TIMSS 1999 Technical Report

edited by:

Michael O. Martin
Kelvin D. Gregory
Steven E. Stemler

with contributors:

Pierre Foy
Robert Garden
Eugenio J. Gonzalez
Dirk Hastedt
Marc Joncas
Edward Kulik
Barbara Malak
Ina V.S. Mullis
Kathleen M. O'Connor
Teresa A. Smith
Kentaro Yamamoto

International Study Center Boston College • Chestnut Hill, Massachusetts

TIMSS 1999 Technical Report / edited by
Michael O. Martin, Kelvin D. Gregory and
Steven E. Stemler

Publisher: International Study Center
Lynch School of Education
Boston College

Library of Congress Catalog Card Number:
00-111879

ISBN 1-889938-17-3

For more information about TIMSS contact:

The International Study Center
Lynch School of Education
Boston College
Chestnut Hill, MA 02467
United States

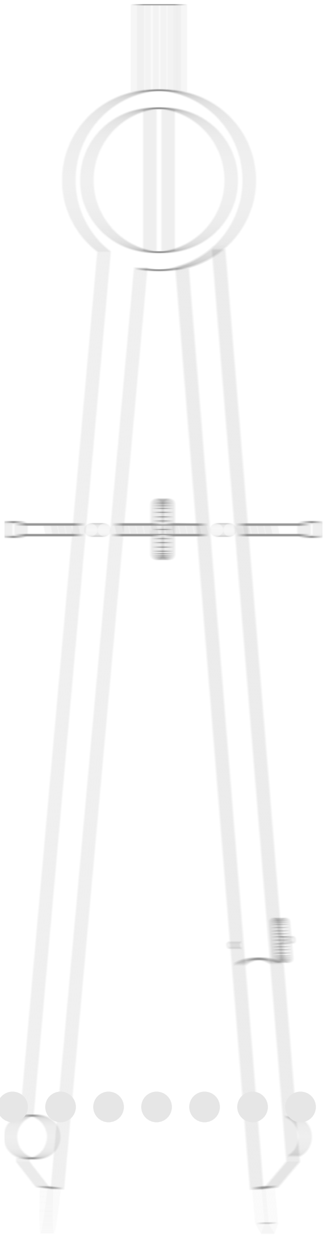
For information on ordering this report, write to the
above address or call +1-617-552-1600

This report also is available on the
World Wide Web: http://www.timss.bc.edu

Funding for the international coordination of TIMSS 1999 was provided by the National Center for Education Statistics of the U.S. Department of Education, the U.S. National Science Foundation, the World Bank, and participating countries. Each participating country was responsible for funding national project costs and implementing TIMSS 1999 in accordance with the international procedures.

Boston College is an equal opportunity,
affirmative action employer.

Printed and bound in the United States

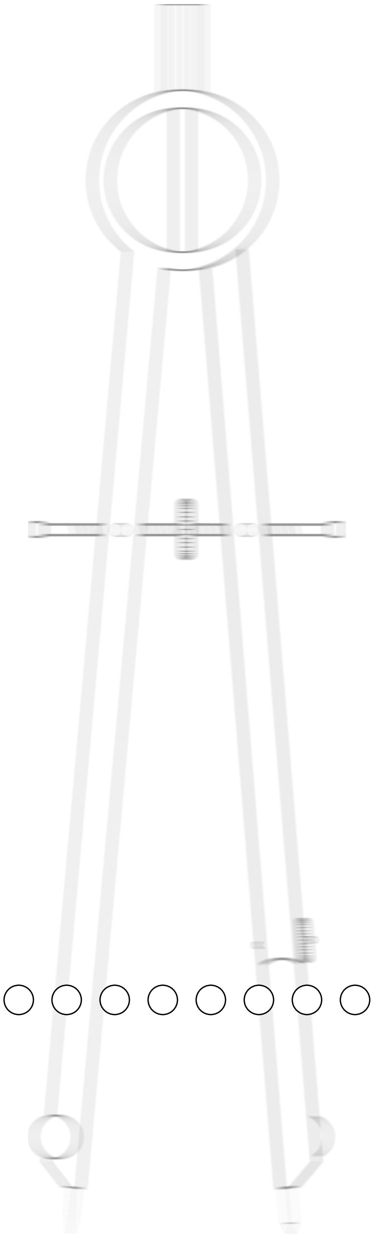


Contents

1

TIMSS 1999: an Overview

Michael O. Martin
Ina V.S. Mullis



1 TIMSS 1999: an Overview

Michael O. Martin
Ina V.S. Mullis

1.1 Introduction

TIMSS 1999 represents the continuation of a long series of studies conducted by the International Association for the Evaluation of Educational Achievement (IEA). Since its inception in 1959, the IEA has conducted more than 15 studies of cross-national achievement in the curricular areas of mathematics, science, language, civics, and reading. IEA conducted its First International Science Study (FISS) in 1970-71 and the Second International Science Study (SISS) in 1983-84. The First and Second International Mathematics Studies (FIMS and SIMS) took place in 1964 and 1980-82, respectively. The Third International Mathematics and Science Study (TIMSS), conducted in 1995-1996, was the largest and most complex IEA study to date, and included both mathematics and science at third and fourth grades, seventh and eighth grades, and the final year of secondary school.

In 1999, TIMSS again assessed eighth-grade students in both mathematics and science to measure trends in student achievement since 1995. This study was also known as TIMSS-Repeat, or TIMSS-R.

The results of TIMSS 1999 were published in two companion volumes, *TIMSS 1999 International Mathematics Report* (Mullis, Martin, Gonzalez, Gregory, Garden, O'Connor, Chrostowski, and Smith, 2000) and *TIMSS 1999 International Science Report* (Martin, Mullis, Gonzalez, Gregory, Smith, Chrostowski, Garden, and O'Connor, 2000). This volume, the *TIMSS 1999 Technical Report,* describes the technical aspects of the study and summarizes the main activities involved in the development of the data collection instruments, the data collection itself, and the analysis and reporting of the data.

1.2 Participants in TIMSS 1999

Of the 42 countries that participated in TIMSS[1] at the eighth grade in 1995, 26 availed themselves of the opportunity to measure changes in the achievement of their students by also taking part in 1999 (see Exhibit 1.1). Twelve additional countries participated in 1999, for a total of 38 countries. Of those taking part in 1999, 19 had also participated in 1995 at the fourth grade.[2] Since fourth-grade students in 1995 were in eighth grade in 1999, these countries can compare their eighth-grade performance with their performance at the fourth grade, as well as with the eighth-grade performance of students in other countries.

○○○

[1] Results for 41 countries are reported in the 1995 international reports; Italy also completed the 1995 testing, but too late to be included. It is counted as a 1995 country in this report and included in all trend exhibits in the 1999 international reports. Unweighted data for the Philippines were reported in an appendix to the international reports in 1995. These data were not included in trend exhibits for 1999.

[2] Two of the 19 countries with fourth-grade data from 1995 (Israel and Thailand) did not satisfy guidelines for sampling procedures at the classroom level and were not included in the comparisons for fourth and eighth grade.

Exhibit 1.1 Countries Participants in TIMSS 1999 and 1995

Country	TIMSS 1999	TIMSS 1995 (Grade 8)	TIMSS 1995 (Grade 4)
Australia	●	●	●
Austria		●	●
Belgium (Flemish)	●	●	
Belgium (French)		●	
Bulgaria	●	●	
Canada	●	●	●
Chile	●		
Chinese Taipei	●		
Colombia		●	
Cyprus	●	●	●
Czech Republic	●	●	●
Denmark		●	
England	●	●	●
Finland	●		
France		●	
Germany		●	
Greece		●	●
Hong Kong, SAR	●	●	●
Hungary	●	●	●
Iceland		●	●
Indonesia	●		
Iran, Islamic Republic	●	●	●
Ireland		●	●
Israel	●	●	●
Italy	●	●	●
Japan	●	●	●
Jordan	●		
Korea, Republic of	●	●	●
Kuwait		●	●
Latvia	●	●	●
Lithuania	●	●	
Macedonia, Republic of	●		
Malaysia	●		
Moldova	●		
Morocco	●		
Netherlands	●	●	●
New Zealand	●	●	●
Norway		●	●
Philippines	●		
Portugal		●	●
Romania	●	●	
Russian Federation	●	●	
Scotland		●	●
Singapore	●	●	●
Slovak Republic	●	●	
Slovenia	●	●	●
South Africa	●	●	
Spain		●	
Sweden		●	
Switzerland		●	
Thailand	●	●	●
Tunisia	●		
Turkey	●		
United States	●	●	●

1.3 The Student Population

TIMSS in 1995 had as its target population students enrolled in the two adjacent grades that contained the largest proportion of 13-year-old students at the time of testing, which were seventh- and eighth-grade students in most countries. TIMSS in 1999 used the same definition to identify the target grades, but assessed students in the upper of the two grades only, the eighth grade in most countries.

1.4 Survey Administration Dates

Since school systems in countries in the Northern and Southern Hemispheres do not have the same school year, TIMSS 1999 had to operate on two schedules. The Southern Hemisphere countries administered the survey from September to November, 1998, while the Northern Hemisphere countries did so from February to May, 1999.

1.5 The TIMSS 1999 Assessment Framework

IEA studies have the central aim of measuring student achievement in school subjects, with a view to learning more about its nature and extent and the context in which it occurs. The goal is to isolate the factors directly relating to student learning that can be manipulated through policy changes in, for example, curricular emphasis, allocation of resources, or instructional practices. Clearly, an adequate understanding of the influences on student learning can come only from careful study of the nature of student achievement and the characteristics of the learners themselves, the curriculum they follow, the teaching methods of their teachers, and the resources in their classrooms and their schools. Such school and classroom features are of course embedded in the community and the education system, which in turn are aspects of society in general.

The designers of TIMSS in 1995 chose to focus on curriculum as a broad explanatory factor underlying student achievement (Robitaille and Garden, 1996). From that perspective, curriculum was considered to have three manifestations: what society would like to see taught (the intended curriculum), what is actually taught (the implemented curriculum), and what the students learn (the attained curriculum). This view was first conceptualized for the IEA's Second International Mathematics Study (Travers and Westbury, 1989).

The three aspects of the curriculum bring together three major influences on student achievement. The intended curriculum states society's goals for teaching and learning. These goals reflect the ideals and traditions of the greater society and are constrained by the resources of the education system. The implemented curriculum is what is taught in the classroom. Although presumably inspired by the intended curriculum, actual classroom events are usually determined in large part by the teacher, whose behavior may be greatly influenced by his or her own education, training, and experience, by the nature and organizational structure of the school, by interaction with teaching colleagues, and by the composition of the student body. The attained curriculum is what the students actually learn. Student achievement depends partly on the implemented curriculum and its social and educational context, and to a large extent on the characteristics of individual students, including ability, attitude, interests, and effort.

Since TIMSS 1999 essentially replicated the eighth-grade part of the 1995 study, much of the conceptual underpinning of the 1999 study was derived from the three-strand model of curriculum. The organization and coverage of the intended curriculum were investigated through curriculum questionnaires that were completed by National Research Coordinators (NRCs) and their curriculum advisors. Although more modest in scope than the extensive curriculum analysis component of the 1995 study (Schmidt et al., 1997a; 1997b), the TIMSS 1999 questionnaires yielded valuable information on the curricular intentions of participating countries.

Data on the implemented curriculum were collected as part of the TIMSS 1999 survey of student achievement. Questionnaires completed by the mathematics and science teachers of the students in the survey, and by the principals of their schools, provided information about the topics in mathematics and science that were taught, the instructional methods used in the classroom, the organizational structures that supported teaching, and the factors that were seen to facilitate or inhibit teaching and learning.

The student achievement survey provided data for the study of the attained curriculum. The wide-ranging mathematics and science tests that were administered to nationally representative samples of students provided not only a sound basis for international comparisons of student achievement, but a rich resource for the study of the attained curriculum in each country. Information about students' characteristics, and about their attitudes, beliefs, and experiences, was collected from each participating student. This information was used to identify the student characteristics associated with learning and provide a context for the study of the attained curriculum.

1.6 Developing the TIMSS 1999 Achievement Tests

The TIMSS curriculum framework underlying the mathematics and science tests was developed for TIMSS in 1995 by groups of mathematics educators with input from the TIMSS National Research Coordinators (NRCs). As shown in Exhibit 1.2, the curriculum framework contains three dimensions or aspects. The *content* aspect represents the subject matter content of school mathematics and science. The *performance expectations* aspect describes, in a non-hierarchical way, the many kinds of performance or behavior that might be expected of students in school mathematics and science. The *perspectives* aspect focuses on the development of students' attitudes, interest, and motivation in the subjects. Because the frameworks were developed to include content, performance expectations, and perspectives for the entire span of curricula from the beginning of schooling through the completion of secondary school, not all aspects are reflected in the eighth-grade TIMSS assessment.[3] Working within the framework, mathematics test specifications for TIMSS in 1995 included items representing a wide range of mathematics topics and eliciting a range of skills from the students. The 1995 tests were developed through an international consensus process involving input from experts in mathematics, science, and measurement, ensuring that the tests reflected current thinking and priorities in mathematics and science education.

○○○

[3] The complete TIMSS curriculum frameworks can be found in Robitaille et al. (1993).

Exhibit 1.2 The Three Aspects and Major Categories of the Mathematics and Science Frameworks

Subject	Content	Performance Expectations	Perspectives
Mathematics	Numbers	Knowing	Attitudes
	Measurement	Using Routine Procedures	Careers
	Geometry	Investigating and Problem Solving	Participation
	Proportionality	Mathematical Reasoning	Increasing Interest
	Functions, Relations, and Equations	Communicating	Habits of Mind
	Data Representation		
	Probability and Statistics		
	Elementary Analysis, Validation and Structure		
Science	Earth Science	Understanding	Attitudes
	Life Sciences	Theorizing, Analyzing, and Solving Problems	Careers
	Physical Science	Using Tools, Routine Procedures and Science Processes	Increasing Interest
	History of Science and Technology	Investigating the Natural World	Safety
	Environmental and Resource Issues	Communicating	Habits of Mind
	Nature of Science		
	Science and Other Disciplines		

About one-third of the items in the 1995 assessment were kept secure to measure trends over time; the remaining items were released for public use. An essential part of the development of the 1999 assessment, therefore, was to replace the released items with items of similar content, format, and difficulty. With the assistance of the Science and Mathematics Item Replacement Committee, a group of internationally prominent mathematics and science educators nominated by participating countries to advise on subject matter issues in the assessment, over 300 mathematics and science items were developed as potential replacements. After an extensive process of review and field testing, 114 items were selected as replacements in the 1999 mathematics assessment.

Exhibit 1.3 presents the five content areas included in the 1999 mathematics test and the six content areas in science, together with the number of items and score points in each area. Distributions are also included for the five performance categories derived from the performance expectations aspect of the curriculum framework. About one-fourth of the items were in the free-response format, requiring students to generate and write their own answers. Designed to take about one-third of students' test time, some free-response questions asked for short answers while others required extended responses with students showing their work or providing explanations for their answers. The remaining questions were in the multiple-choice format. Correct answers to most questions were worth one point. Consistent with longer response times for the constructed-response questions, however, responses to some of these questions (particularly those requiring extended responses) were evaluated for partial credit, with a fully correct answer being awarded two points. The number of score points available for analysis thus exceeds the number of items.

Exhibit 1.3 Number of Test Items and Score Points by Reporting Category TIMSS 1999

Reporting Category	Total Number of Score Points	Score Points
Mathematics		
Fractions and Number Sense	61	62
Measurement	24	26
Data Representation, Analysis and Probability	21	22
Geometry	21	21
Algebra	35	38
Total	162	169
Science		
Earth Science	22	23
Life Science	40	42
Physics	39	39
Chemistry	20	22
Environmental and Resource Issues	13	14
Scientific Inquiry and the Nature of Science	12	13
Total	146	153

1.7 TIMSS Test Design

Not all of the students in the TIMSS assessment responded to all of the mathematics items. To ensure broad subject matter coverage without overburdening students, TIMSS used a rotated design that included both the mathematics and science items (Adams and Gonzalez, 1996). Thus, the same students were tested in both mathematics and science. As in 1995, the 1999 assessment consisted of eight booklets, each requiring 90 minutes of response time. Each participating student was assigned one booklet only. In accordance with the design, the mathematics and science items were assembled into 26 clusters (labeled A through Z). The secure trend items were in clusters A through H, and items replacing the released 1995 items in clusters I through Z. Eight of the clusters were designed to take 12 minutes to complete; 10 clusters, 22 minutes; and 8 clusters, 10 minutes. In all, the design provided 396 testing minutes, 198 for mathematics and 198 for science. Cluster A was a core cluster assigned to all booklets. The remaining clusters were assigned to the booklets in accordance with the rotated design so that representative samples of students responded to each cluster.

1.8 Background Questionnaires

TIMSS in 1999 administered a broad array of questionnaires to collect data on the educational context for student achievement and to measure trends since 1995. *National Research Coordinators*, with the assistance of their curriculum experts, provided detailed information on the organization, emphases, and content coverage of the mathematics and science curriculum. The *students* who were tested answered questions pertaining to their attitude towards mathematics and science, their academic self-concept, classroom activities, home background, and out-of-school activities. A special version of the student questionnaire was prepared for countries where earth science, physics, chemistry, and biology are taught as separate subjects. Although not strictly related to the question of what students have learned in mathematics or science, characteristics of pupils can be important correlates for understanding educational processes and attainments. Therefore, students also provided general home and demographic information.

The mathematics and science *teachers* of sampled students each completed a teacher questionnaire. These had two sections. The first section covered general background information on preparation, training, and experience, and about how teachers spend their time in school, and probed their views on mathematics and science. The second section related to instructional practices in the class selected for TIMSS 1999 testing. To obtain information about the implemented curriculum, teachers were asked how many periods the class spent on a range of mathematics and science topics, and about the instructional strategies used in the class, including the use of calculators and computers. Teachers also responded to questions about teaching emphasis on the topics in the curriculum frameworks.

The heads of *schools* responded to questions about school staffing and resources, mathematics and science course offerings, and support for teachers.

1.9 Translation and Verification

The TIMSS instruments were prepared in English and translated into 33 languages, with 10 of the 38 countries collecting data in two languages. In addition, the international versions sometimes needed to be modified for cultural reasons, even in the nine countries that tested in English. This process represented an enormous effort for the national centers, with many checks along the way. The translation effort included (1) developing explicit guidelines for translation and cultural adaptation; (2) translation of the instruments by the national centers in accordance with the guidelines, using two or more independent translator; (3) consultation with subject matter experts on cultural adaptations to ensure that the meaning and difficulty of items did not change; (4) verification of translation quality by professional translators from an independent translation company; (5) corrections by the national centers in accordance with the suggestions made; (6) verification by the International Study Center that corrections were made; and (7) a series of statistical checks after the testing to detect items that did not perform comparably across countries.

1.10 Data Collection

Each participating country was responsible for carrying out all aspects of the data collection, using standardized procedures developed for the study. Training manuals were created for school coordinators and test administrators that explained procedures for receipt and distribution of materials as well as for the activities related to the testing sessions. These manuals covered procedures for test security, standardized scripts to regulate directions and timing, rules for answering students' questions, and steps to ensure that identification on the test booklets and questionnaires corresponded to the information on the forms used to track students.

Each country was responsible for conducting quality control procedures and describing this effort in the NRC's report documenting procedures used in the study. In addition, the International Study Center considered it essential to monitor compliance with the standardized procedures. NRCs were asked to nominate one or more persons unconnected with their national center, such as retired school teachers, to serve as quality control monitors for their countries. The International Study Center developed manuals for the monitors and briefed them in two-day training sessions about TIMSS, the responsibilities of the national centers in conducting the study, and their own roles and responsibilities. In all, 71 quality control monitors participated in this training.

The quality control monitors interviewed the NRCs about data collection plans and procedures. They also visited a sample of 15 schools where they observed testing sessions and interviewed school coordinators. Quality control monitors interviewed school coordinators in all 38 countries, and observed a total of 550 testing sessions.

The results of the interviews indicate that, in general, NRCs had prepared well for data collection and, despite the heavy demands of the schedule and shortages of resources, were able to conduct the data collection efficiently and professionally. Similarly, the TIMSS tests appeared to have been administered in compliance with international procedures, including the activities before the testing session, those during testing, and the school-level activities related to receiving material from the national centers, distributing it, and returning it.

1.11 Scoring the Free-Response Items

Because about one-third of the test time was devoted to free-response items, TIMSS needed to develop procedures for reliably evaluating student responses within and across countries. Scoring used two-digit codes with rubrics specific to each item. The first digit designates the correctness level of the response. The second digit, combined with the first, represents a diagnostic code identifying specific types of approaches, strategies, or common errors and misconceptions. Although not used in this report, analyses of responses based on the second digit should provide insight into ways to help students better understand mathematics concepts and problem-solving approaches. Because of the burden of maintaining scoring consistency across time, no free-response items were used to measure trends from 1995 to 1999. However, samples of student responses from each country to selected items in 1999 have been scanned using advanced imaging technology in preparation for studying trends to 2003 and beyond.

To ensure reliable scoring procedures based on the TIMSS rubrics, the International Study Center prepared detailed guides containing the rubrics and explanations of how to use them, together with example student responses for each rubric. These guides, along with training packets containing extensive examples of student responses for practice in applying the rubrics, served as a basis for intensive training in scoring the free-response items. The training sessions were designed to help representatives of national centers who would then be responsible for training personnel in their countries to apply the two digit codes reliably.

1.12 Data Processing

To ensure the availability of comparable, high-quality data for analysis, TIMSS took rigorous quality control steps to create the international database. TIMSS prepared manuals and software for countries to use in entering their data, so that the information would be in a standardized international format before being forwarded to the IEA Data Processing Center in Hamburg for creation of the international database. Upon arrival at the Data Processing Center, the data underwent an exhaustive cleaning process. This involved several iterative steps and procedures designed to identify, document, and correct deviations from the international instruments, file structures, and coding schemes. The process also emphasized consistency of information within national data sets and appropriate linking among the many student, teacher, and school data files.

Throughout the process, the data were checked and double-checked by the IEA Data Processing Center, the International Study Center, and the national centers. The national centers were contacted regularly and given multiple opportunities to review the data for their countries. In conjunction with the IEA Data Processing Center, the International Study Center reviewed item statistics for each cognitive item in each country to identify poorly performing items. Usually the poor statistics (negative point-biserials for the key, large item-by-country interactions, and statistics indicating lack of fit with the model) were due to translation, adaptation, or printing deviations.

1.13 IRT Scaling and Data Analysis

The reporting of the TIMSS achievement data was based primarily on item response theory (IRT) scaling methods. The mathematics results were summarized using a family of 2-parameter and 3-parameter IRT models for dichotomously scored items (right or wrong), and generalized partial credit models for items with 0, 1, or 2 available score points. The IRT scaling method produces a score by averaging the responses of each student to the items in the student's test booklet in a way that takes into account the difficulty and discriminating power of each item. The method used in TIMSS includes refinements that enable reliable scores to be produced even though individual students responded to relatively small subsets of the total mathematics item pool. Achievement scales were produced for each of the five mathematics content areas (fractions and number sense, measurement, data representation, analysis, and probability, geometry, and algebra), as well as for mathematics overall.

The IRT method was preferred for developing comparable estimates of performance for all students, since students answered different test items depending upon which of the eight test booklets they received. IRT analysis provides a common scale on which performance can be compared across countries. Scale scores are a basis for estimating mean achievement, permit estimates of how students within countries vary, and give information on percentiles of performance. For a reliable measure of student achievement in both 1999 and 1995, the overall mathematics scale was calibrated using students from the countries that participated in both years. When all countries participating in 1995 at the eighth grade are treated equally, the TIMSS scale average over those countries is 500 and the standard deviation is 100. Since the countries vary in size, each country was weighted to contribute equally to the mean and standard deviation of the scale. The average and standard deviation of the scale scores are arbitrary and do not affect scale interpretation. When the metric of the scale had been established, students from the countries that tested in 1999 but not 1995 were assigned scores based on the new scale.

IRT scales were also created for each of the five mathematics and six science content areas for the 1999 data. However, insufficient items were used both in 1995 and in 1999 to establish reliable IRT content area scales for trend purposes. The trend exhibits presented in Chapter 3 of the international reports were based on the average percentage of students responding correctly to the common items in each content area.

To allow more accurate estimation of summary statistics for student subpopulations, the TIMSS scaling made use of plausible-value technology, whereby five separate estimates of each student's score were generated on each scale, based on the responses to the items in the student's booklet and the student's background characteristics. The five score estimates are known as "plausible values," and the variability between them encapsulates the uncertainty inherent in score estimation.

1.14 Management and Operations

Like all previous IEA studies, TIMSS 1999 was essentially a cooperative venture among independent research centers around the world. While country representatives came together to work on instruments and procedures, they were each responsible for conducting TIMSS 1999 in their own country, in accordance with the international standards. Each national center provided its own funding and contributed to the support of the international coordination of the study. A study of the scope and magnitude of TIMSS 1999 offers a tremendous operational and logistic challenge. In order to yield comparable data, the achievement survey must be replicated in each participating country in a timely and consistent manner. This was the responsibility of the NRC in each country. Among the major responsibilities of NRCs in this regard were the following:

- Meeting with other NRCs and international project staff to review data collection instruments and procedures
- Defining the school populations from which the TIMSS 1999 samples were to be drawn, selecting the sample of schools using an approved random sampling procedure, contacting the school principals and securing their agreement to participate in the study, and selecting the classes to be tested, again using an approved random sampling procedure
- Translating all of the tests, questionnaires, and administration manuals into the language of instruction of the country (and sometimes into more than one language), and adapting them where necessary prior to data collection
- Assembling, printing, and packaging the test booklets and questionnaires, and shipping the survey materials to the participating schools
- Ensuring that the tests and questionnaires were administered in participating schools, either by teachers in the school or by an external team of test administrators, and that the completed test protocols were returned to the TIMSS 1999 national center
- Conducting a quality assurance exercise in conjunction with the test administration, whereby some testing sessions were observed by an independent observer to confirm that all specified procedures were followed

- Recruiting and training individuals to score the free-response questions in the achievement tests, including a sample that was rescored independently to assess the reliability of the coding procedure
- Recruiting and training data entry personnel for keying the responses of students, teachers, and principals into computerized data files, and conducting the data entry operation, using the software provided
- Checking the accuracy and integrity of the data files prior to shipping them to the IEA Data Processing Center in Hamburg.

In addition to their role in implementing the TIMSS 1999 data collection procedures, NRCs were responsible for conducting analyses of their national data, and for reporting on the results of TIMSS 1999 in their own countries.[4]

The TIMSS 1999 International Study Directors, Michael O. Martin and Ina V.S. Mullis, were responsible for the direction and coordination of the project. The TIMSS International Study Center, located at Boston College in the United States, was responsible for managing all aspects of the design and implementation of the study at the international level. This included the following:

- Planning, conducting, and coordinating all international TIMSS 1999 activities, including meetings of the Project Management Team, NRCs, and advisory committees
- Development, including field testing, of all data collection instruments
- Devising sampling procedures for efficiently selecting representative samples of students in each country, and monitoring sampling operations to ensure that they conformed to TIMSS 1999 requirements
- Developing and documenting operational procedures to ensure efficient collection of all data
- Designing and implementing a quality assurance program encompassing all aspects of the data collection, including monitoring of test administration sessions in participating countries

[4] A list of the TIMSS 1999 National Research Coordinators is provided in Appendix A.

- Supervising the checking and cleaning of the data from the participating countries, and constructing the TIMSS 1999 international database, including the computation of sampling weights and the scaling of the achievement data
- Analysis of international data, and writing and dissemination of international reports.

Several important TIMSS functions, including test and questionnaire development, translation checking, sampling, data processing, and scaling, were conducted by centers around the world, under the direction of the TIMSS International Study Center. In particular, the following centers have played important roles in TIMSS 1999.

- The IEA Secretariat, based in Amsterdam, the Netherlands, coordinated the verification of each country's translations and organized the visits of the international quality control monitors.
- The IEA Data Processing Center (DPC), located in Hamburg, Germany, was responsible for checking and processing data and for constructing the international database. The DPC also worked with Statistics Canada to develop software to facilitate the within-school sampling activities.
- Statistics Canada, located in Ottawa, Canada, was responsible for advising NRCs on their sampling plans, for monitoring progress in all aspects of sampling, and computing the sampling weights.
- Educational Testing Service, located in Princeton, New Jersey, conducted psychometric analyses of the field-test data, and was responsible for scaling the achievement data from the main data collection.

As Sampling Referee, Keith Rust of WESTAT, Inc. (United States), worked with Statistics Canada and the NRCs to ensure that sampling plans met the TIMSS 1999 standards, and advised the International Study Directors on all matters relating to sampling.

The Project Management Team, consisting of the International Study Directors and representatives of each of the above organizations, met regularly throughout the study to plan major activities and to monitor progress.

1.15 Summary of the Report

Pierre Foy and Marc Joncas describe in Chapter 2 the student population for TIMSS 1999, and the design chosen to sample this population. They pay particular attention to the coverage of the target population, and to identifying those subgroups of the population (e.g., mentally handicapped students) that were to be excluded from testing. The authors present the sampling precision requirements of TIMSS 1999, and show how these were used to determine sample size in the participating countries. They describe the use of stratification and multistage sampling, and illustrate the method used in sampling schools in TIMSS (the sampling of classrooms is described in Chapter 7 on field operations).

In Chapter 3, Robert Garden and Teresa Smith (subject matter coordinators in mathematics and science, respectively) describe the TIMSS 1999 test development process, including the construction of the replacement items and scoring guides, the item review process, field testing and item analysis, the selection of the final item set, and the test design for the main data collection.

Ina Mullis, Michael Martin, and Steven Stemler in Chapter 4 provide an overview of the background questionnaires used in TIMSS 1999. This chapter describes the conceptual framework and research questions that guided development of the questionnaires, and details the contents of the curriculum, school, teacher, and student questionnaires used in the TIMSS 1999 data collection.

In order to conduct the study in the 38 participating countries, it was necessary to translate the English versions of the achievement tests, the student, teacher, and school questionnaires, and the manuals and tracking forms into the language of instruction. In all, the TIMSS 1999 instruments were translated into 33 languages. Even where the language of testing was English, adaptations had to be made to suit national language usage. In Chapter 5, Kathleen O'Connor and Barbara Malak describes the procedures that were used to ensure that the translations and cultural adaptations made in each country produced local versions that corresponded closely in meaning to the international versions, and in particular that the items in the achievement tests were not made easier or more difficult through translation.

All of the TIMSS 1999 data collection instruments and procedures were subjected to a full-scale field test in the early part of 1998. The field test, which is described in Chapter 6 by Kathleen O'Connor, provided information to help select the replacement items used in the main data collection, and gave TIMSS NRCs an opportunity to try out all field operations procedures before the main data collection.

As a comparative sample survey of student achievement conducted simultaneously in 38 countries, TIMSS 1999 depended crucially on its data collection procedures to obtain high-quality data. In Chapter 7, Eugenio Gonzalez and Dirk Hastedt describe the procedures developed to ensure that the TIMSS data were collected in a timely and cost-effective manner while meeting high standards of survey research. The authors outline the extensive list of procedural manuals that describe in detail all aspects of the TIMSS field operations, and describe the software systems that were provided to participants to help them conduct their data collection activities.

A major responsibility of the TIMSS International Study Center was to ensure that all aspects of the study were carried out to the highest standards. In Chapter 8, Kathleen O'Connor and Steven Stemler describe the TIMSS 1999 program of site visits to each participating country. As part of this program, TIMSS recruited and trained a team of international quality control monitors who visited the national research centers and interviewed the NRCs about all aspects of the implementation of TIMSS 1999. They also visited a sample of 15 of the schools taking part in the study to interview the School Coordinator and Test Administrator and to observe the test administration.

The selection of valid and efficient samples was vital to the quality and success of TIMSS 1999. In consultation with the TIMSS sampling referee, staff from Statistics Canada reviewed the national sampling plans, sampling data, sampling frames, and sample execution to evaluate the quality of the national samples. In Chapter 9, Pierre Foy describes the implementation of the TIMSS sampling design in participating countries, including the grades tested, population coverage, exclusion rates, and sample sizes. Participation rates for schools and students are also documented, as is the particular design for each country (e.g., the use of stratification variables).

To ensure the availability of comparable, high-quality data for analysis, TIMSS took rigorous quality control steps to create the international database. Upon arrival at the IEA Data Processing Center, the data from each country underwent an exhaustive cleaning process. That process involved several iterative steps and procedures designed to identify, document, and correct deviations from the international instruments, file structures, and coding schemes. Following data cleaning and file restructuring, sampling weights and scale scores were merged into the international database by the DPC. Throughout, the International Study Center monitored the process and managed the flow of data. In Chapter 10, Dirk Hastedt and Eugenio Gonzalez describe the procedures for cleaning and verifying the TIMSS data and for constructing the database.

The complex multistage sampling design used in TIMSS 1999 required the use of sampling weights to account for differential probabilities of student selection and to adjust for non-participation in order to compute accurate estimates of student achievement. Statistics Canada was responsible for computing the sampling weights for the TIMSS countries. In Chapter 11, Pierre Foy describes the derivation of TIMSS school, classroom, and student weights, and the adjustments for non-participation that were applied.

Because the statistics presented in the TIMSS 1999 reports are estimates of national performance based on samples of students, rather than the values that could be calculated if every student in every country had answered every question, it is important to have measures of the degree of uncertainty of the estimates. TIMSS used the jackknife procedure to estimate the standard errors associated with each statistic presented in the international reports. In Chapter 12, Eugenio Gonzalez and Pierre Foy describe the jackknife technique and its application to the TIMSS data in estimating the variability of the sample statistics.

Before the achievement data were scaled, the TIMSS 1999 item results were thoroughly checked by the IEA Data Processing Center, the International Study Center, and the national centers. The national centers were contacted regularly and given repeated opportunities to review the data for their countries. The International Study Center reviewed item statistics for every mathematics and science item in each country to identify poorly performing items. In Chapter 13, Ina Mullis and Michael Martin describe the procedures used to ensure that the achievement data included in the scaling and the international database were comparable across countries.

The complexity of the TIMSS test design and the requirement to make comparisons between countries and between 1995 and 1999 led TIMSS to use item response theory in the analysis of the achievement results. In Chapter 14, Kentaro Yamamoto and Ed Kulick describe the scaling method and procedures Educational Testing Service used to produce the TIMSS 1999 achievement scores, including the estimates of international item parameters and the derivation and use of plausible values to provide estimates of student proficiency.

TIMSS identified the 90th, 75th, 50th, and 25th international percentiles as benchmarks with which student performance could be compared. In Chapter 15, Kelvin Gregory and Ina Mullis outline the scale anchoring procedure undertaken by TIMSS 1999 to provide detailed descriptions of what mathematics and science students scoring at these international benchmarks know and can do.

TIMSS reported student achievement in mathematics and science in a number of ways. Mean achievement and percentiles of distribution were reported for each country, together with tests of statistical significance adjusted for multiple comparisons. TIMSS presented mean achievement for girls and boys separately, with indications of significant differences between the genders. TIMSS also contrasted performance at the fourth grade in 1995 with performance at the eighth grade in 1999 to show the change in relative performance for that cohort of students. In Chapter 16, Eugenio Gonzalez and Kelvin Gregory describe the analyses undertaken to present the achievement data in the international reports, and describe how trends in achievement in mathematics and science content areas were analyzed using average percent correct technology.

TIMSS 1999 collected an enormous amount of data on educational context from students, teachers, and school principals, as well as information about the intended curriculum. In Chapter 17, Teresa Smith describes the analysis and reporting of the background data in the international reports - the development of the plans for the international reports, the construction of composite indices, the consensus and review procedures, and special issues in reporting, such as response rates and reporting teacher data.

1.16 Summary

This report provides an overview of the main features of the TIMSS 1999 project and summarizes the technical background of the study. The development of the achievement tests and questionnaires, the sampling and operations procedures, the procedures for data collection and quality assurance, the construction of the international database, including sampling weights and proficiency scores, and the analysis and reporting of the results are all described in sufficient detail to enable the reader of the international reports to have a good understanding of the technical and operational underpinning of the study.

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.

Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., Gregory, K.D., Smith, T.A., Chrostowski, S.J., Garden, R.A., & O'Connor, K.M. (2000). *TIMSS 1999 International Science Report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade.* Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., Gregory, K.D., Garden, R.A., O'Connor, K.M., Chrostowski, S.J., & Smith, T.A. (2000). *TIMSS 1999 International Mathematics Report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade.* Chestnut Hill, MA: Boston College.

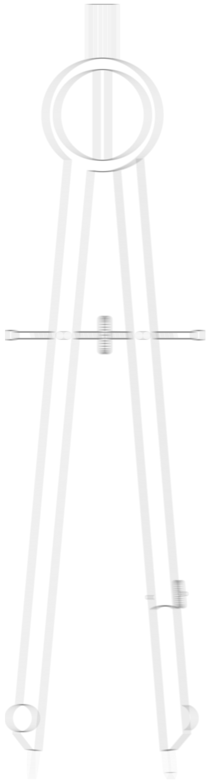
Robitaille, D.F. & Garden, R.A. (1996). Design of the Study in D.F. Robitaille & R.A. Garden (Eds.), *TIMSS Monograph No. 2: Research Questions & Study Design.* Vancouver, Canada: Pacific Educational Press.

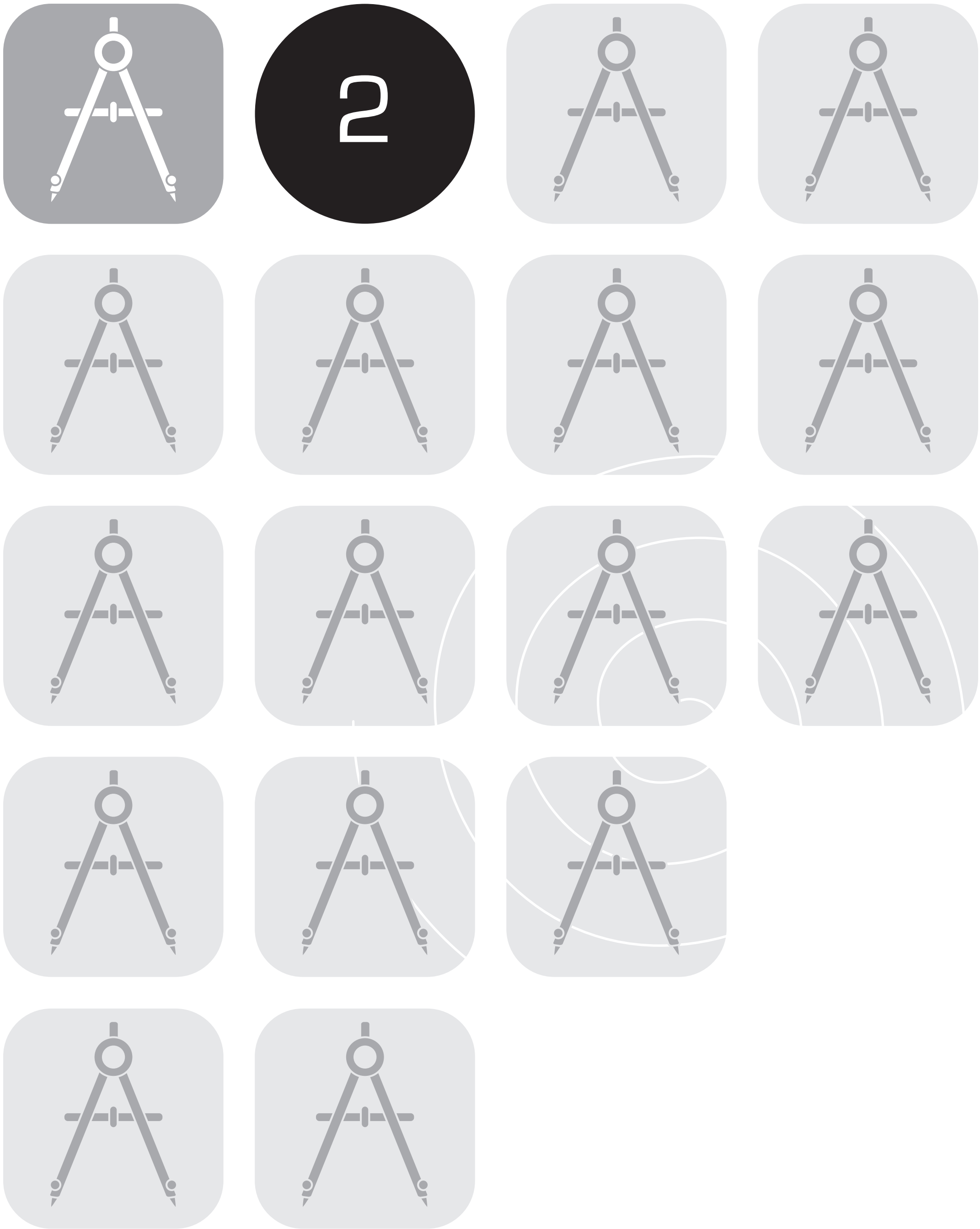
Robitaille, D.F., Schmidt, W.H., Raizen, S.A., McKnight, C.C., Britton, E., & Nicol, C. (1993). *TIMSS Monograph No. 1: Curriculum Frameworks for Mathematics and Science.* Vancouver, Canada: Pacific Educational Press.

Schmidt, W.H., McKnight, C.C., Valverde, G.A., Houang, R.T., & Wiley, D.E. (1997a). *Many Visions, Many Aims: A Cross-National Investigation of Curricular Intentions in School Mathematics.* Norwell, MA: Kluwer Academic Press.

Schmidt, W.H., Raizen, S.A., Britton, E.D., & Bianchi, L.J. (1997b). *Many Visions, Many Aims: A Cross-National Investigation of Curricular Intentions in School Science.* Norwell, MA: Kluwer Academic Press.

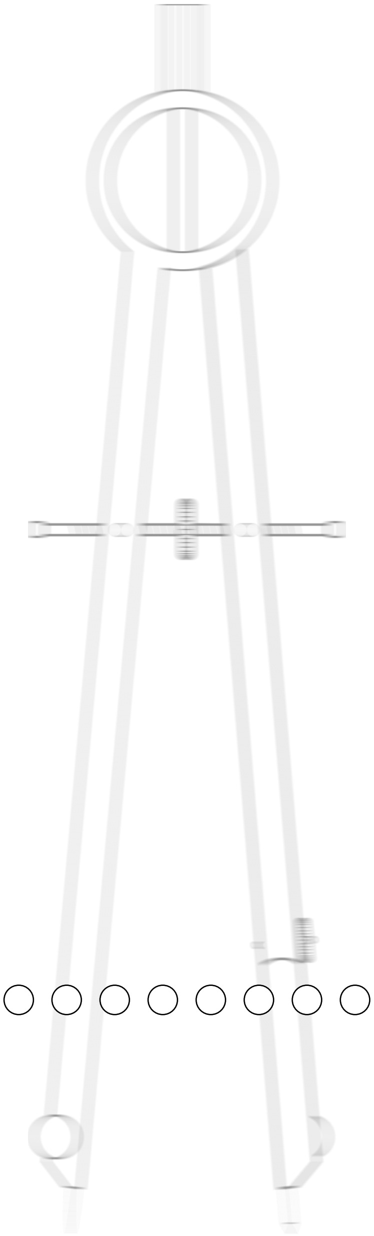
Travers, K.J., & Westbury, I. (1989). *The IEA Study of Mathematics I: Analysis of Mathematics Curricula.* Oxford: Pergamon Press.





TIMSS Sample Design

Pierre Foy,
Marc Joncas



2 TIMSS Sample Design

Pierre Foy
Marc Joncas

2.1 Overview

This chapter describes the procedures developed to ensure proper sampling of the student populations in each participating country. To be acceptable for TIMSS 1999, national sample designs had to result in probability samples that gave accurately weighted estimates of population parameters, and for which estimates of sampling variance could be computed. The TIMSS 1999 sample design was very similar to that of its predecessor, TIMSS 1995, with minor refinements made as a result of the 1995 sampling. The TIMSS design was chosen so as to balance analytical requirements and operational constraints, while keeping it simple enough for all participants to implement. Representative and efficient samples in all countries were crucial to the success of the project. The quality of the samples depends on the sampling information available at the design stage, and particularly on the sampling procedures.

The National Research Coordinators (NRCs) were aware that in a study as ambitious as TIMSS 1999 the sample design and sampling procedures would be complex, and that gathering the required information about the national education systems would place considerable demands on resources and expertise. At the same time, those directing and coordinating the project realized that the national centers had only limited numbers of qualified sampling personnel. Keeping the procedures as simple as possible, especially the sample selection within schools, was thus a major consideration.

The international project management provided manuals and expert advice to help NRCs adapt the TIMSS 1999 sample design to their national system and to guide them through the phases of sampling. The TIMSS 1999 *School Sampling Manual* (TIMSS, 1997) described how to implement the international sample design and offered advice on planning, working within constraints, establishing appropriate sample selection procedures, and fieldwork. The *Survey Operations Manual* (TIMSS, 1998a) and *School Coordinator Manual* (TIMSS, 1998b) discussed sample selection and execution within schools, the assignment of test booklets to selected students, and administration and monitoring procedures used to identify and track respondents and nonrespondents. NRCs also received software designed to automate the sometimes complex within-school sampling procedures.

In addition, NRCs had access to expert support. Statistics Canada, in consultation with the TIMSS 1999 sampling referee, reviewed and approved the national sampling plans, sampling data, sampling frames, and sample selection. Statistics Canada also assisted nearly half of the TIMSS 1999 participants in drawing national school samples.

NRCs were allowed to adapt the basic TIMSS sample design to the needs of their education system by using more sampling information or more sophisticated designs and procedures. These adjustments, however, had to be approved by the International Study Center at Boston College and monitored by Statistics Canada.

2.2 Target Populations and Exclusions

In IEA studies, the target population for all countries is known as the *international desired population*. The international desired population for TIMSS 1999 was as follows:

- All students enrolled in the upper of the two adjacent grades that contain the largest proportion of 13-year-olds at the time of testing.

The TIMSS 1999 target grade was the upper grade of the TIMSS 1995 population 2 definition[1] and was expected to be the eighth grade in most countries. This would allow countries participating in both TIMSS 1995 and TIMSS 1999 to establish a trend line of comparable achievement data.

2.2.1 School and Within-School Exclusions

TIMSS 1999 expected all participating countries to define their *national desired population* to correspond as closely as possible to its definition of the international desired population. Sometimes, however, NRCs had to make changes. For example, some countries had to restrict geographical coverage by excluding remote regions; or to exclude a segment of their education system. The international reports document any deviations from the international definition of the TIMSS 1999 target population.

○○○

[1] For the TIMSS 1995 Population definition, see Foy, Rust, & Schleicher (1996).

Using their national desired population as a basis, participating countries had to operationally define their population for sampling purposes. This definition, known in IEA terminology as the *national defined population*, is essentially the sampling frame from which the first stage of sampling takes place. The national defined population could be a subset of the national desired population. All schools and students from the former excluded from the latter are referred to as the *excluded population.*

TIMSS 1999 participants were expected to keep the excluded population to no more than 10% of the national desired population. Exclusions could occur at the school level, within schools, or both. Because the national desired population was restricted to schools that contained the target grade, schools not containing this grade were considered to be outside the scope of the sampling frame, and not part of the excluded population. Participants could exclude schools from the sampling frame for the following reasons:

- They were in geographically remote regions.
- They were of extremely small size.
- They offered a curriculum, or school structure, that was different from the mainstream education system(s).
- They provided instruction only to students in the exclusion categories defined as "within-sample exclusions."

Within-sample exclusions were limited to students who, because of some disability, were unable to take the TIMSS 1999 tests. NRCs were asked to define anticipated within-sample exclusions. Because these definitions can vary internationally, NRC's were also asked to follow certain rules adapted to their jurisdictions. In addition, they were to estimate the size of such exclusions so that compliance with the 10% rule could be gauged in advance.

The general TIMSS 1999 rules for defining within-school exclusions included:

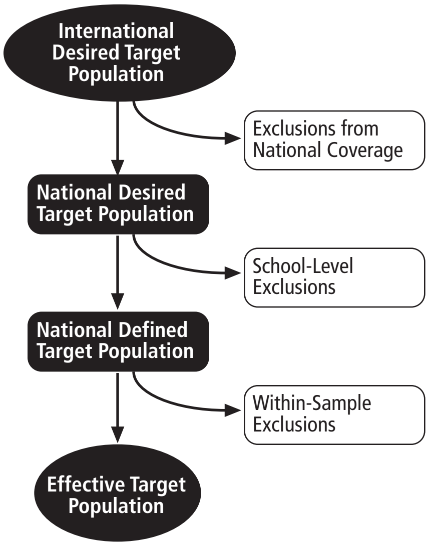
- **Educable mentally disabled students**. These are students who were considered, in the professional opinion of the school principal or other qualified staff members, to be educable mentally disabled, or students who had been so diagnosed by psychological tests. This included students who were emotionally or mentally unable to follow even the general instructions of the TIMSS 1999 test. It did not include students who merely exhibited poor academic performance or discipline problems.

- **Functionally disabled students**. These are students who were permanently physically disabled in such a way that they could not perform in the TIMSS 1999 tests. Functionally disabled students who could perform were included in the testing.
- **Non-native-language speakers**. These are students who could not read or speak the language of the test and so could not overcome the language barrier of testing. Typically, a student who had received less than one year of instruction in the language of the test was excluded, but this definition was adapted in different countries.

The stated objective in TIMSS 1999 was that the effective target population, the population actually sampled by TIMSS 1999, be as close as possible to the international desired population. Exhibit 2.1 illustrates the relationship between the desired populations and the excluded populations. Any exclusion of eligible students from the international desired population had to be accounted for, both at the school level and within samples.

The size of the excluded population was documented and served as an index of the coverage and representativeness of the selected samples.

Exhibit 2.1 Relationship Between the Desired Populations and Exclusions


International Desired Target Population
Exclusions from National Coverage
National Desired Target Population
School-Level Exclusions
National Defined Target Population
Within-Sample Exclusions
Effective Target Population

2.3 Sample Design

The basic sample design for TIMSS 1999 is generally referred to as a two-stage stratified cluster sample design. The first stage consisted of a sample of schools[2], which may be stratified; the second stage consisted of a single mathematics classroom selected at random from the target grade in sampled schools. It was also permissible to add a third stage, in which students could be sampled within classrooms. This design lent itself to the many analytical requirements of TIMSS 1999.

2.3.1 Units of Analysis and Sampling Units

The TIMSS 1999 analytical focus was both on the cumulative learning of students and on the instructional characteristics affecting learning. The sample design, therefore, had to address the measurement both of characteristics thought to influence cumulative learning and of specific characteristics of instruction. Because schools, classrooms, and students were all considered potential units of analysis, they had to be considered as sampling units. This was necessary in order to meet specific requirements for data quality and sampling precision at all levels.

○○○

[2] In some very large countries, it was necessary to include an extra preliminary stage in which school districts were sampled first, and then schools.

