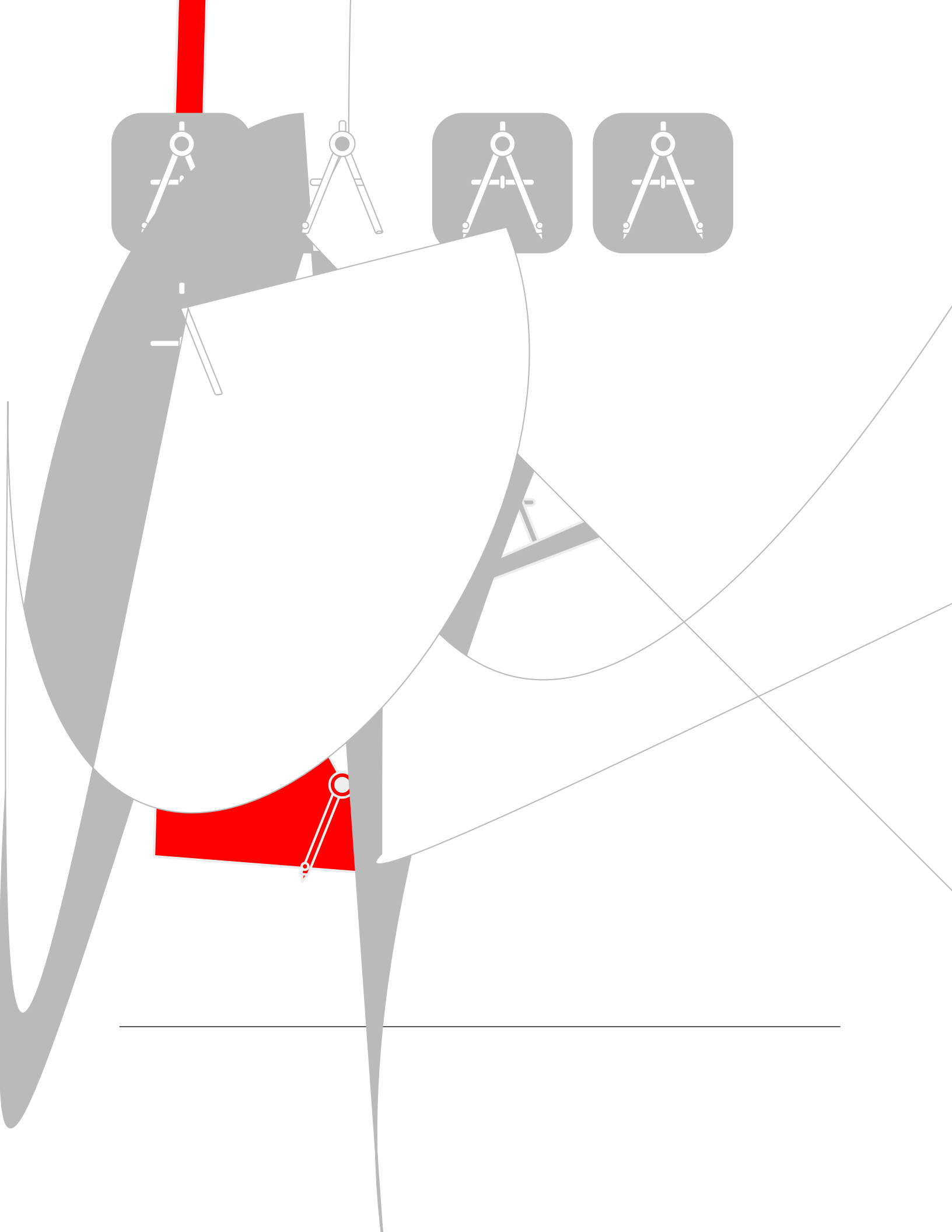

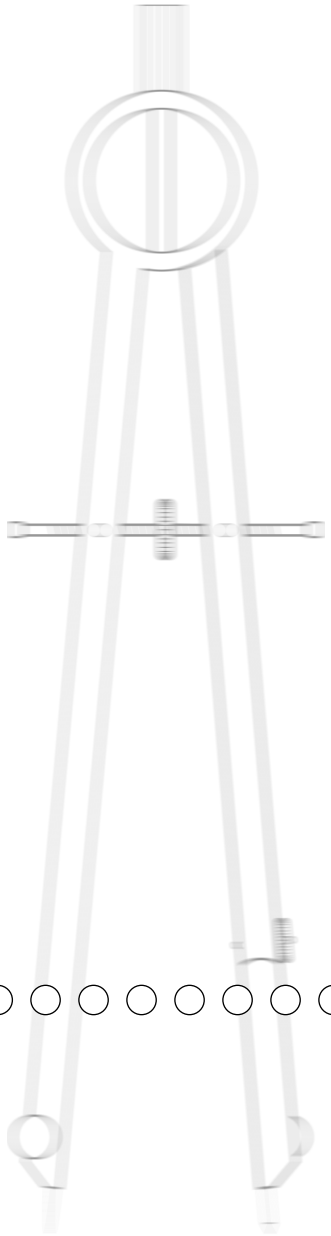

6 TIMSS Field Test

Kathleen M. O'Connor

6.1 Overview

Although TIMSS 1995 set new standards for quality in data collection in international studies, there were areas in sampling, translation verification, and survey operations where improvements could be made. TIMSS 1999 built on the tradition of high-quality data collection established by TIMSS 1995, and sought to achieve even greater compliance with international procedures among participating countries. An essential step towards achieving this goal was to conduct a full-scale field test of all instruments and operational procedures under conditions approximating as closely as possible those present during the main survey data collection. By encouraging countries to participate fully in the field test, TIMSS 1999 sought to anticipate and eliminate as many potential survey operations problems as possible during the main survey.

In addition to improving survey operations in TIMSS 1999, the field test was crucial to the development of the instruments for the main survey, particularly the achievement tests. As part of the dissemination of the TIMSS 1995 results, about two-thirds of the achievement items were released into the public domain so that readers of the international reports could develop a good appreciation of the nature and coverage of the tests. In planning for TIMSS 1999, therefore, a major task was to replace the released items with newly developed items that were comparable in terms of content, format, and difficulty.[1] An essential aspect of this item development was that potential replacement items be tried out in schools, so that the psychometric characteristics of these items could be thoroughly investigated, and the best possible replacements selected. Although the school, teacher, and student questionnaires were adapted from TIMSS 1995 without major redesign, there were a number of additions and refinements made for TIMSS 1999, and it was necessary to field test these as well.[2]

○○○

1. See Chapter 3 for a description of the TIMSS test development.
2. See Chapter 4 for a description of the TIMSS questionnaires.

Therefore, the field test had two major purposes: (i) To ensure that all survey operations procedures could be implemented efficiently in all participating countries; and (ii) To ensure that the items on the achievement tests and questionnaires were appropriate for the measurement purposes for which they were designed.

6.2 Design of the Field Test

The field test was designed to be an integral part of the TIMSS 1999 study, and to mirror as closely as possible the activities of the main survey. The major parameters of the field test were as follows:

