exhibit 3.9 Number of Science Items in Each Booklet by Subject Matter Content Category - TIMSS 1999

Content Category	Booklet							
	1	2	3	4	5	6	7	8
Earth Science	7	7	6	6	5	6	8	6
Life Sciences	8	10	9	14	7	12	8	9
Physics	12	12	10	10	9	11	9	11
Chemistry	3	4	4	4	5	9	4	4
Environmental and Resource Issues	3	8	3	3	3	3	7	5
Scientific Inquiry and the Nature of Science	2	2	2	2	1	2	2	2
Total	35	43	34	39	30	43	38	37

The corresponding maximum number of score points in each booklet by mathematics and science reporting categories is shown in Exhibits 3.10 and 3.11.

Exhibit 3.10 Maximum Number of Mathematics Score Points in Each Booklet by Subject Matter Content Category - TIMSS 1999

Content Category	Booklet							
	1	2	3	4	5	6	7	8
Fractions and Number Sense	16	12	16	12	16	12	14	18
Measurement	9	5	11	4	9	4	3	4
Data Representation, Analysis and Probability	5	4	4	6	8	6	8	5
Geometry	5	6	6	3	5	4	5	5
Algebra	12	6	9	9	9	7	11	9
Total	47	33	46	34	47	33	41	41

Exhibit 3.11 Maximum Number of Science Score Points in Each Booklet by Subject Matter Content Category - TIMSS 1999

Content Category	Booklet							
	1	2	3	4	5	6	7	8
Earth Science	7	7	6	6	5	7	8	6
Life Science	8	10	9	15	7	13	8	8
Physics	12	12	10	10	9	11	9	11
Chemistry	3	4	4	4	6	8	4	4
Environmental and Resource Issues	3	6	3	3	3	3	5	4
Scientific Inquiry and the Nature of Science	2	3	2	3	1	2	2	2
Total	35	42	34	41	31	44	36	35

References

Adams, R.J., & Gonzalez, E.J. (1996). The TIMSS Test Design, in M.O. Martin & D.L. Kelly (Eds.), *Third International Mathematics and Science Study Technical Report Volume I: Design and Development.* Chestnut Hill, MA: Boston College.

Lie, S., Taylor, A., & Harmon, M. (1996). Scoring Techniques and Criteria, in M.O. Martin & D.L. Kelly (Eds.), *Third International Mathematics and Science Study Technical Report Volume I: Design and Development.* Chestnut Hill, MA: Boston College.

Mullis, I.V.S., & Smith, T.A. (1996). Quality Control Steps for Free-Response Scoring, in M.O. Martin and I.V.S. Mullis (Eds.), *Third International Mathematics and Science Study: Quality Assurance in Data Collection.* Chestnut Hill, MA: Boston College.

Robitaille, D.F., Schmidt, W.H., Raizen, S., McKnight, C., Britton, E., & Nicol, C. (1993). *Curriculum Frameworks for Mathematics and Science.* Vancouver: Pacific Educational Press.