Although in the second sampling stage the sampling units were intact mathematics classrooms, the ultimate sampling elements were students. Consequently, it was important that each student from the target grade be a member of one and only one of the mathematics classes in a school from which the sampled classes were to be selected. In most education systems, the mathematics class coincided with a student homeroom or science class. In some systems, however, mathematics and science classes did not coincide. In any case, participating countries were asked to define the classrooms on the basis of mathematics instruction. If not all students in the national desired population belonged to a mathematics class, then an alternative definition of the classroom was required for ensuring that the non-mathematics students had an opportunity to be selected.

2.3.2 Sampling Precision and Sample Size

Sample sizes for TIMSS 1999 had to be specified so as to meet the analytic requirements of the study. Since students were the principal units of analysis, the ability to produce reliable estimates of student characteristics was important. The TIMSS 1999 standard for sampling precision required that all population samples have an effective sample size of at least 400 students for mathematics and science achievement. In other words, the samples should have sampling errors no greater than those that would be obtained from a simple random sample of 400 students.

An effective sample size of 400 students results in the following 95% confidence limits for sample estimates of population means, percentages, and correlation coefficients.

- Means: m ± 0.1s (where m is the mean estimate and s is the estimated standard deviation for students)
- Percentages: p ± 5.0% (where p is a percentage estimate)
- Correlations: r ± 0.1 (where r is a correlation estimate)

Furthermore, since TIMSS 1999 was designed to allow for analyses at the school and classroom levels, at least 150 schools were to be selected from the target population. A sample of 150 schools results in 95% confidence limits for school-level and classroom-level mean estimates that are precise to within ± 16% of their standard deviations. To ensure sufficient sample precision for these units of analysis, some participants had to sample more schools than they would have selected otherwise.

The precision of multistage cluster sample designs are generally affected by the so-called clustering effect. A classroom as a sampling unit constitutes a cluster of students who tend to be more like each other than like other members of the population. The *intraclass correlation* is a measure of this similarity. Sampling 30 students from a single classroom, when the intraclass correlation is positive, will yield less information than a random sample of 30 students spread across all classrooms in a school. Such sample designs are less efficient, in terms of information per sampled student, than a simple random sample of the same size. This clustering effect had to be considered in determining the overall sample size for TIMSS 1999.

The magnitude of the clustering effect is determined by the size of the cluster (classroom) and the size of the intraclass correlation. For planning the sample size, therefore, each country had to choose a value for the intraclass correlation, and a value for the expected cluster size (this was known as the minimum cluster size). The intraclass correlation for each country was estimated from past studies, such as TIMSS 1995, or from national assessments. In the absence of such sources, an intraclass correlation of 0.3 was assumed. Since all participants chose to test intact classrooms, the minimum cluster size was in fact the average classroom size. The specification of the minimum cluster size affected not only the number of schools sampled, but also the way in which small schools and small classrooms were treated.

Sample-design tables were produced and included in the TIMSS 1999 School Sampling Manual (see Exhibit 2.2 for an example). These tables illustrated the number of schools that had to be sampled to meet the TIMSS sampling precision requirements for a range of values of intraclass correlation and minimum cluster sizes. TIMSS 1999 participants could use these tables to determine how many schools they should sample. For example, an examination of Exhibit 2.2 shows that a participant whose intraclass correlation was expected to be 0.6 and whose average classroom size was 30 needed to sample a minimum of 248 schools. Whenever the estimated number of schools to sample fell below 150, participants were asked to sample at least 150 schools.

The sample-design tables could be used also to determine sample sizes for more complex designs. For example, a number of strata could be constructed for which different minimum cluster sizes could be specified, thereby refining the national sample design in a way that might avoid special treatment of small schools (See section 2.3.6, Small Schools).

Exhibit 2.2: Sample-Design Table* (95%Confidence Limits For Means ±0.1s / Percentages ±5.0)

MCS**		Intraclass Correlation								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
5	a	150	157	189	221	253	285	317	349	381
	n	750	785	945	1 105	1 265	1 425	1 585	1 745	1 905
10	a	150	150	155	191	227	263	299	335	371
	n	1 500	1 500	1 550	1 910	2 270	2 630	2 990	3 350	3 710
15	a	150	150	150	180	218	255	292	330	367
	n	2 250	2 250	2 250	2 700	3 270	3 825	4 380	4 950	5 505
20	a	150	150	150	175	213	251	289	327	365
	n	3 000	3 000	3 000	3 500	4 260	5 020	5 780	6 540	7 300
25	a	150	150	150	172	211	249	287	326	364
	n	3 750	3 750	3 750	4 300	5 275	6 225	7 175	8 150	9 100
30	a	150	150	150	170	209	248	286	325	364
	n	4 500	4 500	4 500	5 100	6 270	7 440	8 580	9 750	10 920
35	a	150	150	150	169	208	246	285	324	363
	n	5 250	5 250	5 250	5 915	7 280	8 610	9 975	11 340	12 705
40	a	150	150	150	168	207	246	285	324	363
	n	6 000	6 000	6 000	6 720	8 280	9 840	11 400	12 960	14 520
45	a	150	150	150	167	206	245	284	323	362
	n	6 750	6 750	6 750	7 515	9 270	11 025	12 780	14 535	16 290
50	a	150	150	150	166	205	245	284	323	362
	n	7 500	7 500	7 500	8 300	10 250	12 250	14 200	16 150	18 100

a = number of sampled schools
n = number of sampled students in target grade
*Minimum school sample required = 150
**MCS is the number of students selected in each sampled school (generally the average classroom size).

2.3.3 Stratification

Stratification is the grouping of sampling units (e.g., schools) in the sampling frame according to some attribute or variable prior to drawing the sample. It is generally used for the following reasons:

- To improve the efficiency of the sample design, thereby making survey estimates more reliable
- To apply different sample designs, or disproportionate sample-size allocations, to specific groups of schools (such as those within certain states or provinces)
- To ensure adequate representation in the sample of specific groups from the target population.

Examples of stratification variables for school samples are geography (such as states or provinces), school type (such as public and private schools), and level of urbanization (such as rural and urban). Stratification variables in the TIMSS 1999 sample design could be used explicitly, implicitly, or both.

Explicit stratification consists of building separate school lists, or sampling frames, according to the stratification variables under consideration. Where, for example, geographic regions were an explicit stratification variable, separate school sampling frames were constructed for each region. Different sample designs, or different sampling fractions, could then be applied to each school-sampling frame to select the sample of schools. In practice, the main reason for considering explicit stratification in TIMSS 1999 was disproportionate allocation of the school sample across strata. For example, a country might require an equal number of schools from each stratum, regardless of the relative size of each stratum.

Implicit stratification makes use of a single school sampling frame, but sorts the schools in this frame by a set of stratification variables. This is a simple way of ensuring proportional sample allocation without the complexity of explicit stratification. Implicit stratification can also improve the reliability of survey estimates, provided the variables are related to school mean student achievement in mathematics and science.

2.3.4 Replacement Schools

Although TIMSS participants placed great emphasis on securing school participation, it was anticipated that a 100% participation rate would not be possible in all countries. To avoid losses in sample size, a mechanism was instituted to identify, a priori, two replacement schools for each sampled school. The use of implicit stratification variables and the subsequent ordering of the school sampling frame by size ensured that any sampled school's replacement would have similar characteristics. Although this approach was not guaranteed to avoid response bias, it would tend to minimize the potential for bias. Furthermore, it was deemed more acceptable than over-sampling to accommodate a low response rate.

2.3.5 First Sampling Stage

The sample-selection method used for the first-stage of sampling in TIMSS 1999 made use of a systematic probability-proportional-to-size (PPS) technique. Use of this method required some measure of size (MOS) of the sampling units. Ideally this was the number of sampling elements within the unit (e.g., number of students in the target grade in the school). If this information was unavailable, some other highly correlated measure, such as total school enrollment, was used.

The schools in each explicit stratum were listed in order of the implicit stratification variables, together with the MOS for each school. They were further sorted by MOS within variable. The measures of size were accumulated from school to school, and the running total (the cumulative MOS) was listed next to each school (see Exhibit 2.3). The cumulative MOS was a measure of the size of the population of sampling elements; dividing it by the number of schools sampled gives the *sampling interval.*

The first school was sampled by choosing a random number in the range between 1 and the sampling interval. The school whose cumulative MOS contained the random number was the sampled school. By adding the sampling interval to that first random number, a second school was identified. This process of consistently adding the sampling interval to the previous selection number resulted in a PPS sample of the required size.

As each school was selected, the next school in the sampling frame was designated as a replacement school for use should the sampled school not participate in the study, and the next after that as a second replacement, for use should neither the sampled school nor its replacement participate.

Two of the many benefits of the PPS sample selection method are that it is easy to implement, and that it is easy to verify that it was implemented properly. The latter was critical since one of TIMSS 1999's major objectives was to be able to verify that a sound sampling methodology had been used.

Exhibit 2.3 illustrates the PPS systematic sampling method applied to a fictitious sampling frame. The first three sampled schools are shown, as well as their corresponding first and second replacements (R1 and R2).

Exhibit 2.3: Application of the PPS Systematic Sampling Method

Total MOS:	392154	Sampling Interval:	2614.3600
School Sample:	150	Random Start:	1135.1551

School Identification Number	Measure of Size (MOS)	Cumulative MOS	Sampled and Replacement Schools
172989	532	532	
976181	517	1049	
564880	487	1536	S
387970	461	1997	R1
483231	459	2456	R2
550766	437	2893	
228699	406	3299	
60318	385	3684	
201035	350	4034	S
107346	341	4375	R1
294968	328	4703	R2
677048	311	5014	
967590	299	5313	
644562	275	5588	
32562	266	5854	
194290	247	6101	
129135	215	6316	
1633	195	6511	S
256393	174	6685	R1
754196	152	6837	R2
750793	133	6970	
757843	121	7091	
743500	107	7198	
84930	103	7301	
410355	97	7398	

S = Sampled School
R1, R2 = Replacement Schools

2.3.6 Small Schools

Small schools tend to be problematic in PPS samples because students sampled from these schools get disproportionately large sampling weights, and when the school size falls below the minimum cluster size, it reduces the overall student sample size. A school was deemed small in TIMSS 1999 if it was smaller than the minimum cluster size. Thus, if the minimum cluster size for a country was set at 20, then a school with fewer than 20 students in the target grade was considered a small school.

In TIMSS 1999, small schools were handled differently than in TIMSS 1995. The 1999 approach for dealing with them consisted of two steps

- **Extremely small schools**. Extremely small schools were defined as schools with fewer students than half the minimum cluster size. For example, if the minimum cluster size was set at 20, then schools with fewer than 10 students in the target grade were considered extremely small schools. If student enrollment in these schools was less than 2% of the eligible population, they were excluded, provided the overall exclusion rate did not exceed the 5% criterion (see Section 2.3).
- **Explicit stratum of small schools**. If fewer than 10% of eligible students were enrolled in small schools, then no additional action was required. If, however, more than 10% of eligible students were enrolled in small schools, then an explicit stratum of small schools was required. The number of schools to sample from this stratum remained proportional to the stratum size, but all schools had an equal probability of selection. This action ensured greater stability in the resulting sampling weights.

2.3.7 Optional Preliminary Sampling Stage

Some very large countries chose to introduce a preliminary sampling stage before sampling schools. This consisted of a PPS sample of geographic regions. A sample of schools was then selected from each sampled region. This design was used mostly as a cost-reduction measure where the construction of a comprehensive list of schools would have been either impossible or prohibitively expensive. Also, this additional sampling stage reduced the dispersion of the school sample, thereby potentially reducing travel costs. Sampling guidelines were put in place to ensure that an adequate number of units were sampled from this preliminary stage. The sampling frame had to consist of at least 80 primary sampling units, of which at least 40 had to be sampled at this stage.

2.3.8 Second Sampling Stage

The second sampling stage consisted of selecting classrooms within sampled schools. As a rule, one classroom per school was sampled, although some participants opted to sample two classrooms. Classrooms were selected either with equal probabilities or with probabilities proportional to their size. Participants who opted to test all students in selected classrooms sampled classrooms with equal probabilities. This was the method of choice for most participants. A procedure was also available whereby NRCs could choose to sub-sample students within selected classrooms using PPS.

2.3.9 Small Classrooms

Generally, classes in an education system tend to be of roughly equal size. Occasionally, however, small classes are devoted to special activities, such as remedial or accelerated programs. These can become problematic, since they can lead to a shortfall in sample size and thus introduce some instability in the resulting sampling weights when classrooms are selected with PPS.

In order to avoid these problems, the classroom sampling procedure specified that any classroom smaller than half the minimum cluster size be combined with another classroom from the same grade and school. For example, if the minimum cluster size was set at 30, then any classroom with fewer than 15 students was combined with another. The resulting pseudo-classroom then constituted a sampling unit.

2.4 Participation Rates

Weighted and unweighted response rates were computed for each participating country at the school level and at the student level. The basic formulae for response rates are provided in this section. More elaborate treatment of participation rates, including adjustments for non-participation, may be found in Chapter 11.

2.4.1 School-Level Participation Rates

The minimum acceptable school-level participation rate, before the use of replacement schools, was set at 85%. This criterion was applied to the unweighted school response rate. School response rates were computed and reported both weighted and unweighted, with and without replacement schools. The general formula for computing weighted school-level response rates is shown in the following equation:

$$R_{wgt}(sch) = \frac{\sum_{part} MOS_i/\pi_i}{\sum_{elig} MOS_i/\pi_i}$$

For each sampled school, the ratio of its measure of size (MOS) to its selection probability (π_i) is computed. The weighted school-level participation rate is the sum of the ratios for all participating schools divided by the sum of the ratios for all eligible schools. The unweighted school-level participation rates are computed in a similar way, with all school ratios set to unity. This becomes simply the number of participating schools in the sample divided by the number of eligible schools in the sample. Since in most cases, in selecting the sample, the value of π_i was set proportional to MOS_i within each explicit stratum, weighted and unweighted rates are generally similar.

2.4.2 Student-Level Participation Rates

Like the school-level participation rate, the minimum acceptable student-within-school participation rate was set at 85%. This criterion was applied to the unweighted student-level participation rate. Both weighted and unweighted student participation rates were computed and reported. The general formula for computing student-level participation rates is shown in the following equation:

$$R_{wgt}(std) = \frac{\sum_{part} 1/p_j}{\sum_{elig} 1/p_j}$$

where p_j denotes the probability of selection of the student, incorporating all stages of selection. Thus the weighted student-level participation rate is the sum of the inverse of the selection probabilities for all participating students divided by the sum of the inverse of the selection probabilities for all eligible students. The unweighted student participation rates were computed in a similar way, but with each student contributing equal weight.

2.4.3 Overall Participation Rates

The minimum acceptable overall response rate was set at 75%. This rate was calculated as the product of the weighted school-level participation rate without replacement schools and the weighted student-level participation rate. Weighted overall participation rates were computed and reported both with and without replacement schools.

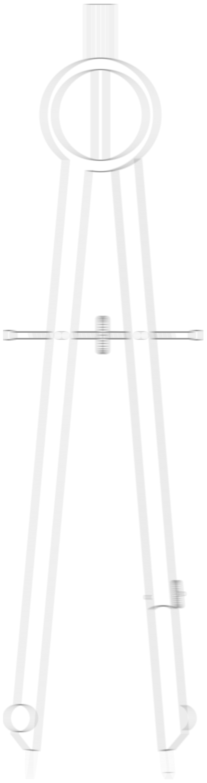
References

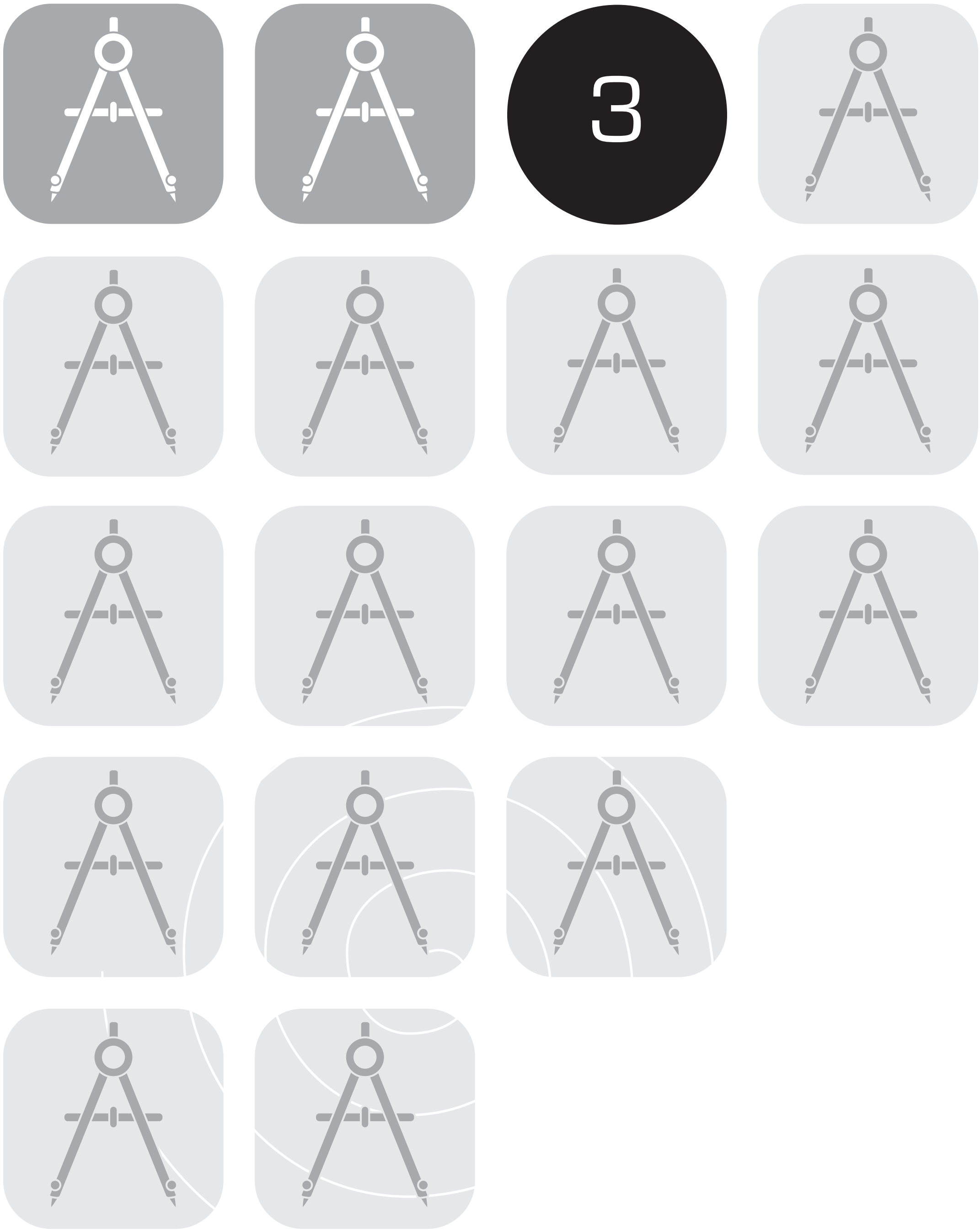
Foy, P., Rust, K., & Schleicher, A. (1996). Sample Design in M.O. Martin and D.L. Kelly (Eds.), *Third International Mathematics and Science Study Technical Report Volume I: Design and Development.* Chestnut Hill, MA: Boston College.

TIMSS (1997). *TIMSS 1999 School Sampling Manual–Version 2* (Doc. Ref.: TIMSS 1999 97-0012). Prepared by Pierre Foy, Statistics Canada. Chestnut Hill, MA: Boston College.

TIMSS (1998a). *Survey Operations Manual–Main Survey* (Doc. Ref.: TIMSS 1999 98-0026). Prepared by the International Study Center. Chestnut Hill, MA: Boston College.

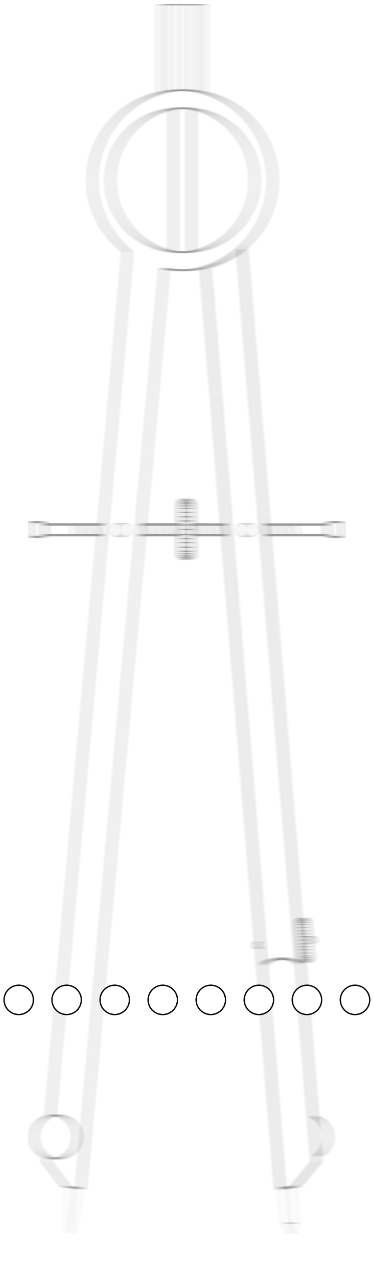
TIMSS (1998b). *School Coordinator Manual–Main Survey* (Doc. Ref.: TIMSS 1999 98-0024). Prepared by the International Study Center. Chestnut Hill, MA: Boston College.





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 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 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

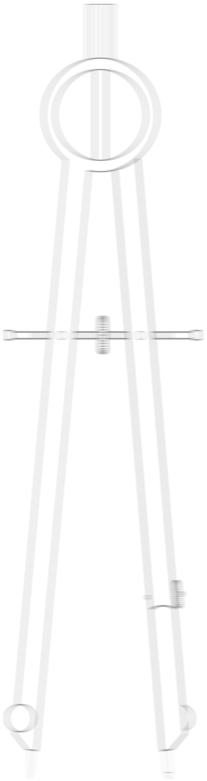
References

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.

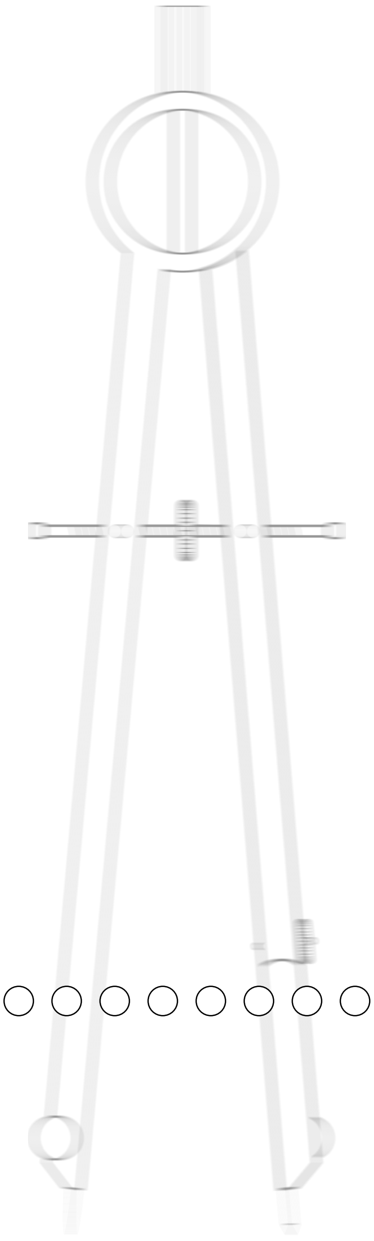
Robitaille, D.F., Schmidt, W.H., Raizen, S., McKnight, C., Britton, E., & Nicol, C. (1993). *Curriculum Frameworks for Mathematics and Science.* Vancouver: Pacific Educational Press.



4

TIMSS Questionnaire Development

Ina V.S. Mullis
Michael O. Martin
Steven E. Stemler



4 TIMSS Questionnaire Development

Ina V.S. Mullis
Michael O. Martin
Steven E. Stemler

4.1 Overview

TIMSS 1999 was designed to measure trends in student achievement over time by building on the data collected from the Third International Mathematics and Science Study of 1995. Consequently, it was important not just to have measures of student achievement that linked the two assessments, but also background questionnaires that had much in common. Four background questionnaires were used to gather information at various levels of the educational system: curriculum questionnaires addressed issues of curriculum design and emphasis in mathematics and science; a school questionnaire asked school principals to provide information about school staffing and facilities, as well as curricular and instructional arrangements; teacher questionnaires asked mathematics and science teachers about their backgrounds, attitudes, and teaching activities and approaches; and a questionnaire for students sought information about their home backgrounds and attitudes, and their experiences in mathematics and science classes.

The approach to questionnaire development adopted for TIMSS 1999 was to retain the parts of the 1995 questionnaires that were found to be most valuable in analysis and reporting and to concentrate development efforts on areas needing expansion or refinement. Each of the questionnaires went through an exhaustive review process prior to the field test, and was reviewed again in light of the field-test data. Items retained for the final versions of the questionnaires were judged to yield the maximum amount of information with the least respondent burden. This chapter begins with an overview of the conceptual framework and research questions that guided the development of the questionnaires and goes on to present the main issues addressed by each questionnaire.

4.2 Conceptual Framework

The conceptual framework for TIMSS was greatly influenced by IEA's Second International Mathematics Study (SIMS), which focused on the curriculum as a major explanatory factor for international variation in student achievement. In the SIMS model, the curriculum was viewed as having three aspects: the *intended* curriculum, the *implemented* curriculum, and the *attained* curriculum.

- The **intended curriculum** refers to the curricular goals of the education system and the structures established to achieve them.
- The **implemented curriculum** refers to the practices, activities, and institutional arrangements within the school and classroom that are designed to implement the goals of the system.
- The **attained curriculum** refers to the products of schooling – what students actually gained from their educational experience.

Building on this view of the educational process, TIMSS in 1995 sought to assess, through context questionnaires, the factors likely to influence students' learning of mathematics and the sciences at the national (or regional), school, classroom, and student level (Schmidt and Cogan, 1996).

4.3 Research Questions

TIMSS in 1995 posed four general research questions to guide the development of the tests and questionnaires and to provide a focus for the analysis and reporting of results: What kinds of mathematics and science are students expected to learn? Who provides the instruction? How is instruction organized? What have students learned? These questions were also the focus of TIMSS in 1999. The question of what students are expected to learn was addressed using questionnaires that were distributed to mathematics and science curriculum experts in participating countries. The question about the characteristics and preparation of mathematics and science teachers was addressed using questionnaires that were distributed to school principals and teachers. The third question, on instructional approaches to the teaching of mathematics and science, was also addressed through questionnaires to principals and teachers, as well as to students. The fourth question was measured by performance on the TIMSS 1999 achievement tests.

The research questions cast a broad net for exploring associations with achievement in mathematics and science. For example, in attempting to answer the question "Who provides the instruction?" the questionnaires tapped characteristics of the person providing instruction, such as gender, age, years of experience, attitude towards the subject, and time spent preparing lessons. The background questionnaires allow researchers to investigate the most influential characteristics of the people, practices, and policies affecting student achievement.

4.4 Curriculum Questionnaires

The TIMSS 1999 study included curriculum questionnaires that were not available for the 1995 survey. These were designed to collect basic information about the organization of the mathematics and science curriculum in each country, and about the topics intended to be covered up to the eighth grade. The National Research Coordinator (NRC) in each country was asked to complete one questionnaire about the mathematics curriculum and one about the science curriculum, drawing on the expertise of mathematics and science specialists in the country as necessary.

The curriculum questionnaires had two parts. The first part sought information about the organization and structure of the curriculum. The second part asked whether a wide range of detailed topics in mathematics and science were in the intended curriculum. In addition, the questionnaires asked what percentage of the eighth-grade student body was exposed to each of the topics in the intended curriculum.

Because there was just one mathematics and one science curriculum questionnaire from each country, it was possible to conduct follow-up interviews with NRCs to resolve ambiguities and develop a clear understanding of each country's curriculum. Several important research questions addressed by the questionnaires were:

- Is there a country-level curriculum? If so, how is implementation monitored?
- What is the nature of country-level assessments, if there are any?
- What content is emphasized in the national curriculum?

The complete contents of the mathematics and science curriculum questionnaires are described further in Exhibits 4.1 and 4.2.

4.5 School Questionnaire

The school questionnaire was completed by the school principal and was designed to elicit information concerning some of the major factors thought to influence student achievement. Several important research questions addressed by the school questionnaire were:

- What staffing and resources are available at each school?
- What are the roles and responsibilities of the teachers and staff?
- How is the mathematics curriculum organized?
- How is the science curriculum organized?
- What is the school climate?

The TIMSS 1999 school questionnaire was very similar to the 1995 version. Four questions about scheduled time for teachers were removed, since they seemed more appropriate to the teacher questionnaires. Questions on computer availability were revised and extended to include access to the Internet for instructional or educational purposes. Finally, questions dealing with provisions for students of different abilities were extensively revised, since responses to the original questions were not as informative as expected.

The complete contents of the school questionnaire are described further in Exhibit 4.3.

4.6 Teacher Questionnaires

In each participating school, a single mathematics class was sampled for the TIMSS 1999 testing. The mathematics teacher of that class was asked to complete a questionnaire that sought information on the teacher's background, beliefs, attitudes, educational preparation, and teaching load, as well as details of the instructional approach used in teaching mathematics to the sampled class. The science teacher (or teachers) of the students in that class was asked to complete another questionnaire, which in many respects paralleled that for the mathematics teachers. Although the general background questions were the same for the two versions, questions pertaining to instructional practices, content coverage, classroom organization, teachers' perceptions about teaching, and views of subject matter were geared towards mathematics or science. Many questions, such as those related to classroom characteristics, activities and homework practices were answered with respect to the specific mathematics and science classes of the sampled TIMSS students.

Like the school questionnaire, the teacher questionnaires were carefully constructed to elicit information on variables thought to be associated with student achievement. Some of the important research questions addressed by the teacher questionnaires were:

- What are the characteristics of mathematics and science teachers?
- What are teachers' perceptions about mathematics and science?
- How do teachers spend their school-related time?
- How are mathematics and science classes organized?
- What activities do students do in their mathematics and science lessons?
- How are calculators and computers used?
- How much homework are students assigned?
- What assessment and evaluation procedures do teachers use?

Several changes were made in the mathematics and science teacher questionnaires for the 1999 assessment. The originals were judged to be too lengthy by most NRCs, and some of the questions needed revision. The first section of the teacher questionnaires dealt with teacher background, experience, attitudes, and teaching load. The 1999 version omitted questions about grades taught, and added several questions on teacher education and preparation for teaching. The review of the descriptive statistics and the error diagnostics produced from the field test also revealed some problems associated with filter questions that were resolved prior to the administration of the questionnaires for the main survey.

The second section of the teacher questionnaires dealt with teaching mathematics or science to the class sampled for TIMSS 1999 testing. This section was shortened, mainly by omitting a set of questions on teaching activities in a recent lesson. A lengthy set of questions on the coverage of mathematics and science topics in class was also simplified and shortened considerably. Additions to the teacher questionnaires for 1999 included questions on subject matter emphasis in class, use of computers and the Internet in class, and teacher activities in class. Two further sections of the original questionnaires, dealing with opportunity to learn and pedagogical approach, were judged by NRCs to be too lengthy; these were omitted from the field-test versions, and consequently also from the TIMSS 1999 final questionnaires.

The complete contents of the mathematics and science teacher questionnaires are described further in Exhibit 4.4.

4.7 Student Questionnaire

Each student in the sampled class was asked to complete a student questionnaire, which sought information about the student's home background, attitudes and beliefs about mathematics and science, and experiences in mathematics and science class. As in 1995, two versions of the questionnaire were used:

- *General science version*: intended for systems where science is taught as a single integrated subject
- *Separate science subject version*: intended for systems where science is taught as separate subjects (e.g., biology, chemistry, earth science, and physics)

Countries administered the version of the student questionnaire that was consistent with the way in which science instruction was organized at the target grade. Although the two versions differed with respect to the science questions, the general background and mathematics-related questions were identical across the two forms. In the general science version, science-related questions pertaining to students' attitudes and classroom activities were based on single questions asking about "science," to which students were to respond in terms of the "general or integrated science" course they were taking. In the separate science subject version, several questions were asked about each science subject area, and students were to respond with respect to each science course they were taking. This structure accommodated the diverse systems that participated in TIMSS.

Consistent with the other questionnaires, the student questionnaires were designed to elicit information on some of the major factors thought to influence student achievement. Several important research questions addressed by the student background questionnaires were:

- What educational resources do students have in their homes?
- What are the academic expectations of students, their families, and their friends?
- How do students spend their out-of-school time during the school week?
- How do students perceive success in mathematics and science?
- What are students' attitudes towards mathematics and science?

Five questions from the 1995 TIMSS student questionnaire that were considered to be of lesser importance were moved from the body of the questionnaire to the "international option" section at the end. Questions added to the TIMSS 1999 questionnaire dealt with the following topics:

- Student self-concept in mathematics and science
- Internet access and use for mathematics and science activities
- Instructional activities in mathematics and science class

Experience with the 1995 TIMSS video study helped frame the questions on activities in mathematics and science class.

The complete contents of the student questionnaires are described further in Exhibit 4.5.

4.8 Summary

The school, teacher, and student questionnaires used in the TIMSS 1999 field test were modified versions of the 1995 questionnaires. The curriculum questionnaire, however, was a new addition to the study. Since TIMSS 1999 was intended to build on TIMSS 1995 in order to track trends in student achievement in mathematics and science, it was important to retain in the questionnaires those elements essential to reporting trends. Consequently, questions that were reported in the international reports were used in their original form, without modification. Not all items in the 1995 TIMSS questionnaires were used in the international reports, largely because of problems with the wording of the questions. Questions with identifiable difficulties were either revised to resolve the problem or eliminated. Occasionally new questions were introduced, either as replacements for eliminated items or to provide extra information in areas considered important to the study. In many cases, questions that were originally dichotomous were expanded to include a range of responses. In general, every effort was made to shorten and streamline the questionnaires in order to reduce the burden on respondents.

Exhibit 4.1 Contents of the Mathematics Curriculum Questionnaire

Question Number	Item Content	Description
PART I: Structure of the Curriculum		
1	National / Regional Curriculum	Identifies countries with a national vs. regional curriculum in mathematics, year the curriculum was introduced, and whether revisions are underway.
2	Standards	Provides information on whether achievement standards are incorporated into the curriculum.
3	Supporting and Monitoring Curriculum Implementation	Identifies steps taken to support and monitor implementation of the national curriculum (e.g., teacher training, school inspections).
4	Examinations and Assessments	Provides information on which countries have public examinations and/or assessments in mathematics, whether they are sample-based, and the grades at which they are administered.
5	Specialist Teachers	Identifies the grade level at which mathematics is first taught by specialist mathematics teachers.
6	Instructional Time	Describes the amount of instructional time expected to be devoted to mathematics instruction at grades 4, 6, and 8 as dictated by the curriculum.
7	Organization of the Curriculum	Identifies the underlying organizational structure of the curriculum (e.g., by subject area).
8	Differentiation of Curriculum	Provides information on whether the curriculum is designed to deal with students of different ability levels (e.g., different curricula for different groups, same curriculum for all groups).
9	Curricular Emphasis	Identifies the extent to which the curriculum emphasizes each of several approaches / processes (e.g., mastering basic skills, solving non-routine problems).
10	Calculator Use	Identifies the policy on calculator use in grade 8 mathematics.
11	Computer Use	Identifies the policy on computer use in grade 8 mathematics.
PART II: Emphasis on Mathematics Topics		
12a	Fractions and Number Sense (15 subtopics)	Identifies the percentage of students expected to have been taught specific Fractions and Number Sense topics (e.g., understanding and representing decimal fractions) up to and including grade 8.
12b	Measurement (9 subtopics)	Identifies the percentage of students expected to have been taught specific Measurement topics (e.g., converting units of measurement).
12c	Geometry (13 subtopics)	Identifies the percentage of students expected to have been taught specific Geometry topics (e.g., angles, Pythagorean theorem).
12d	Proportionality (3 subtopics)	Identifies the percentage of students expected to have been taught specific Proportionality topics (e.g., rate problems, ratios).
12e	Algebra (11 subtopics)	Identifies the percentage of students expected to have been taught specific Algebra topics (e.g., simple algebraic expressions, solving simultaneous equations with two variables).
12f	Data Representation, Analysis, and Probability (5 subtopics)	Identifies the percentage of students expected to have been taught specific Data Representation, Analysis, and Probability topics (e.g., graphing data, simple probabilities).

Exhibit 4.2 Contents of the Science Curriculum Questionnaire

Question Number	Item Content	Description
PART I: Structure of the Curriculum		
1	National / Regional Curriculum	Identifies countries with a national vs. regional curriculum in science, year the curriculum was introduced, and whether revisions are underway.
2	Science Subjects Offered	Provides information on the science courses offered up to an including grade 8 (e.g., biology, chemistry, physics).
3	Standards	Provides information on whether achievement standards are incorporated into the curriculum.
4	Supporting and Monitoring Curriculum Implementation	Identifies the steps taken to support and monitor implementation of the national curriculum (e.g., teacher training, school inspections).
5	Examinations and Assessments	Provides information on which countries have public examinations and/or assessments in science, whether they are sample-based, and the grades at which they are administered.
6	Specialist Teachers	Identifies the grade level at which science is first taught by specialist science teachers.
7	Instructional Time	Describes the amount of instructional time expected to be devoted to science instruction at grades 4, 6, and 8 as dictated by the curriculum.
8	Organization of the Curriculum	Identifies the underlying organizational structure of the curriculum (e.g., by subject area).
9	Differentiation of Curriculum	Provides information on whether the curriculum is designed to deal with students of different ability levels (e.g., different curricula for different groups, same curriculum for all groups).
10	Curricular Emphasis	Identifies the extent to which the curriculum emphasizes each of several approaches / processes (e.g., knowing basic science facts, performing science experiments).
11	Computer Use	Identifies the policy on computer use in grade 8 science.
PART II: Emphasis on Science Topics and Skills		
12a	Earth Science (4 subtopics)	Identifies the percentage of students expected to have been taught specific Earth Science topics (e.g., Earth's atmosphere, Earth in the solar system).
12b	Biology (7 subtopics)	Identifies the percentage of students expected to have been taught specific Biology topics (e.g., human bodily processes, biology of plant and animal life).
12c	Chemistry (12 subtopics)	Identifies the percentage of students expected to have been taught specific Chemistry topics (e.g., classification of matter, chemical reactivity and transformations).
12d	Physics (10 subtopics)	Identifies the percentage of students expected to have been taught specific Physics topics (e.g., physical properties and physical changes of matter, forces and motion).
12e	Environmental and Resource Issues (3 subtopics)	Identifies the percentage of students expected to have been taught specific Environmental and Resources Issues topics (e.g., pollution, conservation of natural resources).
12f	Nature of Science and Scientific Inquiry Skills (6 subtopics)	Identifies the percentage of students expected to have been taught specific Nature of Science and Scientific Inquiry Skills topics (e.g., scientific method, experimental design).

Exhibit 4.3 **Contents of the School Questionnaire**

Question Number	Item Content	Description
1	Community	Situates the school within a community of a specific type.
2-4	Staff	Describes the school's professional full and part-time staff and the percentage of teachers at the school for 5 or more years.
5	Years Students Stay with Teacher	Indicates the number of years students typically stay with the same teacher.
6	Collaboration Policy	Identifies the existence of a school policy promoting teacher cooperation and collaboration.
7	Principal's Time	Indicates the amount of time the school's lead administrator typically spends on particular roles and functions.
8	School Decisions	Identifies who has the responsibility for various decisions for the school.
9	Curriculum Decisions	Identifies the amount of influence various individuals and educational and community groups have on curriculum decisions.
10	Formal Goals Statement	Indicates the existence of school-level curriculum goals for mathematics and science.
11-12	Instructional Resources	Provides a description of the material factors limiting the school's instructional activities.
13	Students in the school	Provides total school enrollment and attendance data.
14	Students in the target grade	Provides target grade enrollment and attendance data, student's enrollment in mathematics and science courses, and typical class sizes.
15	Number of Computers	Provides the number of computers for use by students in the target grade, by teachers, and in total.
16	Internet Access	Identifies whether the school has Internet access as well as identifying whether the school actively posts any school information on the world wide web.
17	Student Behaviors	Provides a description of the frequency with which schools encounter various unacceptable student behaviors.
18	Instructional Time	Indicates the amount of instructional time scheduled for the target grade, according to the school's academic calendar.
19	Instructional Periods	Indicates the existence and length of weekly instructional periods for the target grade.
20	Organization of Mathematics Instruction	Describes the school's provision for students with different ability levels in mathematics (e.g., setting/streaming, tracking, and remedial/enrichment programs).
21	Program Decision Factors in Mathematics	Indicates how important various factors are in assigning students to different educational programs or tracks in mathematics.
22	Organization of Science Instruction	Describes the school's provision for students with different ability levels in science (e.g., setting/streaming, tracking, and remedial/enrichment programs).
23	Program Decision Factors in Science	Indicates how important various factors are in assigning students to different educational programs or tracks in science.
24	Admissions	Describes the basis on which students are admitted to the school.
25	Parental Involvement	Describes the kinds of activities in which parents are expected to participate (e.g., serve as teacher's aids, fundraising).

Exhibit 4.4 Contents of the Teacher Questionnaires

Question Number	Item Content	Description
Section A		
1-2	Age and Sex	Identifies teacher's sex and age range.
3	Teaching Experience	Describes the teacher's number of years of teaching experience.
4-5	Instructional Time	Identifies the number of hours per week the teacher devotes to teaching mathematics, science, and other subjects.
6	Administrative Tasks	Identifies the number of hours per week spent on administrative tasks such as student supervision and counseling.
7	Other Teaching-Related Activities	Describes the amount of time teachers are involved in various professional responsibilities *outside* the formally-scheduled school day.
8	Teaching Activities	Describes the total number of hours per week spent on teaching activities.
9	Meet with Other Teachers	Describes the frequency with which teachers collaborate and consult with their colleagues.
10	Teacher's Influence	Describes the amount of influence that teachers perceive they have on various instructional decisions.
11	Being Good at Mathematics / Science	Describes teacher's beliefs about what skills are necessary for students to be good at mathematics / science.
12	Ideas about Mathematics / Science	Describes teacher's beliefs about the nature of mathematics / science and how the subject should be taught.
13	Document Familiarity	Describes teacher's knowledge of curriculum guides, teaching guides, and examination prescriptions (country-specific options).
14	Mathematics / Science Topics Prepared to Teach	Provides an indication of teacher's perceptions of their own preparedness to teach the TIMSS 1999 in-depth topic areas in mathematics or science.
15-18	Formal Education and Teacher Training	Describes the highest level of formal education completed by the teacher, the number of years of teacher training completed, and the teacher's major area of study.
International Options		
19-20	Career Choices	Identifies whether teaching was a first choice and if the teacher would change careers if given the opportunity.
21	Social Appreciation	Describes whether teachers believe society appreciates their work.
22	Student Appreciation	Describes whether teachers believe students appreciates their work.
23	Books in Home	Provides an indicator of teacher's cultural capital.

Exhibit 4.4 Contents of the Teacher Questionnaires (continued)

Question Number	Item Content	Description
Section B		
1	Target Class	Identifies the number of students in the TIMSS 1999 tested class, by gender.
2	Instructional Emphasis	Identifies the subject matter emphasized most in the target mathematics / science class.
3	Instructional Time	Identifies the number of minutes per week the class is taught.
4	Textbook Use	Identifies whether textbook is used in mathematics / science class as well as the approximate percentage of weekly instructional time that is based on the textbook.
5-7	Calculators	Describes the availability of calculators and how they are used in the target class.
8	Computers	Describes the availability of computers and whether they are used to access the internet.
9	Planning Lessons	Identifies the extent to which a teacher relies on various sources for planning lessons (e.g., curriculum guides, textbooks, exam specifications).
10	Tasks Students are Asked to Do	Describes the frequency with which teachers ask students various types of questions and ask students to perform various mathematics / science activities during lessons.
11	Student's Work Arrangements	Describes how often students work in various group arrangements.
12	Time Allocation	Describes the percentage of time spent on each of several activities associated with teaching (e.g., homework review, tests).
13	Mathematics / Science Topic Coverage	Indicates the extent of teacher's coverage in target class of mathematics / science topics included in the assessment.
14	Classroom Factors	Identifies the extent to which teachers perceive that various factors limit classroom instructional activities.
15-16	Amount of Homework Assigned	Describes the frequency and amount of homework assigned to the target class.
17-18	Type and Use of Homework	Describes the homework assignments and how the homework is used by the teacher.
19-20	Assessment	Describes the kind and use of various forms of student assessment in the target class.

Exhibit 4.5 Contents of the Student Questionnaires

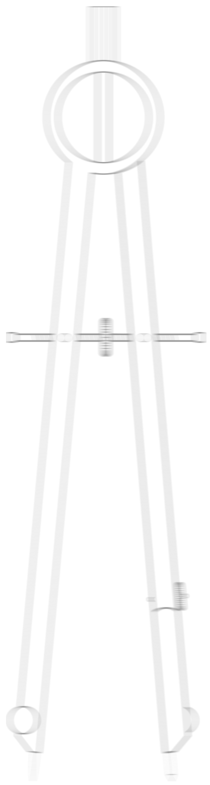
Question Number		Item Content	Description
General Version	Separate Science Version		
1-4	1-4	Student Demographics	Provides basic demographic information such as age, sex, language of the home, whether born in country and if not how long he/she has lived in country.
5	5	Academic Activities Outside of School	Provides information on student activities that can affect their academic achievement (e.g., extra lessons, science club).
6	6	Time Spent Outside of School	Provides information about the amount of time student spends on homework and leisure activities on a normal school day.
7	7	Parents' Education	Provides information about the educational level of the student's mother and father. Used as an indicator of the home environment and socioeconomic status.
8	8	Student's Future Educational Plans	Identifies the student's plans for further education.
9	9	Parents' Country of Birth	Provides information regarding immigrant status.
10	10	Books in the home	Provides information about the number of books in the home. Used as an indicator of the home environment and socioeconomic status.
11	11	Possessions in the home	Provides information about possessions found in the home (e.g., calculator, computer, study desk, country-specific items). Used as an indicator of academic support in the home environment as well as an indicator of socioeconomic status.
12	12	Mother's Values	Provides information about the student's perception of the degree of importance his/her mother places on academics and other activities. Used as an indicator of the home environment and general academic press
13	13	Student's Behavior in Mathematics Class	Provides a description of typical student behavior during mathematics lessons.
14	14	Peers' Values	Provides information about the student's perception of the degree of importance his/her peers place on academics and other activities. Used as an indicator of peers' values and student's social environment.
15	15	Student's Values	Provides information about the degree of importance the student places on academics and other activities. Used as an indicator of student's values.
16	16	Competence in Mathematics / Science	Provides an indication of student's self-description of academic competence in mathematics and science (specialized version asks about biology, earth science, chemistry, and physics separately).
17	17	Difficulty of Mathematics	Provides a description of student's perception of the difficulty level of mathematics.
18	18	Doing Well in Mathematics	Identifies student's attributions for doing well in mathematics.
19	19-22	Difficulty of Science	Provides a description of student's perception of the difficulty level of science (specialized version asks about biology, earth science, chemistry, and physics separately).
20	23	Doing Well in Science	Identifies student's attributions for doing well in science.

Exhibit 4.5 Contents of the Student Questionnaire (continued)

Question Number		Item Content	Description
General Version	Separate Science Version		
21	24	Liking Mathematics / Science	Identifies how much students like mathematics and science; a key component of student motivation (specialized version asks about biology, earth science, chemistry, and physics separately).
22	25	Liking Computers for Mathematics / Science	Identifies how much students like using computers to learn mathematics and science.
23	26	Internet Access	Identifies whether students are accessing the Internet and for what purposes they are using it.
24	27	Interest, Importance, & Value of Mathematics	Provides a description of student's interest, importance rating, and value attributed to mathematics.
25	28	Reasons to Do Well in Mathematics	Provides the extent to which students endorse certain reasons they need to do well in mathematics.
26	29	Classroom Practices in Mathematics	Provides a description of student's perceptions of classroom practices in mathematics instruction.
27	30	Beginning a New Mathematics Topic	Describes the frequency with which specific strategies are used in the classroom to introduce a new mathematics topic.
28	31	Taking Science Class(es)	Identifies whether or not the student is enrolled in science classes this year (specialized version asks about biology, earth science, chemistry, and physics separately).
29	32, 36, 40, 44	Interest, Importance, & Value of Science	Provides a description of student's interest, importance rating, and value attributed to science (specialized version asks about biology, earth science, chemistry, and physics separately).
30	33, 37, 41, 45	Reasons to Do Well in Science	Provides the extent to which students endorse certain reasons they need to do well in science (specialized version asks about biology, earth science, chemistry, and physics separately).
31	34, 38, 42, 46	Classroom Practices in Science	Provides a description of student's perceptions of classroom practices in science instruction (specialized version asks about biology, earth science, chemistry, and physics separately).
32	35, 39, 43, 47	Beginning a New Science Topic	Describes the frequency with which specific strategies are used in the classroom to introduce a new science topic (specialized version asks about biology, earth science, chemistry, and physics separately).
International Options			
33-34	48-49	People Living in the Home	Provides information about the home environment as an indicator of academic support and economic capital.
35-36	50-51	Cultural Activities	Provides a description of student's involvement in cultural events or programming such as plays or concerts.
37	52	Report on Student Behaviors	Provides an indication of the student's perspective of the existence of specific problematic student behaviors at school.
38	53	Environmental Issues	Provides and indication of student's beliefs about how much the application of science can help in addressing environmental issues.
39	54	Science Use in a Career	Identifies preference for sciences in careers.

References

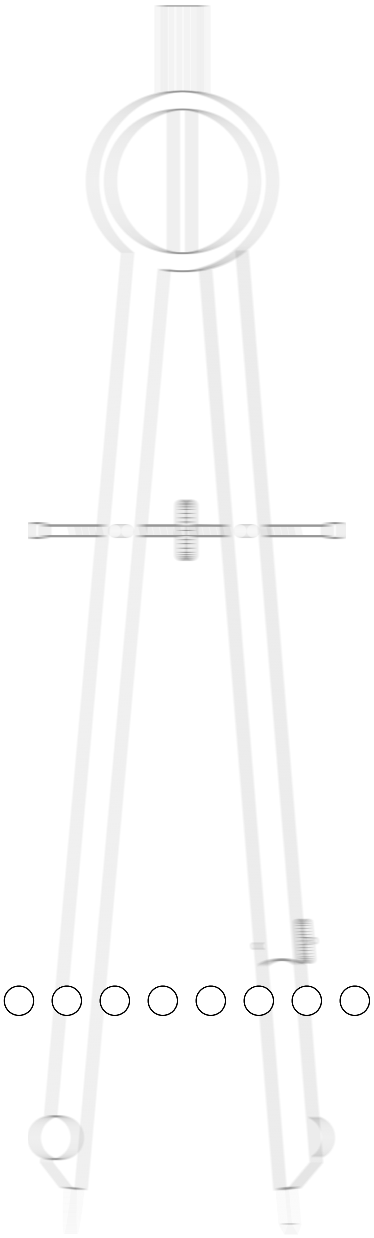
Schmidt, W. & Cogan, L. (1996). Development of the TIMSS Context Questionnaires, in M.O. Martin & D.L. Kelly (Eds.), *Third International Mathematics and Science Study Technical Report, Volume 1*. Chestnut Hill, MA: Boston College.