- The field test in each country was to be conducted in a random sample of 25 schools. These schools were to be sampled in conjunction with the sampling for the main survey to avoid overlap. Approximately 40 eighth-grade students were to be sampled from each school. This could be accomplished by sampling one or more eighth-grade mathematics classes, or by sampling eighth-grade students directly within the school.
- The achievement test items were grouped into 5 distinct booklets for the field test (there were 8 booklets in the main survey). Booklets were to be distributed among students in sampled classes using the procedure prescribed for the main survey. Each student was to respond to just one booklet. Approximately 200 students per booklet (1000 students in total) were required for each country. The amount of testing time was 90 minutes, the same as for the main survey.
- Each student also was asked to respond to a student questionnaire. There was a school questionnaire for the school principal, as well as teacher questionnaires for the mathematics and science teachers of the sampled students.
- Participating countries were responsible for translating the survey instruments into the local language of instruction. Completed translations were sent to the IEA Secretariat for verification by the translation verification company.
- Participants were expected to comply with all internationally agreed upon procedures for instrument translation and adaptation, test administration, scoring, data entry, and data processing. Field-test data were to be sent to the IEA Data Processing Center (DPC) in Hamburg, Germany.

6.2.1 Survey Operations

In the operational arena there were three general areas where TIMSS 1999 aimed to make improvements and for which the field test was a vital component: (i) Sampling operations, (ii) Translation verification, and (iii) General field operations, including data processing procedures.

The plans to improve sampling operations for TIMSS 1999 involved an emphasis on the early development of national sampling plans for the main survey, and the integration of the sampling plan for the field test with that for the main survey. Getting started early with sampling ensured more time to deal with unexpected problems, and more time to secure high participation rates from schools and teachers. Integrating the sampling plans for the field test and main survey ensured that the sampling activities for the main survey were tried out in a realistic situation, and also ensured that the field-test samples were, in most instances, properly constituted random samples of the target population in each country.