5

Translation and Cultural Adaptation of the TIMSS Instruments

Kathleen M. O'Connor
Barbara Malak



5 Translation and Cultural Adaptation of the TIMSS Instruments

Kathleen M. O'Connor
Barbara Malak

5.1 Overview

The TIMSS 1999 data-collection instruments (achievement tests and background questionnaires) were prepared in English and translated into 33 languages. Ten of the thirty-eight participating countries collected data in two languages. The most common languages of testing were English (nine countries) and Arabic (four countries).

For the TIMSS 1999 main survey, each country had to translate the following instruments:

- Eight booklets of mathematics and science achievement items (Test Booklets 1-8)
- One Student Questionnaire
- One Mathematics Teacher Questionnaire
- One Science Teacher Questionnaire
- One School Questionnaire

The translation process was designed to ensure standard instruments across countries. National Research Coordinators (NRCs) received guidelines for translating the testing instruments into their national languages and cultural context (TIMSS, 1998a). After the translation was completed, the translated instruments were checked by an international translation company against the TIMSS 1999 international version to assess the faithfulness of translation. The NRC then received feedback from the translation company and the International Study Center suggesting additional revisions. After these had been made, the final version was checked by the International Study Center at Boston College.

5.2 Translation of Instruments

The TIMSS 1999 survey translation guidelines called for two independent translations of each test instrument from English into the target language. A translation review team then compared the two translations to arrive at a final version. Any deviation from the international version of the instrument and all cultural adaptations made were reported on a *Translation Deviation Form* (blank form is provided in Appendix B of this report).

The translation procedure at the National Research Centers included the following steps:

- Identify the test language
- Identify translators for two independent translations
- Translate instruments and adapt as necessary
- Confer and reconcile the two independent translations
- Document all translation deviations and cultural adaptations

5.2.1 Identifying the Test Language

Each NRC identified the language or languages to be used in testing and the geographical or political areas associated with them. Most countries tested in just one language, but 9 tested in two languages (see Exhibit 5.1).

If a single translation was prepared within a country, translators needed to ensure that it was acceptable to all of the dialects of the language in which the assessment was to be administered. Professionals in these dialects were to be involved in adapting the instruments and testing materials. The language of the test in each country is presented in Exhibit 5.1.

Exhibit 5.1 Language of Testing in Each Country

Country	Language(s) of Test	Country	Language(s) of Test
Australia	English	Latvia	Latvian
Belgium (Flemish)	Flemish	Lithuania	Lithuanian
Bulgaria	Bulgarian	Macedonia, Rep. of	Macedonian and Albanian
Canada	English and French	Malaysia	Malay
Chile	Spanish	Moldova	Moldavian and Russian
Chinese Taipei	Chinese	Morocco	Arabic
Cyprus	Greek	Netherlands	Dutch
Czech Republic	Czech	New Zealand	English
England	English	Philippines	English and Filipino
Finland	Finnish and Swedish	Romania	Romanian
Hong Kong, SAR	Chinese and English	Russian Federation	Russian
Hungary	Hungarian	Singapore	English
Indonesia	Indonesian	Slovak Republic	Slovak
Iran, Islamic Rep.	Farsi	Slovenia	Slovenian
Israel	Hebrew and Arabic	South Africa	English and Afrikaans
Italy	Italian and German*	Thailand	Thai
Japan	Japanese	Tunisia	Arabic and French**
Jordan	Arabic	Turkey	Turkish
Korea, Republic of	Korean	United States	English

* Italy did not have the German version of the items and student questionnaire verified. Less than 1% of the population took the assessment and student questionnaire in German.

** Tunisia translated only the Teacher Questionnaires into French.

5.2.2 Identifying Translators for Two Independent Translations

Translators were expected to have an excellent knowledge of both English and the target language, experience with eighth-grade students, and experience in the subject matter and test development.

For the achievement tests, four translators were required for each target language, two each with expertise in mathematics education and in science education. Where subject-matter experts were not available as translators, the translators were expected to work closely with subject-matter experts to ensure that the content and difficulty of the items did not change in translation.

Translators of general text materials (school, teacher, and student questionnaires and manuals) did not need to be subject-matter specialists, so only two translators were necessary for these documents.

5.2.3 Translation and Cultural Adaptation of Instruments

Translators were given guidelines and procedures to follow in translating the data collection instruments and adapting them to their national cultural context. The guidelines were designed to yield translations that were as close as possible to the international (English) versions in style and meaning, while allowing for cultural adaptations where necessary. Translators were cautioned not to change the meaning or the difficulty level of an item.

The translators' tasks included:

- Identifying and minimizing cultural differences
- Finding equivalent words and phrases
- Ensuring that the reading level was the same in the target language as in the international version (English)
- Ensuring that the essential meaning of the text did not change
- Ensuring that the difficulty level of achievement items did not change
- Being aware of possible changes in the instrument layout due to translation

Translators were permitted to adapt the text as necessary to make unfamiliar contextual terms culturally appropriate. Acceptable adaptations included changes in the names of seasons, people, places, animals, plants, currencies, and the like. Exhibit 5.2 shows a list provided to translators detailing the types of adaptations that were acceptable.

Exhibit 5.2 Types of Acceptable Cultural Adaptations

Type of Change	Specific Change from	Specific Change to
Punctuation/Notation	decimal point	decimal comma
	place value comma	space
Units	centimeters	inches
	liters	quarts
	ml	mL
Proper nouns	Ottawa	Oslo
	Mary	Maria
Common nouns	robin	kiwi
	elevator	lift
Spelling	center	centre
Verbs (not related to content)	skiing	sailing
Usage	Bunsen burner	hot plate

Translators were allowed to change terms and expressions that were not common to their national culture. It was important, however, that their changes did not affect the following:

- The meaning of the question
- The reading level of the text
- The difficulty level of the item
- The likelihood of another possible correct answer for the test item

Although item writers and reviewers attempted to write and select items that would readily translate into the language of the participating countries, occasionally an item proved problematic for translators. In those instances, the International Study Center was notified and a corresponding statement included in the NRC survey activities report.

5.2.4 Reviewing Independent Translations for Consensus

The two completed translations were compared item by item, and any differences reconciled. In most cases, by discussing the differences in the translations of a particular item, the translators were able to agree on the version appropriate for the study. A third translation expert was consulted if any disagreement remained.

After a single translation had been agreed upon, the translation deviation form was used to record all deviations in test and questionnaire items. Translators documented all changes in vocabulary and content not authorized in the translation guidelines. The description of each deviation included the English term, the translated term, and an explanation of why that term was selected. Translators also noted any other changes or problems with the translation. This record of deviations was used during translation verification and during the item analysis and review.

5.3 Verification of Instruments

Each country's translated documents went through a rigorous verification process that included statistical verification of the item translations at the national centers, verification by an international translation company, a review by the International Study Center, and a check by quality control monitors.

5.3.1 Verification of Translations at National Centers

The results of item analyses from the field test were reviewed by each country. Since unusual results for an item could indicate errors in translation, each NRC was asked to check the results to identify items that might have been mistranslated. NRCs were notified of any potentially problematic items and asked to check whether the translation was sound.

5.3.2 Submission of Instruments for External Verification

Once the final translated version of each instrument was agreed upon, the translation was checked through an external verification process. NRCs were required to send (no later than 6 weeks before printing) the following material to the IEA Secretariat in preparation for external translation verification:

- One copy of the test item clusters (A through Z) and the accompanying instructions for students
- One set of test booklets (1 through 8)
- One copy of the school questionnaire, student questionnaire, and teacher questionnaires

All 38 countries that participated in the TIMSS 1999 main survey submitted national versions of instruments for translation verification (See Appendix B for a list of instruments submitted). Three countries deviated in some way from the formal verification process. Italy verified the Italian versions of the instruments but did not verify the German versions, which were administered to less than 1% of the sample. In the case of South Africa and Singapore, the results of verification were not obtained before test administration. The review and documentation of the translation deviations for both countries took place after the test had been conducted. The verifiers did, however, find that South African and Singaporean cultural adaptations of English, as well as the South African translation into Afrikaans, were of high quality.

5.3.3 International Verification of the Translations

The IEA Secretariat, which organized and managed the translation-verification process, enlisted Berlitz, an international translating company with a reputation for excellence, to check the quality of the translations. Berlitz staff were to document all errors and omissions, and make suggestions for improvements so that National Research Coordinators could review and revise their instruments.

The translators Berlitz chose as translation verifiers for TIMSS 1999 were required to have the target language as their first language, to have formal credentials as translators working in English, and to be living and working in the target country. Verifiers received general information about the study and the design of the instruments. They also received materials describing the translation procedures used by the national centers along with detailed instructions for reviewing the instruments (TIMSS, 1998b). They were asked to recommend improvements in the translation, when necessary, and to alert the national centers to any deviation in the layout of the test instruments. Each verifier received a package consisting of:

- The international version of each survey instrument
- A set of translated instruments to be verified
- A copy of the instructions given to the translators in their country
- Instructions for verifying the layout of the survey instruments
- Instructions for verifying the content of the survey instruments
- Instructions for verifying the instructions to students
- Translation verification control forms to be completed for each document
- Translation verification report forms

The main task of the translation verifiers was to evaluate the accuracy of the translation and the comparability of layout of the survey instruments. The verification guidelines emphasized the importance of maintaining the meaning, difficulty level, and format of each item while allowing for cultural adaptations as necessary.

For TIMSS 1999 countries that also participated in 1995, verifiers were responsible for ensuring that the translated version of the trend items was identical to that administered in 1995. Accordingly, verifiers reviewing instruments for trend countries also received the following:

- A set of trend item clusters A through H (1995 version used in that country)
- A trend item verification form

5.3.4 Translation Verification Reports

The translation verifier prepared two types of reports to document the verification process. First, the verifier completed a translation verification control form for each instrument. Its cover sheet served as a summary and indicated whether or not deviations were found. If the translated version was judged to be equivalent to the international version, no further entry needed to be made in the form. Second, for each translated version of an item that differed in any way from the international version, an entry was made in the translation verification report form giving:

- The location of the deviation (item #)
- The severity of the deviation (using the severity code below)
- A description of the change
- A suggested alternative translation

These records are used to document the quality of the translations and the comparability of the testing materials across countries. The *severity codes* ranged from 1 (serious error) to 4 (acceptable adaptation)[1] as follows:

○○○

[1] When in doubt as to the severity of the deviation, verifiers used code 1.

Code 1 - Major Change or Error: Examples include incorrect ordering of choices in a multiple-choice item; omission of a graph; omission of an item; incorrect translation of text such that the answer is indicated by the question; an incorrect translation that changes the meaning or difficulty of the question; incorrect ordering of the items or placement of the graphics.

Code 2 - Minor Change or Error: Examples include spelling errors that do not affect comprehension; misalignment of margins or tabs; incorrect font or font size; discrepancies in the headers or footers of the document.

Code 3 - Suggestions for Alternative: The translation may be adequate, but the verifier suggests a different wording for the item.

Code 4 - Acceptable Changes: The verifier identifies changes that are acceptable and appropriate, for example, a reference to winter that is changed from January to July for the Southern Hemisphere.

The layout of the documents was also reviewed during verification for any changes or deviations. Exhibit 5.3 details the layout issues to be considered and checked for each survey instrument.

Exhibit 5.3: Layout Issues Considered in Verification

Layout Issues	Verification Details
Instructions	Test items should not have been visible when the test booklet was opened to the Instructions
Items	All items should have been included in the same order and location as in the international version
Response options	Response options should have appeared in the same order as in the international version
Graphics	All graphics should have been in the same order and modifications should have been limited to necessary translation of text or labels
Font	Font and font size should have been consistent with the international version
Word emphasis	Word emphasis should have remained the same as in the international version; if the form of emphasis was not appropriate for the given language, an acceptable alternate form should have been used (e.g., italics instead of capital letters)
Shading	Items with shading should have been clear and text legible
Page and item identification	Headers and footers that include booklet and page identification as well as item identification should have been present
Pagination	Page breaks should have corresponded with the international version of the instruments

If the layout of an instrument differed in any way from the original international version, an entry was made in the translation verification report form indicating the location and severity of the deviation and describing the change. If necessary and appropriate, a suggestion for improving the layout was included. In the case of TIMSS 1995 participants, any differences between the 1995 and 1999 versions of test items were entered in the trend item verification form, and the nature of the change was described.

The completed translation verification forms were sent to NRCs and to the International Study Center at Boston College. The NRCs were responsible for reviewing the report forms and revising the instruments based on the translation verifiers' suggestions.

5.3.5 International Study Center Item Review

As a final review, when NRCs had acted upon the suggestions of the verifiers, they submitted a print-ready copy of the achievement test booklets and questionnaires to the International Study Center at Boston College. These were reviewed by the International Study Center primarily to identify issues such as misplaced graphics, improper format, and inconsistent text.

For all countries, items were compared with the international version to identify any changes in text, graphics, and format. For languages in which the reviewers were not fluent, items were reviewed for format and similarity of words used in the stem and options.

For trend countries, each item in Clusters A-H was compared with the 1995 translated version to note whether it had been changed. When the reviewer was not familiar with the language of these items, the NRC was asked about any apparent changes.

NRCs were given a list of any deviations identified by the International Study Center that went beyond those recorded in the translation deviation forms and translation verification forms. NRCs used these comments to correct errors prior to printing. Deviations that were not corrected before the final printing of the test booklets were noted in the database and used when reviewing the item data after the data collection.

5.3.6 Quality Control Monitor Item Review

As part of an ambitious quality control program, Quality Control Monitors (QCMs) were hired to document the quality of the TIMSS 1999 assessment in each country (see Chapter 8 for a description of the work of the Quality Control Monitors). An important task for the QCMs was to review the translation of the test items. QCMs reviewed the translation verification reports for each test language, verified whether the suggested changes were made in the final document, and noted these changes on a copy of the translation verification report.

5.4 Summary

The rigorous procedures for translation, cultural adaptations, translation verification, and review of the instruments put in place for TIMSS 1999 provided for comparable translations across participating countries. The verification process of internal statistical review, external translation verification by bilingual judges, and review by the International Study Center and Quality Control Monitors proved to be a comprehensive program for verifying and documenting deviations. The thorough documentation allowed for an informative review of anomalies, further ensuring accuracy in the analysis and reporting of the main survey data.

References

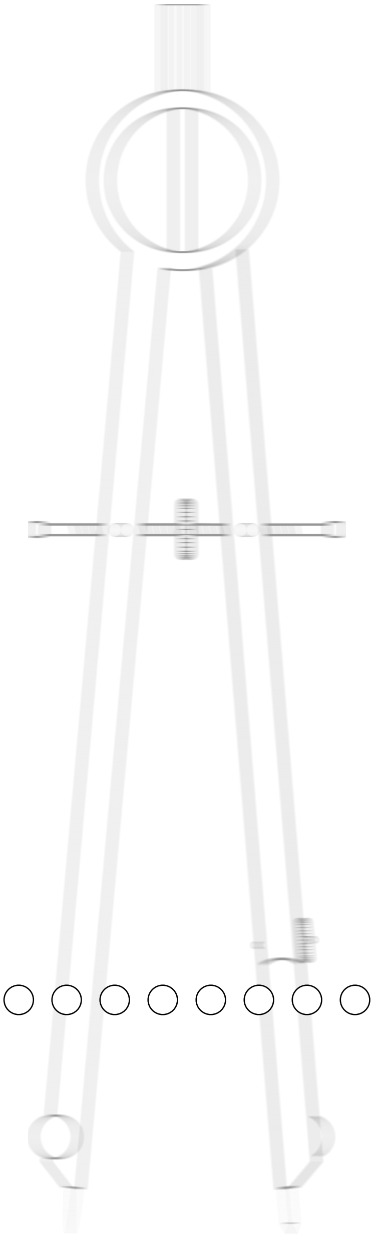
TIMSS (1998a). *Survey Operations Manual* (Doc. Ref. No. 98-0026). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

TIMSS (1998b). *Guidelines for the Translation Verification of the TIMSS-R Main Survey Instruments* (Doc. Ref. No. 98-0042). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

6

TIMSS Field Test

Kathleen M. O'Connor



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 DataEntry-Manager (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

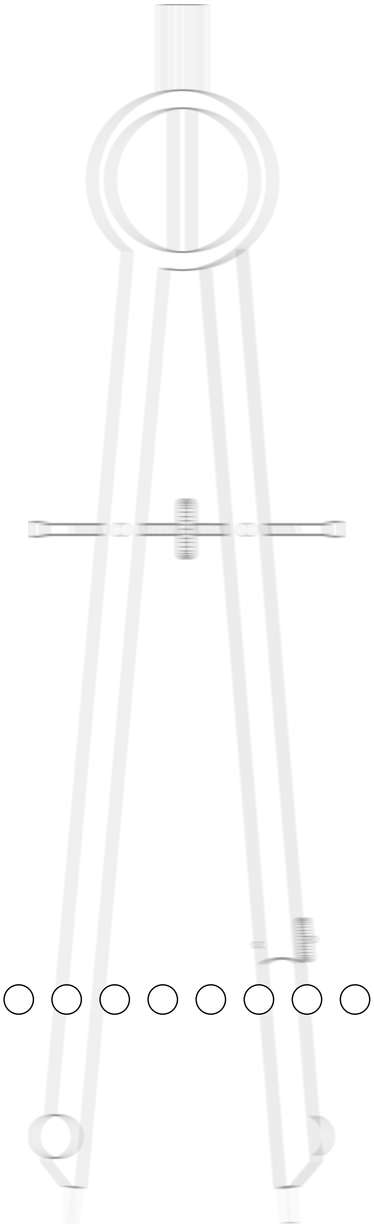
TIMSS (1998a). *Survey Operations Manual* (Doc. Ref. No. 98-0026). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

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.

TIMSS Field Operations and Data Preparation

Eugenio J. Gonzalez
Dirk Hastedt



7 TIMSS Field Operations and Data Preparation

Eugenio J. Gonzalez
Dirk Hastedt

7.1 Overview

The TIMSS 1999 data collection in each country was a very demanding exercise, requiring close cooperation between the National Research Coordinator (NRC) and school personnel - principals and teachers - and students. The first part of this chapter describes the field operations necessary to collect the data, including the responsibilities of the NRC, the procedure for sampling classrooms within schools and tracking students and teachers, and the steps involved in administering the achievement tests and background questionnaires. The second part describes the activities involved in preparing the data files at the national center, particularly the procedures for scoring the free-response items, creating and checking data files for achievement test and questionnaire responses, and dispatching the completed files to the IEA Data Processing Center in Hamburg.

7.2 TIMSS 1999 Field Operations

The TIMSS 1999 field operations were designed by the International Study Center at Boston College, the IEA Data Processing Center, and Statistics Canada. They were based on procedures used successfully in TIMSS 1995 and other IEA studies, and refined on the basis of experience with the TIMSS 1999 field test.

7.2.1 Responsibilities of the National Research Coordinator

In each country, the National Research Coordinator was the key person in conducting the field operations. The NRC was responsible for collecting the data for the TIMSS assessment according to internationally agreed procedures and preparing the data according to international specifications. Earlier chapters of this report have outlined the tasks of the NRC with regard to choosing a sample of schools and translating the achievement tests and questionnaires.[1] This section focuses on NRC activities with regard to administering the assessment in participating schools.

○○○

[1] See Chapter 2 for information about sampling schools, and Chapter 5 for details of the translation task.

Specifically it describes the procedures for sampling classes within schools, for tracking classes, teachers, and students in the sampled schools, and for organizing the administration of the achievement tests and questionnaires.

7.2.2 Documentation and Software

NRCs were provided with a comprehensive set of procedural manuals detailing all aspects of the data collection.

- The *Survey Operations Manual* (TIMSS, 1997a) was the essential handbook of the National Research Coordinator, and described in detail all of the activities and responsibilities of the NRC, from the moment the TIMSS instruments arrived at the national center to the moment the cleaned data files and accompanying documentation were submitted to the IEA Data Processing Center.
- The *TIMSS-R School Sampling Manual* (TIMSS, 1997b) defined the TIMSS 1999 target population and sampling goals and described the procedures for the sampling of schools.
- The *School Coordinator Manual* (TIMSS, 1997c) described the activities of the school coordinator (the person in the school responsible for organizing the TIMSS test administration), from the time the testing materials arrived at the school to the time the completed materials were returned to the national TIMSS center.
- The *Test Administrator Manual* (TIMSS, 1997d) described in detail the procedures for administering the TIMSS tests and questionnaires, from the beginning of the test administration to the return of the testing materials to the school coordinator.
- The *Scoring Guides for Mathematics and Science Free-Response Items* (TIMSS, 1998a) contained instructions for scoring the short-answer and extended-response test items.
- The *Manual for Entering the TIMSS-R Data* (TIMSS, 1998b) provided the NRC with instructions for coding, entering, and verifying the data. The manual included the codebook, which defined the variables and file formats in the data files.
- The *Manual for National Quality Control Observers* (TIMSS, 1998c) provided instructions for conducting classroom observations in a sample of participating schools.

Additionally, two software packages were supplied by the IEA Data Processing Center to assist NRCs in the main study

- The within-school sampling software (W3S), a computer program designed to help NRCs select the within-school sample, prepare the survey tracking forms, and assign test booklets to students was supplied along with its corresponding manual.
- The DataEntryManager, a computer program for data entry and data verification was supplied along with its corresponding manual.

In addition to the manuals and software, NRCs received hands-on training in the procedures and use of the software from staff of the International Study Center, the IEA Data Processing Center, and Statistics Canada.

7.2.3 Within-School Sampling Procedures

The study design anticipated relational analyses between student achievement and teacher-level data at the class level. For field operations, this meant that intact classes had to be sampled, and that for each sampled class the mathematics and science teachers had to be tracked and linked to their students. Although intact classes were the unit to be sampled in each school, the ultimate goal was a nationally representative sample of students. Consequently, in each country a classroom organization had to be chosen that ensured that every student in the school was in one class or another, and that no student was in more than one class. Such an organization is necessary for a random sample of classes to result in a representative sample of students. In most countries at the eighth grade, mathematics classes serve this purpose well, and so were chosen as the sampling units. In countries where students attended different classes for mathematics and science, classrooms were defined on the basis of mathematics instruction for sampling purposes.[2]

The TIMSS design required that for each student in each sampled class, all eighth-grade mathematics and science teachers be identified and asked to complete a teacher questionnaire.

[2] For countries where a suitable configuration of classes for sampling purposes could not be identified, TIMSS also provided a procedure for sampling individual students directly from the eighth grade.

When sampling mathematics classes in a school, the procedure was as follows:

- The NRC asked the school coordinator for a list of all mathematics classes in the target (eighth) grade along with the names of their mathematics teachers.
- The school coordinator sent the requested list to the NRC.
- The NRC transcribed the information onto a document known as a *Class Sampling Form* and applied a prescribed sampling algorithm to select one or more classes.
- For each sampled class, the NRC prepared a *Teacher-Student Linkage Form* designed to link the students in the class to each of their eighth-grade mathematics and science teachers. The form was then sent to the school coordinator to be completed.
- The school coordinator completed the Teacher-Student Linkage Form by listing all of the students in the class (name or identification number, date of birth, and sex), together with their mathematics and science teachers and classroom identifiers as necessary, and returned it to the NRC.
- From the information provided in the Teacher-Student Linkage Form, the NRC produced a *Student Tracking Form,* which listed all students in the class to be tested with their TIMSS identification numbers and booklet assignments, and a *Teacher Tracking Form,* which listed all mathematics and science teachers of the students in the class, their student-teacher link numbers, and their questionnaire assignments. These forms were sent to the school coordinator along with the test instruments.
- During the test administration, the test administrator and school coordinator used the tracking forms to record student and teacher participation, and returned them to the NRC after the test administration together with the completed test booklets and questionnaires.

7.2.4 Excluding Students from Testing

Although all students enrolled in the target grade were part of the target population and were eligible to be selected for testing, TIMSS recognized that some students in every school would be unable to take part in the 1999 assessment because of some physical or mental disability. Accordingly, the sampling procedures provide for the exclusion of students with any of several disabilities (see Chapter 2). Countries were required to track and account for all excluded students, and were cautioned that excluding an excessive proportion would lead to their results being annotated in international reports. It was important that the conditions under which students could be excluded be carefully delineated, because the definition of "disabled" students varied considerably from country to country.

7.2.5 Survey Tracking Forms

As is evident from the description of the within-school sampling procedure provided earlier, TIMSS 1999 relied on a series of "tracking forms" to implement and record the sampling of classes, teachers, and students. It was essential that the tracking forms were completed accurately, since they made explicit exactly who should be given which instruments, and recorded what happened in each school. In addition to facilitating the data collection, the tracking forms provided essential information for the computation of sampling weights and for evaluating the quality of the sampling procedures All tracking forms were retained for review by staff of the International Study Center.

Survey tracking forms were provided for sampling classes and students; for tracking schools, classes, teachers, and students; for linking students and teachers; and for recording information during test administration. Each of these forms is described below.

7.2.6 Linking Students, Teachers, and Classes

Within each school, a class identification number (ID) was assigned to each class in the target grades listed on the class tracking form. The class ID consisted of the three-digit school ID plus a two-digit identification number for the class within the school.

Each student listed on the student tracking form was assigned a student identification number. This was a seven-digit number consisting of the five-digit class ID plus a two-digit number corresponding to the student's sequential position in the student tracking form. All students listed on the student tracking form, including those marked for exclusion, had to be assigned a student ID.

All mathematics and science teachers of the selected classes (those listed on the teacher tracking form) were assigned a teacher ID that consisted of the three-digit school ID plus a two-digit number of the teacher within the school. Since a teacher could be teaching both mathematics and science to some or all of the students in a class, it was necessary to have a unique identification number for each teacher/class and teacher/subject combination. This was achieved by adding a two-digit link number to the five digits of the teacher ID, giving a unique seven-digit teacher/class identification number. Careful implementation of these procedures was necessary so that later each class could be linked to a teacher, and student outcomes could be analyzed in relation to teacher-level variables.

7.2.7 Assigning Testing Materials to Students and Teachers

Eight different test booklets were distributed to the students in each sampled class. Each student was required to complete one booklet, and the student questionnaire. Booklets were assigned to students by the NRC using a random assignment procedure, after which the assignment was recorded on the student tracking form.

Each teacher listed on the teacher tracking form was assigned a mathematics or a science teacher questionnaire. Where teachers taught both mathematics and science to the class, every effort was made to collect information about both. However, NRCs had the final decision as to how much response burden to place on such teachers.

7.2.8 Administering the Test Booklets and Questionnaires

The school coordinator was the person in the school responsible for organizing the administration of the TIMSS 1999 tests. This could be the principal, the principal's designee, or an outsider appointed by the NRC with the approval of the principal. The NRC was responsible for ensuring that the school coordinators were familiar with their responsibilities.

The major responsibilities of the school coordinators are detailed in the school coordinator manual (TIMSS, 1997c). Prior to the test administration the tasks for the school coordinator included:

- Providing the NRC with all information necessary to complete the various tracking forms
- Checking the testing materials when they arrived in the school to ensure that everything was in order
- Ensuring that the testing materials were kept in a secure place before and after the test administration
- Arranging the dates of the test administration with the national center

- Arranging for a test administrator and giving a briefing on the TIMSS 1999 study, the testing materials, and the testing sessions
- Working with the school principal, the test administrator, and the teachers to plan the testing day; this involved arranging rooms, times, classes and materials.

The Test Administrator was responsible for administering the TIMSS tests and student questionnaires. Specific responsibilities were described in the test administrator manual (TIMSS, 1997d), and included:

- Ensuring that each student received the correct testing materials which were specially prepared for him or her
- Administering the test in accordance with the instructions in the manual
- Ensuring the correct timing of the testing sessions by using a stopwatch and recording the time when the various sessions started and ended on the test administration form
- Recording student participation on the student tracking form.

The responsibilities of the school coordinator after the test administration included:

- Ensuring that the test administrator returned all testing materials, including the completed student tracking form, the test administration form, and any unused materials
- Calculating the student response rate and arranging for makeup sessions if it was below 90%
- Distributing the teacher questionnaires to the teachers listed on the teacher tracking form, ensuring that the questionnaires were returned completed, and recording teacher participation information on the teacher tracking form
- Preparing a report for the NRC about the test administration in the school
- Returning both completed and unused test materials and all tracking forms to the NRC.

The NRC prepared two packages for each sampled class. One contained the test booklets for all students listed on the student tracking form and the other the student questionnaires. For each participating school, the test booklets and student questionnaires were bundled together with the teacher tracking form and teacher questionnaires, the school questionnaire, and the materials prepared for briefing school coordinators and test administrators, and were sent to the school coordinator. A set of labels and prepaid envelopes addressed to the NRC was included to facilitate the return of testing materials.

7.3 National Quality Control Program

The International Study Center implemented an international quality control program whereby international quality control monitors visited a sample of 15 schools in each country and observed the test administration. In addition, NRCs were expected to organize a national quality control program, based upon the international model. This national program required Quality Control Observers to document data collection activities in their country. They visited a 10% sample of TIMSS 1999 schools, observed actual testing sessions, and recorded compliance of the test administration with prescribed procedures.

The International Study Center prepared *The Manual for National Quality Control Observers* (TIMSS, 1998c) which contained information about TIMSS 1999, and detailed the role and responsibilities of the National Quality Control Observers.

7.4 Data Preparation

In the period immediately following the administration of the TIMSS 1999 tests, the major tasks for the NRC included retrieving the materials from the schools; recruiting and training scorers to score the free-response items; scoring these items, including double scoring a 25% reliability sample; entering the data from the achievement tests and background questionnaires; submitting the data files and materials to the IEA Data Processing Center; and preparing a report on survey activities.

When the testing materials were received back from the schools, NRCs were to do the following:

- Check that the appropriate testing materials were received for every student listed on the student tracking form
- Verify all identification numbers on all instruments that were not precoded at the national center
- Check that the participation status recorded on the tracking forms matched the information on the test instruments

- Follow up on schools that did not return the testing materials or for which forms were missing, incomplete, or inconsistent.

NRCs then organized the tests for scoring and data entry. The procedures involved were designed to maintain identification information that linked students to schools and teachers, minimize the time and effort spent handling the booklets, ensure reliability in the free-response coding, and document the reliability of the coding.

7.4.1 Scoring the Free-Response Items

Reliable application of the scoring guides to the free-response questions, and empirical documentation of the reliability of the scoring process, were critical to the success of TIMSS 1999. The survey operations manual (TIMSS, 1997a) contained information about arranging for staff and facilities for the free-response scoring effort required for the TIMSS 1999 main survey; for effective training of the scorers; and for distributing booklets to scorers to accomplish the scoring for the main data set. Countries were to double score a 25% sample to document scoring reliability.

For most countries, the scope of the free-response scoring effort was substantial. The main survey contained 68 free-response questions. Each of the 8 booklets had between 9 and 14 free-response questions. On average, each country had to score about 50,000 student responses.

To ascertain the staff requirements for free-response scoring, it was necessary to estimate the amount of scoring to be done and the amount of time available to do it, and also to make provision for staff training and for clerical and quality control throughout the operation. The International Study Center recommended at least one half-day of training on each of the 8 booklets, for a total of about a week for training activities.

In scoring the free-response items it was vital that scoring staff apply the scoring rules consistently and in the same way in all participating countries. Hence, in selecting those who were to do the scoring, NRCs took care to arrange for persons who were conscientious and attentive to detail, knowledgeable in mathematics and science, and willing to apply the scoring guides as stated, even if they disagreed with a particular definition or category. Preference was given to individuals with educational backgrounds in the mathematics and science curriculum areas or who had taught at the middle school level. Good candidates for scoring included teachers, retired teachers, college or graduate students, and staff of education agencies or ministries and research centers.

7.4.2 Preparing Materials to Train the Scorers

The success of assessments containing free-response questions depends upon the reliability of scoring student responses. In TIMSS 1999, reliability was assured through the provision of scoring guides (manuals), extensive training in their use, and monitoring of the quality of the work. In addition, TIMSS 1999 provided training packets for training in selected questions, and practice papers to help scorers achieve a consistent level of scoring.

Each scorer received a copy of the *TIMSS 1999 Main Survey Scoring Guides for Mathematics and Science Free-Response Items* (TIMSS, 1998a). This document explained the TIMSS scoring system, which was designed to produce a rich and varied profile of the range of students' competencies in mathematics and science.[3]

At the international scoring training meetings, NRCs received training packets containing example responses and practice papers to help them achieve accuracy and consistency in scoring. For scoring guides that were difficult, example responses were selected to illustrate the scoring categories. The scores on these responses were explained and attached to the scoring guides. Practice sets were created for the more difficult guides. These papers illustrated a range of responses, beginning with several clear-cut examples. About 10 to 15 responses were enough for most guides, but sometimes more practice was necessary.

7.4.3 Documenting the Reliability of the Free-Response Scoring

In order to demonstrate the quality of the TIMSS 1999 data, it was important to document the agreement between scorers. To establish the scoring reliability, NRCs were required to have a 25% random sample of each booklet type independently scored by two scorers. The degree of agreement between the two scores assigned was a measure of the reliability of the scoring process. The two scorers did not know the scores assigned by the others.

○○○

[3] The TIMSS scoring scheme for free-response items is described in Chapter 3

Since the purpose of the double scoring was to document the consistency of scoring, the procedure used for scoring the booklets in the reliability sample had to be as close as possible to that used for scoring the booklets in general. The recommended procedure was designed to blend the scoring of the sample in with the normal scoring activity, to take place throughout the scoring process, and to be systematically implemented across student responses and scorers.

7.4.4 Scoring the Free-Response Items

TIMSS 1999 recommended that scorers be organized into teams of about six, headed by a team leader. The leader's primary responsibility was to monitor scoring reliability by continually checking and rechecking the scores that scorers had assigned. This process, known as back-reading, was essential for identifying scorers who did not understand particular guides or categories. Early detection of any misunderstandings permitted clarification and rectification of mistakes before too many responses had been scored. The back-reading systematically covered the daily work of each scorer. If a particular scorer appeared to have difficulty, however, then the percentage of back-reading for that scorer was increased. Any errors discovered were brought to the attention of the scorer responsible and corrected immediately. If a scorer was found to have been consistently making an error, then all of the booklets scored by that person were checked and any errors corrected.

In scoring the booklets for the main data set, scorers entered their scores directly into the student booklets. Therefore, in order that the reliability scoring be done "blind" (i.e., so that the two scorers did not know each other's scores), the reliability scoring had to be done before the scoring for the main data, and the reliability scores had to be recorded on a separate scoring sheet, and not in the booklets.

To implement the scoring plan effectively it was necessary that the scorers be divided between two equivalent teams (Team A and Team B), and that booklets be divided into two equivalent sets (Set A and Set B). The scorers in Team A scored 25% of the booklets in Set B and all the booklets in Set A, while the scorers in Team B scored 25% of the booklets in Set A and all of the booklets in Set B. Each team, therefore, handled both sets of booklets. For the set it handled first, the team scored every fourth booklet and recorded the results on a separate answer sheet (this was the reliability sample). In the other set, the team scored all booklets and wrote the scores directly into the booklets.

Periodically during the day, the Team B scorers scored the reliability sample (every fourth booklet) in the Set A batches, while the Team A scorers scored the reliability sample in the Set B batches. It was important that every fourth booklet be scored, and not just the top quarter in the set. When the reliability scoring was finished, Team B scorers marked it as completed and forwarded the batch to the Team A scorers. Similarly, the Team A scorers forwarded their scored reliability booklets from Set B to the Team B scorers. Once the booklets from Set A had been distributed to Team A scorers and the Set B booklets to the Team B scorers, all the free-response items were scored, and the scores were entered directly into the booklets.

7.5 Data Entry

The DPC provided an integrated computer program for data entry and data verification known as the DataEntryManager (DEM). This program worked on all IBM-compatible personal computers running under DOS, OS/2 or Windows 3.x, 95 or 98. It facilitated data entry directly from the tracking forms and test instruments and provided a convenient checking and editing mechanism. DEM also offered data and file management capabilities, interactive error detection, reporting, and quality control procedures. Detailed information and operational instructions were provided in the DataEntryManager manual. Since DEM incorporated the international codebooks describing all variables, use of the software ensured that the data files were produced according to the TIMSS 1999 rules and standards for data entry.

Although use of DEM for all data entry tasks was strongly recommended, NRCs were permitted to use their own procedures and computer programs, as long as all data files conformed to the specifications of the international codebooks. NRCs who chose not to use DEM were responsible for ensuring that all data files were delivered to the DPC in the international format.

Even if NRCs did not use the DEM program for data entry, they still had to apply the data verification options of this program to verify their data before sending them to the DPC. The DEM data-checking facility could: (i) identify a range of problems in the identification variables, and invalid codes; and (ii) identify problems in the structure of the data files, which could then be fixed before submission to the NRC.

Data files were regarded as having been satisfactorily checked only if the reports generated by the DEM program indicated no errors.

During the TIMSS 1999 main survey operations, data were gathered from several sources, including students, teachers, and principals, as well as from a range of tracking forms. Before beginning data entry, the NRC needed to ensure that the corresponding tracking forms and instruments had been completed and sorted correctly. The data were entered into one of six data files, as follows:

- The **school background file** contained information from the school background questionnaire
- The **mathematics teacher background file** contained information from the mathematics teacher questionnaire
- The **science teacher background file** contained information from the science teacher questionnaire
- The **student background file** contained data from the student background questionnaire
- The **student achievement file** contained the achievement test booklet data
- The **free-response scoring reliability file** contained the reliability data from the scoring of the free-response items

When all data files had passed the DEM quality control checks, they were dispatched to the IEA Data Processing Center in Hamburg for further checking and processing.

7.5.1 Survey Activities Report

NRCs were requested to maintain a record of their experiences during the TIMSS 1999 data collection and to send a report to the International Study Center when data-collection activities were completed. This should describe any problems or unusual occurrences in selecting the sample or securing school participation, translating or preparing the data-collection instruments, administering the tests and questionnaires in the schools, scoring the free-response items, or creating and checking the data files.

7.6 Summary

This chapter has summarized the design and implementation of the TIMSS 1999 field operations from the point of first contact with the sampled schools to the return of the cleaned data files to the IEA Data Processing Center. Although the procedures were sometimes complex, each step was clearly documented in the TIMSS operations manuals and supported by training sessions at the NRC meetings. NRC reports indicated that the field operations went well, and that the TIMSS data were of high quality.

References

TIMSS (1997a). *Survey Operations Manual* (TIMSS 1999 Doc. Ref. No. 98-0026). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

TIMSS (1997b). *TIMSS-R School Sampling Manual* (TIMSS 1999 Doc. Ref. 97-0012). Prepared by Pierre Foy, Statistics Canada. Chestnut Hill, MA: Boston College.

TIMSS (1997c). *School Coordinators Manual* (TIMSS 1999 Doc. Ref. 98-0024). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

TIMSS (1997d). *Test Administrator Manual* (TIMSS 1999 Doc. Ref. 98-0025). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.

TIMSS (1998a). *Scoring Guides for Mathematics and Science Free-Response Items* (TIMSS 1999 Doc. Ref. 98-0049). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

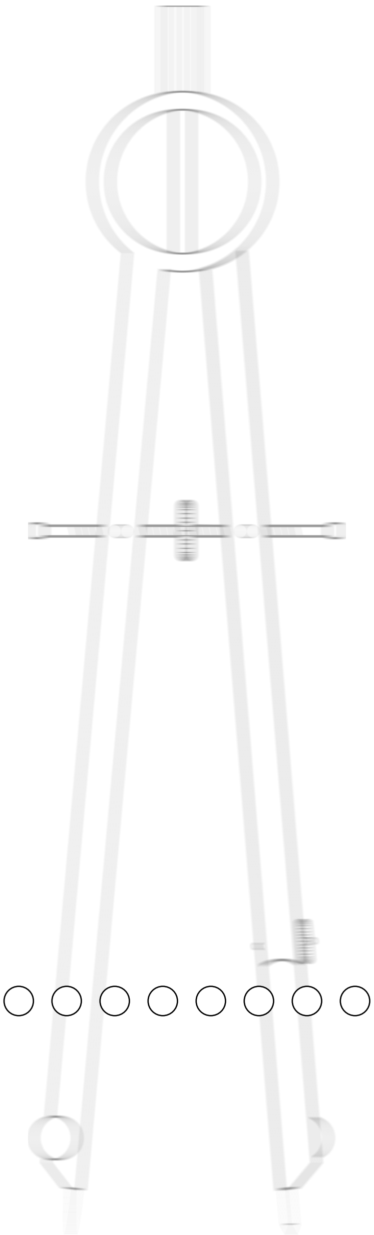
TIMSS (1998b). *Manual for Entering the TIMSS-R Data* (TIMSS 1999 Doc. Ref. 98-0028). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

TIMSS (1998c). *Manual for National Quality Control Observers* (TIMSS 1999 Doc. Ref. 98-0044). Prepared joined by the International Study Center at Boston College and the IEA's Data Processing Center. Chestnut Hill, MA: Boston College.

8

Quality Control in the TIMSS Data Collection

Kathleen M. O'Connor
Steven E. Stemler



8 Quality Control in the TIMSS Data Collection

Kathleen M. O'Connor
Steven E. Stemler

8.1 Overview

To verify that standardized procedures were followed across all participating countries, the International Study Center (ISC) instituted a program for quality assurance in data collection. In collaboration with the IEA Secretariat, one or more international Quality Control Monitors (QCMs) were recruited in each country to document data collection procedures at both the national and the school level.

Quality Control Monitors had two major responsibilities: to visit a sample of 15 schools in their country so as to observe the test administration, and to interview the National Research Coordinator (NRC) about the survey operations and activities. QCMs completed a *Classroom Observation Record* for each testing session that they observed, and an *Interview with the NRC Form* to record those interviews.

Monitors were trained in survey operations procedures and documenting findings at two-day training sessions conducted by the staff of the International Study Center. Three training sessions were held, with more than 20 monitors attending each session. During the training, QCMs were given an overview of the TIMSS 1999 survey operations procedures and instructed how to conduct their quality-control task. They received a *Manual for International Quality Control Monitors* (TIMSS, 1998), which explained their duties in detail, as well as copies of the TIMSS survey operations manual and manuals for school coordinators and test administrators.

In total, 71 Quality Control Monitors were recruited and trained. These monitors observed a total of 550 testing sessions and conducted interviews with the national research coordinator in each of the 38 participating countries.

8.2 Observing the TIMSS Test Administration

The classroom observation record was designed to allow the Quality Control Monitor to keep a simple and accurate record of the major activities relating to the test administration. The record had four sections:

1. Preliminary activities of the test administrator
2. Test session activities
3. General impressions
4. Interview with the school coordinator

8.2.1 Preliminary Activities of the Test Administrator

Section A of the classroom observation record dealt with preparations for the testing session. Monitors were asked to note whether the test administrator had checked the testing materials, read the administration script, organized space for the session, and arranged for the necessary equipment (pencils, timers, etc.).

Exhibit 8.1 summarizes the results for this section. It shows that in almost all cases, the preparatory testing procedures were followed. In the rare instances where deviations occurred, reasonable explanations were given. For example, of the 21 cases where QCMs reported that seals on student booklets were not intact, 15 were due to the use of plastic instead of a seal, a minor deviation that did not breach the security of the exam. In the few cases where it was reported that there was not enough room for students, QCMs noted that this was due to unavoidable circumstances (e.g., the test was administered in a small classroom in a very old school, the desks were too narrow, the room was configured such that students sat at round tables, etc.).

The absence of a visible wall clock was also considered more of an environmental limitation than a limitation of the implementation of the testing procedures. While more than half of the classrooms were missing a wall clock, QCMs frequently reported that students had their own watches or that the TA wrote the time remaining on the test on the board so that all students were aware of it. In general, QCMs observed no procedural deviations in preparations for the testing that were severe enough to compromise the integrity of the test administration.

Exhibit 8.1 Preliminary Activities of the Test Administrator

Question	Yes	No	N/A
Had the test administrator verified adequate supplies of the test booklets?	530*	16**	4
Had the test administrator familiarized himself or herself with the script prior to testing?	522*	24**	4
Were all the seals intact on the test booklets prior to distribution?	338	21	191+
Did the Student Identification information on test booklet correspond with the Student Tracking Form?	525	18	7
Was there adequate seating space for the students to work without distractions?	530	19	1
Was there adequate room for the test administrator to move about the room during testing?	540	0	1
Did the test administrator have a stopwatch or timer for accurately timing testing sessions?	500	43	7
Did the test administrator have an adequate supply of pencils and other materials?	474	70	4
Was there a wall clock visible for the students to check their timing during the testing?	219	329	2

+ Seals were not used on the booklets in these countries

* Represents the number of respondents answering either Definitely Yes or Probably Yes

** Represents the number of respondents answering either Definitely No or Probably No

8.2.2 Test Session Activities

Section B of the classroom observation record dealt with the test session activities themselves. These activities included the extent to which the test administrator followed the script, how the test booklets were distributed and collected, and the various announcements made during the testing session.

The achievement test was administered in two sessions, with a short break between. Exhibit 8.2 documents the activities associated with the first testing session and shows that at least 80% of the test administrators followed their script exactly when preparing the students, distributing the test materials, and beginning testing. In the rare instances where changes were made, they tended to be additions to the script.

Further examination of Exhibit 8.2 shows that only in about 60% of the sessions did the test administrator collect booklets one at a time at the end of the session, as prescribed in the directions. While this may seem surprising, it turns out that when the booklets were not collected singly from each student, students were instructed to close their test booklets and leave them on their desk during the break. The room was then either secured or supervised during the break. In some instances, the administrators even gave the students adhesive tape and instructed them to apply the tape to the test before leaving the room.

When asked whether the break between sessions was 20 minutes long, QCMs tended to interpret the question quite literally. As a result only 41% of classrooms reported starting the test after a break that was "exactly" 20 minutes. The rest reported having breaks that lasted within a few minutes in either direction (break time ranged from 10 to 20 minutes).

Exhibit 8.2 Testing Session 1

Question	Yes	No	N/A
Did the test administrator follow the test administrator's script exactly in each of the following sections?			
Preparing the students	444	92 (minor) 11 (major)	3
Distributing the materials	492	46 (minor) 7 (major)	5
Beginning testing	485	56 (minor) 4 (major)	5
If the test administrator made changes to the script, how would you describe them?			
Additions	81	140	329
Revisions	59	148	343
Deletions	26	173	351
Did the test administrator distribute test booklets one at a time to students?	527	18	5
Did the test administrator distribute the test booklets according to the booklet assignment on the *Student Tracking Form*?	537	8	5
Did the test administrator record attendance correctly on the *Student Tracking Form*?	534	8	8
Did the total testing time for session 1 equal the time allowed?	510	37	3
Did the test administrator announce "you have 10 minutes left" prior to the end of session 1?	507	41	2
Were any other "time remaining" announcements made during session 1?	48	458	8
At the end of session 1, did the test administrator collect the test booklets one at a time from students?	337	207	6
Was the total time for the break between session 1 and session 2 equal to 20 minutes?	229	309	12

Exhibit 8.3 summarizes QCMs' observations from the second testing session. Here too, QCMs took the question about time for restarting literally. In about half of the sessions, the time spent to restart the testing session was 5 minutes. For the rest, the session took up to 10 minutes longer to restart. More important, in the large majority of sessions, the test administrator kept to the time limits prescribed in the directions. Exhibit 8.3 also reveals that about 65% of the test administrators stuck to the testing script for signaling a break. Of those who did make changes, most made additions or other minor changes such as paraphrasing the directions.

A final statistic from Exhibit 8.3 worth noting is that almost two-thirds of students requested additional time to complete the student questionnaire. In most cases, this request was granted.

Exhibit 8.3 Testing Session 2

Question	Yes	No	N/A
Was the time spent to restart the testing in session 2 equal to 5 minutes?	270	270	10
Did the total testing time for session 2 equal the time allowed?	506	42	2
Did the test administrator announce "you have 10 minutes left" prior to the end of session 2?	502	45	3
Were any other "time remaining" announcements made during session 2?	62	481	
At the end of session 2, did the test administrator collect the test booklets one at a time from the students?	495	51	4
When the test administrator read the script for the end of testing session 2, did he or she announce a break to be followed by the *Student Questionnaire*?	421	89	40
How accurately did the test administrator follow the script to end the testing and signal a break?	359 (no changes)	112 (minor) 12 (major)	67
If there were any changes, how would you describe them?			
Additions	33	139	378
Some minor changes	94	100	356
Omissions	33	138	379
At the end of the break, did the test administrator distribute the student questionnaires and give directions as specified in the script?	449	38	63
Did the students ask for additional time to complete the questionnaire?	338	156	56
At the end of the session, prior to dismissing the students, did the test administrator thank the students for participating in the study?	454	42	54

Exhibit 8.4 presents the results of the remaining questions asked about the test session activities. These questions dealt with topics such as student compliance with instructions, and the alignment between scripted instructions and their implementation.

The results show that in almost all of the sessions, the students complied well or very well with the instructions to stop testing. In addition, in about 70% of the sessions, breaks were conducted exactly or nearly the same as directed in the script. Where this was not the case, it was mostly due to differences in the amount of time allocated for the break.

Exhibit 8.4 Test Session Activities

Question	Very well	Well	Fairly well	Not well	N/A
When the test administrator ended session 1, how well did the students comply with the instructions to "stop work"?	446	90	9	0	5
When the test administrator ended session 2, how well did the students comply with the instructions to "stop work"?	463	74	9	1	3
	Exactly	**Nearly the same**	**Somewhat differently**	**Not well**	**N/A**
Was the first break conducted as directed by the script?	382	95	53	1	19
Was the second break conducted as directed by the script?	347	44	41	21	97
	Exactly	**Longer**	**Shorter**	**N/A**	
How did the actual break time compare with the recommended time in the script?	258	75	102	115	
How does the total time allocated for the administration of the *Student Questionnaire* compare with the time specified in the script?	152	307	13	78	
	Very orderly	**Somewhat orderly**	**Not orderly at all**	**N/A**	
How orderly was the dismissal of students?	361	132	12	45	

8.2.3 General Impressions

Section C of the QCM survey dealt with the Quality Control Monitor's general observations and overall impressions of the test administration. It covered topics such as how well the test administrator monitored the behavior of the students during the testing, and any unusual circumstances that may have come up during the session (e.g., cheating, emergency situations, student refusal to participate, defective instrumentation).

Examination of the results presented in Exhibit 8.5 shows that in almost all sessions, the testing took place without any problems. In the few sessions where problems arose due to defective instrumentation, the instruments were replaced appropriately by the test administrator about half of the time.