Since the international version of the survey instruments had to be translated by each country into the local language before data could be collected, the translation process was of central importance. Recognizing this, TIMSS 1995 instituted a rigorous procedure for verifying the translations of the achievement tests produced by each participant. TIMSS 1999 further expanded the translation verification process to include not only the achievement tests but also the school, teacher, and student questionnaires. Furthermore, to achieve the best possible verification, TIMSS 1999 retained an internationally renowned translation company to conduct the translation verification. Given the expanded nature of the verification process and the need to work with an unfamiliar translation company, it was beneficial to work through all procedures prior to the main survey.

Although many of the TIMSS 1999 countries had taken part in TIMSS 1995 and were already familiar with the operational procedures, there were also many that had no previous TIMSS experience (or no large-scale assessment experience at all). Even among those with previous experience, there were countries that were unable, for one reason or another, to comply fully with the prescribed 1995 TIMSS procedures, either because of the enormous burden of the data collection, or because they found the procedures unduly complex.

Considerable effort was expended in simplifying and streamlining the survey operations for TIMSS 1999 so that countries would find the data collection easier and more efficient this time around. It was necessary that participants worked through the 1999 procedures in a realistic field test to ensure that they were feasible and effective, and that all participants were comfortable in using them.

6.2.2 Achievement Tests

As described in Chapter 3, approximately one-third of the 1995 achievement items were kept secure for use in 1999, and the remaining items were released for public use. The secure items were those in item clusters A through H, and the released items were in item clusters I through Z. The 1999 test development effort was designed to replace the released items with items of similar content coverage and expectations for student performance. As an integral part of the test development process, the field test was designed to try out the replacement items with representative samples of eighth-grade students before finalizing the tests for the main survey data collection.

For field-test purposes, the 1995 clusters I through Z were replaced with 1999 replacement clusters rI through rZ. As in 1995, clusters were classified as either *breadth clusters* or *free-response clusters*. Breadth clusters (rI - rR) consisted of multiple-choice and short-answer questions, designed to ensure broad subject-matter coverage; free-response clusters (rS through rZ) consisted largely of extended-response questions. Clusters rI through rZ contained the items considered by the item developers to be most likely to be selected as replacements for the main data collection (these were known as the "preferred" replacement items – see Chapter 3).

In addition to field testing one preferred replacement item for each released item, the field-test design provided for testing a set of alternate items. The alternate items were available for use as replacement items in the main survey in cases where the preferred replacement item did not perform well in the field test. Approximately 40% of the preferred replacement items had an alternate item in the field test. Alternate items were placed in item clusters a01 through a12. Clusters a01 through a06 were

mathematics or science breadth clusters; clusters a07 through a12 were mainly free-response clusters. Each field-test item was assigned to one cluster only. Exhibit 6.1 summarizes the organization of preferred and alternate field-test items into clusters.

Exhibit 6.1 TIMSS 1999 Field-Test Clusters

Selection	Cluster ID	Cluster Type	Subject(s)	Time per Cluster
Preferred	rI - rR	breadth	mathematics/science	22 minutes
	rS - rV	free-response	mathematics	10 minutes
	rW - rZ	free-response	science	10 minutes
Alternate	a01, a03, a05	breadth	science	16 minutes
	a02, a04, a06	breadth	mathematics	16, 16, 10 min. respectively
	a07, a09, a11	free-response	science	10 minutes
	a08, a10, a12	free-response	mathematics	10 minutes

Since the field-test clusters contained far too many items for any one student to answer in a single testing session, it was necessary to package the clusters in a way that kept student response burden to a minimum, while keeping the testing conditions as close as possible to those in the main data collection. Accordingly, in the field test the item clusters were distributed across five booklets, with each student responding to one booklet only. Each booklet contained a unique set of multiple-choice and free-response items in mathematics and science, requiring 90 minutes of testing time. Exhibit 6.2 details the cluster allocation scheme for the five field-test booklets.