It is worth noting that in roughly 10% of sessions, QCMs reported seeing evidence of students attempting to cheat on the test. However, when asked to expand on this, QCMs generally indicated that students were merely looking around at their neighbors to see whether their test booklets were indeed different. Because the TIMSS test design involves 8 different booklets distributed among the students, students usually did not have the same booklet as their neighbors, so any students who may have tried to copy a neighbor's answers would have been frustrated by the test design. The QCMs reported that on the rare occasions when they observed serious efforts to cheat, the test administrator intervened in the situation and prevented cheating.

Exhibit 8.5 Summary Observations of the QCM

Question	Yes	No	N/A
During the testing situation did the test administrator walk around the room to be sure students were working on the correct section of the test and/or behaving properly?	530	13	7
In your opinion, did the test administrator address students' questions appropriately?	532	7	11
Did you see any evidence of students attempting to cheat on the tests (e.g., by copying from a neighbor)?	45	503	2
Were any defective booklets detected and replaced before the testing began?	8	540	2
Were any defective booklets detected and replaced after the testing began?	9	521	20
If any defective test booklets were replaced, did the test administrator replace them appropriately?	14	13	523
Did any students refuse to take the test either prior to the testing or during the testing?	15	526	9
If a student refused, did the test administrator accurately follow the instructions for excusing the student (collect the test booklet and record the incident on the *Student Tracking Form*)?	16	4	527
Did any students leave the room for an "emergency" during the testing?	42	500	8
If yes, did the test administrator address the situation appropriately (collect the booklet, and if the student was readmitted, return the test booklet and record time out of the room on the test booklet)?	19	20	508

Finally, Exhibit 8.6 indicates that in almost all of the testing sessions, QCMs found the behavior of students to be orderly and cooperative. Where it was less than perfect, the test administrator was almost always able to control the students and the situation. For the great majority of sessions, QCMs reported that the overall quality of the sessions was either excellent or very good.

Exhibit 8.6 Summary Observations of Student Behavior

Question	Extremely	Moderately	Somewhat	Hardly at all	N/A
To what extent would you describe the students as orderly and cooperative?	395	136	14	4	1
	No, no late students	**No, not admitted**	**Yes, before testing began**	**Yes, after testing began**	**N/A**
Were any late students admitted to the testing room?	485	11	23	23	8
	Excellent	**Very good**	**Good**	**Fair**	**Poor**
In general, how would you describe the overall quality of the testing session?	280	184	66	14	3
	Definitely Yes	**Some effort was made**	**Hardly any effort was made**	**N/A**	
If the students were not cooperative and orderly, did the test administrator make an effort to control the students and the situation?	151	19	1	379	

8.2.4 Interview with the School Coordinator

In Section D of the classroom observation record the QCM recorded details of the interview with the school coordinator. Issues addressed included shipping of assessment materials, satisfaction with arrangements for the test administration, the responsiveness of the NRC to queries, necessity for make-up sessions, and, as a check on within-school sampling activities, the organization of classes in the school.

The results presented in Exhibit 8.7 show that TIMSS 1999 was an administrative success in the eyes of the school coordinators. In 80% or more of the cases, school officials received the correct shipment of the test materials. Mistakes that did occur tended to be minor and could be remedied prior to testing. Furthermore, more than 80% of school coordinators reported that the NRCs were responsive to their questions or concerns, and that relations were cordial and cooperative.

About half of the school coordinators reported that they were able to collect the completed teacher questionnaires prior to student testing. Of the rest, the vast majority reported that they were missing only one or two questionnaires and were expecting them to be turned in shortly.

It was estimated that the teacher questionnaires would take about 60 minutes to complete. About 55% of the school coordinators indicated that the estimate of 60 minutes was about right, while about 30% reported that the questionnaires took longer and about 15% that they took less time to complete.

Finally, it is worth noting that in about 53% of the cases, school coordinators indicated that students were given special instructions, motivational talks, or incentives prior to testing. A more in-depth analysis of the results revealed that of the 292 cases involved, 208 reported that students received motivational talks by a school official, 64 reported that students received special instructions about the test beyond what was written in the TIMSS 1999 manuals (sometimes in the form of test preparation based on the released items from TIMSS 1995), and just 13 indicated that students received incentives for participation such as extra credit in their classes.

Exhibit 8.7 Interview with the School Coordinator

Question	Yes	No	N/A
Prior to the test day did you have time to check your shipment of materials from your TIMSS national coordinator?	453	66	31
Did you receive the correct shipment of the following items?			
Test booklets	473	50	27
Test administrator manual	454	50	46
School coordinator manual	433	59	58
Student tracking forms	466	37	47
Student questionnaires	472	51	27
Teacher questionnaires	507	15	28
School questionnaires	506	17	27
Test administration form	465	50	35
Teacher tracking form	363	117	70
Student-teacher linkage form (if applicable)	163	197	190
Envelopes or boxes addressed to the national center for the purpose of returning the materials after the assessment	371	126	52
Was the national coordinator responsive to your questions or concerns?	452	33	64
Were you able to collect completed *Teacher Questionnaires* prior to the test administration?	267	252	30
It was expected that the *Teacher Questionnaire* would require about 60 minutes to complete. In your opinion, was that estimate correct?	259	141 (longer) 68 (less time)	82
Were you satisfied with the accommodations (testing room) you were able to arrange for the testing?	522	20	8
Do you anticipate that makeup sessions will be required at your school?	36	485	29
If yes, do you intend to conduct one?	55	91	404
Did the students receive any special instructions, motivational talk, or incentives to prepare them for the assessment?	292	242	16
Were students given any opportunity to practice on questions like those in the tests before the testing session?	45	494	11
Is this a complete list of the mathematics classes in this grade in this school?	453	45	52
To the best of your knowledge, are there any students in this grade level who are *not* in any of these mathematics classes?	27	469	54
To the best of your knowledge, are there any students in this grade level in more than one of these mathematics classes?	21	477	52
If there were another international assessment, would you be willing to serve as a school coordinator?	492	42	16

Perhaps the biggest tribute to the successful planning and implementation of TIMSS 1999 was the fact that about 90% of respondents said that if there were to be another international assessment that they would be willing to serve as the school coordinator. Furthermore, the results shown in Exhibit 8.8 suggest that practically all of the school coordinators thought the testing sessions went well, and that many thought that staff members in their school felt positive about the TIMSS 1999 testing.

Exhibit 8.8 Interview with the School Coordinator (continued)

Question	Very well	Satisfactory	Unsatisfactory	N/A
Overall, how would you say the session went?	449	86	3	12
	Positive	**Neutral**	**Negative**	**N/A**
Overall, how would you rate the attitude of the other school staff members towards the TIMSS testing?	375	150	17	8
	Worked well	**Needs improvement**	**N/A**	
Overall, do you feel the TIMSS 1999 school coordinator manual worked well or does it need improvement?	436	31	83	

8.3 Interview with the National Research Coordinator

In addition to visiting a sample of schools to observe the testing sessions, Quality Control Monitors held face-to-face interviews with the national research coordinator for their country. In countries with more than one QCM, interviews with the NRC were either conducted by the QCMs jointly, or one QCM was chosen. Interviews with NRCs were conducted in 36 of the 38 countries that participated in TIMSS 1999. The results are summarized in the following section.

The interview probed NRC's experiences in preparing for and conducting the TIMSS data collection, focusing on sampling activities, working with school coordinators, translating the instruments, assembling and printing the test materials, packing and shipping the test materials, scoring free-response questions, data entry and verification, choosing quality assurance samples, and suggestions for improvement in the process.

8.3.1 Sampling

Section A of the NRC interview related to sampling. Topics covered included the extent to which the NRCs used the manuals and sampling software provided by the International Study Center as well as the complexity of the task.

Exhibit 8.9 shows that only four countries did not use the sampling manuals provided, mainly because they dealt directly with Statistics Canada. In one case, no sampling was necessary because the sample was the population. Just over half of the NRCs used the within-school sampling software provided by the IEA DPC to select classes. In the cases where the sampling software was not used, it was generally because the country started the process late and did not have time to learn the software, and occasionally because of software incompatibility.

A number of NRCs encountered organizational constraints in their systems that necessitated deviations from the basic sample design. In each case, a sampling expert was consulted to ensure that the altered design remained compatible with the TIMSS standards. Most NRCs reported that the sampling procedures were not unduly difficult to implement, although some found the process very difficult. Nevertheless, all NRCs managed to deliver school and student samples of high quality for the data collection.

Exhibit 8.9 Interview with the NRC – Sampling

Question	Yes	No	N/A
Did you use the manuals provided by the International Study Center to select a sample of schools and students within schools?	32	4	0
Did you use the within-school sampling software provided by the International Study Center to select classes or students?	19	17	0
Were there any conditions or organizational constraints that necessitated deviations from the basic TIMSS 1999 sampling design?	6	29	1
	Very difficult	**Somewhat difficult**	**Not difficult**
In terms of the complexity of the procedures and number of personnel needed, how would you describe the process of sample selection?	4	10	21

8.3.2 Working with the School Coordinators

Section B of the NRC interview asked about working with the school coordinators, specifically about contacting them, shipping materials, and training.

Exhibit 8.10 Interview with the NRC – School Coordinator

Question	Yes	No	N/A
Have all the school coordinators for your sample been contacted?	33	3	0
If yes, have you sent them materials about the testing procedures?	25	7	4
Did you have formal training sessions for the school coordinators?	19	14	3

At the time the interviews were conducted, almost all of the NRCs had already selected the school coordinators for their sample, and most of them had been sent the appropriate materials on the testing procedures. Where this was not the case, it was often because the schools were on break, or because a meeting had been set up but not yet held. About half the NRCs conducted formal training sessions for school coordinators prior to the test administration.

8.3.3 Translating the Instruments

Section C of the NRC interview dealt with the difficulty of translating and adapting the assessment instruments and manuals.

Exhibit 8.11 shows that most NRCs reported little difficulty in translating and adapting the test booklets, questionnaires, or manuals, but that the scoring guides for the free-response items proved more problematic.

In translating the test booklets, NRCs generally reported using their own staff or a combination of staff and outside experts. All NRCs reported that they had submitted the achievement test booklets to the translation verification program at the International Study Center. At the time of the interview, almost all of them had received a translation verification report back.

Exhibit 8.11 Interview with the NRC – Translation

Question	Very difficult	Somewhat difficult	Not difficult
How difficult was it to translate and/or adapt the test booklets?	3	18	14
How difficult was it to adapt the questionnaires?	2	20	13
How difficult was it to adapt the *Test Administrator Manual*?	0	10	26
How difficult was it to adapt the *School Coordinator Manual*?	1	8	25
Did you translate or do you plan to translate the *Scoring Guide for Mathematics and Science Free Response Items*?	18	18	0

8.3.4 Assembling and Printing the Test Materials

Section D of the NRC survey dealt with assembling and printing the test materials. It included quality control issues related to the accuracy of the materials and the security of their storage.

The results from Exhibit 8.12 indicate that NRCs were able to assemble the test booklets according to the instructions provided, and that almost all conducted the recommended quality control checks during the process. In the rare instances where the NRCs did not check the test booklets during the printing process, it was because of a shortage of time.

All NRCs reported having followed procedures to protect the security of the tests during assembly and printing. The one case where an NRC reported a breach of security was in fact simply a potential breach, that never materialized. Most countries elected to send their test booklets and questionnaires to an external printer, but more often printed their manuals in-house.

Exhibit 8.12 Interview with the NRC – Test Materials

Question	Yes	No	N/A
Were you able to assemble the test booklets according to the instructions provided by the International Study Center?	36	0	0
Did you conduct the quality assurance procedures for checking the test booklets during the printing process?	33	3	0
Were any errors detected during the printing process?	14	19	3
Poor print quality	6	7	23
Pages missing	3	11	22
Page order	5	9	22
Upside down pages	5	9	22
Did you follow procedures to protect the security of the tests during the assembly and printing process?	36	0	0
Did you discover any breaches of security?	1	35	0
	In-House	**External**	**Combination**
Where did you print the test booklets?	8	24	3
Where did you print the questionnaires?	8	20	7
Where did you print the manuals?	21	12	2

8.3.5 Packing and Shipping the Testing Materials

Section E of the NRC interview dealt with the extent to which NRCs discovered errors in the testing materials as they were packed for shipping to school coordinators. Exhibit 8.13 shows that overall, very few errors were found in any of the materials. The rare errors detected before distribution were remedied.

In addition, about two-thirds of NRCs reported having established procedures requiring schools to confirm receipt of the testing materials and verification of the contents. In most countries, NRCs reported that the deadline for return of materials from the schools was within a day or two of testing. All NRCs reported that the deadline was within two weeks of testing.

Exhibit 8.13 Interview with the NRC – Test Materials

Question	No Errors	Errors found before distribution	Errors found after distribution
In packing the assessment materials for shipment to schools, did you detect any errors in any of the following items?			
Supply of test booklets	27	1	8
Supply of student questionnaires	27	1	8
Student tracking forms	34	1	1
Teacher tracking forms	35	1	0
Student-teacher linkage forms	28	0	0
Test administrator manual	35	1	0
School coordinator manual	35	1	0
Supply of teacher questionnaires	36	0	0
School questionnaire	36	0	0
Test booklet ID labels	34	2	0
Sequencing of booklets or questionnaires	35	1	0
Return labels	36	0	0
Self-addressed post-cards for test dates	36	0	0

8.3.6 Scoring Free-Response Questions

The TIMSS 1999 assessment contained a significant proportion of free-response items that needed to be scored by specially trained individuals. The scoring process was a considerable undertaking in each country, requiring the recruitment and training of scoring staff and an ambitious scoring effort that included double scoring 25% of the student responses as a check on reliability.

Exhibit 8.14 shows that, at the time of the NRC interview, more than two-thirds of the NRCs had selected their scoring staff, and roughly half of those had already begun the training process. All NRCs reported that they understood the procedures for scoring the 25% reliability sample as explained in the survey operations manual.

Exhibit 8.14 Interview with the NRC – Scoring

Question	Yes	No	N/A
Have you selected your scorers for the free-response questions?	25	11	
If yes, have you trained the scorers?	12	13	11
Have you scheduled the scoring sessions for the free-response questions?	30	6	0
Do you understand the procedure for scoring the 25% reliability sample as explained in the survey operations manual?	36	0	0

8.3.7 Data Entry and Verification

Section G of the NRC interview dealt with preparations for data entry and verification. Again, at the time of the interviews about two-thirds of the NRCs had selected their data entry staff and about half of those had conducted training sessions.

By way of quality assurance, about 80% of the NRCs reported that they planned to enter the data from part of the test booklets twice as a verification procedure. The proportion of booklets that was being entered twice ranged from 5% to 30%, with one country reporting that it planned to reenter 100% of the data.

Exhibit 8.15 Interview with the NRC – Data Entry and Verification

Question	Yes	No	N/A
Have you selected the data entry staff?	22	13	1
If yes, have you conducted training sessions for the data entry staff?	12	15	9
Do you plan to key enter a percentage of test booklets twice as a verification procedure?	28	7	1
Have you established a secure storage area for the returned tests after coding and until the original documents can be discarded?	36	0	0

8.3.8 Quality Assurance Sample

As part of their national quality assurance activities, NRCs were required to conduct site visits to a 10% sample of the TIMSS schools to observe the test administration and document compliance with prescribed procedures. These site visits were additional to the visits to 15 schools conducted by the international Quality Control Monitors, and summarized in the first part of this chapter.

At the time of the NRC interviews, 70% of the NRCs had selected their 10% quality assurance sample for on-site classroom observations. Three NRCs reported that an external agency would conduct the observations, 16 NRCs reported that a member of their staff would do so, and 9 NRCs reported that a combination of staff and external agency people would conduct the observations.

8.3.9 The Survey Activities Report

The final section of the NRC interview asked for suggestions for improving the assessment process. The major problem that most NRCs faced was shortage of time to accomplish all that had to be done in order to keep to the very demanding TIMSS schedule. Most NRCs who commented expressed a desire for more time, particularly for translation and instrument preparation.

8.4 Summary

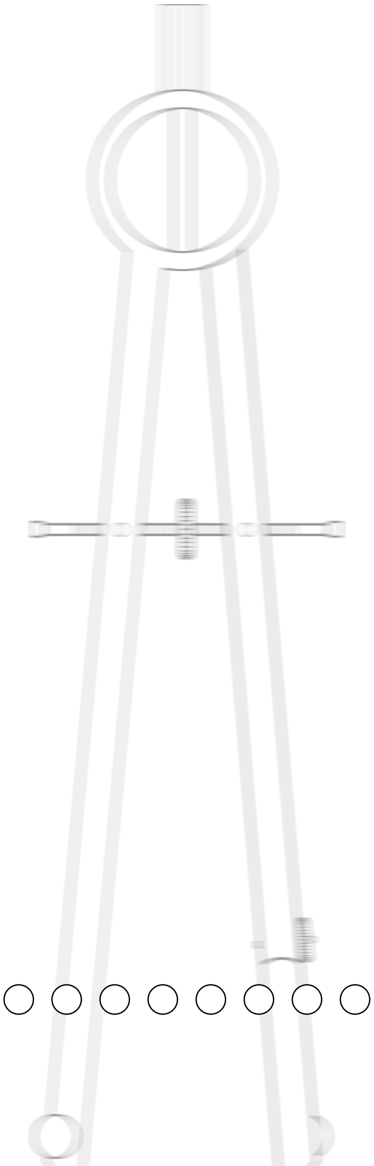
In summary, the observations by the Quality Control Monitors and the interviews with the National Research Coordinators indicate that the data collected in the TIMSS 1999 study met high quality standards, and that as a result there can be a high level of confidence in the findings.

References

TIMSS (1998). *Manual for International Quality Control Monitors* (Doc. Ref: 98-0023). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

Implementation of the Sample Design

Pierre Foy
Marc Joncas



9 Implementation of the Sample Design

Pierre Foy
Marc Joncas

9.1 Overview

The selection of valid and efficient samples is crucial to the quality and success of an international comparative study. The accuracy of the survey results depends on the quality of the sampling information available when planning the sample, and on the care with which the sampling activities themselves are conducted. For TIMSS 1999, National Research Coordinators (NRCs) worked on all phases of sampling with staff from Statistics Canada. NRCs were trained in how to select the school and student samples and how to use the sampling software. In consultation with the TIMSS 1999 sampling referee (Keith Rust, Westat), staff from Statistics Canada reviewed the national sampling plans, sampling data, sampling frames, and sample selection. This documentation was used by the International Study Center (ISC) jointly with Statistics Canada, the sampling referee, and the Project Management Team (PMT) to evaluate the quality of the samples. Summaries of the sample design for each country, including details of population coverage and exclusions, stratification variables, and participation rates, are provided in Appendix C.

9.2 TIMSS 1999 Target Population

In IEA studies, the target population for all countries is known as the *international desired population*. The international desired population for TIMSS 1999 was the following[1]:

- All students enrolled in the upper of the two adjacent grades that contain the largest proportion of 13-year-olds at the time of testing.

The TIMSS 1999 target grade was intended to be the upper grade of the TIMSS 1995 population 2 definition and was expected to be the eighth grade in most countries. This would allow countries participating in both TIMSS 1995 and TIMSS 1999 to establish a trend line of comparable achievement data.

[1] See Chapter 2 for more information the TIMSS 1999 sample design.

Exhibit 9.1 summarizes the grades identified as the target grade in all participating countries. For most countries, the target grade did indeed turn out to be the eighth grade.[2] Only in Finland, Morocco, and some states in the Russian Federation was the seventh grade the target grade. In parts of Australia and New Zealand, the target grade was the ninth grade. Average student ages ranged from 13.8 in Finland and New Zealand to 15.5 in South Africa.

Exhibit 9.1 National Grade Definitions

Country	Country's Name for Grade Tested	Years of Formal Schooling	Mean Age of Students Tested
Australia	8 or 9	8 or 9	14.3
Belgium (Flemish)	2A & 2P	8	14.1
Bulgaria	8	8	14.8
Canada	8	8	14.0
Chile	8	8	14.4
Chinese Taipei	2nd Grade Junior High School	8	14.2
Cyprus	8	8	13.8
Czech Republic	8	9	14.4
England	Year 9	9	14.2
Finland	7	7	13.8
Hong Kong, SAR	Secondary 2	8	14.2
Hungary	8	8	14.4
Indonesia	2nd Grade Junior High School	8	14.6
Iran, Islamic Rep.	9	8	14.6
Israel	9	8	14.1
Italy	3rd Grade Middle School	8	14.0
Japan	2nd Grade Lower Secondary	8	14.4
Jordan	8	8	14.0
Korea, Rep. of	2nd Grade Middle School	8	14.4
Latvia (LSS)	8	8	14.5
Lithuania	9	8.5	15.2
Macedonia, Rep. of	8	8	14.6
Malaysia	Form 2	8	14.4
Moldova	8	9	14.4
Morocco	7	7	14.2
Netherlands	Secondary 2	8	14.2
New Zealand	Year 9	8.5 to 9.5	14.0
Philippines	1st Year High School	7	14.1
Romania	8	8	14.8
Russian Federation	8	7 or 8	14.1
Singapore	Secondary 2	8	14.4
Slovak Republic	8	8	14.3
Slovenia	8	8	14.8
South Africa	8	8	15.5
Thailand	Secondary 2	8	14.5
Tunisia	8	8	14.8
Turkey	8	8	14.2
United States	8	8	14.2

[2] In TIMSS in 1995, Romania and Slovenia selected the eighth grade as the upper of their target grades. Subsequently, analysis of the age distributions in those countries showed that there students were older, on average, than students in most other countries. Both countries chose to test the same grade again in 1999 in order to have comparable trend data.

9.2.1 Coverage And Exclusions

Exhibit 9.2 summarizes national coverage and exclusions in the TIMSS 1999 target populations. National coverage of the international desired target population was generally comprehensive. Only Latvia and Lithuania chose a national desired population less than the international desired population.[3] Because coverage of the international desired population fell below 65% for Latvia, the Latvian results have been labelled "Latvia (LSS)," for Latvian-Speaking Schools. Coverage was more inclusive in Lithuania, but since it was less than 100%, the Lithuanian results were footnoted to reflect this situation. The Lithuanian results were also footnoted to indicate that although Lithuania tested the same cohort of students as other countries, it did so later in 1999, at the beginning of the next school year.

School-level exclusions generally consisted of schools for the disabled and very small schools; however, there were some national deviations that are documented in Appendix C. Within-school exclusions generally consisted of disabled students and students that could not be assessed in the language of the test. Only in Israel did the level of excluded students exceed the TIMSS maximum of 10%, and this was reflected in a footnote in the international reports. A few countries had no within-school exclusions.

○○○

[3] The Latvian population was restricted to schools catering to Latvian-speaking students only, and the Lithuanian population to schools catering to Lithuanian-speaking students only.

Exhibit 9.2 National Coverage and Overall Exclusion Rates

	International Desired Population		National Desired Population		Overall
	Coverage	Notes on Coverage	School-Level Exclusions	Within-Sample Exclusions	Overall Exclusions
Australia	100%		1%	1%	2%
Belgium (Flemish)	100%		1%	0%	1%
Bulgaria	100%		5%	0%	5%
Canada	100%		4%	2%	6%
Chile	100%		3%	0%	3%
Chinese Taipei	100%		1%	1%	2%
Cyprus	100%		0%	1%	1%
Czech Republic	100%		5%	0%	5%
England	100%		2%	3%	5%
Finland	100%		3%	0%	4%
Hong Kong, SAR	100%		1%	0%	1%
Hungary	100%		4%	0%	4%
Indonesia	100%		0%	0%	0%
Iran, Islamic Rep.	100%		4%	0%	4%
Israel	100%		8%	8%	16%
Italy	100%		4%	2%	7%
Japan	100%		1%	0%	1%
Jordan	100%		2%	1%	3%
Korea, Rep. of	100%		2%	2%	4%
Latvia	61%	Latvian-speaking students only	4%	0%	4%
Lithuania	87%	Lithuanian-speaking students only	5%	0%	5%
Macedonia, Rep. of	100%		1%	0%	1%
Malaysia	100%		5%	0%	5%
Moldova	100%		2%	0%	2%
Morocco	100%		1%	0%	1%
Netherlands	100%		1%	0%	1%
New Zealand	100%		2%	1%	2%
Philippines	100%		3%	0%	3%
Romania	100%		4%	0%	4%
Russian Federation	100%		1%	1%	2%
Singapore	100%		0%	0%	0%
Slovak Republic	100%		7%	0%	7%
Slovenia	100%		3%	0%	3%
South Africa	100%		2%	0%	2%
Thailand	100%		3%	0%	3%
Tunisia	100%		0%	0%	0%
Turkey	100%		2%	0%	2%
United States	100%		0%	4%	4%

9.3 Sampling of Schools and Students

9.3.1 General Sample Design

The basic sample design used in TIMSS 1999 was a two-stage stratified cluster design.[4] The first stage consisted of a sample of schools and the second stage of samples of intact mathematics classrooms from the target grade in the sampled schools.

[4] The TIMSS sample design is described in Chapter 2.

The TIMSS 1999 design allowed countries to stratify the school sampling frame to improve the precision of survey results. Some countries used an explicit stratification procedure, whereby schools were categorized according to some criterion (e.g., regions of the country). This allowed them to ensure that a predetermined number of schools were selected from each explicit stratum. Countries also used an implicit stratification procedure, whereby the school sampling frame was sorted according to a set of stratification variables prior to sampling. This approach provided a convenient method of allocating the school sample in proportion to the size of the implicit stratum when used in conjunction with a systematic PPS method.

Most countries sampled approximately 150 schools and one intact classroom (with all of its students) within each school. Countries that selected larger school samples included large countries, such as the United States and the Russian Federation, and countries such as Australia, Canada, and Turkey that required accurate survey estimates for regions or provinces. Schools were selected with probability proportional to size, and classrooms with equal probabilities.[5] Some countries chose to sample more than one classroom per selected school. Details of the sampling of schools and students for each country are provided in Appendix C.

9.3.2 Target Population Sizes

Exhibit 9.3 summarizes the number of schools and students in each country's target population, as well as the sample sizes of schools and students that participated in the study. Most of the target population sizes are derived from the sampling frames from which the TIMSS samples were drawn. The school and student population sizes for Turkey, however, were estimated from the number of students in the primary sampling units (provinces) that Turkey sampled. In addition, the school and student population sizes for the United States and the Russian Federation were not computed from the sampling frame, but were provided by their respective NRC. Using the sampling weights computed for each country (see Chapter 11), TIMSS derived an estimate of the student population size, which matched closely the student population size from the sampling frame (see Exhibit 9.3).

○○○

[5] Because of large class sizes, Morocco chose a sub-sample of students from each sampled classroom.

Exhibit 9.3 Population and Sample Sizes

Country	Population		Sample		
	Schools	Students	Schools	Students	Est. Pop.
Australia	2072	255648	170	4032	260130
Belgium (Flemish)	697	67765	135	5259	65539
Bulgaria	2160	85066	163	3272	88389
Canada	5925	395960	385	8770	371061
Chile	4044	238894	185	5907	208910
Chinese Taipei	758	342753	150	5772	310428
Cyprus	61	9862	61	3116	9785
Czech Republic	1606	124583	142	3453	119462
England	3784	566590	128	2960	552231
Finland	649	64386	159	2920	59665
Hong Kong SAR	408	79397	137	5179	79097
Hungary	2693	114156	147	3183	111298
Indonesia	18565	2167498	150	5848	1956221
Iran Islamic Rep.	24560	1576860	170	5301	1655741
Israel	834	95031	139	4195	81486
Italy	5488	582110	180	3328	548711
Japan	10102	1449671	140	4745	1411038
Jordan	1276	100176	147	5052	89171
Korea Rep. of	2504	635080	150	6114	609483
Latvia	586	19663	145	2873	18122
Lithuania	954	41824	150	2361	40452
Macedonia Rep. of	355	30387	149	4023	30280
Malaysia	1642	378762	150	5577	397762
Moldova	1216	64241	150	3711	59956
Morocco	1094	330186	173	5402	347675
Netherlands	730	175513	126	2962	198144
New Zealand	379	51716	152	3613	51553
Philippines	5001	1233150	150	6601	1078093
Romania	6691	258833	147	3425	259621
Russian Federation	58595	2100000	189	4332	2057412
Singapore	145	41700	145	4966	41346
Slovak Republic	1392	76790	145	3497	72521
Slovenia	434	24645	149	3109	23514
South Africa	7234	968857	194	8146	844705
Thailand	7839	790788	150	5732	727087
Tunisia	533	140580	149	5051	139639
Turkey	6531	636242	204	7841	618058
United States	41499	3464627	221	9072	3336295

9.3.3 Participation Rates

Weighted school, student, and overall participation rates were computed for each participating country using the procedures documented in Chapter 11. Countries understood that the goal for sampling participation was 100% for all sampled schools and students, and that the guidelines established by TIMSS in 1995 for reporting achievement data for countries securing less than full participation also would be applied in 1999.

According to TIMSS, countries would be assigned to one of three categories on the basis of their sampling participation (Exhibit 9.4). Countries in Category 1 were considered to have met the TIMSS sampling requirements and to have an acceptable participation rate. Countries in Category 2 met the sampling requirements only after including replacement schools. Countries that failed to meet the participation requirements even with the use of replacement schools were assigned to Category 3. One of the main goals for quality data in TIMSS 1999 was to have as many countries as possible achieve Category 1 status, and to have no countries in Category 3.

Exhibit 9.4 Categories of Sampling Participation

Category 1	Acceptable sampling participation rate **without** the use of replacement schools. In order to be placed in this category, a country had to have: • An **unweighted** school response rate **without** replacement of at least 85% (after rounding to nearest whole percent) AND an **unweighted** student response rate (after rounding) of at least 85% OR • A **weighted** school response rate **without** replacement of at least 85% (after rounding to nearest whole percent) AND a **weighted** student response rate (after rounding) of at least 85% OR • The product of the (unrounded) **weighted** school response rate **without** replacement and the (unrounded) **weighted** student response rate of at least 75% (after rounding to the nearest whole percent). Countries in this category appeared in the tables and figures in international reports without annotation ordered by achievement as appropriate.
Category 2	Acceptable sampling participation rate **only when replacement schools were included**. A country was placed in category 2 if: • It failed to meet the requirements for Category 1 but had either an unweighted or weighted school response rate **without** replacement of at least 50% (after rounding to the nearest percent) AND HAD EITHER • An **unweighted** school response rate **with** replacement of at least 85% (after rounding to nearest whole percent) AND an **unweighted** student response rate (after rounding) of at least 85% OR • A **weighted** school response rate **with** replacement of at least 85% (after rounding to nearest whole percent) AND a **weighted** student response rate (after rounding) of at least 85% OR • The product of the (unrounded) **weighted** school response rate **with** replacement and the (unrounded) **weighted** student response rate of at least 75% (after rounding to the nearest whole percent). Countries in this category were annotated in the tables and figures in international reports and ordered by achievement as appropriate.
Category 3	Unacceptable sampling response rate even when replacement schools are included. Countries that could provide documentation to show that they complied with TIMSS sampling procedures and requirements but did not meet the requirements for Category 1 or Category 2 were placed in Category 3. Countries in this category would appear in a separate section of the achievement tables, below the other countries, in international reports. These countries were presented in alphabetical order.

Exhibits 9.5 through 9.8 present the school, student, and overall participation rates and achieved sample sizes for each participating country. As can be seen from these exhibits, all TIMSS 1999 countries except England met the requirements for category 1. England had an unweighted school participation rate before including replacement schools of 51%. With replacement this increased to 85%, which meant that England belonged in category 2. Accordingly the results for England were annotated with an obelisk in the achievement exhibits in the international reports. In TIMSS 1999, no country was assigned to category 3.

9.4 Summary

Population coverage and sampling participation rates were good for all countries that participated in TIMSS 1999. Unlike the situation in 1995 when a number of countries had difficulty securing acceptable participation rates or complying fully with sampling guidelines, all countries met the standards for compliance in 1999 and had acceptable participation rates (although one country had to rely on replacement schools). Full details of the outcome of the TIMSS sampling in each country is presented in Appendix C.

Exhibit 9.5 School Participation Rates & Sample Sizes

Country	School Participation Before Replacement (Weighted Percentage)	School Participation After Replacement (Weighted Percentage)	Number of Schools in Original Sample	Number of Eligible Schools in Original Sample	Number of Schools in Original Sample That Participated	Number of Replacement Schools That Participated	Total Number of Schools That Participated
Australia	83%	93%	184	182	152	18	170
Belgium (Flemish)	72%	89%	150	150	106	29	135
Bulgaria	97%	97%	172	169	163	0	163
Canada	92%	95%	410	398	376	9	385
Chile	98%	100%	186	185	181	4	185
Chinese Taipei	100%	100%	150	150	150	0	150
Cyprus	100%	100%	61	61	61	0	61
Czech Republic	94%	100%	150	142	136	6	142
England	49%	85%	150	150	76	52	128
Finland	97%	100%	160	160	155	4	159
Hong Kong, SAR	75%	76%	180	180	135	2	137
Hungary	98%	98%	150	150	147	0	147
Indonesia	84%	100%	150	150	132	18	150
Iran, Islamic Rep.	96%	100%	170	170	164	6	170
Israel	98%	100%	150	139	137	2	139
Italy	94%	100%	180	180	170	10	180
Japan	93%	93%	150	150	140	0	140
Jordan	99%	100%	150	147	146	1	147
Korea, Rep. of	100%	100%	150	150	150	0	150
Latvia	96%	98%	150	148	143	2	145
Lithuania	100%	100%	150	150	150	0	150
Macedonia, Rep. of	99%	99%	150	150	149	0	149
Malaysia	99%	100%	150	150	148	2	150
Moldova	96%	100%	150	150	145	5	150
Morocco	99%	99%	174	174	172	1	173
Netherlands	62%	85%	150	148	86	40	126
New Zealand	93%	97%	156	156	145	7	152
Philippines	98%	100%	150	150	148	2	150
Romania	98%	98%	150	150	147	0	147
Russian Federation	98%	100%	190	190	186	3	189
Singapore	100%	100%	145	145	145	0	145
Slovak Republic	95%	96%	150	150	143	2	145
Slovenia	98%	99%	150	150	147	2	149
South Africa	85%	91%	225	219	183	11	194
Thailand	93%	100%	150	150	143	7	150
Tunisia	84%	100%	150	149	126	23	149
Turkey	99%	100%	204	204	202	2	204
United States	83%	90%	250	246	202	19	221

Exhibit 9.6 Student Participation Rates & Sample Sizes

Country	Within School Student Participation (Weighted Percentage)	Number of Sampled Students in Participating Schools	Number of Students Withdrawn from Class/ School	Number of Students Excluded	Number of Students Eligible	Number of Students Absent	Number of Students Assessed
Australia	90%	4600	96	53	4451	419	4032
Belgium (Flemish)	97%	5387	12	0	5375	116	5259
Bulgaria	96%	3461	63	0	3398	126	3272
Canada	96%	9490	84	245	9161	391	8770
Chile	96%	6283	119	18	6146	239	5907
Chinese Taipei	99%	5889	30	42	5817	45	5772
Cyprus	97%	3296	38	32	3226	110	3116
Czech Republic	96%	3640	24	0	3616	163	3453
England	90%	3400	27	115	3258	298	2960
Finland	96%	3060	17	13	3030	110	2920
Hong Kong SAR	98%	5310	18	1	5291	112	5179
Hungary	95%	3350	0	0	3350	167	3183
Indonesia	97%	6162	106	1	6055	207	5848
Iran Islamic Rep.	98%	5497	104	0	5393	92	5301
Israel	94%	4670	29	187	4454	259	4195
Italy	97%	3531	23	86	3422	94	3328
Japan	95%	4996	15	12	4969	224	4745
Jordan	99%	5300	130	42	5128	76	5052
Korea Rep. of	100%	6285	29	128	6128	14	6114
Latvia	93%	3128	16	4	3108	235	2873
Lithuania	89%	2668	0	0	2668	307	2361
Macedonia Rep. of	98%	4096	0	0	4096	73	4023
Malaysia	99%	5713	98	0	5615	38	5577
Moldova	98%	3824	23	0	3801	90	3711
Morocco	92%	5841	42	0	5799	397	5402
Netherlands	95%	3099	12	0	3087	125	2962
New Zealand	94%	3966	96	22	3848	235	3613
Philippines	92%	7591	461	0	7130	529	6601
Romania	98%	3514	36	0	3478	53	3425
Russian Federation	97%	4557	48	34	4475	143	4332
Singapore	98%	5100	37	0	5063	97	4966
Slovak Republic	98%	3695	149	0	3546	49	3497
Slovenia	95%	3287	0	4	3283	174	3109
South Africa	93%	9071	256	0	8815	669	8146
Thailand	99%	5831	59	0	5772	40	5732
Tunisia	98%	5189	45	0	5144	93	5051
Turkey	99%	7972	49	0	7923	82	7841
United States	94%	9981	115	142	9724	652	9072

Exhibit 9.7 **Unweighted Participation Rates**

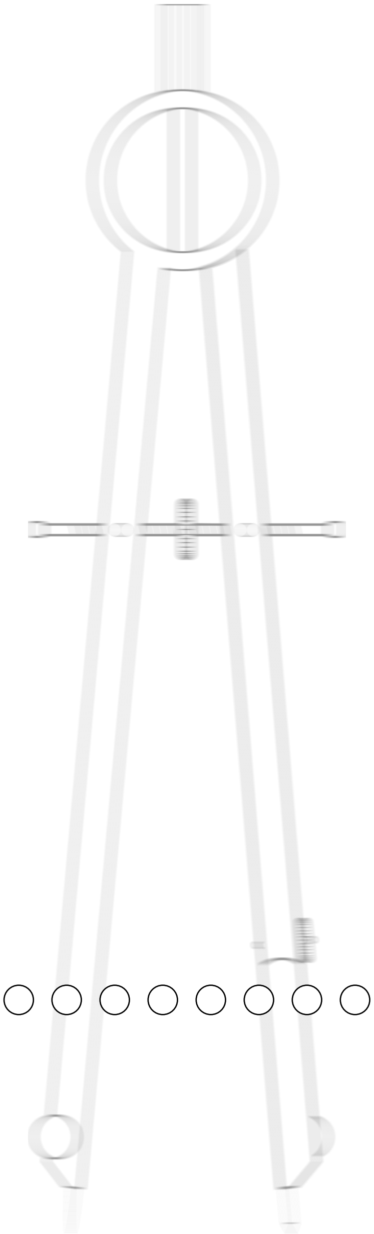
Country	School Participation Before Replacement	School Participation After Replacement	Student Participation	Overall Participation Before Replacement	Overall Participation After Replacement
Australia	84%	93%	91%	76%	85%
Belgium (Flemish)	71%	90%	98%	69%	88%
Bulgaria	96%	96%	96%	93%	93%
Canada	94%	97%	96%	90%	93%
Chile	98%	100%	96%	94%	96%
Chinese Taipei	100%	100%	99%	99%	99%
Cyprus	100%	100%	97%	97%	97%
Czech Republic	96%	100%	95%	91%	95%
England	51%	85%	91%	46%	78%
Finland	97%	99%	96%	93%	96%
Hong Kong, SAR	75%	76%	98%	73%	75%
Hungary	98%	98%	95%	93%	93%
Indonesia	88%	100%	97%	85%	97%
Iran, Islamic Rep.	96%	100%	98%	95%	98%
Israel	99%	100%	94%	93%	94%
Italy	94%	100%	97%	92%	97%
Japan	93%	93%	95%	89%	89%
Jordan	99%	100%	99%	98%	99%
Korea, Rep. of	100%	100%	100%	100%	100%
Latvia	97%	98%	92%	89%	91%
Lithuania	100%	100%	88%	88%	88%
Macedonia, Rep. of	99%	99%	98%	98%	98%
Malaysia	99%	100%	99%	98%	99%
Moldova	97%	100%	98%	94%	98%
Morocco	99%	99%	93%	92%	93%
Netherlands	58%	85%	96%	56%	82%
New Zealand	93%	97%	94%	87%	91%
Philippines	99%	100%	93%	91%	93%
Romania	98%	98%	98%	97%	97%
Russian Federation	98%	99%	97%	95%	96%
Singapore	100%	100%	98%	98%	98%
Slovak Republic	95%	97%	99%	94%	95%
Slovenia	98%	99%	95%	93%	94%
South Africa	84%	89%	92%	77%	82%
Thailand	95%	100%	99%	95%	99%
Tunisia	85%	100%	98%	83%	98%
Turkey	99%	100%	99%	98%	99%
United States	82%	90%	93%	77%	84%

Exhibit 9.8 Weighted Participation Rates

Country	School Participation Before Replacement	School Participation After Replacement	Student Participation	Overall Participation Before Replacement	Overall Participation After Replacement
Australia	83%	93%	90%	75%	84%
Belgium (Flemish)	72%	89%	97%	70%	87%
Bulgaria	97%	97%	96%	93%	93%
Canada	92%	95%	96%	88%	92%
Chile	98%	100%	96%	94%	96%
Chinese Taipei	100%	100%	99%	99%	99%
Cyprus	100%	100%	97%	97%	97%
Czech Republic	94%	100%	96%	90%	96%
England	49%	85%	90%	45%	77%
Finland	97%	100%	96%	93%	96%
Hong Kong, SAR	75%	76%	98%	74%	75%
Hungary	98%	98%	95%	93%	93%
Indonesia	84%	100%	97%	81%	97%
Iran, Islamic Rep.	96%	100%	98%	95%	98%
Israel	98%	100%	94%	93%	94%
Italy	94%	100%	97%	91%	97%
Japan	93%	93%	95%	89%	89%
Jordan	99%	100%	99%	98%	99%
Korea, Rep. of	100%	100%	100%	100%	100%
Latvia	96%	98%	93%	89%	91%
Lithuania	100%	100%	89%	89%	89%
Macedonia, Rep. of	99%	99%	98%	98%	98%
Malaysia	99%	100%	99%	98%	99%
Moldova	96%	100%	98%	94%	98%
Morocco	99%	99%	92%	91%	92%
Netherlands	62%	85%	95%	59%	81%
New Zealand	93%	97%	94%	87%	91%
Philippines	98%	100%	92%	91%	92%
Romania	98%	98%	98%	97%	97%
Russian Federation	98%	100%	97%	95%	97%
Singapore	100%	100%	98%	98%	98%
Slovak Republic	95%	96%	98%	93%	94%
Slovenia	98%	99%	95%	93%	94%
South Africa	85%	91%	93%	79%	84%
Thailand	93%	100%	99%	93%	99%
Tunisia	84%	100%	98%	82%	98%
Turkey	99%	100%	99%	98%	99%
United States	83%	90%	94%	78%	85%

Data Management and Database Construction

Dirk Hastedt
Eugenio J. Gonzalez



10 Data Management and Database Construction

Dirk Hastedt
Eugenio J. Gonzalez

10.1 Overview

The TIMSS 1999 data were processed in close cooperation among the TIMSS International Study Center at Boston College, the IEA Data Processing Center, the Educational Testing Service, Statistics Canada, and the national research centers of the participating countries. Under the direction of the International Study Center, each institution was responsible for specific aspects of the data processing.

Data processing consisted of six general tasks: data entry, creation of the international database, calculation of sampling weights, scaling of achievement data, analysis of the background data, and creation of the exhibits for the international reports. Each task was crucial to ensuring the quality and accuracy of the TIMSS results. This chapter describes the data entry task undertaken by each national research coordinator,[1] the data checking and database creation that was implemented by the IEA Data Processing Center, and the steps taken to ensure the quality and accuracy of the international database.[2] It discusses the responsibilities of each participant in creating the international database; the flow of the data files among the centers involved in the data processing; the structure of the data files submitted by each country for processing, and the resulting files that are part of the international database; the rules, methods, and procedures used for data verification and manipulation; the data products created during data cleaning and provided to the national centers; and the computer software used in that process.

10.2 Data Flow

The data collected in the TIMSS 1999 survey were entered into data files with a common international format at the national research centers of the participating countries. These data files were then submitted to the IEA Data Processing Center for clean-

○○○

[1] Further information about the role of the national research coordinator in entering and checking the TIMSS data may be found in Chapter 7, TIMSS Field Operations.

[2] The TIMSS weighting, scaling, and analysis procedures are described in Chapters 11, 14, and 15, respectively.

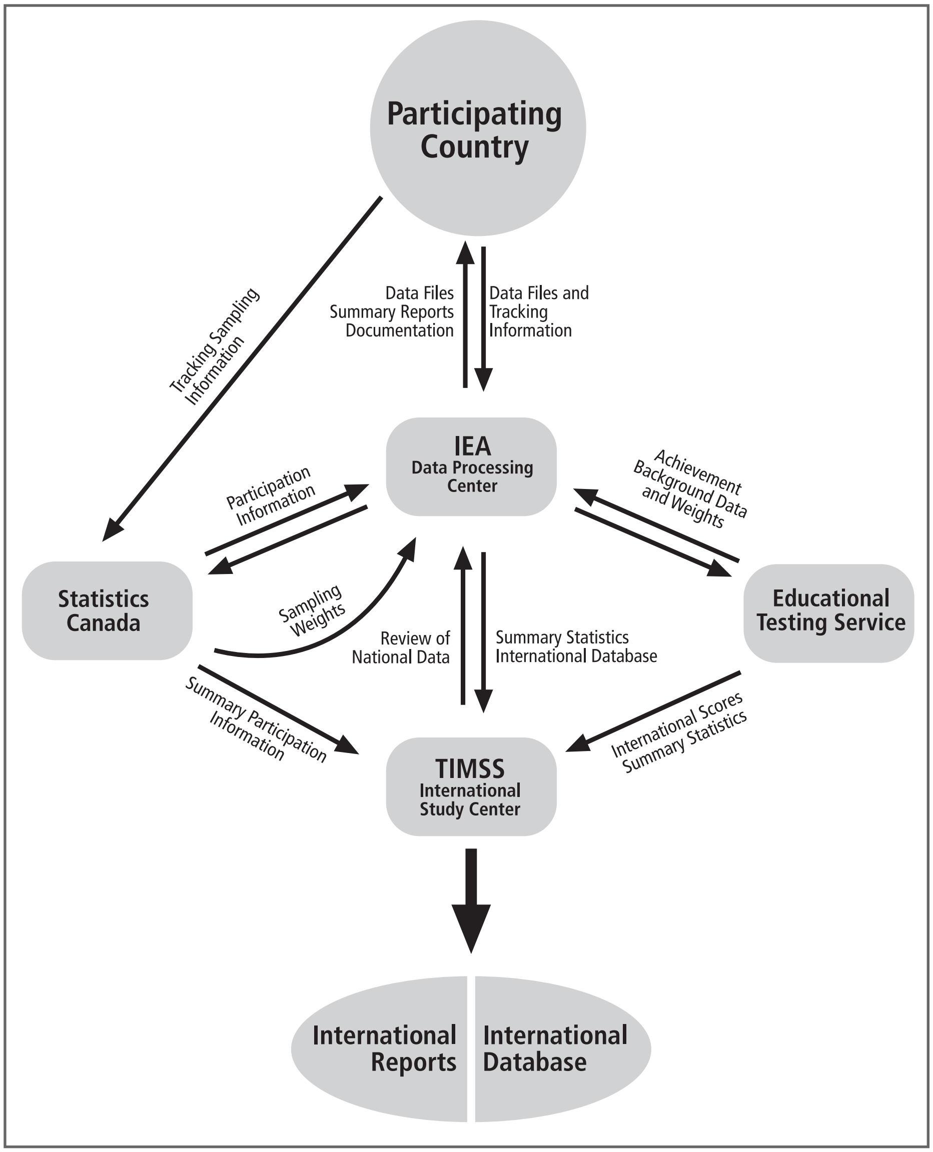
ing and verification. The major responsibilities of the Data Processing Center were to check that the data files received matched the international standard and to make modifications where necessary; to apply standard cleaning rules to the data to verify their consistency and accuracy; to interact with the national research coordinators (NRCs) to ensure their accuracy; produce summary statistics of the background and achievement data for review by the TIMSS International Study Center; and finally, upon feedback from the individual countries and the TIMSS International Study Center, to construct the international database. The IEA Data Processing Center was also responsible for distributing the national data files to each of the participating countries.

Once verified and in the international file format, the achievement data were sent to the International Study Center where basic item statistics were produced and reviewed.[3] At the same time the Data Processing Center sent data files containing information on the participation of schools and students in each country's sample to Statistics Canada. This information, together with data provided by the national research coordinator from tracking forms and the within-school sampling software, was used by Statistics Canada to calculate sampling weights, population coverage, and school and student participation rates. The sampling weights were sent to the TIMSS International Study Center for verification and then forwarded to Educational Testing Service for use in scaling the achievement data. When the review of the item statistics was completed and the Data Processing Center had updated the database, the student achievement files were sent to Educational Testing Service who conducted the scaling and generated proficiency scores in mathematics and science for each participating student. Once the sampling weights and the proficiency scores were verified at the International Study Center, they were sent to the Data Processing Center for inclusion in the international database and then distributed to the national research centers. The International Study Center prepared the exhibits for the international reports and published the results of the study. Exhibit 10.1 is a pictorial representation of the flow of the data files.

○○○

[3] The item review process is described in Chapter 13.

Exhibit 10.1 Flow of Data Files


Participating Country
Data Files
Summary Reports
Documentation
Data Files and Tracking Information
Tracking Sampling Information
IEA
Data Processing Center
Participation Information
Achievement Background Data and Weights
Statistics Canada
Educational Testing Service
Sampling Weights
Review of National Data
Summary Statistics
International Database
Summary Participation Information
International Scores Summary Statistics
TIMSS
International Study Center
International Reports
International Database

10.3 Data Entry at the National Research Centers

Each TIMSS 1999 national research center was responsible for transcribing the information from the achievement booklets and questionnaires into computer data files. Data-entry software, DATAENTRYMANAGER (DEM), adapted specifically for TIMSS 1999 was provided to each participating country, together with the *Manual for Entering the TIMSS-R Data* (TIMSS, 1998) and codebooks describing the layout of the data. The codebooks contained information about the variable names used for each variable in the survey instruments, and about field length, field location, labels, valid ranges, default values, and missing codes.

The codebooks were designed to be an integral part of the DEM system for data entry, although they could also be used in conjunction with other data-entry systems. Although DEM was the recommended software, some of the participating countries elected to use a different data entry system. Data files were accepted from the countries provided they conformed to the specifications given in the international codebooks.

In order to facilitate data entry, the codebooks and data files were structured to match the tests and questionnaires. This meant that there was a data file for each survey instrument. Each country was responsible for submitting six data files: Student Background, Student Achievement, Scoring Reliability, Mathematics Teacher Background, Science Teacher Background, and School Background. Each file had its own codebook.

The following files were used during data entry

- The Student Background Data File contained one record for every student listed on the *Student Tracking Form* for each sampled class, including students that did not participate in the testing session and students that were excluded. There are two versions of the student questionnaire, one tailored to countries where the sciences are taught as a general integrated subject and the other for countries where biology, physics, and chemistry are taught as separate subjects. Each version of the questionnaire had its own codebook and data file layout. The data for the student background data file came from the *Student Tracking Form*, the *Teacher Student Linkage Form*, and the *Test Administration Form*, as well as from the *Student Background Questionnaire*.
- The Student Achievement Data File contained a record for each student that completed an achievement test booklet. Each record contains the student's responses to whichever of the 8 test booklets was assigned to the student.
- In order to check the reliability of the free-response item scoring, the free-response items in a random sample of 25 percent of booklets were scored independently by a second scorer. The responses from these booklets were stored in a Scoring Reliability File, which contained one record for each student in the reliability sample.

- The Teacher Questionnaire Data Files contained one record for every entry on the *Teacher Tracking Form*, including teachers who did not complete a teacher questionnaire. There were separate data files for mathematics and science teachers. The data for this file came from the mathematics and science teacher questionnaires and from the teacher tracking form.
- The School Data file contained one record for each school sampled to take part in the study, whether the school participated or not. The file also contained a record for each replacement school in which data collection took place. The data for this came from the school questionnaire and the school tracking form.

10.3.1 Data Files Received by the IEA Data Processing Center

In addition to the data files, countries were also required to submit documentation supporting their field procedures and copies of their national translated tests and questionnaires. The documentation included a report of their survey activities, a series of data management forms with clear indications of any changes made in the tests or the layout of the data files, and copies of all sampling and tracking forms. These materials were archived at the IEA Data Processing Center and kept for reference during data processing.

Each country was provided with a program called LINKCHK to carry out checks on the data files before submitting them to the IEA Data Processing Center. The program was designed to help NRCs perform an initial check of the system of student, teacher, and school identification numbers after data entry, both within and across files. The reports produced by the LINKCHK program allowed countries to correct problems in the identification system before transferring the data to the IEA Data Processing Center.

10.4 Data Cleaning at the IEA Data Processing Center

Once the data were entered into data files at the national research center, the data files were submitted to the IEA Data Processing Center for checking and input into the international database. This process is generally referred to as data cleaning. The goals of the TIMSS 1999 data cleaning were to identify, document, and, where necessary and possible, correct deviations from the international file structure, and to correct key punch errors, systematic deviations from the international data formats, problems in linking observations across files, inconsistent tracking information across and within files, and inconsistencies within and across observations. The main objective of the process was to ensure that the data adhered to international formats and accurately and consistently reflected the information collected within each country.

Data cleaning involved several steps. Some of these were repeated until satisfactory results were achieved. During the first step, all incoming data files were checked and reformatted as necessary so that their file structure conformed to the international format. As a second step, all problems with identification variables, linkage across files, codes used for different groups of variables, and participation status were detected and corrected. The distribution for each variable was examined, with particular attention to variables that presented implausible or inconsistent distributions based on the information from the country involved and on other answers in the questionnaires.

During this stage, a series of data summary reports were generated for each country. The reports contained listings of codes used for each variable and pointed to outliers and changes in the structure of the data file. They also contained univariate statistics. The reports were sent to each participating country, and the NRC was asked to review the data and advise how best to resolve inconsistencies. In many cases the NRC was asked to go back to the original booklets from which the data had been entered.

In data cleaning two main procedures were used to make necessary changes in the data. Inconsistencies that could unambiguously be solved were corrected automatically by a program using standard cleaning routines. Errors that could not be solved with these routines had to be solved case by case by the DPC staff. In either case, all changes made in the data were documented. A database was created in which each change was recorded, and it was possible to reconstruct the original database received from a country.

10.4.1 Standardization of the National File Structure

The first step in the data processing at the international level was to verify the compatibility of the national datasets with the international file structure as defined in the TIMSS 1999 international codebook. This was necessary before the standard cleaning with the Data Processing Center cleaning software could be performed.

Although the TIMSS 1999 international codebooks distributed with the data entry software gave clear and detailed instructions about the structure and format of the files each country was to submit to the IEA Data Processing Center, some countries opted to enter and submit their data files in other formats, using structures different from the international standard. For the most part, these differences were due to specific national circumstances. The manual for entering the TIMSS 1999 data asked countries to prepare and send their data files using the DEM software, which produces an extended dBase format file. Some data files, however, were received in ASCII fixed format (raw data), SPSS format, and SAS format.

After the national files were converted into dBase format, the structure of the files was inspected and deviations from the international file structure were identified. A custom-designed program scanned the file structure of the files for each country and identified the following deviations:

- International variables omitted
- National variables added
- Different variable length or number of decimal positions
- Different coding schemes or out-of-range values
- Specific national variables
- Gang-punched variables

Together with the inspection of the national data files, the data management and tracking forms submitted by each NRC were reviewed. As a result of this initial review, the Data Processing Center outlined and implemented necessary changes in the national data to make the files compatible with the international format. In most cases programs had to be prepared to fit the file structures and particularities of each country.

As part of the standardization process, the file structure was rearranged to facilitate data analysis, since direct correspondence between the files and the data-collection instruments was no longer necessary. At this stage also, the Student Background file and Student Achievement files were merged to form a single student file. Variables created during data entry for verification only were omitted from all files at this time, and new variables were added (i.e., reporting variables, derived variables, sampling weights, and achievement scores).

10.4.2 Cleaning Rules and Procedures

After the data files received from the countries were transformed into the international format, certain standard checks and cleaning rules were applied to each file.

The first step was to check for deviations from international standards in both data-collection instruments and data files. Instruments were checked for missing questions or items, changes in the number of answer categories, alterations in coding schemes, and other national adaptations. Data files were examined for missing variables, changes in field length and number of decimal places, modifications of coding schemes, and additional national variables.

After all deviations from the international standard had been identified, a cleaning program was run on each file to make a set of standard changes. This was to facilitate the application of more specific cleaning rules at the next stage. After this step, each data file matched the international standard as specified in the international codebook. Among changes made at this time were adjustments to the hierarchical identification number system, differentiation between ‘Not Applicable’, ‘Missing’, and ‘Not Administered’ codes, adding omitted variables and coding them as ‘Not administered’, and recoding systematic deviations from the international coding scheme.

The TIMSS 1999 cleaning program labelled each problem with an identification number, a description of the problem, and the action taken by the program. All problems were recorded in an error database containing one error file for each file that was checked. As problems were identified that could not be automatically rectified, they were reported to the responsible NRC so that data-collection instruments and tracking forms could be checked to trace the source of the errors. Wherever possible, staff at the IEA Data Processing Center suggested a remedy, and asked the NRCs to either accept it or propose an alternative. The data files were updated as solutions to problems were found. Where problems could not be solved by the NRC by inspecting the instruments or forms, they were rectified by applying a general cleaning program.

After all automatic updates had been applied, remaining corrections to the data files were applied directly, or manually, using a specially developed editing program. The manual corrections took into account country-specific information that could not be used by the cleaning program.

10.4.3 National Cleaning Documentation

National research coordinators received a detailed report of all problems identified in their data, and of the steps taken to correct them. These included:

- A record of all deviations from the international data-collection instruments and the international file structure
- Documentation of the data problems uncovered by the cleaning program and the steps taken to resolve them
- A list of all manual corrections made in each data file

In addition to documentation of data errors and updates, the IEA Data Processing Center provided each NRC with new data files that incorporated all agreed updates. The data files were transformed from the standard layout designed to facilitate data entry to a new format oriented more toward data analysis. The updated files included a range of new variables that could be used for analytic purposes. For example, the student files included nationally standardized scores in mathematics and science that could be used in national analyses to be conducted before the international database became available.

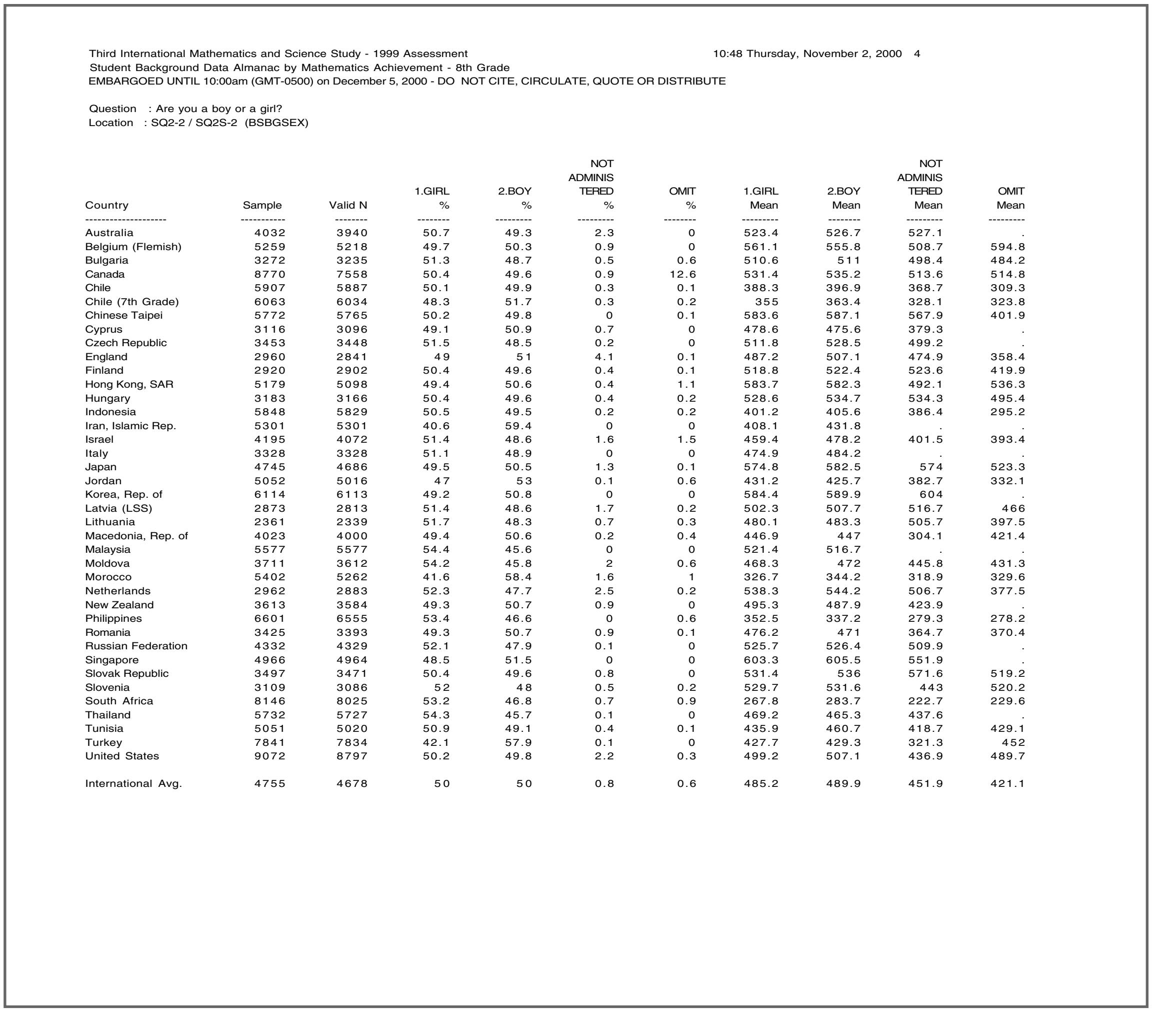
10.5 Data Products

Data products sent by the IEA Data Processing Center to NRCs included both data almanacs and data files.

10.5.1 Data Almanacs

Each country received a set of data almanacs, or summaries, produced by the TIMSS International Study Center. These contained weighted summary statistics for each participating country on each variable included in the survey instruments. There were two types of display. The display for categorical variables included an estimate of the size of the student population, the sample size, the weighted percentage of students who were not administered the question, the percentage of students choosing each of the options on the question, and the percentage of students who chose none of the valid options. The percentage of students to whom the question did not apply was also presented. For continuous variables the display included an estimate of the size of the student population, the sample size, the weighted percentage of students who were not administered the question, the percentage who did not respond, the percentage to whom the question did not apply, the mean, mode, minimum, maximum, and the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles. An example of such data displays is presented in Exhibits 10.2 and 10.3. These data almanacs were sent to the participating countries for review. When necessary, they were accompanied by specific questions about the data presented in them. They were also used by the TIMSS International Study Center during the data review and in the production of the reporting exhibits.

Exhibit 10.2 Example Data Almanac Display for a Categorical Student Background Variable

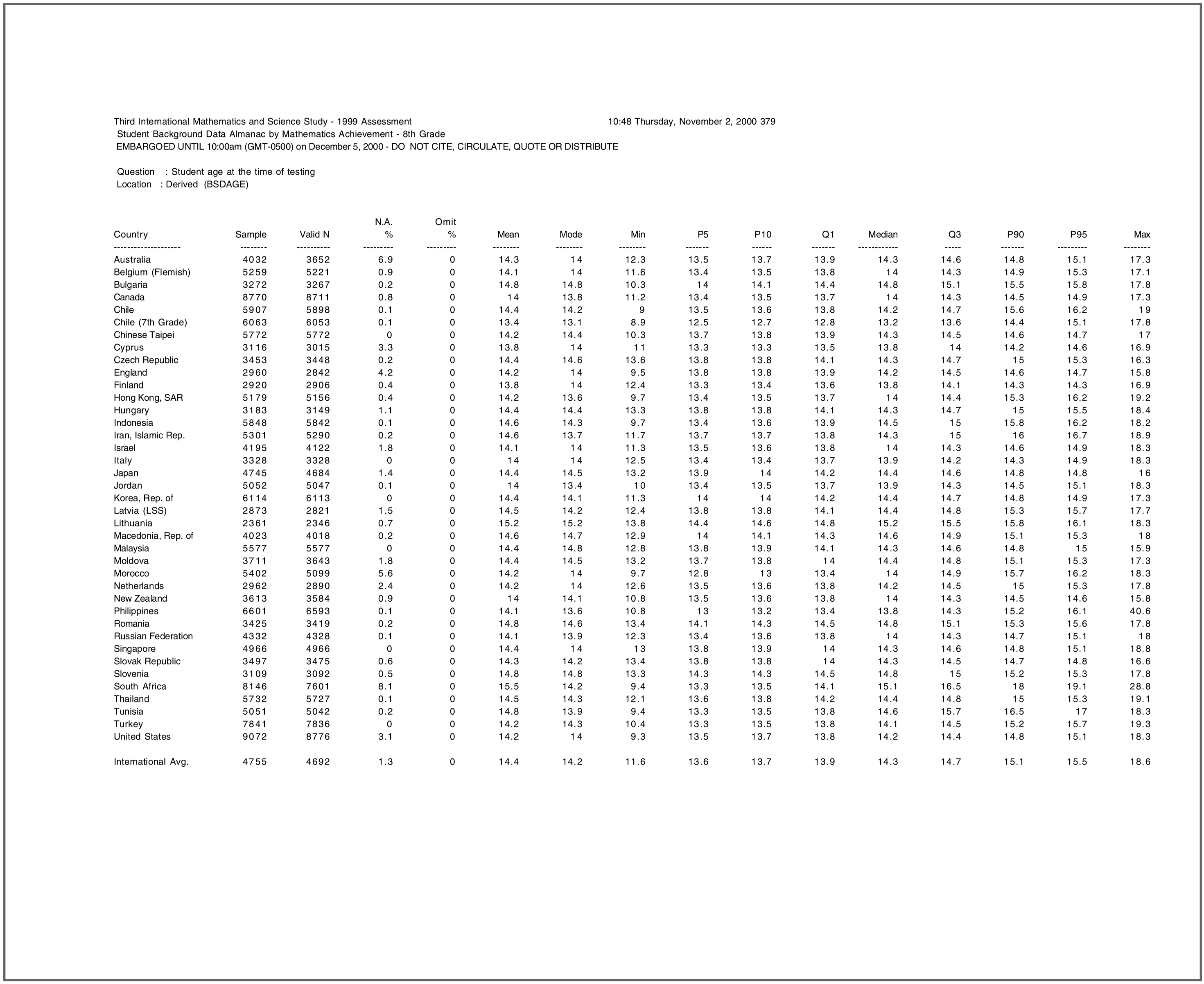


Third International Mathematics and Science Study - 1999 Assessment 10:48 Thursday, November 2, 2000 4
Student Background Data Almanac by Mathematics Achievement - 8th Grade
EMBARGOED UNTIL 10:00am (GMT-0500) on December 5, 2000 - DO NOT CITE, CIRCULATE, QUOTE OR DISTRIBUTE

Question : Are you a boy or a girl?
Location : SQ2-2 / SQ2S-2 (BSBGSEX)

					NOT				NOT	
					ADMINIS				ADMINIS	
			1.GIRL	2.BOY	TERED	OMIT	1.GIRL	2.BOY	TERED	OMIT
Country	Sample	Valid N	%	%	%	%	Mean	Mean	Mean	Mean
Australia	4032	3940	50.7	49.3	2.3	0	523.4	526.7	527.1	.
Belgium (Flemish)	5259	5218	49.7	50.3	0.9	0	561.1	555.8	508.7	594.8
Bulgaria	3272	3235	51.3	48.7	0.5	0.6	510.6	511	498.4	484.2
Canada	8770	7558	50.4	49.6	0.9	12.6	531.4	535.2	513.6	514.8
Chile	5907	5887	50.1	49.9	0.3	0.1	388.3	396.9	368.7	309.3
Chile (7th Grade)	6063	6034	48.3	51.7	0.3	0.2	355	363.4	328.1	323.8
Chinese Taipei	5772	5765	50.2	49.8	0	0.1	583.6	587.1	567.9	401.9
Cyprus	3116	3096	49.1	50.9	0.7	0	478.6	475.6	379.3	.
Czech Republic	3453	3448	51.5	48.5	0.2	0	511.8	528.5	499.2	.
England	2960	2841	49	51	4.1	0.1	487.2	507.1	474.9	358.4
Finland	2920	2902	50.4	49.6	0.4	0.1	518.8	522.4	523.6	419.9
Hong Kong, SAR	5179	5098	49.4	50.6	0.4	1.1	583.7	582.3	492.1	536.3
Hungary	3183	3166	50.4	49.6	0.4	0.2	528.6	534.7	534.3	495.4
Indonesia	5848	5829	50.5	49.5	0.2	0.2	401.2	405.6	386.4	295.2
Iran, Islamic Rep.	5301	5301	40.6	59.4	0	0	408.1	431.8	.	.
Israel	4195	4072	51.4	48.6	1.6	1.5	459.4	478.2	401.5	393.4
Italy	3328	3328	51.1	48.9	0	0	474.9	484.2	.	.
Japan	4745	4686	49.5	50.5	1.3	0.1	574.8	582.5	574	523.3
Jordan	5052	5016	47	53	0.1	0.6	431.2	425.7	382.7	332.1
Korea, Rep. of	6114	6113	49.2	50.8	0	0	584.4	589.9	604	.
Latvia (LSS)	2873	2813	51.4	48.6	1.7	0.2	502.3	507.7	516.7	466
Lithuania	2361	2339	51.7	48.3	0.7	0.3	480.1	483.3	505.7	397.5
Macedonia, Rep. of	4023	4000	49.4	50.6	0.2	0.4	446.9	447	304.1	421.4
Malaysia	5577	5577	54.4	45.6	0	0	521.4	516.7	.	.
Moldova	3711	3612	54.2	45.8	2	0.6	468.3	472	445.8	431.3
Morocco	5402	5262	41.6	58.4	1.6	1	326.7	344.2	318.9	329.6
Netherlands	2962	2883	52.3	47.7	2.5	0.2	538.3	544.2	506.7	377.5
New Zealand	3613	3584	49.3	50.7	0.9	0	495.3	487.9	423.9	.
Philippines	6601	6555	53.4	46.6	0	0.6	352.5	337.2	279.3	278.2
Romania	3425	3393	49.3	50.7	0.9	0.1	476.2	471	364.7	370.4
Russian Federation	4332	4329	52.1	47.9	0.1	0	525.7	526.4	509.9	.
Singapore	4966	4964	48.5	51.5	0	0	603.3	605.5	551.9	.
Slovak Republic	3497	3471	50.4	49.6	0.8	0	531.4	536	571.6	519.2
Slovenia	3109	3086	52	48	0.5	0.2	529.7	531.6	443	520.2
South Africa	8146	8025	53.2	46.8	0.7	0.9	267.8	283.7	222.7	229.6
Thailand	5732	5727	54.3	45.7	0.1	0	469.2	465.3	437.6	.
Tunisia	5051	5020	50.9	49.1	0.4	0.1	435.9	460.7	418.7	429.1
Turkey	7841	7834	42.1	57.9	0.1	0	427.7	429.3	321.3	452
United States	9072	8797	50.2	49.8	2.2	0.3	499.2	507.1	436.9	489.7
International Avg.	4755	4678	50	50	0.8	0.6	485.2	489.9	451.9	421.1

Exhibit 10.3 Example Data Almanac Display for a Numerical School Background Variable



Third International Mathematics and Science Study - 1999 Assessment 10:48 Thursday, November 2, 2000 379
Student Background Data Almanac by Mathematics Achievement - 8th Grade
EMBARGOED UNTIL 10:00am (GMT-0500) on December 5, 2000 - DO NOT CITE, CIRCULATE, QUOTE OR DISTRIBUTE

Question : Student age at the time of testing
Location : Derived (BSDAGE)

Country	Sample	Valid N	N.A. %	Omit %	Mean	Mode	Min	P5	P10	Q1	Median	Q3	P90	P95	Max
Australia	4032	3652	6.9	0	14.3	14	12.3	13.5	13.7	13.9	14.3	14.6	14.8	15.1	17.3
Belgium (Flemish)	5259	5221	0.9	0	14.1	14	11.6	13.4	13.5	13.8	14	14.3	14.9	15.3	17.1
Bulgaria	3272	3267	0.2	0	14.8	14.8	10.3	14	14.1	14.4	14.8	15.1	15.5	15.8	17.8
Canada	8770	8711	0.8	0	14	13.8	11.2	13.4	13.5	13.7	14	14.3	14.5	14.9	17.3
Chile	5907	5898	0.1	0	14.4	14.2	9	13.5	13.6	13.8	14.2	14.7	15.6	16.2	19
Chile (7th Grade)	6063	6053	0.1	0	13.4	13.1	8.9	12.5	12.7	12.8	13.2	13.6	14.4	15.1	17.8
Chinese Taipei	5772	5772	0	0	14.2	14.4	10.3	13.7	13.8	13.9	14.3	14.5	14.6	14.7	17
Cyprus	3116	3015	3.3	0	13.8	14	11	13.3	13.3	13.5	13.8	14	14.2	14.6	16.9
Czech Republic	3453	3448	0.2	0	14.4	14.6	13.6	13.8	13.8	14.1	14.3	14.7	15	15.3	16.3
England	2960	2842	4.2	0	14.2	14	9.5	13.8	13.8	13.9	14.2	14.5	14.6	14.7	15.8
Finland	2920	2906	0.4	0	13.8	14	12.4	13.3	13.4	13.6	13.8	14.1	14.3	14.3	16.9
Hong Kong, SAR	5179	5156	0.4	0	14.2	13.6	9.7	13.4	13.5	13.7	14	14.4	15.3	16.2	19.2
Hungary	3183	3149	1.1	0	14.4	14.4	13.3	13.8	13.8	14.1	14.3	14.7	15	15.5	18.4
Indonesia	5848	5842	0.1	0	14.6	14.3	9.7	13.4	13.6	13.9	14.5	15	15.8	16.2	18.2
Iran, Islamic Rep.	5301	5290	0.2	0	14.6	13.7	11.7	13.7	13.7	13.8	14.3	15	16	16.7	18.9
Israel	4195	4122	1.8	0	14.1	14	11.3	13.5	13.6	13.8	14	14.3	14.6	14.9	18.3
Italy	3328	3328	0	0	14	14	12.5	13.4	13.4	13.7	13.9	14.2	14.3	14.9	18.3
Japan	4745	4684	1.4	0	14.4	14.5	13.2	13.9	14	14.2	14.4	14.6	14.8	14.8	16
Jordan	5052	5047	0.1	0	14	13.4	10	13.4	13.5	13.7	13.9	14.3	14.5	15.1	18.3
Korea, Rep. of	6114	6113	0	0	14.4	14.1	11.3	14	14	14.2	14.4	14.7	14.8	14.9	17.3
Latvia (LSS)	2873	2821	1.5	0	14.5	14.2	12.4	13.8	13.8	14.1	14.4	14.8	15.3	15.7	17.7
Lithuania	2361	2346	0.7	0	15.2	15.2	13.8	14.4	14.6	14.8	15.2	15.5	15.8	16.1	18.3
Macedonia, Rep. of	4023	4018	0.2	0	14.6	14.7	12.9	14	14.1	14.3	14.6	14.9	15.1	15.3	18
Malaysia	5577	5577	0	0	14.4	14.8	12.8	13.8	13.9	14.1	14.3	14.6	14.8	15	15.9
Moldova	3711	3643	1.8	0	14.4	14.5	13.2	13.7	13.8	14	14.4	14.8	15.1	15.3	17.3
Morocco	5402	5099	5.6	0	14.2	14	9.7	12.8	13	13.4	14	14.9	15.7	16.2	18.3
Netherlands	2962	2890	2.4	0	14.2	14	12.6	13.5	13.6	13.8	14.2	14.5	15	15.3	17.8
New Zealand	3613	3584	0.9	0	14	14.1	10.8	13.5	13.6	13.8	14	14.3	14.5	14.6	15.8
Philippines	6601	6593	0.1	0	14.1	13.6	10.8	13	13.2	13.4	13.8	14.3	15.2	16.1	40.6
Romania	3425	3419	0.2	0	14.8	14.6	13.4	14.1	14.3	14.5	14.8	15.1	15.3	15.6	17.8
Russian Federation	4332	4328	0.1	0	14.1	13.9	12.3	13.4	13.6	13.8	14	14.3	14.7	15.1	18
Singapore	4966	4966	0	0	14.4	14	13	13.8	13.9	14	14.3	14.6	14.8	15.1	18.8
Slovak Republic	3497	3475	0.6	0	14.3	14.2	13.4	13.8	13.8	14	14.3	14.5	14.7	14.8	16.6
Slovenia	3109	3092	0.5	0	14.8	14.8	13.3	14.3	14.3	14.5	14.8	15	15.2	15.3	17.8
South Africa	8146	7601	8.1	0	15.5	14.2	9.4	13.3	13.5	14.1	15.1	16.5	18	19.1	28.8
Thailand	5732	5727	0.1	0	14.5	14.3	12.1	13.6	13.8	14.2	14.4	14.8	15	15.3	19.1
Tunisia	5051	5042	0.2	0	14.8	13.9	9.4	13.3	13.5	13.8	14.6	15.7	16.5	17	18.3
Turkey	7841	7836	0	0	14.2	14.3	10.4	13.3	13.5	13.8	14.1	14.5	15.2	15.7	19.3
United States	9072	8776	3.1	0	14.2	14	9.3	13.5	13.7	13.8	14.2	14.4	14.8	15.1	18.3
International Avg.	4755	4692	1.3	0	14.4	14.2	11.6	13.6	13.7	13.9	14.3	14.7	15.1	15.5	18.6

10.5.2 Versions of the National Data Files

Building the international database was an iterative process. The IEA Data Processing Center provided NRCs with a new version of their country's data files whenever a major step in data processing was completed. This also guaranteed that the NRCs had a chance to review their data and run their own checks to validate the data files.

Three versions of the data files were sent out to the countries before the TIMSS international database was made available. Each country received its own data only. The first version was sent to the NRC as soon as that country's data had been cleaned. These files contained nationally standardized achievement scores calculated by the Data Processing Center using a Rasch-based scaling method. Documentation, with a list of the cleaning checks and all corrections made in the data, was included to enable the NRC to review the cleaning process. Univariate statistics for the background data and item statistics for the achievement data were also provided for statistical review. A second version of the data files was sent to the NRCs when the weights and the international achievement scores were available and had been merged with the files. A third version was sent together with the data almanacs after final updates had been made, to enable the NRCs to validate the results presented in the first international reports.

10.5.3 Reports

Several reports were produced during data processing at the IEA Data Processing Center to inform and assist the NRCs, the TIMSS International Study Center, and other institutions involved in TIMSS 1999. The NRCs were provided with diagnostic reports and univariate statistics to help them check their data. The TIMSS International Study Center and ETS were provided with international item statistics. The International Study Center also received international coding reliability statistics and international univariate statistics. A report was made to the International Study Center and the TIMSS 1999 Project Management Committee about the status of each country's data, any problems encountered in the data cleaning, and general statistics about the number of observations per file and preliminary student response rates.

10.6 Computer Software

dBase was used as the standard database program for handling the incoming data. Tools for precleaning and programs such as LINKCHCK (described earlier) and MANCORR and CLEAN (described below) were developed for manipulating the data. Statistical analyses (e.g., univariate statistics) for data cleaning and review were carried out with SAS. The final data sets were also created using SAS. For item statistics, the Data Processing Center used the QUEST software (Adams and Khoo, 1993).

The main programs that were developed by the Data Processing Center and used for TIMSS 1999 are described below. Most of the programs that were written for country-specific cleaning needs are not listed. The programming resources in the main cleaning process were spent largely in developing this set of programs.

10.6.1 MANCORR

The most time-consuming and error-prone part of data cleaning is the direct or 'manual' editing of errors uncovered by the review process. Based on the Data Processing Center's experience in the IEA Reading Literacy Study, TIMSS 1995, and the pilot phases of TIMSS 1999, the data-editing program MANCORR was developed. It is easy to use and generates automatic reports of all data manipulation. Its main advantage compared with other editors is that all changes in the data are documented in a log database, from which reports can be generated. As updated data were received from countries, the time-intensive manual changes could be automatically repeated. An 'Undo' function allowed the restoration of original values that had been modified with the MANCORR program. The report on which changes were made in the data, by whom, and when was important for internal quality control and review. The MANCORR program was designed using CLIPPER in order to manipulate DATAENTRYMANAGER files.

10.6.2 CLEAN

The main software instrument for data cleaning in TIMSS 1999 was the diagnostic program CLEAN. This program was derived from earlier versions used in the IEA Reading Literacy Study and TIMSS 1995. It was used to check all the TIMSS 1999 files individually, the linkages across files, and all between-file comparisons. An important feature of the program is that it could be used on a data file as often as necessary. It could first be used to make automatic corrections, and subsequently for creating a report only, without making corrections. Thus it was possible to run a check on the files at all stages of work until the end, when the file format was changed to the SAS format. This meant that the program was used not only for initial checks but also to check the work done at the Data Processing Center.

A feature of the TIMSS 1999 data cleaning tools is that all deviations are reported to a database, so that reports can be generated by type of problem or by record. Reports previously generated by the program could be compared automatically with newer reports to see which problems had been solved, and even more important, whether additional deviations were introduced during manual correction. The databases were used to generate the final reports to be sent to the countries. These reports showed which deviations were initially in the data, which were solved automatically, which were solved manually, and which remained unchanged.

10.6.3 Programs Creating Meta Databases

Using SAS, several programs were developed by the Data Processing Center for reviewing and analyzing both the background data and the test items. For the background data, a meta database containing information provided by the initial analysis and by the international codebook was created. Another meta database containing the relevant item parameters was created for the achievement test items. Later, all statistical checks and reports used these databases instead of running the statistics over all data sets again and again. If the data for one country were changed then statistics had to be recalculated for that country only. This reduced the computing time for certain procedures from hours to a few minutes. Both databases are the base sources of several reports produced at both the national and international levels (e.g., for the univariate and item analysis reports).

10.6.4 Export Programs

As mentioned above, SAS was the main program for analyzing the data. Using SAS, export programs were developed and tested to create output data sets for data distribution that are readable by either SAS or SPSS.

10.7 Summary

The structures and procedures designed for processing the TIMSS 1999 data were found to be very successful. In planning for the TIMSS data processing, the major problems were anticipated and provision for dealing with them was incorporated into the data processing system. The IEA Data Processing Center was closely involved in the planning phase of the study. The study thus benefited from the Center's knowledge and experience in data processing. TIMSS 1999 also benefited from the experience gained with TIMSS in 1995. Procedures and practices developed in response to problems encountered in the earlier study were refined and made even more effective in 1999. In particular, since the work of checking and cleaning the TIMSS database required a vast amount of interaction between NRCs in each country, the IEA Data Processing Center in Hamburg, the TIMSS International Study Center in Boston, Statistics Canada in Ottawa, and Educational Testing Service in Princeton, improvements in communications and data transmission greatly facilitated the entire enterprise.

References

Adams, R.J., & Khoo, S. (1993). *Quest: The interactive test analysis system.* Melbourne: Australian Council for Educational Research.

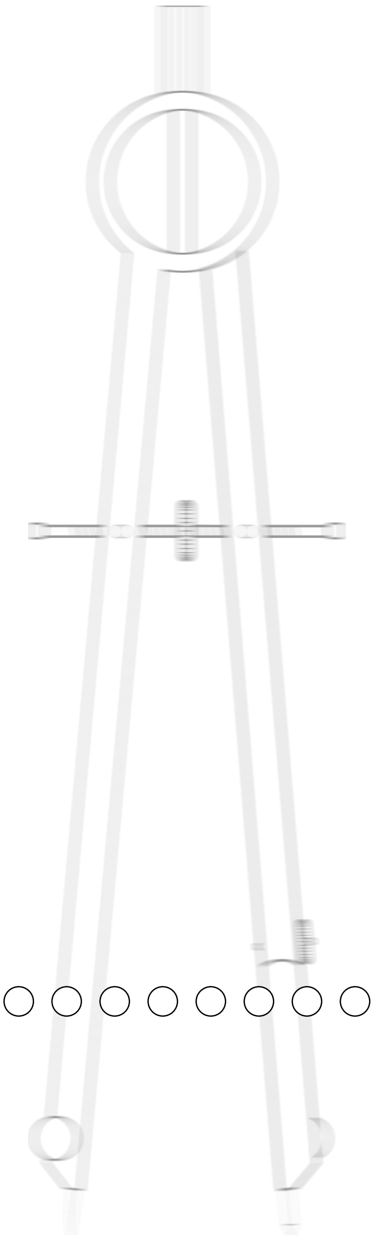
Gonzalez, E.J., & Smith, T.A., Eds. (1997). *User Guide for the TIMSS international database: Primary and middle school years – 1995 assessment.* Chestnut Hill, MA: Boston College.

TIMSS (1998). *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.

11

Sampling Weights

Pierre Foy



11 Sampling Weights

Pierre Foy

11.1 Overview

The basic sample design used in TIMSS 1999 was a two-stage stratified cluster design, with schools as the first stage and classrooms as the second. The design required schools to be sampled using a probability proportional to size (PPS) systematic method and classrooms to be sampled with equal probabilities[1]. TIMSS participants adapted the basic design to the requirements of their education systems, with guidance from the TIMSS sampling consultants at Statistics Canada and the sampling referee. Very large countries could add an extra, preliminary stage, where districts or regions were sampled first, and then schools within districts.[2] Also, countries where classes were usually very large could select a subsample of students from sampled classes.[3] Participants used stratification to improve the precision of their samples where appropriate. These adaptations could be quite complex, as may be seen from the information in Appendix D showing how the TIMSS design was implemented in each country.

While the TIMSS multistage stratified cluster design provided very economical and effective data collection in a school environment, it results in differential probabilities of selection for the ultimate sampling elements, the students. Consequently, one student in the assessment does not necessarily represent the same proportion of students in the population as another, as would be the case with a simple random sampling approach. To account for differential probabilities of selection due to the design and to ensure proper survey estimates, TIMSS computed a sampling weight for each participating student. Just as in TIMSS 1995, the ability to provide proper sampling weights was an essential characteristic of an acceptable sample design in TIMSS 1999, since appropriate sampling weights were essential for the computation of accurate survey estimates. This chapter describes the procedures for calculating sampling weights for the TIMSS 1999 data.

[1] The TIMSS sample design is presented in Chapter 2.

[2] For example, the United States sampled school districts as primary sampling units and then schools within the school districts.

[3] Morocco was the only country to exercise this option in 1999.

11.2 Weighting Procedure

The weighting procedure required three steps, reflecting the TIMSS sample design. The first step consisted of calculating a school weight; this also incorporated weighting factors from any additional front-end sampling stages such as districts or regions. A school-level participation adjustment was then made in the school weight to compensate for any sampled schools that did not participate. That adjustment was calculated independently for each explicit stratum.

In the second step a classroom weight was calculated. No classroom-level participation adjustment was necessary, since in most cases a single classroom was sampled in each school. If a school agreed to take part in the study but the sampled classroom refused to participate, the non-participation adjustment was made at the school level. If one of two selected classrooms in a school did not participate, the classroom weight was calculated as though a single classroom had been selected in the first place. The classroom weight was calculated independently for each school.

The third and final step consisted of calculating a student weight. A non-participation adjustment was made to compensate for students that did not take part in the testing. The student weight was calculated independently for each sampled classroom. The basic sampling weight attached to each student record was the product of the three intermediate weights: the first stage (school) weight, the second stage (classroom) weight, and the third stage (student) weight. The overall student sampling weight was the product of these three weights and the two non-participation adjustments, school-level and student-level.

11.2.1 The First Stage (School) Weight

The first stage weight represented the inverse of the first stage selection probability assigned to a sampled school. The TIMSS 1999 sample design required that school selection probabilities be proportional to the school size (PPS) school size being enrolment in the target grade. The basic first stage weight for the i^{th} sampled school was thus defined as

$$BW_{sc}^{i} = \frac{M}{n \bullet m_i}$$

where n was the number of sampled schools, m_i was the measure of size for the i^{th} school, and

$$M = \sum_{i=1}^{N} m_i$$

where N was the total number of schools in the explicit stratum.

The basic first stage weight also incorporated weighting factors resulting from any additional front-end sampling stages that were applied. The calculation of such weighting factors was similar to that of the first stage weight, since geographical regions were also sampled PPS. The resulting first stage weight in such cases was simply the product of the “region” weight and the first stage weight, as described earlier.

In some countries, schools were selected with equal probabilities. This generally occurred when no reliable measure of school size was available. In some countries also, explicit strata were defined to deal with very large schools or with small schools, and equal probability sampling was necessary in these strata.

Under equal probability sampling, the basic first stage weight for the i^{th} sampled school was defined as

$$BW_{sc}^{i} = \frac{N}{n}$$

where n was the number of sampled schools and N was the total number of schools in the explicit stratum. The basic weight for all sampled schools in an explicit stratum was identical in this context.

11.2.2 School Non-Participation Adjustment

First stage weights were calculated for all sampled schools and replacement schools that participated. A school-level participation adjustment was required to compensate for schools that were sampled but did not participate and were not replaced. Sampled schools that were found to be ineligible[4] were removed from the calculation of this adjustment. The school-level participation adjustment was calculated separately for each explicit stratum.

The adjustment was calculated as follows:

$$A_{sc} = \frac{n_s + n_{r1} + n_{r2} + n_{nr}}{n_s + n_{r1} + n_{r2}}$$

[4] A sampled school was ineligible if it was found to contain no eligible (i.e., eighth-grade students). Such schools usually were in the sampling frame by mistake, and included schools that had recently closed, or amalgamated with another school.

where n_s was the number of originally sampled schools that participated, n_{r1} and n_{r2} the number of first and second replacement schools, respectively, that participated, and n_{nr} the number of schools that did not participate.

The final first stage weight for the ith school, corrected for nonparticipating schools, thus became:

$$FW_{sc}^{i} = A_{sc} \cdot BW_{sc}^{i}$$

11.2.3 The Second Stage (Classroom) Weight

The second stage weight represented the inverse of the second stage selection probability assigned to a sampled classroom. Although almost all TIMSS 1999 participants sampled intact classrooms using equal probability sampling, it also was permissible to subsample students within classes using PPS techniques. Procedures for calculating sampling weights are presented below for both approaches.

Equal Probability Weighting: For the ith school, let C^i be the total number of classrooms and c^i the number of sampled classrooms. Using equal probability sampling, the final second stage weight assigned to all sampled classrooms in the ith school was

$$FW_{cl1}^{i} = \frac{C^i}{c^i}$$

As a rule, c^i took the values 1 or 2 and remained fixed for all sampled schools. In those cases where c^i took the value 2 and only one of the sampled classrooms participated, the second stage weight was adjusted by multiplying it by 2.

Probability Proportional to Size Weighting: For the ith school, let $k^{i,j}$ be the size of the jth classroom. Using PPS sampling, the final second stage weight assigned to the jth sampled classroom in the ith school was

$$FW_{cl2}^{i,j} = \frac{K^i}{c^i \cdot k^{i,j}}$$

where c^i was the number of sampled classrooms in the ith school, as defined earlier, and

$$K^i = \sum_{j=1}^{c^i} k^{i,j}$$

Again, usually c^i took the values 1 or 2 and remained fixed for all sampled schools. In those cases where c^i took the value 2 and only one of the sampled classrooms participated, the second stage weight was adjusted by multiplying it by 2.

11.2.4 The Third Stage (Student) Weight

The third stage weight represented the inverse of the third stage selection probability attached to a sampled student. Although almost all participants sampled intact classrooms where all eligible students were to be tested, some countries with large classes took a subsample of students from within the sampled classes. Procedures for calculating weights are presented below for both sampling approaches. The third stage weight was calculated independently for each sampled classroom.

Sampling Intact Classrooms: If intact classrooms were sampled, then the basic third stage weight for the j^th^ classroom in the i^th^ school was simply

$$BW_{st1}^{i,j} = 1.0$$

Although in the standard TIMSS data collection each student was assigned one of 8 achievement test booklets[5], countries were permitted to add a further national booklet as required. Where a country chose to add a national booklet, the basic third stage weight was adjusted to reflect the change in the fraction of students responding to each booklet. The basic third stage weight thus became

$$BW_{st1}^{i,j} = \frac{k_{TIMSS\ 1999}^{i,j} + k_{natl}^{i,j}}{k_{TIMSS\ 1999}^{i,j}}$$

where

$k_{TIMSS\ 1999}^{i,j}$ = number of students assigned a TIMSS 1999 booklet in the jth classroom of the ith school,

$k_{natl}^{i,j}$ = number of students assigned a national booklet in the jth classroom of the ith school, and

[5] See Chapter 2 for a description of the TIMSS test design.

$$k^{i,j}_{TIMSS\ 1999} + k^{i,j}_{natl} + k^{i,j}_{ex} = k^{i,j}$$

where $k^{i,j}_{ex}$ was the number of excluded students[6] that were not assigned any booklet. Note that this number could be zero if there were no excluded students in the classroom.

Subsampling Students: If subsampling of students occurred within sampled classrooms, then the basic third stage weight for the jth classroom of the ith school was

$$BW^{i,j}_{st2} = \frac{k^{i,j}}{s^{i,j}}$$

where $k^{i,j}$ was the size of the jth classroom in the ith school, as defined earlier, and $s^{i,j}$ was the number of sampled students per sampled classroom. The latter number usually remained constant for all sampled classrooms.

When a country added a national booklet to the set of TIMSS 1999 booklets, the basic third stage weight was adjusted to reflect this. The basic third stage weight thus became

$$BW^{i,j}_{st2} = \frac{k^{i,j}}{s^{i,j}} \cdot \frac{s^{i,j}_{TIMSS\ 1999} + s^{i,j}_{natl}}{s^{i,j}_{TIMSS\ 1999}}$$

where

$s^{i,j}_{TIMSS\ 1999}$ = number of sub-sampled students assigned a TIMSS 1999 booklet in the jth classroom of the ith school,

$s^{i,j}_{natl}$ = number of sub-sampled students assigned a national booklet in the jth classroom of the ith school, and

$$s^{i,j}_{TIMSS\ 1999} + s^{i,j}_{natl} + s^{i,j}_{ex} = s^{i,j}$$

where $s^{i,j}_{ex}$ was the number of excluded students that were not assigned any type of booklet. Again, this number could be zero if there were no excluded students in the classroom sub-sample.

[6] Criteria for excluding students from the data collection are presented in Chapter 2.

11.2.5 Adjustment for Student Non-Participation

The student non-participation adjustment was calculated separately for each participating classroom as follows:

$$A_{st}^{i,j} = \frac{s_{rs}^{i,j} + s_{nr}^{i,j}}{s_{rs}^{i,j}}$$

This adjustment is the inverse of the unweighted student participation rate, R_{st}, computed for the corresponding classroom:

$$A_{st}^{i,j} = \frac{1}{R_{st}^{i,j}}$$

The third and final stage weight for the jth classroom in the ith school thus became

$$FW_{st1}^{i,j} = A_{st}^{i,j} \cdot BW_{st1}^{i,j}$$

when intact classrooms were sampled, or

$$FW_{st2}^{i,j} = A_{st}^{i,j} \cdot BW_{st2}^{i,j}$$

when sub-sampling of students within sampled classrooms occurred.

11.2.6 Overall Sampling Weights

The overall sampling weight was simply the product of the final first stage weight, the final second stage weight, and the final third stage weight. When intact classrooms were tested the overall sampling weight was

$$W^{i,j} = A_{sc}^{i,j} \cdot BW_{sc}^{i} \cdot FW_{cl1}^{i,j} \cdot A_{st}^{i,j} \cdot BW_{st1}^{i,j}$$

or

$$W^{i,j} = FW_{sc}^{i} \cdot FW_{cl1}^{i,j} \cdot FW_{st1}^{i,j}$$

When students were subsampled within classrooms, the overall sampling weight was

$$W^{i,j} = A_{sc}^{i,j} \cdot BW_{sc}^{i} \cdot FW_{cl2}^{i,j} \cdot A_{st}^{i,j} \cdot BW_{st2}^{i,j}$$

or

$$W^{i,j} = FW_{sc}^{i} \cdot FW_{cl2}^{i,j} \cdot FW_{st2}^{i,j}$$

It is important to note that sampling weights vary by school and classroom, but that students within the same classroom have the same sampling weights.

11.3 Participation Rates

Since lack of participation by sampled schools or students can lead to bias in the results, a variety of participation rates were computed to reveal how successful countries had been in securing participation from their sampled schools. To monitor school participation, three school participation rates were computed: using originally sampled schools only; using sampled and first replacement schools; and using sampled and both first and second replacement schools. Student participation rates were also computed, as were overall participation rates.

11.3.1 Unweighted School Participation Rates

The three unweighted school participation rates that were computed were the following:

R_{unw}^{sc-s} = unweighted school participation rate for originally-sampled schools only,

R_{unw}^{sc-r1} = unweighted school participation rate, including sampled and first replacement schools,

R_{unw}^{sc-r2} = unweighted school participation rate, including sampled, first and second replacement schools.

Each unweighted school participation rate was defined as the ratio of the number of participating schools to the number of originally-sampled schools, excluding any ineligible schools. The rates were calculated as follows:

$$R_{unw}^{sc-s} = \frac{n_s}{n_s + n_{r1} + n_{r2} + n_{nr}}$$

$$R^{sc-r1}_{unw} = \frac{n_s + n_{r1}}{n_s + n_{r1} + n_{r2} + n_{nr}}$$

$$R^{sc-s}_{unw} = \frac{n_s + n_{r1} + n_{r2}}{n_s + n_{r1} + n_{r2} + n_{nr}}$$

11.3.2 Unweighted Student Participation Rate

The unweighted student participation rate was computed as follows:

$$R^{st}_{unw} = \frac{\sum_{i,j} s^{i,j}_{rs}}{\sum_{i,j} s^{i,j}_{rs} + \sum_{i,j} s^{i,j}_{nr}}$$

11.3.3 Unweighted Overall Participation Rates

Three unweighted overall participation rates were computed for each country. They were as follows:

R^{ov-s}_{unw} = unweighted overall participation rate for originally sampled schools only,

R^{ov-r1}_{unw} = unweighted overall participation rate, including sampled and first replacement schools,

R^{ov-r2}_{unw} = unweighted overall participation rate, including sampled, and first and second replacement schools.

For each country, the overall participation rate was defined as the product of the unweighted school participation rate and the unweighted student participation rate. They were calculated as follows:

$$R^{ov-s}_{unw} = R^{sc-s}_{unw} \cdot R^{st}_{unw}$$

$$R^{ov-r1}_{unw} = R^{sc-r1}_{unw} \cdot R^{st}_{unw}$$

$$R^{ov-r2}_{unw} = R^{sc-s}_{unw} \cdot R^{st}_{unw}$$

11.3.4 Weighted School Participation Rates

In TIMSS 1995, the weighted school-level participation rates were computed using school sampling frame information. However, TIMSS 1999 used student-level information instead. The alternate method has two advantages:

- Weighted school participation rates can be easily replicated by all data users since all the required data are available from the international database
- These rates more accurately reflect the current size of the target population since they rely on up-to-date within-school sampling information.

The 1995 method relied on school data as reported on the sampling frame, which often were not up to date with regard to current school enrollment. Conceptually, however, both methods are equivalent when assuming an up-to-date sampling frame, and should yield comparable results in practice.

Three weighted school-level participation rates were computed using the alternate method. They were as follows:

R_{wtd}^{sc-s} = weighted school participation rate for originally-sampled schools only,

R_{wtd}^{sc-r1} = weighted school participation rate, including sampled and first replacement schools,

R_{wtd}^{sc-r2} = weighted school participation rate, including sampled, first and second replacement schools.

The weighted school participation rates were calculated as follows:

$$R_{wtd}^{sc-s} = \frac{\sum_{i,j}^{s} BW_{sc}^{i} \cdot FW_{clx}^{i,j} \cdot FW_{stx}^{i,j}}{\sum_{i,j}^{s+r1+r2} FW_{sc}^{i} \cdot FW_{clx}^{i,j} \cdot FW_{stx}^{i,j}}$$

$$R_{wtd}^{sc-r1} = \frac{\sum_{i,j}^{s+r1} BW_{sc}^{i} \cdot FW_{clx}^{i,j} \cdot FW_{stx}^{i,j}}{\sum_{i,j}^{s+r1+r2} FW_{sc}^{i} \cdot FW_{clx}^{i,j} \cdot FW_{stx}^{i,j}}$$

$$R_{wtd}^{sc-r2} = \frac{\sum_{i,j}^{s+r1+r2} BW_{sc}^{i} \cdot FW_{clx}^{i,j} \cdot FW_{stx}^{i,j}}{\sum_{i,j}^{s+r1+r2} FW_{sc}^{i} \cdot FW_{clx}^{i,j} \cdot FW_{stx}^{i,j}}$$

where both the numerator and denominator were summations over all responding students and the appropriate classroom-level and student-level sampling weights were used. Note that the basic school-level weight appears in the numerator, whereas the final school-level weight appears in the denominator.

The denominator remains unchanged in all three equations and is the weighted estimate of the total enrollment in the target population. The numerator, however, changes from one equation to the next. Only students from originally sampled schools were included in the first equation; students from first replacement schools were added in the second equation; and students from first and second replacement schools were added in the third equation.

11.3.5 Weighted Student Participation Rates

The weighted student response rate was computed as follows:

$$R_{wtd}^{st} = \frac{\sum_{i,j}^{s+r1+r2} BW_{sc}^{i} \cdot FW_{clx}^{i,j} \cdot BW_{stx}^{i,j}}{\sum_{i,j}^{s+r1+r2} BW_{sc}^{i} \cdot FW_{clx}^{i,j} \cdot FW_{stx}^{i,j}}$$

where both the numerator and denominator were summations over all responding students and the appropriate classroom-level and student-level sampling weights were used. Note that the basic student weight appears in the numerator, whereas the final student weight appears in the denominator. Furthermore, the denominator in this formula was the same quantity that appears in the numerator of the weighted school-level participation rate for all participating schools, sampled and replacement.