Exhibit 6.2 TIMSS 1999 Field-Test Booklets

	Contents	Booklet					
Time			**1**	**2**	**3**	**4**	**5**
48 min	Breadth Cluster	22 min	rI	rJ	rK	rL	rM
	Alternate Breadth Cluster Free-Response Cluster	16 min 10 min	a01 rT	a02 a09	a04 a07	a03 rY	a05 rV
Break							
42 min	Breadth Cluster	22 min	rN	rO	rP	rQ	rR
	Free-Response Cluster	10 min	rS	rW	rX	rZ	rU
	Alternate Breadth/Free-Response Cluster	10 min	a06	a08	a12	a10	a11

6.2.3 Questionnaires

As described in Chapter 4, the school, teacher, and student questionnaires used in 1999 were modified versions of the TIMSS 1995 questionnaires. While most of the questions were the same in both assessments, some questions from 1995 were eliminated, and some new questions were introduced in 1999, either as replacements for eliminated items or to provide extra information in areas considered important to the study. In general, every effort was made to shorten and streamline the questionnaires in order to reduce the burden on respondents. Since questionnaire length was particularly problematic for the 1995 teacher questionnaires, the teacher questionnaires for TIMSS 1999 were significantly reduced in length. The field test included full tryouts of each questionnaire, as well as the achievement tests.

6.3 Field Test Participation

Participants were required to sample enough schools, classrooms, and students to ensure that a minimum of 1,000 students would be included in the field test in each country. In general, this meant sampling 25 schools, with two classrooms per school, and testing all students in the sampled classrooms. Variations on the standard design, such as sampling more schools, or sampling more than two classrooms in some schools, were required in some countries when the standard design did not provide participants with the required minimum of 1,000 tested students.

The principal sampling objective for the TIMSS-R field test was to replicate as much as possible all the sampling activities the participants would encounter in the main survey. This included selecting probability samples of schools as well as probability samples of classrooms and students within schools.

6.3.1 Sampling Schools

The selection of the school sample for the field-test was integrated with the selection of the school sample for the main survey. This meant that both school samples were to be drawn simultaneously, thereby avoiding the possibility of the same schools appearing in both samples. This approach had the added benefit that the field-test samples were probability samples, and not convenience samples as is often the case in field trials.

The basic school sampling design for the field test consisted of drawing a sample of 25 schools, using the PPS systematic sampling method. Explicit and implicit stratification was used to optimize the reliability and representation of the resulting samples. For each sampled school, a replacement school was identified a priori, should the sampled school not participate.

6.3.2 Sampling Classrooms

Within each sampled school, all eligible classrooms in the appropriate target grade were listed, and two classrooms were drawn at random from the list (with equal probabilities). Although the sampling design for the main survey was to sample a single classroom in each school, participants in the field test were encouraged to sample two classrooms per school in order to achieve the requisite number of students while keeping the number of schools to a minimum. Some participants preferred to select a single classroom per school, and consequently selected more than the minimum number of schools. There were also some participants who sampled more than two classes per school because they were unable to sample more than 25 schools and class sizes were small or some of the sampled schools had only one eligible classroom. All participants tested all students in the sampled classrooms. Although it was permissible to sub-sample students within classrooms, none of the participants chose to do this.

6.3.3 Field Test Sample Size

Exhibit 6.3 provides summary statistics on the school and student samples for 29 of the 31 countries that participated in the field test.[3] Altogether, field-test data were available for 29,236 students from 724 schools in 29 countries.

Exhibit 6.3 Number of Schools and Students that Participated in the Field Test

Country	Schools sampled	Participating schools			Non-participating schools	Students in sampled classrooms	Student status			
		Sampled	Replaced	Other[a]			Removed	Excluded	Absent	Tested
Australia	—	—	—	19	—	932	12	0	148	772
Belgium (Flemish)	25	21	2	—	2	873	0	0	21	852
Bulgaria	25	25	—	—	—	934	0	0	0	934
Canada	50	47	—	—	3	1237	0	0	48	1189
Chile	25	25	—	—	—	1725	0	3	2	1720
Chinese Taipei	25	25	—	—	—	1204	4	8	11	1181
Czech Republic	25	24	—	—	1	1182	6	0	89	1087
England	25	8	8	7	9	1078	0	0	47	1031
Finland	25	20	4	—	1	931	5	0	75	851
Indonesia	25	NA	NA	NA	NA	NA	NA	NA	NA	NA
Iran Islamic Rep.	34	34	—	—	—	1069	6	0	27	1036
Italy	30	30	—	—	—	1198	9	22	48	1119
Japan	25	21	3	8	1	1185	0	0	57	1128
Jordan	25	24	1	—	—	998	3	0	3	992
Korea Rep. of	25	25	—	1	—	1103	0	0	27	1076
Latvia	25	24	—	—	1	944	0	0	0	944
Lithuania	—	—	—	20	—	691	0	0	54	637
Macedonia Rep. of	25	22	2	1	1	1218	2	0	23	1193
Malaysia	25	22	3	—	—	1089	7	0	16	1066
Morocco	35	35	—	—	—	1176	0	27	25	1124
Netherlands	25	14	5	—	6	907	0	0	49	858
New Zealand	—	—	—	27	—	1116	28	0	52	1036
Philippines	25	NA	NA	NA	NA	NA	NA	NA	NA	NA
Romania	25	24	1	—	—	1060	7	0	19	1034
Russian Federation	24	24	—	—	—	1128	0	0	1	1127
Singapore	15	15	—	—	—	980	14	0	73	893
Slovak Republic	25	14	5	4	6	1090	54	5	55	976
Slovenia	25	22	3	—	—	1122	2	0	57	1063
South Africa	25	23	2	—	—	1016	0	0	64	952
Tunisia	25	22	3	0	0	1009	0	0	48	961
United States	—	—	—	24	—	1280	14	10	80	1176
Total	713	590	42	92	31	30543	161	75	1071	29236