11.3.6 Weighted Overall Participation Rates

Three weighted overall participation rates were computed. They were as follows:

R_{wtd}^{ov-s} = weighted overall participation rate for originally-sampled schools only,

R_{wtd}^{ov-r1} = weighted overall participation rate, including sampled and first replacement schools,

R_{wtd}^{ov-r2} = weighted overall participation rate, including sampled, first and second replacement schools.

Each weighted overall participation rate was defined as the product of the appropriate weighted school participation rate and the weighted student participation rate. They were computed as follows:

$$R_{wtd}^{ov-s} = R_{wtd}^{sc-s} \cdot R_{wtd}^{st}$$

$$R_{wtd}^{ov-r1} = R_{wtd}^{sc-r1} \cdot R_{wtd}^{st}$$

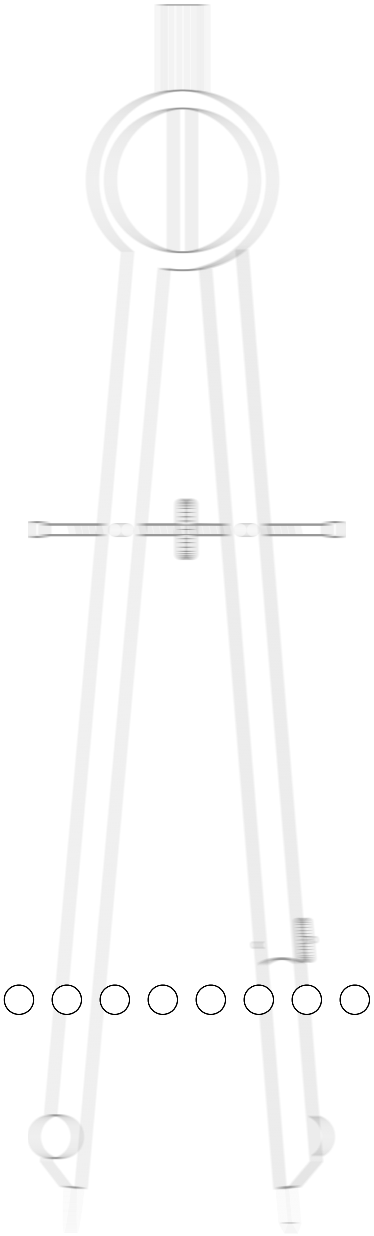
$$R_{wtd}^{ov-r2} = R_{wtd}^{sc-r2} \cdot R_{wtd}^{st}$$

11.4 Summary

The multi-stage nature of the TIMSS sampling design means that student have been sampled with varying probabilities. Consequently, if statistics computed from the sample data are to accurately reflect population values, the TIMSS sampling weights must be used when analyzing the data.

Estimation of Sampling Variance

Eugenio J. Gonzalez
Pierre Foy



12 Estimation of Sampling Variance

Eugenio J. Gonzalez
Pierre Foy

12.1 Overview

To obtain estimates of student proficiency in mathematics and science that were both accurate and cost-effective, TIMSS 1999 made extensive use of probability sampling techniques to sample students from national student populations.[1] Statistics computed from these national probability samples were used as estimates of population parameters. Because some uncertainty is involved in generalizing from samples to populations, the important statistics in the TIMSS 1999 international reports (Mullis et al., 2000; Martin et al., 2000) are presented together with their standard errors, which are a measure of this uncertainty.

The TIMSS 1999 item pool was far too extensive to be administered in its entirety to any one student, and so a complex test design was developed whereby each student was given a single test booklet containing only a part of the entire assessment.[2] The results for all of the booklets were then aggregated using item response theory to provide results for the entire assessment. Thus each student responded to just a few items from each content area, and therefore multiple imputation or "plausible values" had to be used to derive reliable indicators of student proficiency. Since every proficiency estimate incorporates some uncertainty, TIMSS followed the customary procedure of generating five estimates for each student and using the variability among them as a measure of this imputation uncertainty, or error. In the TIMSS 1999 international report the imputation error for each variable has been combined with the sampling error for that variable to provide a standard error incorporating both.

12.2 Estimating Sampling Variance

The TIMSS 1999 sampling design applied a stratified multistage cluster-sampling technique to the problem of selecting efficient and accurate samples of students while working with schools and classes. This design capitalized on the structure of the student population (i.e., students grouped in classes within schools) to

[1] The TIMSS sample design is presented in Chapter 2.

[2] Details of the TIMSS test design may be found in Chapter 3.

derive student samples that permitted efficient and economical data collection. Unfortunately, however, such a complex sampling design complicated the task of computing standard errors to quantify sampling variability.

When, as in TIMSS, the sampling design involves multistage cluster sampling, there are several options for estimating sampling errors that avoid the assumption of simple random sampling (Wolter, 1985). The jackknife repeated replication technique (JRR) was chosen by TIMSS in both 1995 and 1999 because it is computationally straightforward and provides approximately unbiased estimates of the sampling errors of means, totals, and percentages.

The variation on the JRR technique used in TIMSS 1999 is described in Johnson and Rust (1992). It assumes that the primary sampling units (PSUs) can be paired in a manner consistent with the sample design, with each pair regarded as members of a pseudo-stratum for variance estimation purposes. When used in this way, the JRR technique appropriately accounts for the combined effect of the between- and within-PSU contributions to the sampling variance. The general use of JRR entails systematically assigning pairs of schools to sampling zones, and randomly selecting one of these schools to have its contribution doubled and the other to have it zeroed, so as to construct a number of "pseudo-replicates" of the original sample. The statistic of interest is computed once for all of the original sample, and once again for each pseudo-replicate sample. The variation between the estimates for each of the replicate samples and the original sample estimate is the jackknife estimate of the sampling error of the statistic.

12.2.1 Construction of Sampling Zones for Sampling Variance Estimation

To apply the JRR technique used in TIMSS 1999 the sampled schools had to be paired and assigned to a series of groups known as sampling zones. This was done at Statistics Canada by working through the list of sampled schools in the order in which they were selected and assigning the first and second schools to the first sampling zone, the third and fourth schools to the second zone, and so on. In total 75 zones were used, allowing for 150 schools per country. When more than 75 zones were constructed, they were collapsed to keep the total number to 75.

Sampling zones were constructed within design domains, or explicit strata. Where there was an odd number of schools in an explicit stratum, either by design or because of school nonresponse, the students in the remaining school were randomly divided to make up two “quasi” schools for the purposes of calculating the jackknife standard error. Each zone then consisted of a pair of schools or “quasi” schools. Exhibit 12.1 shows the range of sampling zones used in each country.

Exhibit 12.1 Range of Sampling Zones Used in Each Country

Country	Zones	Country	Zones
Australia	75	Latvia (LSS)	73
Belgium (Flemish)	74	Lithuania	75
Bulgaria	75	Macedonia, Rep. of	75
Canada	75	Malaysia	75
Chile	75	Moldova	75
Chinese Taipei	75	Morocco	75
Cyprus	61	Netherlands	63
Czech Republic	71	New Zealand	75
England	64	Philippines	75
Finland	75	Romania	74
Hong Kong, SAR	69	Russian Federation	56
Hungary	74	Singapore	73
Indonesia	75	Slovak Republic	73
Iran, Islamic Rep.	75	Slovenia	75
Israel	70	South Africa	75
Italy	75	Thailand	75
Japan	71	Tunisia	75
Jordan	74	Turkey	62
Korea, Rep. of	75	United States	53

12.2.2 Computing Sampling Variance Using the JRR Method

The JRR algorithm used in TIMSS 1999 assumes that there are H sampling zones within each country, each containing two sampled schools selected independently. To compute a statistic t from the sample for a country, the formula for the JRR variance estimate of the statistic t is then given by the following equation:

$$Var_{jrr}(t) = \sum_{h=1}^{H} [t(J_h) - t(S)]^2$$

where H is the number of pairs in the sample for the country. The term $t(S)$ corresponds to the statistic for the whole sample (computed with any specific weights that may have been used to compensate for the unequal probability of selection of the different elements in the sample or any other post-stratification weight). The element $t(J_h)$ denotes the same statistic using the h^{th} jackknife replicate. This is computed using all cases except those in the h^{th} zone of the sample; for those in the h^{th} zone, all cases associated with one of the randomly selected units of the pair are removed, and the elements associated with the other unit in the zone are included twice. In practice, this is effectively accomplished by recoding to zero the weights for the cases of the element of the pair to be excluded from the replication, and multiplying by two the weights of the remaining element within the h^{th} pair.

The computation of the JRR variance estimate for any statistic in TIMSS 1999 required the computation of the statistic up to 76 times for any given country: once to obtain the statistic for the full sample, and up to 75 times to obtain the statistics for each of the jackknife replicates (J_h). The number of times a statistic needed to be computed for a given country depended on the number of implicit strata or sampling zones defined for that country.

Doubling and zeroing the weights of the selected units within the sampling zones was accomplished effectively by creating replicate weights that were then used in the calculations. This approach requires the user to temporarily create a new set of weights for each pseudo-replicate sample. Each replicate weight is equal to k times the overall sampling weight, where k can take values of 0, 1, or 2 depending on whether the case is to be removed from the computation, left as it is, or have its weight doubled. The value of k for an individual student record for a given replicate depends on the assignment of the record to the specific PSU and zone.

Within each zone the members of the pair of schools are assigned an indicator (u_i), coded randomly to 1 or 0 so that one of them has a value of 1 on the variable u_i, and the other a value of 0. This indicator determines whether the weights for the elements in the school in this zone are to be doubled or zeroed. The replicate weight ($W_h^{g,i,j}$) for the elements in a school assigned to zone h is computed as the product of k_h times their overall sampling weight, where k_h can take values of 0, 1, or 2 depending on whether the school is to be omitted, be included with its usual weight, or have its weight doubled for the computation of the statistic of interest. In TIMSS 1999, the replicate weights were not permanent variables, but were created temporarily by the sampling variance estimation program as a useful computing device.

To create replicate weights, each sampled student was first assigned a vector of 75 weights, $W_h^{g,i,j}$, where h takes values from 1 to 75. The value of $W_0^{g,i,j}$, is the overall sampling weigh*t*, which is simply the product of the final school weight, the appropriate final classroom weight, and the appropriate final student weight, as described in Chapter 11.

The replicate weights for a single case were then computed as

$$W_h^{g,i,j} = W_0^{g,i,j} \cdot k_{hi}$$

where the variable k_h for an individual i takes the value $k_{hi} = 2 * u_i$ if the record belongs to zone h, and $k_{hi} = 1$ otherwise.

In the TIMSS 1999 analysis, 75 replicate weights were computed for each country regardless of the number of actual zones within the country. If a country had fewer than 75 zones, then the replicate weights W_h, where h was greater than the number of zones within the country, were each the same as the overall sampling weight. Although this involved some redundant computation, having 75 replicate weights for each country had no effect on the size of the error variance computed using the jackknife formula, but it facilitated the computation of standard errors for a number of countries at a time.

Although standard errors presented in the international reports were computed using SAS programs developed at the International Study Center, they were also verified against results produced by the WesVarPC software (Westat, 1997) as an additional quality control check.

12.3 Estimating Imputation Variance

The general procedure for estimating the imputation variance using plausible values is the following (Mislevy et al., 1992). First compute the statistic (t), for each set of plausible values (M). The statistics t_m can be anything estimable from the data, such as a mean, the difference between means, percentiles, and so forth. Each of these statistics will be called t_m, where m = 1, 2, …, 5.

Once the statistics are computed, the imputation variance is then computed as:

$$Var_{imp} = \left(1 + \frac{1}{M}\right)Var(t_m)$$

where M is the number of plausible values used in the calculation, and $Var(t_m)$ is the variance of the estimates computed using each plausible value.

12.4 Combining Sampling and Imputation Variance

When reporting standard errors for proficiency estimates using plausible values, it was necessary to combine the sampling and imputation components of the error variance for the estimate. Under ideal circumstances and with unlimited computing resources, the user would compute the imputation variance for the plausible values and the JRR sampling variance for each of the plausible values. This would be equivalent to computing the same statistic up to 380 times (once overall for each of the five plausible values using the overall sampling weights, and then 75 times more for each plausible value using the complete set of replicate weights). An acceptable shortcut, however, is to compute the JRR variance component using one plausible value, and then the imputation variance using the five plausible values. Using this approach, the same statistic needed to be computed only 80 times. With this procedure the error variance component for a statistic was computed using the following formula: $Var \cdot (t_{pv}) = Var_{jrr}(t_1) + Var_{imp}$

where $Var_{jrr}(t_1)$ is the sampling variance for the first plausible value. The User Guide for the TIMSS 1999 International Database will contain programs in SAS and SPSS that compute each of these variance components for the TIMSS 1999 data.

Exhibits 12.2 through 12.14 show basic summary statistics for mathematics and its five content areas: algebra; data representation, analysis and probability; fractions and number sense; geometry; and measurement, and for science and its six content areas: chemistry; earth science; environment and resource issues; life science; physics; and scientific inquiry and the nature of science. Each exhibit presents the student sample size, the mean and standard deviation, averaged across the five plausible values, the jackknife standard error for the mean, and the overall standard errors for the mean including imputation error.

Exhibit 12.2 Summary Statistics and Standard Errors for Mathematics Proficiency

Country	Sample Size	Mean Proficiency[a]	Standard Deviation[a]	Jackknife Sampling Error	Overall Standard Error[b]
Australia	4032	525	80	4.7	4.8
Belgium (Flemish)	5259	558	77	3.1	3.3
Bulgaria	3272	511	86	5.8	5.8
Canada	8770	531	73	2.2	2.5
Chile	5907	392	85	4.1	4.4
Chinese Taipei	5772	585	104	3.9	4.0
Cyprus	3116	476	82	1.6	1.8
Czech Republic	3453	520	79	4.1	4.2
England	2960	496	83	4.1	4.1
Finland	2920	520	65	2.6	2.7
Hong Kong, SAR	5179	582	73	4.2	4.3
Hungary	3183	532	85	3.6	3.7
Indonesia	5848	403	101	4.6	4.9
Iran, Islamic Rep.	5301	422	83	3.2	3.4
Israel	4195	466	96	3.9	3.9
Italy	3328	479	87	3.8	3.8
Japan	4745	579	80	1.5	1.7
Jordan	5052	428	103	3.4	3.6
Korea, Rep. of	6114	587	79	1.7	2.0
Latvia (LSS)	2873	505	78	3.3	3.4
Lithuania	2361	482	78	4.0	4.3
Macedonia, Rep. of	4023	447	93	4.2	4.2
Malaysia	5577	519	81	4.3	4.4
Moldova	3711	469	85	3.8	3.9
Morocco	5402	337	91	1.8	2.6
Netherlands	2962	540	73	6.9	7.1
New Zealand	3613	491	89	5.1	5.2
Philippines	6601	345	97	5.5	6.0
Romania	3425	472	93	5.6	5.8
Russian Federation	4332	526	86	5.9	5.9
Singapore	4966	604	79	6.1	6.3
Slovak Republic	3497	534	75	3.9	4.0
Slovenia	3109	530	83	2.7	2.8
South Africa	8146	275	109	5.8	6.8
Thailand	5732	467	85	4.8	5.1
Tunisia	5051	448	64	2.1	2.4
Turkey	7841	429	86	4.0	4.3
United States	9072	502	88	3.9	4.0

[a] Average across the five plausible values.

[b] Includes error due to sampling and imputation.

Exhibit 12.3 Summary Statistics and Standard Errors for Geometry Proficiency

Country	Sample Size	Mean Proficiency[a]	Standard Deviation[a]	Jackknife Sampling Error	Overall Standard Error[b]
Australia	4032	497	91	3.5	5.7
Belgium (Flemish)	5259	535	101	3.1	4.1
Bulgaria	3272	524	107	4.8	5.9
Canada	8770	507	89	1.5	4.7
Chile	5907	412	102	3.3	5.4
Chinese Taipei	5772	557	104	3.2	5.8
Cyprus	3116	484	90	2.0	4.6
Czech Republic	3453	513	107	3.8	5.5
England	2960	471	86	3.0	4.2
Finland	2920	494	100	3.3	6.0
Hong Kong, SAR	5179	556	88	3.3	4.9
Hungary	3183	489	108	3.5	4.3
Indonesia	5848	441	103	3.7	5.1
Iran, Islamic Rep.	5301	447	93	2.7	2.9
Israel	4195	462	102	4.1	5.4
Italy	3328	482	96	3.0	5.6
Japan	4745	575	98	2.5	5.1
Jordan	5052	449	101	2.6	7.1
Korea, Rep. of	6114	573	98	2.0	3.9
Latvia (LSS)	2873	522	94	2.5	5.6
Lithuania	2361	496	95	3.7	5.8
Macedonia, Rep. of	4023	460	114	3.5	6.1
Malaysia	5577	497	93	3.7	4.4
Moldova	3711	481	112	3.6	5.0
Morocco	5402	407	113	1.9	2.2
Netherlands	2962	515	92	4.9	5.5
New Zealand	3613	478	86	3.6	4.2
Philippines	6601	383	93	3.0	3.4
Romania	3425	487	111	3.9	6.4
Russian Federation	4332	522	113	4.7	6.0
Singapore	4966	560	93	4.9	6.7
Slovak Republic	3497	527	91	3.5	7.3
Slovenia	3109	506	111	3.1	6.2
South Africa	8146	335	106	3.8	6.6
Thailand	5732	484	90	2.8	4.4
Tunisia	5051	484	83	1.7	4.4
Turkey	7841	428	101	4.3	5.7
United States	9072	473	90	2.3	4.4

a. Average across the five plausible values.

b. Includes error due to sampling and imputation.

Exhibit 12.4 Summary Statistics and Standard Errors for Data Representation, Analysis and Probability Proficiency

Country	Sample Size	Mean Proficiency[a]	Standard Deviation[a]	Jackknife Sampling Error	Overall Standard Error[b]
Australia	4032	522	97	4.5	6.3
Belgium (Flemish)	5259	544	103	3.7	3.8
Bulgaria	3272	493	112	5.3	6.1
Canada	8770	521	93	2.5	4.5
Chile	5907	429	90	3.0	3.8
Chinese Taipei	5772	559	108	3.2	5.1
Cyprus	3116	472	94	1.5	4.6
Czech Republic	3453	513	107	3.8	5.9
England	2960	506	94	4.3	8.0
Finland	2920	525	105	2.9	3.8
Hong Kong, SAR	5179	547	89	3.7	5.4
Hungary	3183	520	118	3.9	5.9
Indonesia	5848	423	93	3.1	4.4
Iran, Islamic Rep.	5301	430	89	2.9	6.0
Israel	4195	468	102	3.9	5.1
Italy	3328	484	101	3.8	4.5
Japan	4745	555	89	2.0	2.3
Jordan	5052	436	98	2.5	7.8
Korea, Rep. of	6114	576	98	1.7	4.2
Latvia (LSS)	2873	495	104	3.2	4.8
Lithuania	2361	493	88	3.2	3.6
Macedonia, Rep. of	4023	442	111	3.7	6.2
Malaysia	5577	491	86	3.2	4.0
Moldova	3711	450	104	3.1	5.7
Morocco	5402	383	101	1.8	3.5
Netherlands	2962	538	98	7.1	7.9
New Zealand	3613	497	97	4.5	5.0
Philippines	6601	406	82	2.5	3.5
Romania	3425	453	110	3.8	4.7
Russian Federation	4332	501	110	4.5	4.8
Singapore	4966	562	94	5.6	6.2
Slovak Republic	3497	521	101	4.0	4.6
Slovenia	3109	530	114	2.8	4.2
South Africa	8146	356	94	3.3	3.8
Thailand	5732	476	91	3.6	4.0
Tunisia	5051	446	79	1.6	5.1
Turkey	7841	446	87	2.9	3.3
United States	9072	506	102	3.7	5.2

a. Average across the five plausible values.

b. Includes error due to sampling and imputation.

Exhibit 12.5 Summary Statistics and Standard Errors for Measurement Proficiency

Country	Sample Size	Mean Proficiency[a]	Standard Deviation[a]	Jackknife Sampling Error	Overall Standard Error[b]
Australia	4032	529	84	3.8	4.9
Belgium (Flemish)	5259	549	77	2.9	4.0
Bulgaria	3272	497	96	5.4	6.6
Canada	8770	521	80	2.0	2.4
Chile	5907	412	92	3.3	4.9
Chinese Taipei	5772	566	96	3.1	3.4
Cyprus	3116	471	93	2.2	4.0
Czech Republic	3453	535	83	3.3	5.0
England	2960	507	84	3.7	3.8
Finland	2920	521	74	2.6	4.7
Hong Kong, SAR	5179	567	79	4.0	5.8
Hungary	3183	538	84	2.6	3.5
Indonesia	5848	395	117	4.4	5.1
Iran, Islamic Rep.	5301	401	100	3.5	4.7
Israel	4195	457	97	3.9	5.1
Italy	3328	501	89	3.4	5.0
Japan	4745	558	75	1.7	2.4
Jordan	5052	438	106	3.2	4.4
Korea, Rep. of	6114	571	79	1.9	2.8
Latvia (LSS)	2873	505	89	3.1	3.5
Lithuania	2361	467	81	3.1	4.0
Macedonia, Rep. of	4023	451	101	3.4	5.2
Malaysia	5577	514	86	4.1	4.6
Moldova	3711	479	97	3.5	4.9
Morocco	5402	348	115	2.2	3.5
Netherlands	2962	538	73	5.4	5.8
New Zealand	3613	496	86	4.4	5.3
Philippines	6601	355	104	4.2	6.2
Romania	3425	491	99	4.4	4.9
Russian Federation	4332	527	94	5.5	6.0
Singapore	4966	599	87	5.6	6.3
Slovak Republic	3497	537	77	3.0	3.3
Slovenia	3109	523	94	2.7	3.7
South Africa	8146	329	108	3.7	4.8
Thailand	5732	463	92	4.4	6.2
Tunisia	5051	442	81	2.3	3.1
Turkey	7841	436	93	4.5	6.5
United States	9072	482	92	3.5	3.9

a. Average across the five plausible values.

b. Includes error due to sampling and imputation.

Exhibit 12.6 Summary Statistics and Standard Errors for Algebra Proficiency

Country	Sample Size	Mean Proficiency[a]	Standard Deviation[a]	Jackknife Sampling Error	Overall Standard Error[b]
Australia	4032	520	81	4.1	5.1
Belgium (Flemish)	5259	540	86	3.2	4.6
Bulgaria	3272	512	88	4.8	5.1
Canada	8770	525	73	1.7	2.4
Chile	5907	399	96	3.9	4.3
Chinese Taipei	5772	586	114	4.3	4.4
Cyprus	3116	479	80	1.5	1.6
Czech Republic	3453	514	87	3.8	4.0
England	2960	498	77	3.3	4.9
Finland	2920	498	73	2.3	3.1
Hong Kong, SAR	5179	569	78	3.6	4.5
Hungary	3183	536	94	3.4	4.1
Indonesia	5848	424	104	3.9	5.7
Iran, Islamic Rep.	5301	434	88	2.8	4.9
Israel	4195	479	97	4.1	4.5
Italy	3328	481	84	3.3	3.6
Japan	4745	569	82	1.5	3.3
Jordan	5052	439	108	3.6	5.3
Korea, Rep. of	6114	585	90	1.9	2.7
Latvia (LSS)	2873	499	83	3.0	4.3
Lithuania	2361	487	74	3.4	3.7
Macedonia, Rep. of	4023	465	100	3.8	4.0
Malaysia	5577	505	81	3.8	4.8
Moldova	3711	477	91	3.2	3.7
Morocco	5402	353	111	2.2	4.7
Netherlands	2962	522	77	6.9	7.7
New Zealand	3613	497	81	4.3	4.7
Philippines	6601	345	119	5.2	5.8
Romania	3425	481	99	5.0	5.2
Russian Federation	4332	529	95	4.8	4.9
Singapore	4966	576	81	5.9	6.2
Slovak Republic	3497	525	76	3.6	4.6
Slovenia	3109	525	85	2.7	2.9
South Africa	8146	293	125	6.1	7.7
Thailand	5732	456	91	4.2	4.9
Tunisia	5051	455	74	1.9	2.7
Turkey	7841	432	98	4.3	4.6
United States	9072	506	90	3.4	4.1

a. Average across the five plausible values.

b. Includes error due to sampling and imputation.

Exhibit 12.7 **Summary Statistics and Standard Errors for Fractions and Number Sense Proficiency**

Country	Sample Size	Mean Proficiency[a]	Standard Deviation[a]	Jackknife Sampling Error	Overall Standard Error[b]
Australia	4032	519	78	4.1	4.3
Belgium (Flemish)	5259	557	74	2.8	3.1
Bulgaria	3272	503	97	6.3	6.6
Canada	8770	533	74	1.9	2.5
Chile	5907	403	88	3.6	4.9
Chinese Taipei	5772	576	101	3.8	4.2
Cyprus	3116	481	82	2.0	3.0
Czech Republic	3453	507	90	4.0	4.8
England	2960	497	82	3.7	3.8
Finland	2920	531	75	3.1	3.8
Hong Kong, SAR	5179	579	75	4.0	4.5
Hungary	3183	526	95	3.8	4.2
Indonesia	5848	406	99	3.9	4.1
Iran, Islamic Rep.	5301	437	82	2.8	4.5
Israel	4195	472	93	4.0	4.4
Italy	3328	471	88	3.6	5.0
Japan	4745	570	84	1.6	2.6
Jordan	5052	432	101	2.9	3.2
Korea, Rep. of	6114	570	78	1.9	2.7
Latvia (LSS)	2873	496	89	3.6	3.7
Lithuania	2361	479	84	4.0	4.3
Macedonia, Rep. of	4023	437	100	4.1	4.7
Malaysia	5577	532	83	4.2	4.7
Moldova	3711	465	92	3.7	4.2
Morocco	5402	335	113	1.8	3.6
Netherlands	2962	545	79	6.7	7.1
New Zealand	3613	493	88	4.5	5.0
Philippines	6601	378	97	4.7	6.3
Romania	3425	458	100	5.3	5.7
Russian Federation	4332	513	98	6.1	6.4
Singapore	4966	608	82	5.4	5.6
Slovak Republic	3497	525	81	4.6	4.8
Slovenia	3109	527	90	3.1	3.7
South Africa	8146	300	115	5.2	6.0
Thailand	5732	471	90	4.4	5.3
Tunisia	5051	443	79	2.2	2.8
Turkey	7841	430	88	3.6	4.3
United States	9072	509	88	3.8	4.2

a. Average across the five plausible values.

b. Includes error due to sampling and imputation.

Exhibit 12.8 Summary Statistics and Standard Errors for Science Proficiency

Country	Sample Size	Mean of 5 Plausible Values	S.D.[a]	Error Due to Sampling	S.E.[b]
Australia	4032	540	87	4.3	4.4
Belgium (Flemish)	5259	535	69	2.6	3.1
Bulgaria	3272	518	93	5.3	5.4
Canada	8770	533	78	1.8	2.1
Chile	5907	420	88	3.7	3.7
Chinese Taipei	5772	569	89	3.6	4.4
Cyprus	3116	460	84	1.8	2.4
Czech Republic	3453	539	80	3.7	4.2
England	2960	538	91	4.3	4.8
Finland	2920	535	78	3.0	3.5
Hong Kong, SAR	5179	530	70	3.5	3.7
Hungary	3183	552	84	3.4	3.7
Indonesia	5848	435	84	4.1	4.5
Iran, Islamic Rep.	5301	448	84	3.7	3.8
Israel	4195	468	105	4.4	4.9
Italy	3328	493	87	3.5	3.9
Japan	4745	550	76	1.9	2.2
Jordan	5052	450	103	3.4	3.8
Korea, Rep. of	6114	549	85	1.9	2.6
Latvia (LSS)	2873	503	78	3.1	4.8
Lithuania	2361	488	83	3.8	4.1
Macedonia, Rep. of	4023	458	97	4.3	5.2
Malaysia	5577	492	82	4.2	4.4
Moldova	3711	459	95	3.9	4.0
Morocco	5402	323	102	2.9	4.3
Netherlands	2962	545	77	6.7	6.9
New Zealand	3613	510	93	4.6	4.9
Philippines	6601	345	121	7.2	7.5
Romania	3425	472	97	5.0	5.8
Russian Federation	4332	529	93	6.1	6.4
Singapore	4966	568	97	8.0	8.0
Slovak Republic	3497	535	78	3.0	3.3
Slovenia	3109	533	84	2.9	3.2
South Africa	8146	243	132	7.4	7.8
Thailand	5732	482	73	3.9	4.0
Tunisia	5051	430	67	2.0	3.4
Turkey	7841	433	80	3.5	4.3
United States	9072	515	97	4.4	4.6

a. Standard deviation of the five plausible values

b. Standard error due to imputation

Exhibit 12.9 Summary Statistics and Standard Errors for Life Science Proficiency Sample

Country	Sample Size	Mean of 5 Plausible Values	S.D.[a]	Error Due to Sampling	S.E.[b]
Australia	4032	530	96	4.0	4.4
Belgium (Flemish)	5259	535	89	2.8	4.6
Bulgaria	3272	514	107	5.4	6.9
Canada	8770	523	87	2.1	3.8
Chile	5907	431	88	3.0	3.7
Chinese Taipei	5772	550	96	2.8	3.3
Cyprus	3116	468	94	2.1	3.8
Czech Republic	3453	544	99	3.7	4.1
England	2960	533	97	4.3	6.2
Finland	2920	520	94	2.5	4.0
Hong Kong, SAR	5179	516	84	3.1	5.5
Hungary	3183	535	99	3.3	4.0
Indonesia	5848	448	85	3.1	3.6
Iran, Islamic Rep.	5301	437	92	2.7	3.7
Israel	4195	463	103	3.8	4.0
Italy	3328	488	94	3.3	4.6
Japan	4745	534	90	2.1	5.4
Jordan	5052	448	103	3.3	4.1
Korea, Rep. of	6114	528	93	2.0	3.6
Latvia (LSS)	2873	509	90	3.1	3.9
Lithuania	2361	494	87	3.5	4.6
Macedonia, Rep. of	4023	468	113	4.0	4.9
Malaysia	5577	479	94	4.1	5.4
Moldova	3711	477	109	3.7	3.9
Morocco	5402	347	108	1.9	2.8
Netherlands	2962	536	94	6.0	7.2
New Zealand	3613	501	98	4.5	5.6
Philippines	6601	378	110	5.6	5.7
Romania	3425	475	109	4.7	6.0
Russian Federation	4332	517	114	5.7	6.5
Singapore	4966	541	102	7.1	7.2
Slovak Republic	3497	535	93	3.6	6.2
Slovenia	3109	521	103	2.8	3.9
South Africa	8146	289	123	6.2	7.3
Thailand	5732	508	77	2.7	4.5
Tunisia	5051	441	76	1.7	5.0
Turkey	7841	444	85	3.7	4.5
United States	9072	520	104	3.7	4.1

a. Standard deviation of the five plausible values

b. Standard error due to imputation

Exhibit 12.10 Summary Statistics and Standard Errors for Earth Science Proficiency

Country	Sample Size	Mean Proficiency[a]	Standard Deviation[a]	Jackknife Sampling Error	Overall Standard Error[b]
Australia	4032	519	96	3.9	6.1
Belgium (Flemish)	5259	533	92	2.8	3.5
Bulgaria	3272	520	115	5.4	5.7
Canada	8770	519	92	1.7	3.7
Chile	5907	435	93	3.0	7.0
Chinese Taipei	5772	538	89	2.0	3.0
Cyprus	3116	459	87	1.8	5.4
Czech Republic	3453	533	113	4.7	6.9
England	2960	525	88	3.6	3.9
Finland	2920	520	101	3.0	5.5
Hong Kong, SAR	5179	506	82	2.5	4.3
Hungary	3183	560	119	3.8	3.9
Indonesia	5848	431	99	3.7	6.4
Iran, Islamic Rep.	5301	459	96	2.8	5.2
Israel	4195	472	108	4.4	5.2
Italy	3328	502	103	3.6	5.9
Japan	4745	533	91	2.2	6.2
Jordan	5052	446	92	2.4	3.5
Korea, Rep. of	6114	532	98	2.1	2.7
Latvia (LSS)	2873	495	114	3.8	5.4
Lithuania	2361	476	91	3.2	4.4
Macedonia, Rep. of	4023	464	116	3.9	4.2
Malaysia	5577	491	90	3.4	4.2
Moldova	3711	466	117	3.0	4.2
Morocco	5402	363	112	2.0	3.3
Netherlands	2962	534	94	6.0	7.2
New Zealand	3613	504	90	3.7	5.8
Philippines	6601	390	103	4.9	5.0
Romania	3425	475	128	4.5	5.5
Russian Federation	4332	529	124	4.5	5.1
Singapore	4966	521	91	5.4	7.3
Slovak Republic	3497	537	99	4.0	4.3
Slovenia	3109	541	111	3.6	4.3
South Africa	8146	348	102	3.6	4.8
Thailand	5732	470	95	3.4	3.9
Tunisia	5051	442	89	1.6	2.7
Turkey	7841	435	90	3.6	4.6
United States	9072	504	98	3.4	4.2

a. Average across the five plausible values.

b. Includes error due to sampling and imputation.

Exhibit 12.11 Summary Statistics and Standard Errors for Physics Proficiency

Country	Sample Size	Mean Proficiency[a]	Standard Deviation[a]	Jackknife Sampling Error	Overall Standard Error[b]
Australia	4032	531	90	3.6	6.3
Belgium (Flemish)	5259	530	82	2.0	3.5
Bulgaria	3272	505	109	4.8	5.8
Canada	8770	521	85	2.3	3.8
Chile	5907	428	93	2.6	5.6
Chinese Taipei	5772	552	96	3.0	3.9
Cyprus	3116	459	95	2.0	2.9
Czech Republic	3453	526	99	3.6	4.2
England	2960	528	86	3.7	4.5
Finland	2920	520	103	2.6	4.4
Hong Kong, SAR	5179	523	88	3.4	4.9
Hungary	3183	543	102	3.0	4.3
Indonesia	5848	452	94	3.2	5.5
Iran, Islamic Rep.	5301	445	105	4.0	5.7
Israel	4195	484	102	3.9	5.3
Italy	3328	480	93	3.5	4.1
Japan	4745	544	83	1.7	2.9
Jordan	5052	459	108	3.1	3.6
Korea, Rep. of	6114	544	92	2.3	5.1
Latvia (LSS)	2873	495	95	3.1	3.9
Lithuania	2361	510	85	3.5	4.3
Macedonia, Rep. of	4023	463	107	3.8	6.0
Malaysia	5577	494	89	3.2	4.1
Moldova	3711	457	112	3.9	5.5
Morocco	5402	352	120	2.2	4.2
Netherlands	2962	537	91	6.5	6.5
New Zealand	3613	499	93	3.7	4.7
Philippines	6601	393	107	5.1	6.3
Romania	3425	465	110	4.4	6.8
Russian Federation	4332	529	115	5.9	6.3
Singapore	4966	570	96	6.4	6.7
Slovak Republic	3497	518	91	3.5	4.1
Slovenia	3109	525	102	3.4	4.4
South Africa	8146	308	122	5.9	6.7
Thailand	5732	475	90	4.0	4.2
Tunisia	5051	425	87	2.2	6.3
Turkey	7841	441	93	3.9	4.0
United States	9072	498	97	3.7	5.5

a. Average across the five plausible values.

b. Includes error due to sampling and imputation.

Exhibit 12.12 Summary Statistics and Standard Errors for Chemistry Proficiency

Country	Sample Size	Mean Proficiency[a]	Standard Deviation[a]	Jackknife Sampling Error	Overall Standard Error[b]
Australia	4032	520	101	4.2	5.0
Belgium (Flemish)	5259	508	92	2.4	3.3
Bulgaria	3272	527	115	4.5	5.7
Canada	8770	521	94	2.0	5.4
Chile	5907	435	97	3.2	5.2
Chinese Taipei	5772	563	105	3.0	4.3
Cyprus	3116	470	91	1.7	3.4
Czech Republic	3453	512	108	3.5	5.2
England	2960	524	95	3.8	5.5
Finland	2920	535	101	3.0	4.5
Hong Kong, SAR	5179	515	87	2.6	5.2
Hungary	3183	548	111	3.1	4.7
Indonesia	5848	425	88	3.5	3.9
Iran, Islamic Rep.	5301	487	92	2.4	4.1
Israel	4195	479	107	3.8	4.7
Italy	3328	493	94	3.2	4.8
Japan	4745	530	87	1.8	3.1
Jordan	5052	483	112	3.0	5.5
Korea, Rep. of	6114	523	102	2.8	3.7
Latvia (LSS)	2873	490	104	2.9	3.7
Lithuania	2361	485	95	3.8	4.6
Macedonia, Rep. of	4023	481	113	3.7	6.1
Malaysia	5577	485	91	2.9	3.5
Moldova	3711	451	117	3.7	5.6
Morocco	5402	372	107	1.7	4.8
Netherlands	2962	515	95	5.2	6.4
New Zealand	3613	503	96	3.8	4.9
Philippines	6601	394	100	4.2	6.5
Romania	3425	481	115	4.1	6.1
Russian Federation	4332	523	120	6.8	8.0
Singapore	4966	545	116	7.9	8.3
Slovak Republic	3497	525	101	3.4	4.9
Slovenia	3109	509	112	2.5	5.4
South Africa	8146	350	105	3.1	4.0
Thailand	5732	439	97	4.0	4.3
Tunisia	5051	439	83	1.7	3.7
Turkey	7841	437	98	3.1	5.0
United States	9072	508	110	4.0	4.8

a. Average across the five plausible values.

b. Includes error due to sampling and imputation.

Exhibit 12.13 Summary Statistics and Standard Errors for Scientific Inquiry and the Nature of Science Proficiency

Country	Sample Size	Mean Proficiency[a]	Standard Deviation[a]	Jackknife Sampling Error	Overall Standard Error[b]
Australia	4032	535	93	3.5	4.9
Belgium (Flemish)	5259	526	93	2.7	4.9
Bulgaria	3272	479	121	5.4	5.6
Canada	8770	532	86	1.2	5.1
Chile	5907	441	100	3.3	4.7
Chinese Taipei	5772	540	87	3.0	4.9
Cyprus	3116	467	104	2.1	4.6
Czech Republic	3453	522	108	4.8	5.7
England	2960	538	86	3.2	5.1
Finland	2920	528	101	2.6	4.0
Hong Kong, SAR	5179	531	82	2.3	2.8
Hungary	3183	526	103	2.9	5.9
Indonesia	5848	446	99	2.7	4.3
Iran, Islamic Rep.	5301	446	94	2.3	5.3
Israel	4195	476	112	3.8	8.3
Italy	3328	489	96	2.9	4.6
Japan	4745	543	77	1.8	2.8
Jordan	5052	440	109	2.6	5.5
Korea, Rep. of	6114	545	89	2.1	7.3
Latvia (LSS)	2873	495	104	3.2	4.7
Lithuania	2361	483	99	4.0	6.4
Macedonia, Rep. of	4023	464	117	3.2	3.6
Malaysia	5577	488	84	2.5	4.5
Moldova	3711	471	113	3.3	3.8
Morocco	5402	391	134	2.7	4.2
Netherlands	2962	534	98	5.1	6.5
New Zealand	3613	521	95	3.3	6.8
Philippines	6601	403	108	3.7	5.5
Romania	3425	456	118	3.4	5.5
Russian Federation	4332	491	109	3.3	4.9
Singapore	4966	550	85	4.2	5.9
Slovak Republic	3497	507	85	2.7	3.9
Slovenia	3109	513	107	2.9	4.3
South Africa	8146	329	133	4.8	6.4
Thailand	5732	462	99	3.4	4.2
Tunisia	5051	451	95	2.1	3.4
Turkey	7841	445	104	4.0	6.3
United States	9072	522	92	2.6	4.3

a. Average across the five plausible values.

b. Includes error due to sampling and imputation.

Exhibit 12.14 **Summary Statistics and Standard Errors for Environment and Resources Issues Proficiency**

Country	Sample Size	Mean Proficiency[a]	Standard Deviation[a]	Jackknife Sampling Error	Overall Standard Error[b]
Australia	4032	530	104	3.9	6.3
Belgium (Flemish)	5259	513	98	2.3	3.5
Bulgaria	3272	483	126	5.5	6.4
Canada	8770	521	97	2.5	3.5
Chile	5907	449	97	2.6	4.8
Chinese Taipei	5772	567	101	2.4	4.0
Cyprus	3116	475	92	2.2	4.3
Czech Republic	3453	516	111	3.5	5.7
England	2960	518	108	4.1	5.8
Finland	2920	514	101	2.4	7.1
Hong Kong, SAR	5179	518	91	2.9	4.9
Hungary	3183	501	118	3.6	6.6
Indonesia	5848	489	84	2.2	4.8
Iran, Islamic Rep.	5301	470	86	2.6	5.5
Israel	4195	458	105	3.5	4.0
Italy	3328	491	93	2.5	5.4
Japan	4745	506	89	2.2	5.5
Jordan	5052	476	106	2.7	6.0
Korea, Rep. of	6114	523	96	1.5	4.5
Latvia (LSS)	2873	493	98	3.4	5.2
Lithuania	2361	458	98	3.4	5.1
Macedonia, Rep. of	4023	432	117	3.3	4.2
Malaysia	5577	502	89	3.1	4.4
Moldova	3711	444	127	3.5	6.2
Morocco	5402	396	116	3.1	5.1
Netherlands	2962	526	106	7.1	8.5
New Zealand	3613	503	99	4.4	5.2
Philippines	6601	391	114	5.8	7.6
Romania	3425	473	114	4.4	6.6
Russian Federation	4332	495	118	5.2	6.6
Singapore	4966	577	117	7.9	8.3
Slovak Republic	3497	512	94	2.8	4.5
Slovenia	3109	519	110	3.0	3.4
South Africa	8146	350	118	5.4	8.5
Thailand	5732	507	83	2.2	3.0
Tunisia	5051	462	84	1.7	5.0
Turkey	7841	461	88	2.7	3.6
United States	9072	509	107	3.6	6.4

a. Average across the five plausible values.

b. Includes error due to sampling and imputation.

References

Johnson, E.G., & Rust, K.F. (1992). Population references and variance estimation for NAEP data. *Journal of Educational Statistics, 17*, 175-190.

Martin, M.O., Mullis, I.V.S., Gonzalez, E. J., Gregory, K.D., Smith, T.A., Chrostowski, S.J., Garden, R.A., & O'Connor, K.M. (2000). *TIMSS 1999 International Science Report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade.* Chestnut Hill, MA: Boston College.

Mislevy, R.J., Beaton, A.E., Kaplan, B., & Sheenan, K.M. (1992). Estimating Population Characteristics from Sparse Matrix Samples of Item Responses. *Journal of Educational Measurement, 29*, 133-161.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., Gregory, K.D., Garden, R.A., O'Connor, K.M., Chrostowski, S. J., & Smith, T.A. (2000). *TIMSS 1999 International Mathematics Report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade.* Chestnut Hill, MA: Boston College.

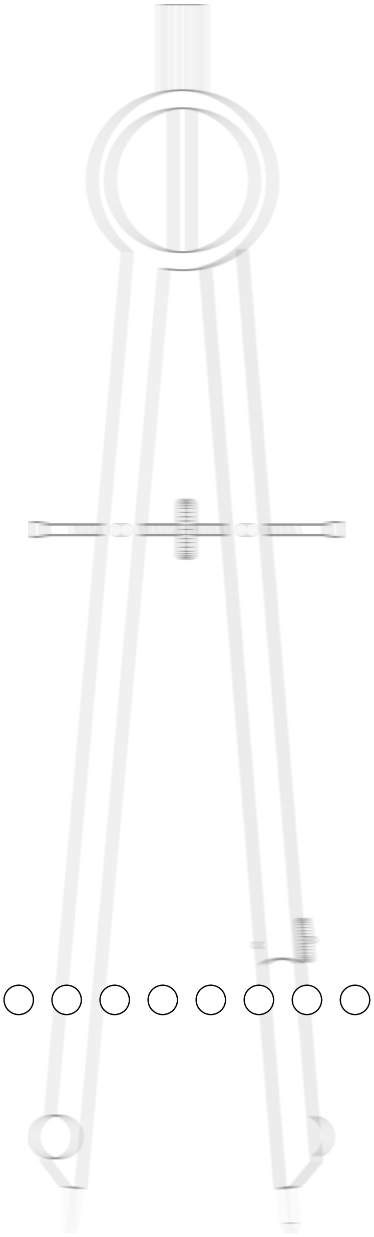
Westat, Inc. (1997). *A user's guide to WesVarPC.* Rockville, MD: Westat, Inc.

Wolter, K.M. (1985). *Introduction to variance estimation.* New York: Springer-Verlag.

13

Item Analysis and Review

Ina V.S. Mullis
Michael O. Martin



13 Item Analysis and Review

Ina V.S. Mullis
Michael O. Martin

13.1 Overview

In order to assess the psychometric properties of the TIMSS 1999 achievement items before proceeding with item response theory (IRT) scaling,[1] TIMSS computed a series of diagnostic statistics for each item in each country. As part of the TIMSS quality assurance process, these statistics were carefully checked for any evidence of unusual item behavior. If an item was uncharacteristically easy or difficult for a particular country, or had unusually low discriminating power, this sometimes suggested a translation or printing problem. On the relatively few occasions that such items were found, the test booklets were examined for flaws, and where necessary the national research coordinator was consulted. Any item that was discovered to have a flaw in a particular country was removed from the database for that country.

13.2 Statistics for Item Analysis

The basic statistics for the item review were calculated at the IEA Data Processing Center and summarized in graphical form for review at the International Study Center. Item statistics were computed for each of the 38 TIMSS 1999 countries, and where countries tested in more than one languages, for each of the languages tested. For each item, the basic item-analysis display presents the number of students that responded in each country, the difficulty level (the percentage of students that answered the item correctly), and the discrimination index (the point-biserial correlation between success on the item and a total score).[2] For multiple-choice items (see Exhibit 13.1 for an example), the display presents the percentage of students that chose each option, including the percentage that omitted or did not reach the item, and the point-biserial correlation between each option and the total score. For free-response items (which could have more than one score level – see Exhibit 13.2 for an example), the display

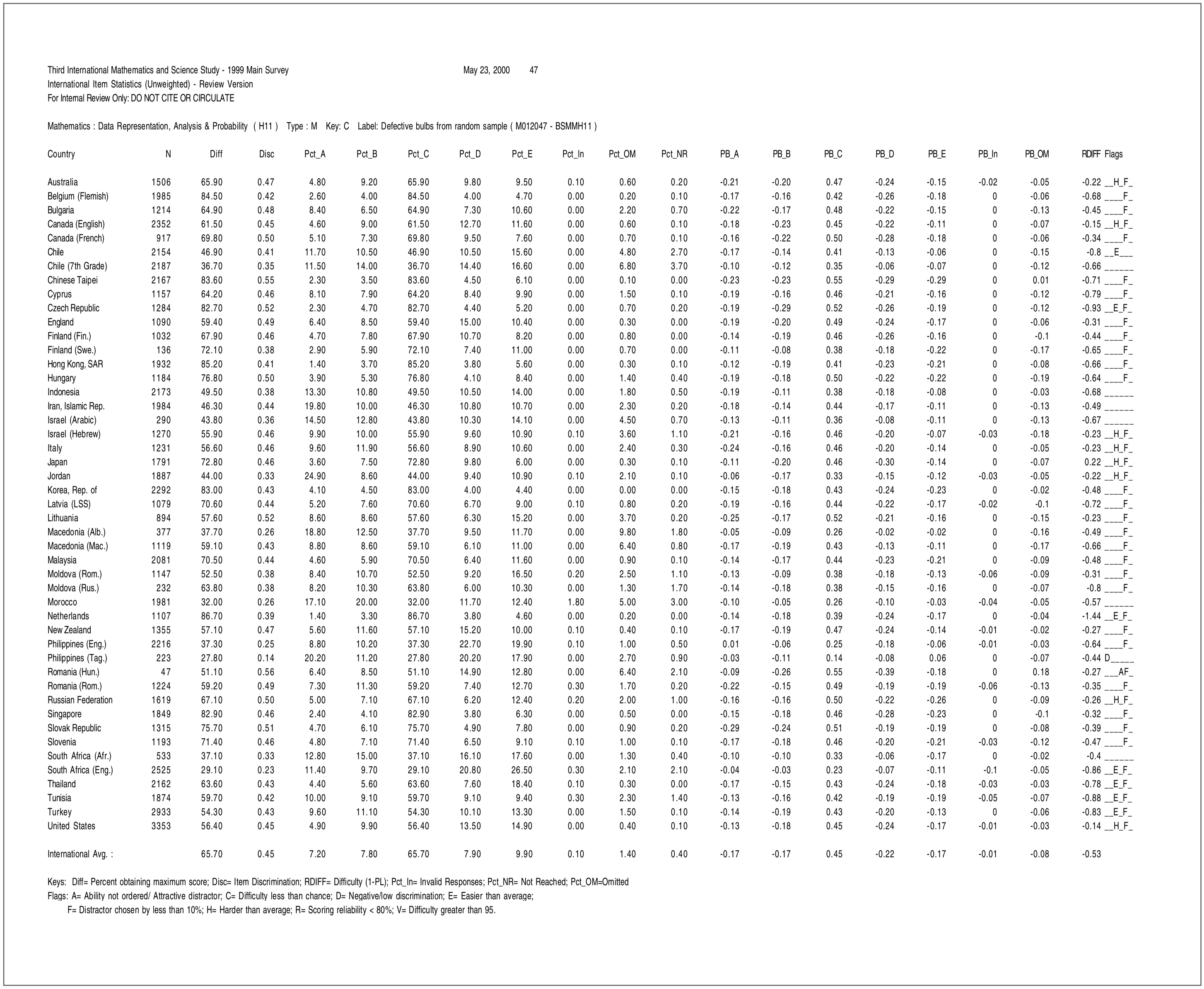
[1] The TIMSS 1999 IRT scaling is described in Chapter 14.

[2] For the purpose of computing the discrimination index, the total score was the percentage of items a student answered correctly.

presents the difficulty and discrimination of each score level. As a prelude to the main IRT scaling, it shows some statistics from a preliminary Rasch analysis, including the Rasch item difficulty for each item and the standard error of this difficulty estimate.

The item-analysis display presents the difficulty level of each item separately for male and female students. As a guide to the overall statistical properties of the item, it also shows the international item difficulty (the mean of the item difficulties across countries) and the international item discrimination (the mean of the item discriminations).

Exhibit 13.1 International Item Statistics for a Multiple-Choice Item



Third International Mathematics and Science Study - 1999 Main Survey May 23, 2000 47
International Item Statistics (Unweighted) - Review Version
For Internal Review Only: DO NOT CITE OR CIRCULATE

Mathematics : Data Representation, Analysis & Probability (H11) Type : M Key: C Label: Defective bulbs from random sample (M012047 - BSMMH11)

Country	N	Diff	Disc	Pct_A	Pct_B	Pct_C	Pct_D	Pct_E	Pct_In	Pct_OM	Pct_NR	PB_A	PB_B	PB_C	PB_D	PB_E	PB_In	PB_OM	RDIFF	Flags
Australia	1506	65.90	0.47	4.80	9.20	65.90	9.80	9.50	0.10	0.60	0.20	-0.21	-0.20	0.47	-0.24	-0.15	-0.02	-0.05	-0.22	__H_F_
Belgium (Flemish)	1985	84.50	0.42	2.60	4.00	84.50	4.00	4.70	0.00	0.20	0.10	-0.17	-0.16	0.42	-0.26	-0.18	0	-0.06	-0.68	____F_
Bulgaria	1214	64.90	0.48	8.40	6.50	64.90	7.30	10.60	0.00	2.20	0.70	-0.22	-0.17	0.48	-0.22	-0.15	0	-0.13	-0.45	____F_
Canada (English)	2352	61.50	0.45	4.60	9.00	61.50	12.70	11.60	0.00	0.60	0.10	-0.18	-0.23	0.45	-0.22	-0.11	0	-0.07	-0.15	__H_F_
Canada (French)	917	69.80	0.50	5.10	7.30	69.80	9.50	7.60	0.00	0.70	0.10	-0.16	-0.22	0.50	-0.28	-0.18	0	-0.06	-0.34	____F_
Chile	2154	46.90	0.41	11.70	10.50	46.90	10.50	15.60	0.00	4.80	2.70	-0.17	-0.14	0.41	-0.13	-0.06	0	-0.15	-0.8	__E___
Chile (7th Grade)	2187	36.70	0.35	11.50	14.00	36.70	14.40	16.60	0.00	6.80	3.70	-0.10	-0.12	0.35	-0.06	-0.07	0	-0.12	-0.66	______
Chinese Taipei	2167	83.60	0.55	2.30	3.50	83.60	4.50	6.10	0.00	0.10	0.00	-0.23	-0.23	0.55	-0.29	-0.29	0	0.01	-0.71	____F_
Cyprus	1157	64.20	0.46	8.10	7.90	64.20	8.40	9.90	0.00	1.50	0.10	-0.19	-0.16	0.46	-0.21	-0.16	0	-0.12	-0.79	____F_
Czech Republic	1284	82.70	0.52	2.30	4.70	82.70	4.40	5.20	0.00	0.70	0.20	-0.19	-0.29	0.52	-0.26	-0.19	0	-0.12	-0.93	__E_F_
England	1090	59.40	0.49	6.40	8.50	59.40	15.00	10.40	0.00	0.30	0.00	-0.19	-0.20	0.49	-0.24	-0.17	0	-0.06	-0.31	____F_
Finland (Fin.)	1032	67.90	0.46	4.70	7.80	67.90	10.70	8.20	0.00	0.80	0.00	-0.14	-0.19	0.46	-0.26	-0.16	0	-0.1	-0.44	____F_
Finland (Swe.)	136	72.10	0.38	2.90	5.90	72.10	7.40	11.00	0.00	0.70	0.00	-0.11	-0.08	0.38	-0.18	-0.22	0	-0.17	-0.65	____F_
Hong Kong, SAR	1932	85.20	0.41	1.40	3.70	85.20	3.80	5.60	0.00	0.30	0.10	-0.12	-0.19	0.41	-0.23	-0.21	0	-0.08	-0.66	____F_
Hungary	1184	76.80	0.50	3.90	5.30	76.80	4.10	8.40	0.00	1.40	0.40	-0.19	-0.18	0.50	-0.22	-0.22	0	-0.19	-0.64	____F_
Indonesia	2173	49.50	0.38	13.30	10.80	49.50	10.50	14.00	0.00	1.80	0.50	-0.19	-0.11	0.38	-0.18	-0.08	0	-0.03	-0.68	______
Iran, Islamic Rep.	1984	46.30	0.44	19.80	10.00	46.30	10.80	10.70	0.00	2.30	0.20	-0.18	-0.14	0.44	-0.17	-0.11	0	-0.13	-0.49	______
Israel (Arabic)	290	43.80	0.36	14.50	12.80	43.80	10.30	14.10	0.00	4.50	0.70	-0.13	-0.11	0.36	-0.08	-0.11	0	-0.13	-0.67	______
Israel (Hebrew)	1270	55.90	0.46	9.90	10.00	55.90	9.60	10.90	0.10	3.60	1.10	-0.21	-0.16	0.46	-0.20	-0.07	-0.03	-0.18	-0.23	__H_F_
Italy	1231	56.60	0.46	9.60	11.90	56.60	8.90	10.60	0.00	2.40	0.30	-0.24	-0.16	0.46	-0.20	-0.14	0	-0.05	-0.23	__H_F_
Japan	1791	72.80	0.46	3.60	7.50	72.80	9.80	6.00	0.00	0.30	0.10	-0.11	-0.20	0.46	-0.30	-0.14	0	-0.07	0.22	__H_F_
Jordan	1887	44.00	0.33	24.90	8.60	44.00	9.40	10.90	0.10	2.10	0.10	-0.06	-0.17	0.33	-0.15	-0.12	-0.03	-0.05	-0.22	__H_F_
Korea, Rep. of	2292	83.00	0.43	4.10	4.50	83.00	4.00	4.40	0.00	0.00	0.00	-0.15	-0.18	0.43	-0.24	-0.23	0	-0.02	-0.48	____F_
Latvia (LSS)	1079	70.60	0.44	5.20	7.60	70.60	6.70	9.00	0.10	0.80	0.20	-0.19	-0.16	0.44	-0.22	-0.17	-0.02	-0.1	-0.72	____F_
Lithuania	894	57.60	0.52	8.60	8.60	57.60	6.30	15.20	0.00	3.70	0.20	-0.25	-0.17	0.52	-0.21	-0.16	0	-0.15	-0.23	____F_
Macedonia (Alb.)	377	37.70	0.26	18.80	12.50	37.70	9.50	11.70	0.00	9.80	1.80	-0.05	-0.09	0.26	-0.02	-0.02	0	-0.16	-0.49	____F_
Macedonia (Mac.)	1119	59.10	0.43	8.80	8.60	59.10	6.10	11.00	0.00	6.40	0.80	-0.17	-0.19	0.43	-0.13	-0.11	0	-0.17	-0.66	____F_
Malaysia	2081	70.50	0.44	4.60	5.90	70.50	6.40	11.60	0.00	0.90	0.10	-0.14	-0.17	0.44	-0.23	-0.21	0	-0.09	-0.48	____F_
Moldova (Rom.)	1147	52.50	0.38	8.40	10.70	52.50	9.20	16.50	0.20	2.50	1.10	-0.13	-0.09	0.38	-0.18	-0.13	-0.06	-0.09	-0.31	____F_
Moldova (Rus.)	232	63.80	0.38	8.20	10.30	63.80	6.00	10.30	0.00	1.30	1.70	-0.14	-0.18	0.38	-0.15	-0.16	0	-0.07	-0.8	____F_
Morocco	1981	32.00	0.26	17.10	20.00	32.00	11.70	12.40	1.80	5.00	3.00	-0.10	-0.05	0.26	-0.10	-0.03	-0.04	-0.05	-0.57	______
Netherlands	1107	86.70	0.39	1.40	3.30	86.70	3.80	4.60	0.00	0.20	0.00	-0.14	-0.18	0.39	-0.24	-0.17	0	-0.04	-1.44	__E_F_
New Zealand	1355	57.10	0.47	5.60	11.60	57.10	15.20	10.00	0.10	0.40	0.10	-0.17	-0.19	0.47	-0.24	-0.14	-0.01	-0.02	-0.27	____F_
Philippines (Eng.)	2216	37.30	0.25	8.80	10.20	37.30	22.70	19.90	0.10	1.00	0.50	0.01	-0.06	0.25	-0.18	-0.06	-0.01	-0.03	-0.64	____F_
Philippines (Tag.)	223	27.80	0.14	20.20	11.20	27.80	20.20	17.90	0.00	2.70	0.90	-0.03	-0.11	0.14	-0.08	0.06	0	-0.07	-0.44	D_____
Romania (Hun.)	47	51.10	0.56	6.40	8.50	51.10	14.90	12.80	0.00	6.40	2.10	-0.09	-0.26	0.55	-0.39	-0.18	0	0.18	-0.27	___AF_
Romania (Rom.)	1224	59.20	0.49	7.30	11.30	59.20	7.40	12.70	0.30	1.70	0.20	-0.22	-0.15	0.49	-0.19	-0.19	-0.06	-0.13	-0.35	____F_
Russian Federation	1619	67.10	0.50	5.00	7.10	67.10	6.20	12.40	0.20	2.00	1.00	-0.16	-0.16	0.50	-0.22	-0.26	0	-0.09	-0.26	__H_F_
Singapore	1849	82.90	0.46	2.40	4.10	82.90	3.80	6.30	0.00	0.50	0.00	-0.15	-0.18	0.46	-0.28	-0.23	0	-0.1	-0.32	____F_
Slovak Republic	1315	75.70	0.51	4.70	6.10	75.70	4.90	7.80	0.00	0.90	0.20	-0.29	-0.24	0.51	-0.19	-0.19	0	-0.08	-0.39	____F_
Slovenia	1193	71.40	0.46	4.80	7.10	71.40	6.50	9.10	0.10	1.00	0.10	-0.17	-0.18	0.46	-0.20	-0.21	-0.03	-0.12	-0.47	____F_
South Africa (Afr.)	533	37.10	0.33	12.80	15.00	37.10	16.10	17.60	0.00	1.30	0.40	-0.10	-0.10	0.33	-0.06	-0.17	0	-0.02	-0.4	______
South Africa (Eng.)	2525	29.10	0.23	11.40	9.70	29.10	20.80	26.50	0.30	2.10	2.10	-0.04	-0.03	0.23	-0.07	-0.11	-0.1	-0.05	-0.86	__E_F_
Thailand	2162	63.60	0.43	4.40	5.60	63.60	7.60	18.40	0.10	0.30	0.00	-0.17	-0.15	0.43	-0.24	-0.18	-0.03	-0.03	-0.78	__E_F_
Tunisia	1874	59.70	0.42	10.00	9.10	59.70	9.10	9.40	0.30	2.30	1.40	-0.13	-0.16	0.42	-0.19	-0.19	-0.05	-0.07	-0.88	__E_F_
Turkey	2933	54.30	0.43	9.60	11.10	54.30	10.10	13.30	0.00	1.50	0.10	-0.14	-0.19	0.43	-0.20	-0.13	0	-0.06	-0.83	__E_F_
United States	3353	56.40	0.45	4.90	9.90	56.40	13.50	14.90	0.00	0.40	0.10	-0.13	-0.18	0.45	-0.24	-0.17	-0.01	-0.03	-0.14	__H_F_
International Avg. :		65.70	0.45	7.20	7.80	65.70	7.90	9.90	0.10	1.40	0.40	-0.17	-0.17	0.45	-0.22	-0.17	-0.01	-0.08	-0.53	

Keys: Diff= Percent obtaining maximum score; Disc= Item Discrimination; RDIFF= Difficulty (1-PL); Pct_In= Invalid Responses; Pct_NR= Not Reached; Pct_OM=Omitted
Flags: A= Ability not ordered/ Attractive distractor; C= Difficulty less than chance; D= Negative/low discrimination; E= Easier than average;
F= Distractor chosen by less than 10%; H= Harder than average; R= Scoring reliability < 80%; V= Difficulty greater than 95.

Exhibit 13.2 International Item Statistics for a Free-Response Item

Third International Mathematics and Science Study - 1999 Main Survey May 23, 2000 60
International Item Statistics (Unweighted) - Review Version
For Internal Review Only: DO NOT CITE OR CIRCULATE

Mathematics : Fractions and Number Sense (J12) Type : S Key: X Label: Division of fractions (M022026 - BSSMJ12)

																	Reliability		
Country	N	Diff	Disc	Pct_0	Pct_1	Pct_2	Pct_3	Pct_OM	Pct_NR	PB_0	PB_1	PB_2	PB_3	PB_OM	RDIFF	Cases	Score	Code	Flags
Australia	494	23.30	0.39	62.30	23.30 .	.		14.40	0.00	-0.22	0.39 .	.		-0.16	2.10	123.00	96.70	92.70	__H___
Belgium (Flem	656	67.20	0.40	27.70	67.20 .	.		5.00	0.00	-0.28	0.40 .	.		-0.29	0.35	90.00	100.00	98.90	______
Bulgaria	411	68.60	0.52	20.00	68.60 .	.		11.40	0.20	-0.29	0.52 .	.		-0.39	-0.60	88.00	100.00	97.70	__E___
Canada (Engli	783	32.30	0.42	57.00	32.30 .	.		10.70	0.10	-0.33	0.42 .	.		-0.10	1.36 .	.	.		__H___
Canada (Frenc	302	33.80	0.32	56.00	33.80 .	.		10.30	0.00	-0.26	0.32 .	.		-0.07	1.44 .	.	.		__H___
Chile	723	26.40	0.48	51.50	26.40 .	.		22.10	0.40	-0.20	0.48 .	.		-0.27	0.26 .	.	.		______
Chile (7th Gra	754	12.60	0.35	56.60	12.60 .	.		30.80	0.70	-0.03	0.35 .	.		-0.20	0.89 .	.	.		______
Chinese Taipe	725	83.70	0.64	13.40	83.70 .	.		2.90	0.00	-0.53	0.64 .	.		-0.33	-0.83	180.00	100.00	100.00	__E___
Cyprus	388	21.40	0.49	54.10	21.40 .	.		24.50	0.00	-0.14	0.49 .	.		-0.31	1.51 .	.	.		__H___
Czech Republi	406	67.70	0.52	27.30	67.70 .	.		4.90	0.00	-0.48	0.52 .	.		-0.14	0.06	121.00	99.20	94.20	______
England	373	4.60	0.26	82.30	4.60 .	.		13.10	0.00	-0.03	0.26 .	.		-0.13	4.12	89.00	100.00	93.30	__H_F_
Finland (Fin.)	344	16.00	0.31	70.90	16.00 .	.		13.10	0.30	-0.16	0.31 .	.		-0.09	2.42 .	.	.		__H___
Finland (Swe.)	44	4.50	0.12	61.40	4.50 .	.		34.10	0.00	0.05	0.12 .	.		-0.10	3.77 .	.	.		D_H_F_
Hong Kong, SA	648	81.60	0.39	14.50	81.60 .	.		3.90	0.20	-0.23	0.39 .	.		-0.34	-0.44	171.00	98.80	97.70	__E___
Hungary	400	68.00	0.46	25.00	68.00 .	.		7.00	0.00	-0.30	0.46 .	.		-0.33	-0.13	98.00	98.00	96.90	__E___
Indonesia	730	24.10	0.59	62.20	24.10 .	.		13.70	0.00	-0.43	0.59 .	.		-0.13	0.61	173.00	98.30	94.80	______
Iran, Islamic Re	668	35.30	0.38	56.70	35.30 .	.		7.90	0.00	-0.30	0.38 .	.		-0.13	0.03	170.00	92.90	73.50	______
Israel (Arabic)	99	16.20	0.50	64.60	16.20 .	.		19.20	0.00	-0.19	0.50 .	.		-0.24	1.09 .	.	.		______
Israel (Hebrev	409	24.00	0.39	46.50	24.00 .	.		29.60	0.70	-0.10	0.39 .	.		-0.25	1.36 .	.	.		__H___
Italy	419	62.10	0.52	24.60	62.10 .	.		13.40	0.20	-0.35	0.52 .	.		-0.29	-0.61	115.00	99.10	96.50	__E___
Japan	589	79.30	0.47	15.60	79.30 .	.		5.10	0.00	-0.34	0.47 .	.		-0.29	-0.32	140.00	100.00	99.30	__E___
Jordan	635	32.00	0.59	57.00	32.00 .	.		11.00	0.00	-0.41	0.59 .	.		-0.23	0.31	153.00	100.00	96.10	______
Korea, Rep. of	765	80.70	0.56	14.80	80.70 .	.		4.60	0.00	-0.40	0.56 .	.		-0.37	-0.28	123.00	97.60	95.10	__E___
Latvia (LSS)	355	67.60	0.44	22.50	67.60 .	.		9.90	0.60	-0.28	0.44 .	.		-0.29	-0.56	97.00	97.90	95.90	__E___
Lithuania	287	65.50	0.43	26.80	65.50 .	.		7.70	0.30	-0.28	0.43 .	.		-0.31	-0.57	68.00	98.50	98.50	__E___
Macedonia (A	119	8.40	0.10	43.70	8.40 .	.		47.90	0.80	0.22	0.10 .	.		-0.27	1.56 .	.	.		D___F_
Macedonia (M	375	28.00	0.45	44.50	28.00 .	.		27.50	0.50	-0.08	0.45 .	.		-0.36	0.93 .	.	.		______
Malaysia	699	54.80	0.61	36.80	54.80 .	.		8.40	0.10	-0.45	0.61 .	.		-0.31	0.37	191.00	100.00	99.00	______
Moldova (Rom	378	55.60	0.53	33.30	55.60 .	.		11.10	0.30	-0.38	0.53 .	.		-0.26	-0.34 .	.	.		__E___
Moldova (Rus.	70	64.30	0.51	28.60	64.30 .	.		7.10	1.40	-0.36	0.51 .	.		-0.34	-0.29 .	.	.		______
Morocco	659	4.60	0.09	61.90	4.60 .	.		33.50	0.90	0.06	0.09 .	.		-0.08	1.46	108.00	95.40	88.90	D_H_F_
Netherlands	370	13.20	0.35	78.40	13.20 .	.		8.40	0.30	-0.16	0.35 .	.		-0.16	3.21	38.00	100.00	92.10	__H___
New Zealand	456	9.90	0.31	73.50	9.90 .	.		16.70	0.00	-0.04	0.31 .	.		-0.20	2.90	114.00	99.10	95.60	__H_F_
Philippines (En	754	12.20	0.41	77.70	12.20 .	.		10.10	0.40	-0.24	0.41 .	.		-0.11	0.94 .	.	.		______
Philippines (Ta	79	2.50	0.26	86.10	2.50 .	.		11.40	0.00	0.01	0.26 .	.		-0.14	2.30 .	.	.		____F_
Romania (Hun	15	53.30	0.72	40.00	53.30 .	.		6.70	0.00	-0.63	0.69 .	.		-0.15	-0.78 .	.	.		______
Romania (Rom	418	69.40	0.58	21.80	69.40 .	.		8.90	0.50	-0.35	0.58 .	.		-0.41	-0.93 .	.	.		__E___
Russian Feder	527	73.60	0.50	19.70	73.60 .	.		6.60	0.00	-0.37	0.50 .	.		-0.29	-0.61	120.00	98.30	96.70	__E___
Singapore	625	83.70	0.48	13.90	83.70 .	.		2.40	0.00	-0.41	0.48 .	.		-0.23	-0.42	100.00	98.00	98.00	__E___
Slovak Republ	430	73.30	0.42	20.90	73.30 .	.		5.80	0.00	-0.32	0.42 .	.		-0.23	-0.15	101.00	99.00	99.00	__E___
Slovenia	386	74.10	0.43	22.30	74.10 .	.		3.60	0.00	-0.33	0.43 .	.		-0.26	-0.67	104.00	100.00	95.20	__E___
South Africa (	170	5.90	0.42	85.30	5.90 .	.		8.80	0.00	-0.22	0.42 .	.		-0.07	2.24 .	.	.		__H_F_
South Africa (	851	4.20	0.32	88.50	4.20 .	.		7.30	0.20	-0.19	0.32 .	.		0.00	1.45 .	.	.		__H_F_
Thailand	711	41.90	0.59	51.10	41.90 .	.		7.00	0.10	-0.52	0.59 .	.		-0.12	0.28	207.00	100.00	100.00	______
Tunisia	625	28.60	0.36	48.30	28.60 .	.		23.00	0.20	-0.21	0.36 .	.		-0.14	0.65	157.00	99.40	96.80	______
Turkey	984	38.60	0.56	48.20	38.60 .	.		13.20	0.20	-0.38	0.56 .	.		-0.26	-0.17	247.00	100.00	98.80	__E___
United States	1128	37.40	0.52	55.30	37.40 .	.		7.30	0.10	-0.41	0.52 .	.		-0.18	0.77	118.00	99.20	97.50	__H___
International Avg. :		49.80	0.46	39.60	49.80 .	.		10.60	0.10	-0.30	0.46 .	.		-0.24	0.42 .		98.80	95.70	

Keys: Diff = Percent obtaining maximum score; RDIFF= Difficulty (1-PL); Pct_In= Invalid Responses; Pct_NR= Not Reached; Pct_OM= Omitted
Flags: A= Ability not ordered/ Attractive distractor; C= Difficulty less than chance; D= Negative/low discrimination; E= Easier than average;
F= Distractor chosen by less than 10%; H= Harder than average; R= Scoring reliability < 80%; V= Difficulty greater than 95.

Exhibits 13.1 and 13.2 contained the statistics described below.

N: This is the number of students to whom the item was administered. If an item was not reached by a student it was considered to be not administered for the purpose of the item analysis.[3]

Diff: The item difficulty was the percentage of students that provided a fully correct response to the item. In the case of free-response items worth more than one point this was the percentage of students achieving the maximum score on the item. When computing this statistic, not reached items were treated as not administered.

Disc: The item discrimination was computed as the correlation between correctly answering the item and total score on all of the items in the subject area in the test booklet.[4] This correlation should be moderately positive for items with good measurement properties.

PCT_A, PCT_B, PCT_C, PCT_D and PCT_E: Used for multiple-choice items only (Exhibit 13.1), these represent the percentage of students choosing each response option for the item. Not reached items were excluded from the denominator for these calculations.

PCT_0, PCT_1, PCT_2 and PCT_3: Used for open-ended items only (Exhibit 13.2), these are the percentage of students scoring at each score level for the item. Not reached items were excluded from the denominator for these calculations.

PCT_IN: Used for multiple-choice items only, this was the percentage of students that provided an invalid response to a multiple-choice item. Invalid responses were generally the result of choosing more than one response option.

PCT_OM: This is the percentage of students that did not provide a response to the item even though the item was administered and they had reached it. Not reached items were excluded from the denominator when calculating this statistic.

○○○

[3] In TIMSS, for the purposes of item analysis and item parameter estimation in scaling, items not reached by a student were treated as if they had not been administered. For purposes of estimating student proficiency, however, not reached items were treated as incorrectly answered.

[4] For free-response items, the discrimination is the correlation between the number of score points and total score.

PCT_NR: This is the percentage of student that did not reach the item. An item was coded as not reached when there was no evidence of a response to any of the items following it in the booklet and the response to the item preceding it was omitted.

PB_A, PB_B, PB_C, PB_D and PB_E: Used for multiple-choice items only, these present the correlation between choosing each of the response options A, B, C, D, or E and the score on the test booklet. Items with good psychometric properties have zero or negative correlations for the distracter options (the incorrect options) and moderately positive correlations for the correct answer.

PB_0, PB_1, PB_2 and PB_3: Used for free-response items only, these present the correlation between the score levels on the item (0,1,2, or 3) and the score on the test booklet. For items with good measurement properties the correlation coefficients should change from negative to positive as the score on the item increases.

PB_OM: This is the correlation between a binary variable - indicating an omitted response to the item - and the score on the test booklet. This correlation should be negative or near zero.

PB_IN: Used for multiple-choice items only, this presents the correlation between an invalid response to the item (usually caused by selecting more than one response option) and the score on the test booklet. This correlation also should be negative or near zero.

RDIFF: This is an estimate of the difficulty item based on a Rasch one-parameter IRT model. The difficulty of the item is expressed in the logit metric (with a positive logit indicating a difficult item) and was scaled so that the average Rasch item difficulty was zero within each country.

Reliability - Cases: It was expected that the free-response items in approximately one-quarter of the test booklets would be scored by two independent scorers. This column indicates the number of times each item was double scored in a country.

Reliability - Score: This column contains the percentage of times the two independent scorers agreed on the score level for the item.

Reliability - Code: This column contains the percentage of times the two scorers agreed on the two-digit code (score and diagnostic code) for the item.

As an aid to reviewers, the item-analysis display includes a series of "flags" signaling the presence of one or more conditions that might indicate a problem with an item. The following conditions are flagged:

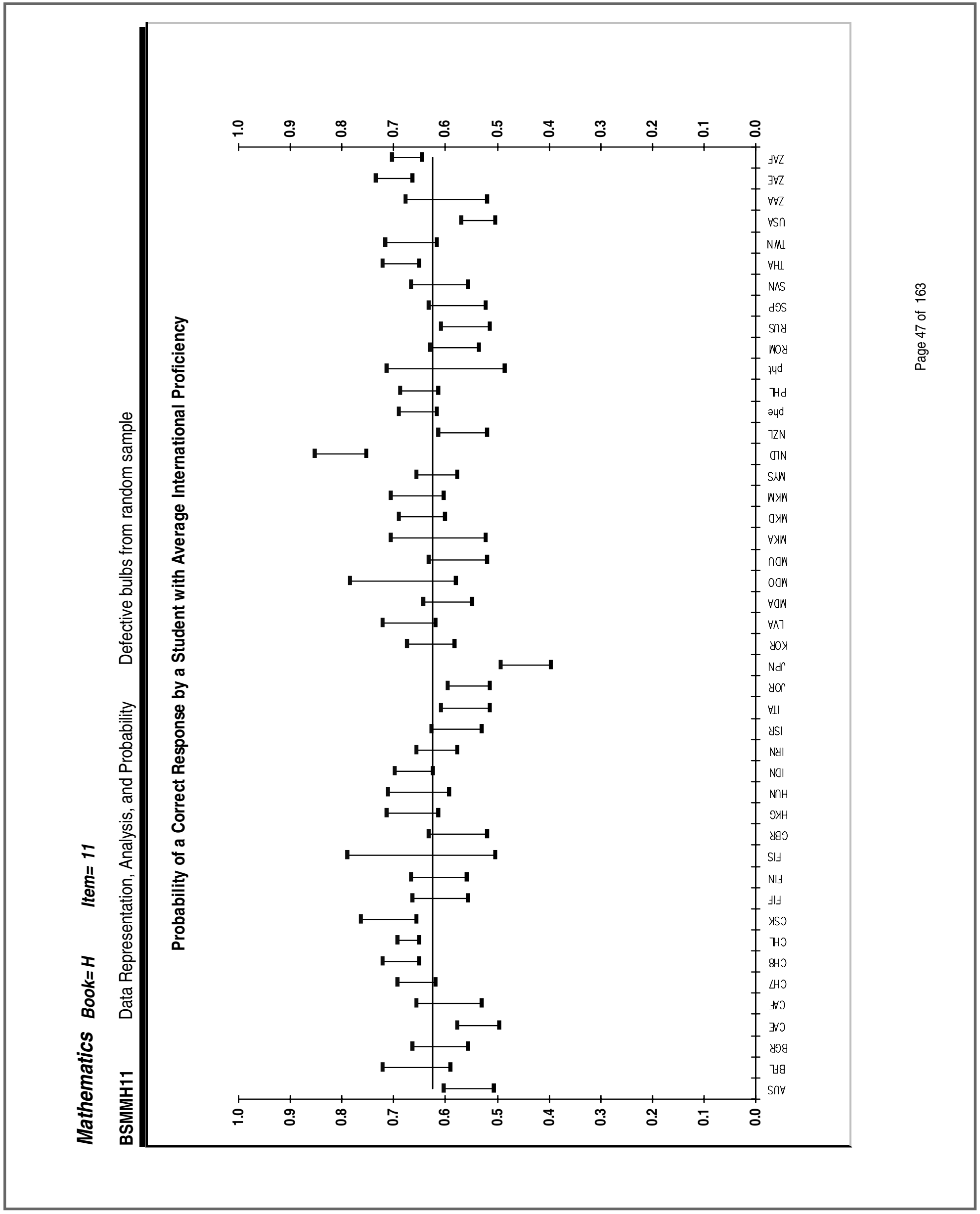
- Item difficulty exceeds 95% in the sample as a whole
- Item difficulty is less than 25% for 4-option multiple-choice items in the sample as a whole (20% for 5-option items)
- Item difficulty exceeds 95% or is less than 25% (20% for 5-option items)
- One or more of the distracter percentages is less than 5%
- One or more of the distracter percentages is greater than the percentage for the correct answer
- Point-biserial correlation for one or more of the distracters exceeds zero
- Item discrimination (i.e., the point-biserial for the correct answer) is less than 0.2
- Item discrimination does not increase with each score level (for an item with more than one score level)
- Rasch goodness-of-fit index is less than 0.88 or greater than 1.12
- Difficulty levels on the item differ significantly for males and females
- Difference in item difficulty levels between males and females diverge significantly from the average difference between males and females across all the items making up the total score

Although not all of these conditions necessarily indicate a problem, the flags are a useful way to draw attention to potential sources of concern. The IEA Data Processing Center also produced information about the inter-rater agreement for the free-response items.

13.2.1 Item-by-Country Interaction

Although there is room for variation across items, in general countries with high average performance on the achievement tests as a whole should perform relatively well on each of the items, and low-scoring countries should do less well on each of items. When this does not occur, i.e., when a high-scoring country has low performance on an item on which other countries are doing well, there is said to be an item-by-country interaction. Since large item-by-country interactions can indicate an item that is flawed in some way, the item review also included this aspect of item performance.

Exhibit 13.3 Example Item-by-Country Interaction Display



To help examine item-by-country interactions, the International Study Center produced a graphical display for each item showing the average probability across all countries of a correct response for a student of average international proficiency, compared with the probability of a correct response by a student of average proficiency in each country (see Exhibit 13.3 for an example). The probability for each country is presented as a 95% confidence interval, which includes a built-in Bonferroni correction for multiple comparisons.

The limits for the confidence interval are computed as follows:

$$UpperLimit = \left(1 - \frac{e^{RDIFF_{ik} + SE_{RDIFF_{ik}} \times Z_b}}{1 + e^{RDIFF_{ik} + SE_{RDIFF_{ik}} \times Z_b}} \right)$$

$$LowerLimit = \left(1 - \frac{e^{RDIFF_{ik} - SE_{RDIFF_{ik}} \times Z_b}}{1 + e^{RDIFF_{ik} - SE_{RDIFF_{ik}} \times Z_b}} \right)$$

where $RDIFF_{ik}$ is the Rasch difficulty of item k within country i; SE$_{RDIFFik}$ is the standard error of the difficulty of item k in country i; and $Z_b$is the critical value from the Z distribution, corrected for multiple comparisons using the Bonferroni procedure.

13.3 Item Checking Procedures

Prior to the IRT scaling of the TIMSS 1999 achievement data by Educational Testing Service, the International Study Center thoroughly reviewed the item statistics for all participating countries to ensure that items were performing comparably across countries. Although only a small number of items were found to be inappropriate for international comparisons, throughout the series of item-checking steps a number of reasons were discovered for differences in items across countries. Most of these were inadvertent changes in the items during printing, such as omitting an item option or misprinting the graphics associated with an item. Differences attributable to translation problems, however, were found for an item or two in several countries.

In particular, items with the following problems were considered for possible deletion from the international database:

- Errors were detected during translation verification but were not corrected before test administration

- Data cleaning revealed more or fewer options than in the original version of the item
- The item-analysis information showed the item to have a negative biserial
- The item-by-country interaction results showed a very large negative interaction for a given country
- The item-fit statistic indicated that the item did not fit the model
- For free-response items, the within-country scoring reliability data showed an agreement of less than 70% for the score level. Also, performance in items with more than one score level was not ordered by score, or correct levels were associated with negative point-biserials.

When the item statistics indicated a problem with an item, the documentation from the translation verification[5] was used as an aid in checking the test booklets and contacting National Research Coordinators (NRCs). If a problem could be detected by the International Study Center (such as a negative point-biserial for a correct answer or too few options for the multiple-choice questions), the item was deleted from the international scaling. If there was a question about potential translation or cultural issues, however, then the NRC was consulted before deciding how the item should be treated. Appendix D provides a list of deleted items as well as a list of recodes made to free-response item codes.

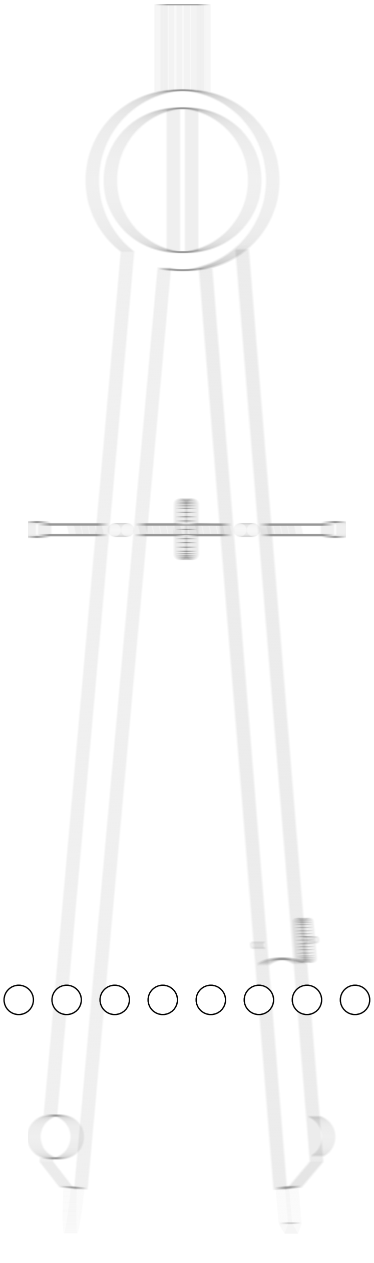
13.4 Summary

Considering that the checking involved more than 300 items for 38 countries (almost 12,000 item-country combinations), very few deviations from the international format were found. Appendix D summarizes the changes that were made to items in the international database before beginning the 1999 IRT scaling.

[5] See Chapter 5 for a description of the translation and verification of the TIMSS data-collection instruments.

Scaling Methodology and Procedures for the TIMSS Mathematics and Science Scales

Kentaro Yamamoto
Edward Kulick



14 Scaling Methodology and Procedures for the TIMSS Mathematics and Science Scales

Kentaro Yamamoto
Edward Kulick

14.1 Overview

The TIMSS achievement test design makes use of matrix-sampling techniques to divide the assessment item pool so that each sampled student responds to just a portion of the items, thereby achieving wide coverage of the mathematics and science subject areas while keeping the response burden on individual students to a minimum.[1] TIMSS relies on a sophisticated form of psychometric scaling known as IRT (Item Response Theory) scaling to combine the student responses in a way that provides accurate estimates of achievement. The TIMSS IRT scaling uses the multiple imputation or “plausible values” method to obtain proficiency scores in mathematics and science and their content areas for all students, even though each student responded to only a part of the assessment item pool.

This chapter first reviews the psychometric models used in scaling the TIMSS 1999 data, and the multiple imputation or “plausible values” methodology that allows such models to be used with sparse item-sampling in order to produce proficiency scale values of respondents. Next, the procedures followed in applying these models to the TIMSS 1999 data are described.

14.2 TIMSS 1999 Scaling Methodology

The psychometric models used in the TIMSS analysis are not new. A similar model has been used in the field of educational measurements since the 1950s and the approach has been even more popular since the 1970s in large-scale surveys, test construction, and computer adaptive testing.[2] (Birnbaum, 1968; Lord and Novick, 1968; Lord, 1980; Van Der Linden and Hambleton,1996).

Three distinct scaling models, depending on item type and scoring procedure, were used in the analysis of the 1999 TIMSS assessment data. Each is a “latent variable” model that describes the probability that a student will respond in a specific way to an

[1] The TIMSS 1999 achievement test design is described in Chapter 2.

item in terms of the respondent's proficiency, which is an unobserved or "latent" trait, and various characteristics (or "parameters") of the item. A three-parameter model was used with multiple-choice items, which were scored as correct or incorrect, and a two-parameter model for free-response items with just two response options, which also were scored as correct or incorrect. Since each of these item types has just two response categories, they are known as dichotomous items. A partial credit model was used with polytomous free-response items, i.e., those with more than two score points.

14.2.1 Two- and Three- Parameter IRT Models for Dichotomous Items

The fundamental equation of the three-parameter (3PL) model gives the probability that a person whose proficiency on a scale *k* is characterized by the unobservable variable θ will respond correctly to item i:

(1)
$$P(x_i = 1 | \theta_k, a_i, b_i, c_i) = c_i + \frac{(1 - c_i)}{1.0 + \exp(-1.7 a_i(\theta_k - b_i))}$$

where

x_i is the response to item i, 1 if correct and 0 if incorrect;

θ_k is the proficiency of a person on a scale k (note that a person with higher proficiency has a greater probability of responding correctly);

a_i is the slope parameter of item i, characterizing its discriminating power;

b_i is its location parameter, characterizing its difficulty;

c_i is its lower asymptote parameter, reflecting the chances of respondents of very low proficiency selecting the correct answer.

[2] Birnbaum, 1968; Lord and Novick, 1968; Lord, 1980; Van Der Linden and Hambleton, 1996. The theoretical underpinning of the imputed value methodology was developed by Rubin (1987), applied to large-scale assessment by Mislevy (1991), and studied further by Mislevy, Johnson and Muraki (1992) and Beaton and Johnson (1992). Other researchers have published widely on related aspects of the methodology; see, for example, Dempster, Laird, and Rubin (1977); Little and Rubin (1983, 1987); Andersen (1980); Engelen (1987); Hoijtink (1991); Laird (1978); Lindsey, Clogg, and Grego (1991); Zwinderman (1991); Tanner and Wong (1987); and Rubin (1987, 1991). The procedures used in TIMSS have also been used in several other large-scale surveys, including the U.S. National Assessment of Educational Progress (NAEP), the U.S. National Adult Literacy Survey (NALS), the International Adult Literacy Survey (IALS), and the International Adult Literacy and Life Skills Survey (IALLS).

The probability of an incorrect response to the item is defined as

$$P_{i0} \equiv P(x_i = 1 \mid \theta_k, a_i, b_i, c_i) = 1 - P_{i1}(\theta_k) \tag{2}$$

The two-parameter (2PL) model was used for the short free-response items that were scored as correct or incorrect. The form of the 2PL model is the same as Equations (1) and (2) with the c_i parameter fixed at zero.

The two- and three-parameter models were used in scaling the TIMSS 1999 data in preference to the one-parameter Rasch model used in TIMSS 1995, primarily because they can more accurately account for the differences among items in their ability to discriminate between students of high and low ability. With the Rasch model, all items are assumed to have the same discriminating power, while the 2PL and 3PL models provide an extra item parameter to account for differences among items in discriminating power. However, the accuracy of representing item response functions by 2PL and 3PL models does not come without cost. Since more item parameters must be estimated, larger amounts of data — and consequently larger sample sizes — are required to obtain the same degree of confidence in the estimated item parameters. However, the TIMSS 1999 database is more than large enough to provide the required level of confidence.

Modeling item response functions as accurately as possible by using 2PL and 3PL models also reduces errors due to model mis-specification. Any mathematical modeling of data without saturated parameters contains errors not accounted for by the model. The error is apparent when the model cannot exactly reproduce or predict the data using the estimated parameters. The difference between the observed data and those generated by the model is directly proportional to the degree of model mis-specification. Current psychometric convention does not allow model mis-specification errors to be represented in the proficiency scores. Instead, once item response parameters are estimated, they are treated as given and model mis-specification is ignored. For that reason it is preferable to use models that characterize the item response function as well as possible.

14.2.2 The IRT Model for Polytomous Items

In TIMSS 1999, free-response items requiring an extended response were scored for partial credit, with 0, 1, and 2 as the possible score levels. These polytomous items were scaled using a generalized partial credit model (Muraki, 1992). The fundamental equation of this model gives the probability that a person with proficiency θ_k on scale k will have, for the i-th item, a response x_i that is scored in the l-th of m_i ordered score categories:

(3)

$$P(x_i = l | \theta_k, a_i, b_i, d_{i,1}, \ldots, d_{i,mi-1}) = \frac{exp\left[\sum_{v=0}^{l} 1.7a_i(\theta_k - b_i + d_{i,v})\right]}{\sum_{g=0}^{m_i-1} exp\left[\sum_{v=0}^{g} 1.7a_i(\theta_k - b_i + d_{i,v})\right]} = P_{il}(\theta_k)$$

where

m_i is the number of response categories for item i;

x_i is the response to item i, possibilities ranging between 0 and m_i-1;

θ_k is the proficiency of person on a scale k;

a_i is the slope parameter of item i, characterizing its discrimination power;

b_i is its location parameter, characterizing its difficulty;

$d_{i,l}$ is category l threshold parameter.

Indeterminacy of model parameters of the polytomous model are resolved by setting $d_{i,0}$ =0 and setting

(4)

$$\sum_{l=1}^{m_i-1} d_{i,l} = 0 .$$

For all of the IRT models there is a linear indeterminacy between the values of item parameters and proficiency parameters, i.e., mathematically equivalent but different values of item parameters can be estimated on an arbitrarily linearly transformed proficiency scale. This linear indeterminacy can be resolved by setting the origin and unit size of the proficiency scale to arbitrary constants, such as mean of 500 with standard deviation of 100. The indeterminacy is most apparent when the scale is set for the first time.

IRT modeling relies on a number of assumptions, the most important being conditional independence. Under this assumption, item response probabilities depend only on θ_κ (a measure of proficiency) and the specified parameters of the item, and are unaffected by the demographic characteristics or unique experiences of the respondents, the data collection conditions, or the other items presented in the test. Under this assumption, the joint probability of a particular response pattern *x* across a set of n items is given by:

(5)

$$P(x|\theta_k, \mathit{item\ parameters}) = \prod_{i=1}^{n} \prod_{l=0}^{m_i-1} P_{il}(\theta_k)^{u_{il}}$$

where $P_{il}(\theta_k)$ is of the form appropriate to the type of item (dichotomous or polytomous), m_i is equal to 2 for the dichotomously scored items, and u_{il} is an indicator variable defined by

(6)

$$U_{il} = \begin{cases} 1 \ \mathit{if\ response}\ x_i\ \mathit{is\ in\ category}\ l \\ 0 \ \mathit{otherwise.} \end{cases}$$

Replacing the hypothetical response pattern with the real scored data, the above function can be viewed as a likelihood function to be maximized by a given set of item parameters. In TIMSS 1999 analyses, estimates of both dichotomous and polytomous item parameters were obtained by the NAEP BILOG/PARSCALE program, which combines Mislevy and Bock's (1982) BILOG and Muraki and Bock's (1991) PARSCALE computer programs. The item parameters in each scale were estimated independently of the parameters of other scales. Once items were calibrated in this manner, a likelihood function for the proficiency θ_k was induced from student responses to the calibrated items. This likelihood function for the proficiency θ_k is called the posterior distribution of the θs for each respondent.

14.2.3 Evaluating Fit of IRT Models to the Data

The fit of the IRT models to the TIMSS 1999 data was examined within each scale by comparing the empirical item response functions with the theoretical item response function curves (see Exhibits 14.1 and 14.2). The theoretical curves are plots of the response functions generated by the model using values of the item parameters estimated from the data. The empirical results are calculated from the posterior distributions of the θs for each respondent who received the item. For dichotomous items the plotted values are the sums of these individual posteriors at each point on the proficiency scale for those students that responded correctly plus a fraction of the omitted responses, divided by the sum of the posteriors of all that were administered the item. For polytomous items, the sums for those who scored in the category of interest is divided by the sum for all those that were administered the item.

Exhibit 14.1 TIMSS 1999 Grade 8 Science Assessment Example Item Response Function Dichotomous Item

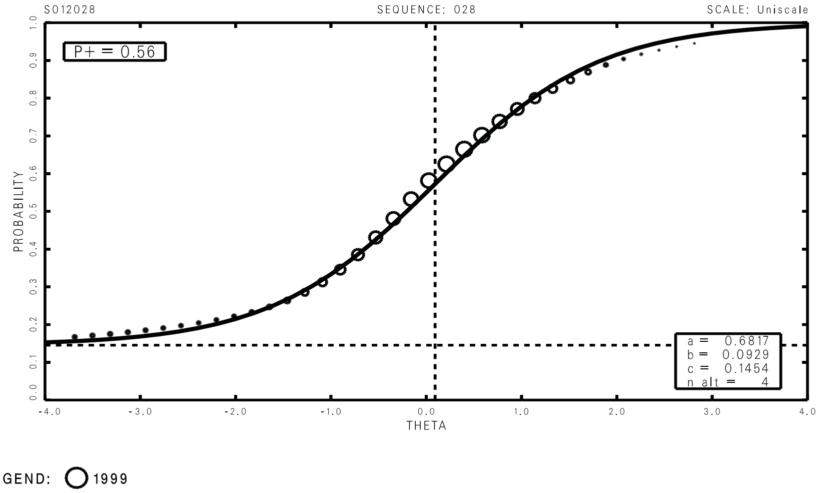


Exhibit 14.2 TIMSS 1999 Grade 8 Science Assessment Example Item Response Function Polytomous Item

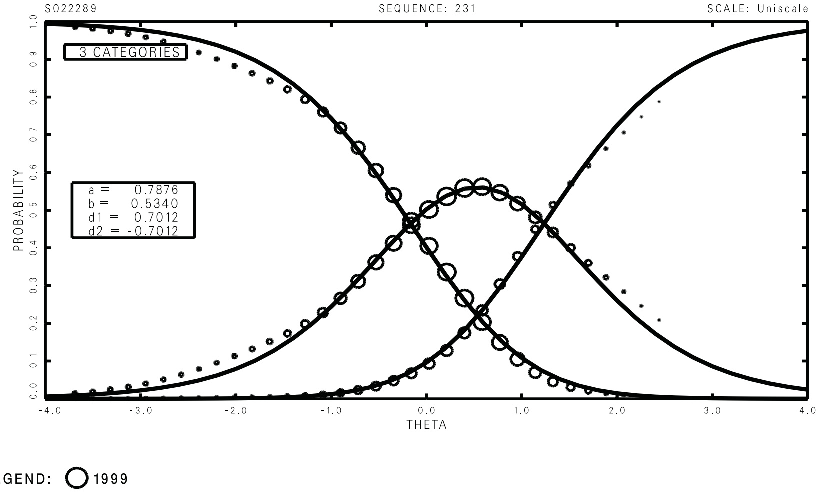


Exhibit 14.1 contains a plot of the empirical and theoretical item response functions for a dichotomous item. In the plot, the horizontal axis represents the proficiency scale, and the vertical axis represents the probability of a correct response. The solid curve is the theoretical curve based on the estimated item parameters. The centers of the small circles represent the empirical proportions correct. The size of the circles is proportional to the sum of the posteriors at each point on the proficiency scale for all of those who received the item; this is related to the number of respondents contributing to the estimation of that empirical proportion correct. Exhibit 14.2 contains a plot of the empirical and theoretical item response functions for a polytomous item. As for the dichotomous item plot above, the horizontal axis represents the proficiency scale, but the vertical axis represents the probability of having a response fall in a given score category. The interpretation of the small circles is the same as in Exhibit 14.1. For items where the model fits the data well, the empirical and theoretical curves are close together.

14.2.4 Scaling Mathematics and Science Domains and Content Areas

In order to estimate student proficiency scores in TIMSS 1999 for the subject domains of mathematics and science, all items in each subject domain were calibrated together. This approach was chosen because it produced the best summary of student proficiency across the whole domain for each subject. Treating the entire mathematics or science item pool as a single domain maximizes the number of items per respondent, and the greatest amount of information possible is used to describe the proficiency distribution. This was found to be a more reliable way to compare proficiency across countries than to make a scale for each of the content areas such as algebra, geometry, etc., and then form a composite measure of mathematics by combining the content area scales. The domain-scaling approach was also found to be more reliable for assessing change from TIMSS 1995 to TIMSS 1999.

A disadvantage of this approach is that differences in content scales may be underemphasized as they tend to regress toward the aggregated scale. Therefore, to enable comparisons of student proficiency on content scales, TIMSS provided separate scale scores of each content area in mathematics and science. If each content area is treated separately when estimating item parameters, differential profiles of content area proficiency can be examined, both across countries and across subpopulations within a country.

14.2.5 Omitted and Not-Reached Responses.

Apart from missing data on that by design were not administered to a student, missing data could also occur because a student did not answer an item, whether because the student did not know the answer, omitted it by mistake, or did not have time to attempt the item. In TIMSS 1999, not-reached items were treated differently in estimating item parameters and in generating student proficiency scores. In estimating the values of the item parameters, items that were considered not to have been reached by students were treated as if they had not been administered. This approach was optimal for parameter estimation. However, since the time allotment for the TIMSS tests was generous, and enough for even marginally able respondents to complete the items, not-reached items were considered to have incorrect responses when student proficiency scores were generated.

14.2.6 Proficiency Estimation Using Plausible Values

Most cognitive skills testing is concerned with accurately assessing the performance of individual respondents for the purposes of diagnosis, selection, or placement. Regardless of the measurement model used, classical test theory or item response theory, the accuracy of these measurements can be improved - that is, the amount of measurement error can be reduced - by increasing the number of items given to the individual. Thus, it is common to see achievement tests designed to provide information on individual students that contain more than 70 items. Since the uncertainty associated with each θ in such tests is negligible, the distribution of θ or the joint distribution of θ with other variables can be approximated using individual θ's.

For the distribution of proficiencies in large populations, however, more efficient estimates can be obtained from a matrix-sampling design like that used in TIMSS 1999. This design solicits relatively few responses from each sampled respondent while maintaining a wide range of content representation when responses are aggregated across all respondents. With this approach, however, the advantage of estimating population characteristics more efficiently is offset by the inability to make precise statements about individuals. The uncertainty associated with individual θ estimates becomes too large to be ignored. In this situation, aggregations of individual student scores can lead to seriously biased estimates of population characteristics (Wingersky, Kaplan, & Beaton, 1987).

Plausible values methodology was developed as a way to address this issue by using all available data to estimate directly the characteristics of student populations and subpopulations, and then generating imputed scores or plausible values from these distributions that can be used in analyses with standard statistical software. A detailed review of plausible values methodology is given in Mislevy (1991)[3].

The following is a brief overview of the plausible values approach. Let y represent the responses of all sampled students to background questions or background data of sampled students collected from other sources, and let θ represent the proficiency of interest. If θ were known for all sampled students, it would be possible to compute a statistic t(θ,y) - such as a sample mean or sample percentile point - to estimate a corresponding population quantity T.

Because of the latent nature of the proficiency, however, θ values are not known even for sampled respondents. The solution to this problem is to follow Rubin (1987) by considering θ as "missing data" and approximate t(θ,y) by its expectation given (x,y), the data that actually were observed, as follows:

$$\text{(7)} \qquad \begin{aligned} t^{*}(x,y) &= E[t(\underline{\theta},\underline{y})|\underline{x},\underline{y}] \\ &= \int t(\underline{\theta},\underline{y})p\ (\underline{\theta}|\underline{x},\underline{y})\,d\theta. \end{aligned}$$

It is possible to approximate t^{*} using random draws from the conditional distribution of the scale proficiencies given the student's item responses x_{j}, the student's background variables y_{j}, and model parameters for the student. These values are referred to as imputations in the sampling literature, and as plausible values in large-scale surveys such as NAEP, NALS, and IALLS. The value of θ for any respondent that would enter into the computation of t is thus replaced by a randomly selected value from his or her conditional distribution. Rubin (1987) proposed repeating this pro-

○○○

[3] Along with theoretical justifications, Mislevy presents comparisons with standard procedures, discusses biases that arise in some secondary analyses, and offers numerical examples.

cess several times so that the uncertainly associated with imputation can be quantified by "multiple imputation." For example, the average of multiple estimates of t, each computed from a different set of plausible values, is a numerical approximation of t^* of the above equation; the variance among them reflects uncertainty due to not observing $\underline{\theta}$. It should be noted that this variance does not include the variability of sampling from the population.