a. All schools that participated in the field test but were not drawn using probability-sampling methods are included.

○○○

3. Indonesia and the Philippines conducted the field test, but their data were not available in time for the field test data analyses.

6.4 Field Test Analysis

After data from the field test had been verified and transformed into the international format, they were sent to the International Study Center at Boston College for further analysis. The purpose of this analysis was to establish empirically the psychometric characteristics of the achievement and questionnaire items so as to inform the item review and selection process. The analyses included the computation of achievement scores as well as an array of descriptive and diagnostic statistics for each item from every country. These computations were used to determine the difficulty of the achievement items, how well items discriminated between high- and low-performing students, and whether there were any biases towards or against any particular country, or in favor of boys or girls. The statistics also described the distribution of responses to the questions in the background questionnaires, and allowed for an analysis of the relationship between questionnaire responses and student achievement in mathematics and science. The results of these analyses were summarized in a series of data almanacs that presented the key statistics for each item from each country. These almanacs were the basic data summaries that were used by the staff of the International Study Center, by expert committees, and by National Research Coordinators and their advisers in assessing the quality of the field-test instruments and in making suggestions for the main survey.

Several data almanacs were produced, summarizing responses to the achievement items as well as to the student, teacher, and school questionnaires. Six different almanacs were produced for mathematics and another six for science, giving twelve different almanacs in total.

Three types of data almanacs were generated for use during achievement item analyses. These almanacs contained basic item analysis statistics for the mathematics and science achievement items for each country, detailed information on the distributions of multiple-choice item response options chosen or free-response item response types given, and information regarding item-by-country interactions for each item. The almanacs containing the item analysis data are listed below:

- International Item Statistics Almanacs - Mathematics and Science
- Percent of Responses by Item Category Almanacs - Mathematics and Science

- Item-by-Country Interaction Almanacs - Mathematics and Science

Three other types of data almanacs were generated to help review the results of the student, teacher, and school background questionnaires. These almanacs displayed descriptive statistics for each questionnaire, including the distributions of responses to questionnaire items and the relationship between student achievement in mathematics and science and the response values for categorical questions. The almanacs including background questionnaire data are listed below:

- Student Background Questionnaire Almanacs - Mathematics and Science
- Teacher Background Questionnaire Almanacs - Mathematics and Science
- School Background Questionnaire Almanacs - Mathematics and Science

The summary statistics presented in the data almanacs were computed through a collaborative effort by the IEA Data Processing Center, Educational Testing Service, and the TIMSS International Study Center. Item statistics, with the exception of indices of differential item functioning (DIF), were the responsibility of the IEA Data Processing Center. The DIF statistics were computed by Educational Testing Service. Summary statistics for the school, teacher, and student questionnaires were computed at the International Study Center.

Since not all of the data were available for analysis at the same time, it was useful to produce draft almanacs using just a few countries initially, both to refine the almanac production procedure and to get started on the item review process. Accordingly, after the data from a subset of 12 countries were processed a set of preliminary almanacs was created and reviewed by the staff at the International Study Center. Shortly afterwards, a second set of almanacs was created with data from 20 countries. This allowed the International Study Center staff to conduct a review of the results from a majority of participating countries before meeting with the advisory committees. A third version of the almanacs, containing data from 21 countries, was created to be reviewed by the Subject Matter Item Replacement Committee and the Questionnaire Item Review Committee, both of which met in London in July 1998. The recommendations of the review committees were based

on the results in this third version. Following the meetings of the review committees and prior to the NRC meeting in Boston in August 1998, a final almanac based on data from 29 countries was created for review by the National Research Coordinators. This final version of the almanacs was reviewed by NRCs at the Boston meeting, and informed their deliberations at that meeting.

6.4.1 Operational Improvements

The level of participation in the TIMSS 1999 field test and the compliance with sampling procedures were remarkably high. There was a high level of participation by sampled schools, and classroom sampling and student tracking were judged to be flawless on the basis of the extensive documentation received from the participants.

No major problems with the within-school sampling software were reported by the countries participating in the field test. Therefore, the general structure of the program and its procedures were kept for the main survey, although a number of improvements were made to the user interface and to the database structure. The user interface and menu structures were modified to make them easier to understand and to use. Changes were undertaken to the database structure, so that all school, class, student, and teacher data could be found within one file. Checks on the hierarchical identification system, on duplicate identification numbers, and on the correctness of data in the files were also improved based on the field test experience.

6.4.2 Translation Verification

The translation verification process for the field test was very successful, and revealed a high standard of translation in most countries. There was, however, considerable variability between countries in how well they accomplished the task. Although most countries had very few translation deviations reported, some had quite a lot. A substantial proportion of the serious deviations reported related to the layout of the cover pages, instructions for students, and headers and footers, and would have had little impact on student results. Of the serious deviations that were directly related to achievement items, about 60% were attributable to just six countries. Five countries had no serious deviations, and 15 countries had very few, accounting for less than 9% of the serious deviations. For all countries, translation problems identified through the field-test translation verification were addressed prior to the main data collection.

The information from each translation verification report form was entered into an electronic database, which was used to inform the item review procedure that followed the analysis of the field test data. Where a country had poor statistics for an item, the verification report for that item was examined to identify any translation problems requiring correction before the main data collection.

As a result of the field-test experience, a number of minor modifications were made to the translation verification procedure, including provision for more direct links between NRCs and Berlitz to speed up communications, and the use of email for information exchange.

6.4.3 Field Operations Procedures

The field test led to some reconfiguration and consolidation of the manuals related to survey operations. Aspects of the *Manual for Checking, Scoring, and Entering the TIMSS 1999 Data* were incorporated in the *Survey Operations Manual* (TIMSS, 1998a) and the *Manual for Entering the TIMSS-R Data (*TIMSS, 1998b). This change reduced the number of manuals necessary for survey operations, while retaining all information necessary for successful project completion. A new manual for countries to use in conducting within-country quality control procedures was created. Entitled the *Manual for National Quality Control Observers* (TIMSS, 1998c), this manual draws on the procedures used in the international quality control activities. In general, NRCs found the survey operations manuals to be clear and helpful in documenting the TIMSS 1999 procedures.

6.4.4 Data Processing Procedures

In general, the data processing procedures at the IEA Data Processing Center (DPC) for the TIMSS 1999 field test were similar to those developed for TIMSS 1995, and were therefore tried and trusted. The field test confirmed that they were effective and appropriate, and consequently no major changes were planned for the main survey.

Some procedures designed to improve communication between the International Study Center and the IEA DPC were implemented in the field test. A new policy of sending data updates to the International Study Center more frequently than before was found to be very useful and helped improve the flow of work. Providing countries with software that enabled them to detect

and correct any problems in their data before sending the data files to the IEA DPC was very successful. Solving identification problems immediately after entering the data and when the testing materials and relevant staff were accessible took significantly less time than the previous procedure of conducting initial data checks at the DPC.

Some countries requested training in the use of the DataEntryManager (DEM) to improve their familiarity with the program, to learn to benefit from its special checking features, to learn how to adapt it to their local needs, and to help convince their data-entry staff to use the program. Accordingly, a special DEM training session was conducted prior to the TIMSS 1999 main survey.

6.5 Summary

The field test for TIMSS 1999 was highly successful at meeting the twin goals of finalizing the instrument development and improving survey operations. Achievement items and questionnaires were developed and revised based on the field-test data, ensuring sound instruments for data collection in the main survey. Conducting a full-scale field test in 31 participating countries provided an opportunity for improving survey operations procedures and identifying potential problems. Based on the results and experience gained in the field test, the TIMSS 1999 participants were able to proceed with confidence into the main-survey data collection.

References

TIMSS (1998a). *Survey Operations Manual* (Doc. Ref. No. 98-0026). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

TIMSS (1998b). *Manual for Entering the TIMSS-R Data* (Doc. Ref. No. 98-0028). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

TIMSS (1998c). *Manual for National Quality Control Observers* (Doc. Ref. No.98-0044). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.