Note that plausible values are not test scores for individuals in the usual sense, but rather are imputed values that may be used to estimate population characteristics correctly. When the underlying model is correctly specified, plausible values will provide consistent estimates of population characteristics, even though they are not generally unbiased estimates of the proficiencies of the individuals with whom they are associated[4].

Plausible values for each respondent j are drawn from the conditional distribution $P(\theta_j|x_j,y_j,\Gamma,\Sigma)$,where Γ is a matrix of regression coefficients for the background variables, and Σ is a common variance matrix for residuals. Using standard rules of probability, the conditional probability of proficiency can be represented as

(8)

$$P(x_{ij} = 1|\theta_k,a_i,b_i,c_i) \;=\; c_i + \frac{(1 - c_i)}{1 + \exp(-1.7\,a_i(\theta_k - b_i))}$$

where θ_j is a vector of scale values, $P(x_j|\theta_j)$ is the product over the scales of the independent likelihoods induced by responses to items within each scale, and $P(\theta_j|y_j,\Gamma,\Sigma)$ is the multivariate joint density of proficiencies of the scales, conditional on the observed value y_j of background responses and parameters Γ and Σ. Item parameter estimates are fixed and regarded as population values in the computations described in this section.

14.2.7 Conditioning

A multivariate normal distribution was assumed for $P(\theta_j|y_j,\Gamma,\Sigma)$, with a common variance, Σ, and with a mean given by a linear model with regression parameters, Γ. Since in large-scale studies like TIMSS there are many hundreds of background variables, it is customary to conduct a principal components analysis to

○○○

[4] For further discussion, see Mislevy, Beaton, Kaplan, and Sheehan (1992).

reduce the number to be used in Γ. Typically, components representing 90% of the variance in the data are selected. These principal components are referred to as the conditioning variables and denoted as y^c. The following model is then fit to the data.

$$\theta = \Gamma' y^c + \varepsilon, \tag{9}$$

where ε is normally distributed with mean zero and variance Σ. As in a regression analysis, Γ is a matrix each of whose columns is the effects for each scale and Σ is the matrix of residual variance between scales.

Note that in order to be strictly correct for all functions Γ of *θ*, it is necessary that $p(\theta|\mathbf{y})$ be correctly specified for all background variables in the survey. In TIMSS 1999, however, principal component scores based on nearly all background variables were used. Those selected variables were chosen to reflect high relevance to policy and to education practices. The computation of marginal means and percentile points of *θ* for these variables is nearly optimal. Estimates of functions Γ involving background variables not conditioned on in this manner are subject to estimation error due to mis-specification. The nature of these errors was discussed in detail in Mislevy (1991).

The basic method for estimating Γ and Σ with the Expectation and Maximization (EM) procedure is described in Mislevy (1985) for a single scale case. The EM algorithm requires the computation of the mean, *θ*, and variance, Σ, of the posterior distribution in (4). For the multiple content area scales of TIMSS 1999, the computer program CGROUP (Thomas, 1993) was used. The program implemented a method to compute the moments using higher-order asymptotic corrections to a normal approximation. Case weights were employed in this step.

14.2.8 Generating Proficiency Scores

After completing the EM algorithm, the plausible values are drawn in a three-step process from the joint distribution of the values of Γ for all sampled. First, a value of Γ is drawn from a normal approximation to $P(\Gamma,\Sigma|x_j,y_j)$ that fixes Σ at the value $\hat{\Sigma}$ (Thomas, 1993). Second, conditional on the generated value of Γ (and the fixed value of $\Sigma=\hat{\Sigma}$), the mean *θ*, and variance Σ_j^p of the posterior distribution in equation (2) are computed using the methods applied in the EM algorithm. In the third step, the proficiency values are drawn independently from a multivariate normal distribution with mean θ and variance Σ_j^p. These three steps are repeated five times, producing five imputations of θ for each sampled respondent.

For respondents with an insufficient number of responses, the Γ and Σs described in the previous paragraph were fixed. Hence, all respondents - regardless of the number of items attempted - were assigned a set of plausible values for the various scales.

The plausible values could then be employed to evaluate equation (1) for an arbitrary function T as follows:

1. Using the first vector of plausible values for each respondent, evaluate T as if the plausible values were the true values of θ. Denote the result T_1.
2. As in step 1 above, evaluate the sampling variance of T, or Var($T_{1,}$), with respect to respondents' first vectors of plausible values. Denote the result Var_1.
3. Carry out steps 1 and 2 for the second through fifth vectors of plausible values, thus obtaining T_u and Var_u for u=2, . . ., M, where M is the number of imputed values.
4. The best estimate of T obtainable from the plausible values is the average of the five values obtained from the different sets of plausible values:

(10)

$$T. = \frac{\sum_u T_u}{5}$$

5. An estimate of the variance of T. is the sum of two components: an estimate of Var(T_u) obtained as in step 4 and the variance among the T_us:

(11)

$$Var(T.) = \frac{\sum_u Var_u}{M} + (1 + M^{-1})\frac{\sum_u (T_u - T.)^2}{M - 1}$$

The first component in *Var(T.)* reflects uncertainty due to sampling respondents from the population; the second reflects uncertainty due to the fact that sampled respondents' θs are not known precisely, but only indirectly through *x* and *y*.

14.2.9 Working with Plausible Values

Plausible values methodology was used in TIMSS 1999 to increase the accuracy of estimates of the proficiency distributions for various subpopulations and for the TIMSS population as a whole. This method correctly retains the uncertainty associated with proficiency estimates for individual respondents by using multiple imputed proficiency values rather than assuming that this type of uncertainty is zero - a more common practice. However, retaining this component of uncertainty requires that additional analytic procedures be used to estimate respondents' proficiencies, as follows.

If θ values were observed for sampled respondents, the statistic $(t-T)/U^{1/2}$ would follow a *t*-distribution with d degrees of freedom. Then the incomplete-data statistic $(t^{*}-T)/(\mathrm{Var}(t^{*}))^{1/2}$ is approximately t-distributed, with degrees of freedom (Johnson & Rust, 1993) given by

(12)
$$v = \frac{1}{\frac{f_M^{\;2}}{M-1} + \frac{(1-f_M)^2}{d}}$$

where *d* is the degrees of freedom, and *f* is the proportion of total variance due to not observing θ values:

(13)
$$f_M = \frac{(1+M^{-1})B_M}{V_M}$$

where B_M is the variance among *M* imputed values and V_M is the final estimate of the variance of T. When B is small relative to U^{*}, the reference distribution for incomplete-data statistics differs little from the reference distribution for the corresponding complete-data statistics. If, in addition, d is large, the normal approximation can be used instead of the t-distribution.

For k-dimensional *t*, such as the *k* coefficients in a multiple regression analysis, each U and U^{*} is a covariance matrix, and B is an average of squares and cross-products rather than simply an average of squares. In this case, the quantity $(T-t^{*})V^{-1}(T-t^{*})'$ is approximately F distributed with degrees of freedom equal to k and ν, with ν defined as above but with a matrix generalization of f_M

(14)
$$f = \frac{(1-M^{-1})\,Trace(BV^{-1})}{k}.$$

A chi-square distribution with k degrees of freedom can be used in place of *f* for the same reason that the normal distribution can approximate the t distribution.

Statistics t*, the estimates of ability conditional on responses to cognitive items and background variables, are consistent estimates of the corresponding population values T, as long as background variables are included in the conditioning variables. The consequences of violating this restriction are described by Beaton and Johnson (1990), Mislevy (1991), and Mislevy and Sheehan (1987). To avoid such biases, the TIMSS 1999 analyses included nearly all background variables.

14.3 Implementing the TIMSS 1999 Scaling Procedures

This section describes how the IRT scaling and plausible value methodology was applied to the TIMSS 1999 data. This consisted of three major tasks, as follows.

Re-scaling of the 1995 TIMSS data. TIMSS in 1995 also made use of IRT scaling with plausible values (Adams, Wu, and Macaskill, 1997). The scaling model, however, relied on the one-parameter Rasch model rather than the more general two- and three-parameter models used in 1999. Since a major goal of TIMSS 1999 was to measure trends since 1995 by comparing results from both data collections, it was important that both sets of data be on the same scale. Accordingly it was decided as a first step to rescale the 1995 data using the scaling models from 1999.

Scaling the 1999 data and linking to the 1995 data. Since the achievement item pools used in 1995 and 1999 had about one-third of the items in common, the scaling of the 1999 data was designed to place both data sets on a common IRT scale. Although the common items administered in 1995 and 1999 formed the basis of the linkage, all of the items used in each data collection were included in the scaling since this increases the information for proficiency estimation and reduces measurement error. Item-level linking of two or more scales in this way is one of the most powerful methods of scale linking and is well suited to IRT methods. This is one of the benefits of using the IRT scaling procedures.

Creating IRT scales for mathematics and science content areas for 1995 and 1999 data. IRT scales were also developed for each of the content areas in mathematics and science for both 1995 and 1999. Because there were few items common to the two assessments, and because of some differences in their composition, the 1995 and 1999 scales were not linked, but rather each was established independently.

14.3.1 Re-scaling of the 1995 TIMSS Data

The re-scaling of 1995 TIMSS followed, as much as possible, the procedures used in the original 1995 analyses, while using two- and three-parameter scaling models in place of the more restrictive one-parameter Rasch model. Item parameter estimates were obtained using an "international calibration sample" that consisted of random samples of 600 eighth-grade students from each of the 37 countries that participated in TIMSS in 1995 (plus 300 from Israel). The calibration samples were drawn with probability proportional to size of sampling weight in each country, so that the sample accurately reflected the distribution of students in the population. The 1995 estimated item parameters for mathematics may be found in Exhibit E.1 in Appendix E and for science in Exhibit E.2.

Using the re-estimated item parameters from the two- and three-parameter and polytomous IRT models, the conditioning analyses were completed, with a conditioning model similar to the one used in 1995. Following that approach, and separately within each country, responses to background variables were summarized through a principal components analysis. Enough principal components were created to account for at least 90% of the variability in the original set of background variables. In addition to the principal components, several background variables were explicitly included in the conditioning model. These included student gender and the school mean on a simple Rasch based measure of student achievement in the subject (mathematics or science) being scaled. Additionally, the conditioning for mathematics included the Rasch score for science, and the conditioning for science, the score for mathematics. Exhibit 14.3 shows the total number of conditioning variables used in the re-scaling for each country.

Exhibit 14.3 Number of Conditioning Variables for TIMSS 1995 Re-scaling

Country	Sample size	Number of Principal Components	Number of conditioning variables
Australia	12852	317	648
Austria	5786	361	735
Belgium (Flemish)	5662	473	952
Belgium (French)	4883	425	858
Bulgaria	3771	2	12
Canada	16581	348	711
Colombia	5304	357	723
Czech Republic	6672	540	1089
Cyprus	5852	358	731
Germany	5763	484	976
Denmark	4370	434	876
Spain	7596	349	707
France	6014	367	745
England	3579	261	527
Greece	7921	556	1119
Hong Kong	6752	319	646
Hungary	5978	524	1057
Ireland	6203	360	726
Iran, Islamic Rep.	7429	328	664
Iceland	3730	492	997
Israel	1415	308	621
Japan	10271	257	520
Korea, Rep. of	5827	346	697
Kuwait	1655	303	610
Lithuania	5056	540	1088
Latvia (LSS)	4976	477	962
Netherlands	4084	451	915
Norway	5736	340	691
New Zealand	6867	352	712
Philippines	11847	379	766
Portugal	6753	413	838
Romania	7471	572	1153
Russian Federation	8160	563	1131
Scotland	5776	220	448
Singapore	8285	334	675
Slovak Republic	7101	521	1050
Slovenia	5606	475	964
Sweden	8855	595	1201
Switzerland	11722	358	727
Thailand	11643	351	710
United States	10973	350	712
South Africa	9792	387	793

Plausible values generated by the conditioning program are initially on the same scale as the item parameters used to estimate them. This scale metric is generally not useful for reporting purposes since it is somewhat arbitrary. Instead, a reporting metric that has desirable properties is usually selected. In the original 1995 scaling, a metric was chosen for reporting TIMSS results such that the combined proficiency distribution for seventh and eighth grade students had a mean of 500 and a standard deviation of 100 (Gonzalez, 1997).

In the re-scaling of the 1995 data, the transformation procedures to establish the reporting metric were slightly different. Since the 1999 assessment consisted of eighth-grade students only (not both seventh- and eighth-grade students as in 1995), and since a major goal of the re-scaling was to establish a trend line to 1999, a metric was chosen for the re-scaled 1995 data that had desirable properties for the proficiency distribution of eighth-grade students. Accordingly, the scale was set so that the distribution of eighth-grade students in 1995 had a mean of 500 and a standard deviation of 100. The same metric transformation was applied to the re-scaled seventh- grade data from 1995. This procedure was followed for both the mathematics and science scales. Extreme scale values were truncated, i.e., plausible values below 5 were set to 5 and plausible values above 995 were set to 995.

Setting the scale metric as described above produces slightly lower means and slightly higher standard deviations than the original 1995 eighth-grade results. This is solely the result of the decision to base the metric on the eighth-grade distribution only rather than on the combined seventh- and eighth-grade distributions. Comparisons between the original and re-scaled 1995 proficiency scores are not appropriate because of this difference in the scale metric.

14.3.2 Scaling the 1999 Data and Linking to the 1995 Data

The linking of the 1995 and 1999 scales was conducted at the mathematics and science domain levels only, since there were not enough common items to enable reliable linking within each mathematics or science content area. As may be seen from Exhibit 14.4, about one-third of the items were common to both assessments (48 items in mathematics and 48 in science), which was enough to provide a reliable link between the 1995 and 1999 assessments.

Exhibit 14.4 Numbers of items Common and Unique to TIMSS 1995 and TIMSS 1999

Subject	Items	TIMSS 1995	TIMSS 1999
Mathematics	Unique to TIMSS 1995	111	
	Unique to TIMSS 1999		115
	Common to both TIMSS 1995 and TIMSS 1999	48	
	Total	159	163
	Grand Total for Mathematics	274	
Science	Unique to TIMSS 1995	94	
	Unique to TIMSS 1999		106
	Common to both TIMSS 1995 and TIMSS 1999	48	
	Total	142	154
	Grand Total for Science	248	

Calibration samples of 1,000 students per country per assessment were selected from each of the 25 countries that participated in both assessments, using the same method as in 1995. All 274 of the mathematics items (common items and items unique to one or the other assessment) were scaled together to provide new item parameter estimates that fit both calibration samples (1995 and 1999). The same procedure was followed for all 248 of the science items. Estimated item parameters from this joint 1995-1999 scaling may be found in Exhibit E.3 in Appendix E for mathematics and in Exhibit E.4 for science.

These item parameters estimates were used to generate plausible values for all of the 38 TIMSS 1999 countries, including those that participated only in 1999.[5] A new set of principal components was calculated for the TIMSS 1999 data in each country for use in conditioning. Exhibit 14.5 shows the total number of conditioning variables used for the TIMSS 1999 for each country. Plausible values were generated for all countries for both assessments using the new, jointly estimated item parameters.

[5] In addition to its eighth-grade sample, Chile also surveyed a seventh-grade sample that was scaled with the 1999 item parameters.

Exhibit 14.5 Number of Variables and Principal Components for Conditioning TIMSS 1999

Country	Sample size	Total number of conditioning variables	Total number of principal components only
Australia	4032	374	348
Belgium (Flemish)	5259	485	479
Bulgaria	3272	582	575
Canada	8770	385	364
Chile	5907	410	405
Chinese Taipei	5772	379	374
Cyprus	3116	394	385
Czech Republic	3453	557	551
England	2960	298	292
Finland	2920	548	538
Hong Kong, SAR	5179	392	384
Hungary	3183	584	578
Indonesia	5848	403	397
Iran, Islamic Rep.	5301	406	400
Israel	4195	405	398
Italy	3328	380	374
Japan	4745	362	354
Jordan	5052	415	409
Korea, Rep. of	6114	408	388
Latvia (LSS)	2873	522	514
Lithuania	2361	342	336
Morocco	5402	681	675
Moldova	3711	599	593
Macedonia, Rep. of	4023	576	569
Malaysia	5577	386	381
Netherlands	2962	437	430
New Zealand	3613	338	332
Philippines	6601	422	415
Romania	3425	597	589
Russian Federation	4332	657	608
Singapore	4966	370	365
Slovak Rep.	3497	408	401
South Africa	8146	441	426
Slovenia	3109	584	578
Thailand	5732	398	390
Tunisia	5051	418	413
Turkey	7841	449	405
United States	9072	409	392

The final step in the scaling of the data was to locate both the 1995 and the 1999 data on the same scale. This was done by calculating transformation constants that matched the means and standard deviations of the re-scaled 1995 plausible values, which were on the required scale, with the means and standard deviations of the jointly-scaled 1995-1999 plausible values for the same set of countries, which were on an independent scale. This procedure was used for the countries that participated in both assessments.[6] The transformation constants, which were applied as $A^*\theta+B$, are shown in Exhibit 14.6 for mathematics and science.

Exhibit 14.6 Transformation Constants for TIMSS 1999 Mathematics and Science Domain Scales

TIMSS 1995 and TIMSS 1999	A	B
Mathematics	99.593	510.169
Science	102.188	508.961

These linear transformations were then applied to the plausible values of the TIMSS 1999 students to place their results on the same scale as the 1995 data. If the transformation is accurate it should produce practically identical means in each country for both the re-scaled 1995 plausible values and the plausible values based on the joint 1995-1999 scaling. Exhibit 14.7 presents a comparison of the results for mathematics from both sets of data, and Exhibit 14.8 shows the same results for science. Both exhibits indicate that the differences between the proficiency means of the re-scaled 1995 data and the jointly-scaled 1995-1999 data are very small for every country. They are on average less than 20% of the standard error of measurement, implying that no systematic errors exist and that the differences can be considered ignorable.

○○○

[6] Because they did not satisfy all sampling guidelines in 1995, Israel, South Africa, and Thailand were omitted from the calculation of transformation constants.

Exhibit 14.7 Comparison of 1995 TIMSS Reanalysis and Univariate Linking Mathematics Scale

Country	Mean from 1995 Re-Scaling	Mean for 1995 from Joint 1995-1999 Scaling	Difference
Singapore	609	609	0.0
Korea	581	581	0.5
Hong Kong, SAR	569	569	0.4
Japan	581	581	0.2
Belgium (Flemish)	550	549	-1.1
Netherlands	529	529	0.0
Hungary	527	526	-0.4
Canada	521	520	-1.0
Slovenia	531	530	-0.9
Russian Federation	524	523	-0.4
Australia	519	518	-0.5
Czech Republic	546	544	-1.1
Bulgaria	527	527	0.0
Latvia (LSS)	488	489	0.5
United States	492	492	-0.4
England	498	496	-1.3
New Zealand	501	501	-0.3
Lithuania	472	472	0.2
Italy	491	492	0.7
Cyprus	468	468	0.2
Romania	474	474	0.6
Iran, Islamic Rep.	418	423	4.3
Mean:	518.750	518.750	
Standard Deviation:	92.363	92.363	

Exhibit 14.8 **Comparison of 1995 TIMSS Reanalysis and Univariate Linking Science Scale**

Country	Mean from 1995 Re-Scaling	Mean for 1995 from Joint 1995-1999 Scaling	Difference
Australia	527	526	0.0
Belgium (Flemish)	533	533	0.0
Bulgaria	545	544	-1.0
Canada	514	516	1.9
Cyprus	452	452	0.2
Czech Republic	555	553	-1.5
England	533	533	-0.8
Hong Kong, SAR	510	512	1.9
Hungary	537	536	-0.5
Iran, Islamic Rep.	463	464	0.7
Italy	497	500	3.2
Japan	554	552	-2.5
Korea, Rep. of	546	545	-0.8
Latvia (LSS)	476	475	-0.8
Lithuania	464	465	1.3
Netherlands	541	542	1.0
New Zealand	511	512	1.2
Romania	471	471	0.3
Russian Federation	523	522	-0.9
Singapore	580	578	-2.4
Slovenia	541	538	-2.9
United States	513	515	2.7
Mean:	517.511	517.511	
Standard Deviation:	91.587	91.587	

14.3.3 Creating IRT Scales for Mathematics and Science Content Areas for 1995 and 1999 Data

The primary function of the IRT scales in the mathematics and science content areas is to portray student achievement in each country in terms of a profile of relative performance in each area. Such profiles should, for example, show countries where performance in algebra was relatively better than in geometry, or in life science than in chemistry. Although it would have been desirable to establish a link from 1995 and 1999 in each content area, there were not enough common items in the two assessments to do this reliably. However, the numbers of items in each content area were considered sufficient to develop content area scales for each assessment separately. The five content areas in mathematics and six areas in science for which scales were developed are presented in Exhibit 14.9.

Exhibit 14.9 Number of Items in Mathematics and Science Content Areas (1995 and 1999 Combined)

Mathematics Content Areas	No. of Items	Science Content Areas	No. of Items
Fractions/Number	104	Earth Science	34
Measurement	39	Life Science	70
Data Representation	33	Physics	64
Geometry	41	Chemistry	39
Algebra	57	Environmental and Resource Issues	17
		Scientific Inquiry and the Nature of Science	12
Total	274	Total	236

The calibration samples used for the joint 1995-1999 scaling were also used to estimate the item parameters for each of the content area scales (shown in Exhibit E.5 in Appendix E for mathematics and in Exhibit E.6 for science). The principal components produced for the conditioning of the joint 1995-1999 mathematics and science domain scales were used for the 1999 content area plausible value analyses as well. Plausible values were generated for all countries for both assessments using the new, jointly estimated item parameters under multivariate conditions.

The indeterminacy of the content area scales in mathematics was resolved by setting the mean of each mathematics content area scale over all of the 38 TIMSS 1999 countries to be the same as the mean of the domain scale for mathematics. The same approach was taken for science. The transformation constants used to do this are presented in Exhibit 14.10.

It should be noted that since there were far fewer items in each content area scale than in the domain scales (for example, 57 algebra items compared with 274 mathematics items), a relatively greater proportion of the variance in the content area scales was due to measurement error. In the scaling, the total variance for content area scales and domain scales was set to be equal, and therefore the measurement error plays a relatively greater role in the variance of the content area scales. This implies that the content area scale means of each country tend to be regressed toward the grand mean, and that the regression is more noticeable for very high- or very low-achieving countries.

Exhibit 14.10 Transformation Constants for TIMSS 1999 Content Area Scales

TIMSS 1999	A	B
Mathematics Scales		
Algebra	82.454	511.536
Data Representation	71.222	506.175
Fractions and Number	85.000	511.931
Geometry	65.933	506.741
Measurement	74.404	511.959
Science Scales		
Chemistry	64.553	505.616
Earth Science	69.688	508.543
Life Science	78.839	507.671
Physics	78.111	507.558
Environmental and Resource Issues	64.201	503.668
Scientific Inquiry and the Nature of Science	48.504	516.944

14.4 Summary

Item Response Theory was used to model the TIMSS achievement data. In order to better monitor trends in mathematics and science achievement, TIMSS used 2- and 3-parameter IRT, and plausible-value technology to re-analyze the 1995 achievement data, and analyze the 1999 achievement data. The procedures used to link the 1995 and 1999 achievement data were described.

References

Adams, R.J., Wu, M.L., & Macaskill, G. (1997). "Scaling Methodology and Procedures for the Mathematics and Science Scales" in M.O. Martin and D. L. Kelly (Eds.), *TIMSS Technical Report Volume II: Implementation and Analysis.* Chestnut Hill, MA: Boston College.

Andersen, E.B. (1980). Comparing latent distributions. *Psychometrika, 45*, 121-134.

Beaton, A.E., & Johnson, E.G. (1990). The average response method of scaling. *Journal of Educational Statistics,* 15, 9-38.

Beaton, A.E., & Johnson, E.G. (1992). Overview of the scaling methodology used in the National Assessment. *Journal of Educational Measurement,* 26(2), 163-175.

Birnbaum, A. (1968). "Some latent trait models and their use in inferring an examinee's ability" in F. M. Lord & M. R. Novick, *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley Publishing.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39, 1-38.

Engelen, R.J.H. (1987). *Semiparametric estimation in the Rasch model.* Research Report 87-1. Twente, the Netherlands: Department of Education, University of Twente.

Gonzalez, E.J. (1997). "Reporting Student Achievement in Mathematics and Science" in M. O. Martin & D. L. Kelly (Eds.) *TIMSS Technical Report Volume II: Implementation and Analysis.* Chestnut Hill, MA: Boston College.

Hoijtink, H. (1991). *Estimating the parameters of linear models with a latent dependent variable by nonparametric maximum likelihood.* Research Bulletin HB-91-1040-EX. Groningen, The Netherlands: Psychological Institute, University of Groningen.

Johnson, E.G., & Rust, K.F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics.*

Laird, N.M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73, 805-811.

Lindsey, B., Clogg, C.C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, 86, 96-107.

Little, R.J.A. & Rubin, D.B. (1983). On jointly estimating parameters and missing data. *American Statistician,* 37, 218-220.

Little, R.J.A. & Rubin, D.B. (1987). *Statistical analysis with missing dat*a. New York, NY: John Wiley and Sons.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum Associates.

Lord, F.M.,& Novick, M.R. (1968). *Statistical theories of mental test scores.* Redding, MA: Addison-Wesley.

Mislevy, R.J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80,* 993-97.

Mislevy, R.J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56,* 177-196.

Mislevy, R.J., Beaton, A., Kaplan, B.A., & Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133-161.

Mislevy, R.J. & Bock, R.D. (1982). *BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Morresville, IN: Scientific Software.

Mislevy, R.J., Johnson, E.G. & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, 17(2), 131-154.

Mislevy, R.J. & Sheehan, K. (1987). "Marginal estimation procedures" in A. E. Beaton (Ed.), *Implementing the new design: The NAEP 1983-84 technical report* (pp. 293-360). (no. 15-TR-20) Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement,* 16(2), 159-176.

Muraki, E., & Bock, R.D. (1991). PARSCALE: *Parameter scaling of rating data.* Chicago, IL: Scientific Software, Inc.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

Rubin, D.B. (1991). EM and beyond. *Psychometrika, 56,* 241-254.

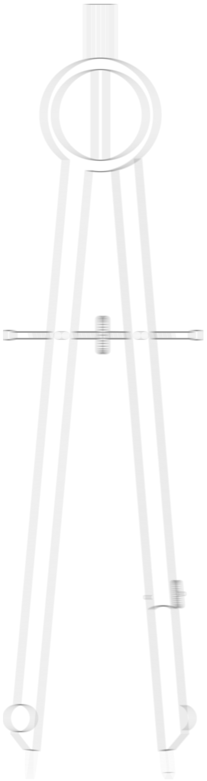
Tanner, M., & Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association, 82,* 528-550.

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics, 2,* 309-22.

Wingersky, M., Kaplan, B.A., & Beaton, A.E. (1987). "Joint estimation procedures" in A. E. Beaton (Ed.), *Implementing the new design: The NAEP 1983-84 technical report* (pp.285-92) (No. 15-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Van Der Linden, W.J. & Hambleton, R. (1996). *Handbook of Modern Item Response Theory.* New York. Springer-Verlag.

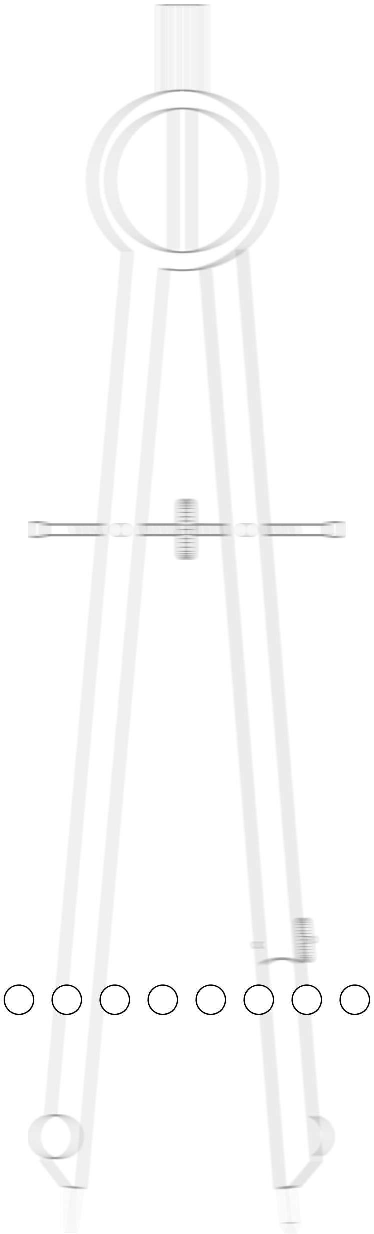
Zwinderman, A.H. (1991). Logistic regression Rasch models. *Psychometrika, 56,* 589-600.



15

Describing International Benchmarks of Student Achievement

Kelvin D. Gregory
Ina V. S. Mullis



15 Describing International Benchmarks of Student Achievement

Kelvin D. Gregory
Ina V. S. Mullis

15.1 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. 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 description, illustrated by example items, of what students scoring at the anchor points are likely to know and be able to do.

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. Second, subject-matter experts examined the knowledge shown by correct responses to the anchor items, summarized student's understandings for each anchor point, and selected example items to support the descriptions.

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, and Martin (2000); the scale anchoring results for science are presented in Smith, Martin, Mullis, and Kelly (2000).[1]

[1] For a discussion of the theoretical underpinnings of scale anchoring and decisions related to the application of scale anchoring to the TIMSS data, see Kelly (1999).

15.2 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-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

15.2.1 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 the anchor point, but 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. The international benchmarks were computed using the combined data from the countries that participated. Exhibit 15.1 shows the scale scores representing the international benchmarks for mathematics and science, respectively.

Exhibit 15.1 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 5 scale points was used. The ranges around the international percentiles and the number of observations within each range are shown in Exhibit 15.2.

Exhibit 15.2 **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

15.3 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 less 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
- Less 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
- Less 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
- Less 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
- Less 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
- Less 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
- Less than 50% of students at the 75th percentile answered the item correctly

To further supplement the pool of items, items that met only the criterion that at least 60% of the students answered correctly (regardless of the performance of students at the next lower point) were identified. The three categories of items were mutually exclusive, and ensured that all of the items were available to inform the descriptions of student achievement at the anchor levels.

15.4 Computing the Item Percent Correct at Each Level

The percentage of students scoring in the range around each anchor point that answered the 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 are 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 0, 1, or 2. Two variables were created:

$v_1 = 1$ if the student receives a 1, or 2, and 0 otherwise

$v_2 = 1$ if the student receives a 2 and 0 otherwise.

The percentage of students receiving a 1 on v_1 and the percentage of those receiving a 1 on v_2 were computed. This yielded the percentage of students receiving at least one point and 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.

15.5 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 15.3 and 15.4 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. Even though these items did not meet the 65%/50% anchoring criteria, they were still items that students scoring at the anchor points had a high probability of answering correctly.

Exhibit 15.3 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 15.4 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

15.6 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 15.5 and 15.6 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. Environmental and resource issues, and scientific inquiry and the nature of science categories had no items that anchored at the 25th percentile, but it should be remembered that these two categories had fewest items.

Exhibit 15.5 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 15.6 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

15.7 The Anchoring Expert Panels

Two panels of expert 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 15.7 and 15.8, respectively. The members have extensive experience in their subject areas and a thorough knowledge of the TIMSS curriculum frameworks and achievement tests.

Exhibit 15.7 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 15.8 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	

15.8 Development of Anchor Level Descriptions

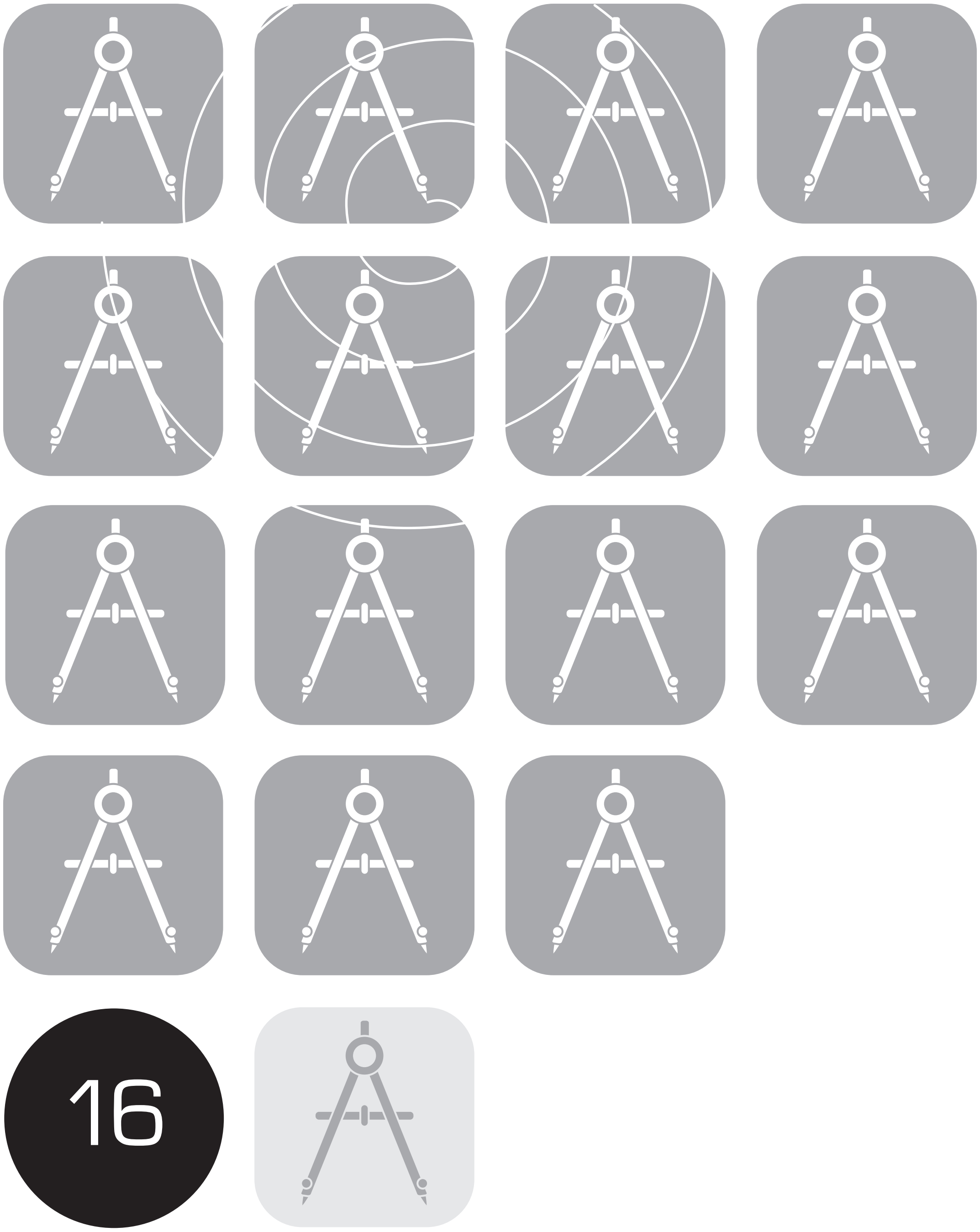
The TIMSS International Study Center convened the two expert panels for a three-day meeting, May 7 to 10, 2000, at Martha's Vineyard, Massachusetts. The panelists' assignment consisted of three tasks: (1) work through each item in each binder and arrive at a short description of the knowledge, understanding, and/or skills demonstrated by students answering the item correctly; (2) based on the items that anchored, almost anchored, and met only the 60-65% 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. Following the meeting, these drafts were edited and revised as necessary, and the panelists reviewed and approved the item descriptions, anchor point descriptions, and selection of example items for use in the TIMSS 1999 International Reports.

References

Kelly, D. L. (1999). *Interpreting the Third International Mathematics and Science Study (TIMSS) Achievement Scales Using Scale Anchoring*. Unpublished doctoral dissertation, Boston College.

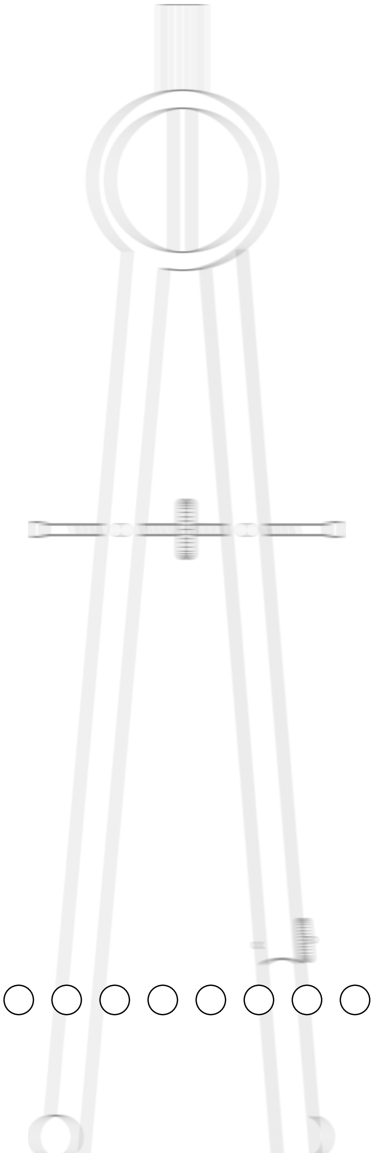
Kelly, D.L., Mullis, I.V.S., & Martin, M.O. (2000). *Profiles of Student Achievement in Mathematics at the TIMSS International Benchmarks: U.S. Performance and Standards in an International Context.* Chestnut Hill, MA: Boston College.

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.



Reporting Student Achievement in Mathematics and Science

Eugenio J. Gonzalez
Kelvin D. Gregory



16 Reporting Student Achievement in Mathematics and Science

Eugenio J. Gonzalez
Kelvin D. Gregory

16.1 Overview

As described in earlier chapters, TIMSS 1999 makes extensive use of imputed student proficiency scores to report achievement in mathematics and science, both in the subjects overall and in content areas. This chapter describes the procedures followed in computing the major statistics used to summarize achievement in the international reports (Mullis et al., 2000; Martin et al., 2000), including country means based on plausible values, Bonferroni adjustments for multiple comparisons, reporting trends in achievement, estimating international benchmarks of achievement, and producing profiles of relative performance in subject matter content areas.

16.2 National and International Student Achievement

The item response theory (IRT) scaling procedure described in Chapter 14 yields five imputed scores or plausible values for each student. A national average for each plausible value was computed as the weighted mean

$$\bar{X}_{pvl} = \frac{\sum_{j=1}^{N} W^{i,j} \cdot pv_{lj}}{\sum_{j=1}^{N} W^{i,j}}$$

where

$\bar{X}_{pvl}$ is the country mean for plausible value l

pv_{lj} is the l-th plausible value for the j-th student

$W^{i,j}$ is the weight associated with the j-th student in class i, described in Chapter 12

N is the number of students in the country's sample.

The country average is the mean of the five national plausible value means.

The international average for each plausible value was computed as the average of the plausible value for each country

$$\bar{X}_{\bullet pvl} = \frac{\sum_{k=1}^{N} \bar{X}_{pvl,\,k}}{N}$$

where

$\bar{X}_{\bullet pvl}$ is the international mean for plausible value l

$\bar{X}_{pvl,\,k}$ is the k-th country mean for plausible value l

and N is the number of countries.

The international average was the average of the five international mean plausible values.

16.3 Achievement Differences Across Countries

An aim of the TIMSS 1999 international reports is to provide fair and accurate comparisons of student achievement across the participating countries. Most of the exhibits in the reports summarize student achievement by means of a statistic such as a mean or percentage, and each statistic is accompanied by its standard error, which is a measure of the uncertainty due to student sampling and the imputation process. In comparisons of performance across countries, standard errors can be used to assess the statistical significance of the difference between the summary statistics.

The multiple comparison charts presented in the TIMSS 1999 international reports allow the comparison of average performance of a country with that of other participating countries. The significance tests reported in these charts include a Bonferroni adjustment for multiple comparisons that holds to 5% the probability of erroneously declaring the mean of one country to be different from that of another country. The Bonferroni adjustment is necessary because that probability greatly increases as the number of simultaneous comparisons increases.

If repeated samples were taken from two populations with the same mean and variance, and in each one the hypothesis that the two means are significantly different at the α= .05 level (i.e., with 95% confidence) was tested, then it would be expected in about 5% of the comparisons significant differences would be found between the sample means even though no difference exists in the populations. The probability of finding significant differences when none exist (the so-called type I error) is given by α= .05. Conversely, the probability of not making such an error is 1 - α, which in the case of a single test is .95. However, comparing the means of three countries involves three tests (country A versus country B, country B versus country C, and country A versus country C). Since these are independent tests, the probability of not making a type I error in any of these tests is the product of the individual probabilities, which is (1 - α) (1 - α) (1 - α). With α= .05, the overall probability of not making a type I error is only .873, which is considerably less than the probability for a single test. As the number of tests increases, the probability of not making a type I error decreases rapidly, and conversely, the probability of making such an error increases.

Several methods can be used to correct for the increased probability of a type I error while making many simultaneous comparisons. Dunn (1961) developed a procedure that is appropriate for testing a set of a priori hypotheses while controlling the probability that the type I error will occur. In this procedure, the value α is adjusted to compensate for the increase in the probability of making the error (the Dunn-Bonferroni procedure for multiple a priori comparisons; Winer, Brown, and Michels, 1991).

The TIMSS 1999 international reports contain multiple comparison exhibits that show the statistical significance of the differences between all possible combinations of the 38 participating countries. There were (38*37)/2 = 703 possible differences. In the Bonferroni procedure the significance level (α)of a statistical test is adjusted by the number of comparisons that are planned and then looking up the appropriate quantile from the normal distribution. In deciding on the appropriate adjustment of the significance level for TIMSS, it was necessary to decide how the multiple comparison exhibits would most likely be used. A very conservative approach would be to adjust the significance level to compensate for all of the 703 possible comparisons among the 38 countries concerned. However, this risks an error of a different kind, that of concluding that a difference in sample means is not significant when in fact there is a difference in the population means.

Since most users are likely to be interested in comparing a single country with all other countries, rather than in making all possible between-country comparisons at once, the more realistic approach of using the number of countries (minus one) to adjust the significance level was adopted. This meant that the number of simultaneous comparisons to be adjusted for was 37 instead of 703. The critical value for a 95% significance test adjusted for 37 simultaneous comparisons is 3.2049, from the appropriate quantiles from the normal (Gaussian) distribution.

Mean proficiencies were considered significantly different if the absolute difference between them, divided by the standard error of the difference, was greater than the critical value. For differences between countries, which can be considered as independent samples, the standard error of the difference in means was computed as the square root of the sum of the squared standard errors of each mean:

$$se_{diff} = \sqrt{se_1^2 + se_2^2}$$

where se_1 and se_2 are the standard errors of the means. Exhibit 16.1 shows the means and standard errors for mathematics and science used in the calculation of statistical significance. By applying the Bonferroni adjustment, it was possible to state that, for any given row or column of the multiple comparison chart, the differences n countries are statistically significant at the 95% level of confidence.

Exhibit 16.1 Means and Standard Errors for Multiple Comparisons Exhibits

Country	Math		Science	
	Mean	S.E.	Mean	SE
Australia	525.080	4.840	540.258	4.395
Belgium (Flemish)	557.958	3.291	534.858	3.074
Bulgaria	510.591	5.850	518.011	5.355
Canada	530.753	2.460	533.082	2.063
Chile	392.494	4.364	420.372	3.720
Chinese Taipei	585.117	4.033	569.076	4.425
Cyprus	476.382	1.792	460.238	2.350
Czech Republic	519.874	4.176	539.417	4.171
England	496.330	4.150	538.468	4.750
Finland	520.452	2.743	535.207	3.471
Hong Kong, SAR	582.056	4.280	529.547	3.655
Hungary	531.601	3.674	552.381	3.693
Indonesia	403.070	4.896	435.472	4.507
Iran, Islamic Rep.	422.148	3.397	448.003	3.765
Israel	466.336	3.932	468.062	4.936
Italy	479.479	3.829	493.281	3.881
Japan	578.604	1.654	549.653	2.227
Jordan	427.664	3.592	450.343	3.832
Korea, Rep. of	587.152	1.969	548.642	2.583
Latvia (LSS)	505.059	3.435	502.693	4.837
Lithuania	481.567	4.281	488.152	4.105
Macedonia, Rep. of	446.604	4.224	458.095	5.240
Malaysia	519.256	4.354	492.431	4.409
Moldova	469.231	3.883	459.137	4.029
Morocco	336.597	2.573	322.816	4.319
Netherlands	539.875	7.147	544.749	6.870
New Zealand	490.967	5.178	509.634	4.905
Philippines	344.905	5.979	345.229	7.502
Romania	472.440	5.787	471.865	5.823
Russian Federation	526.023	5.935	529.220	6.395
Singapore	604.393	6.259	567.894	8.034
Slovak Republic	533.953	3.959	535.009	3.290
Slovenia	530.113	2.777	533.255	3.218
South Africa	274.503	6.815	242.640	7.850
Thailand	467.377	5.088	482.314	3.983
Tunisia	447.925	2.430	429.512	3.436
Turkey	428.606	4.343	432.951	4.268
United States	501.633	3.971	514.915	4.553

16.4 Comparing Achievement with the International Mean

Many of the data exhibits in the TIMSS 1999 international reports show countries' mean achievement compared with the international mean. Since this results in 38 simultaneous comparisons, the critical value was adjusted to 3.2125 using the Dunn-Bonferroni procedure.

When comparing each country's mean with the international average, TIMSS took into account the fact that the country contributed to the international standard error. To correct for this contribution, TIMSS adjusted the standard error of the difference. The sampling component of the standard error of the difference for country *j* was

$$S_{s\_dif\_j} = \frac{\sqrt{((N-1)^2-1)se_j^2 + \sum_{k=1}^{N} se_k^2}}{N}$$

where

$se_{s\_dif\_j}$ is the standard error of the difference due to sampling when country *j* is compared to the international mean

N is the number of countries

se_k^2 is the sampling standard error for country *k*

se_j^2 is the sampling standard error for country *j*.

The imputation component of the standard error was computed by taking the square root of the imputation variance calculated as follows

$$se_{i\_dif\_j} = \sqrt{\frac{6}{5}Var(d_{1\ldots}d_{l\ldots}d_5)}$$

where d_l is the difference between the international mean and the country mean for plausible value *l*.

Finally, the standard error of the difference was calculated as:

$$se_{dif\_j} = \sqrt{se^2{}_{i\_dif\_j} + se^2{}_{s\_dif\_j}}$$

16.5 Trends in Achievement

TIMSS 1999 was designed to enable comparisons between a country's achievement on the 1995 and 1999 assessments. A total of 26 countries participated at the eighth grade in both assessments. Although all countries had acceptable sampling participation in 1999, three countries – Israel, South Africa, and Thailand – failed to meet sampling guidelines in 1995, and were omitted from the calculation of trends.

When assessing whether eighth-grade achievement had significantly changed from 1995 to 1999, TIMSS applied a Bonferroni correction for 23 simultaneous comparisons.

Of the 23 countries with eighth-grade data in both 1995 and 1999, 17 also had fourth-grade data from 1995.To show how countries' relative performance changed from fourth to eighth grade, TIMSS calculated the significance of the difference between each country's mean and the mean across all 17 countries, adjusting for 17 simultaneous comparisons.

The means and standard errors of the 1995 fourth- and eighth-grade students and the 1999 eighth-grade students for countries included in the trend exhibits from the international reports are shown in Exhibit 16.2 and 16.3 for mathematics and science, respectively.

Exhibit 16.2 **Means and Standard Errors for Mathematics Trend Exhibits**

Country	4th Grade 1995		8th Grade 1995		8th Grade 1999	
	Mean	S.E.	Mean	S.E.	Mean	S.E.
Australia	517.190	2.991	518.872	3.803	525.080	4.840
Belgium (Flemish)	--	--	549.679	5.867	557.958	3.291
Bulgaria	--	--	526.780	5.798	510.591	5.850
Canada	505.693	3.385	520.544	2.174	530.753	2.460
Cyprus	474.930	3.221	467.533	2.237	476.382	1.792
Czech Republic	540.503	3.065	545.551	4.521	519.874	4.176
England	483.980	3.345	497.669	2.975	496.330	4.150
Hong Kong, SAR	556.993	3.986	568.886	6.136	582.056	4.280
Hungary	521.326	3.607	526.626	3.182	531.601	3.674
Iran, Islamic Rep.	386.969	4.992	418.450	3.871	422.148	3.397
Israel	--	--	513.315	6.224	481.609	4.706
Italy	510.028	4.681	491.015	3.370	485.411	4.825
Japan	567.219	1.855	581.069	1.575	578.604	1.654
Korea, Rep. of	580.904	1.802	580.720	1.962	587.152	1.969
Latvia (LSS)	498.939	4.557	488.281	3.578	505.059	3.435
Lithuania	--	--	471.839	4.101	481.567	4.281
Netherlands	549.233	2.959	528.843	6.147	539.875	7.147
New Zealand	469.180	4.367	500.944	4.722	490.967	5.178
Romania	--	--	473.729	4.571	472.440	5.787
Russian Federation	--	--	523.618	5.331	526.023	5.935
Singapore	590.187	4.536	608.593	3.978	604.393	6.259
Slovak Republic	--	--	533.991	3.076	533.953	3.959
Slovenia	525.162	3.174	530.953	2.756	530.113	2.777
South Africa	--	--	277.705	9.212	274.503	6.815
Thailand	--	--	516.216	6.050	467.377	5.088
United States	517.847	2.950	492.318	4.746	501.633	3.971
International Avg.	517.428	0.892	519.413	0.861	521.303	0.922

Exhibit 16.3 Means and Standard Errors for Science Trend Exhibits

	4th Grade 1995		8th Grade 1995		8th Grade 1999	
Country	Mean	S.E.	Mean	S.E.	Mean	S.E.
Australia	541.322	3.630	526.502	4.028	540.258	4.395
Belgium (Flemish)	--	--	532.897	6.391	534.858	3.074
Bulgaria	--	--	545.245	5.203	518.011	5.355
Canada	525.343	3.053	513.988	2.638	533.082	2.063
Cyprus	450.029	3.202	452.012	2.091	460.238	2.350
Czech Republic	531.713	3.038	554.955	4.547	539.417	4.171
England	527.670	3.094	533.348	3.570	538.468	4.750
Hong Kong, SAR	507.824	3.330	509.730	5.785	529.547	3.655
Hungary	507.744	3.405	536.754	3.106	552.381	3.693
Iran, Islamic Rep.	380.184	4.553	462.872	3.628	448.003	3.765
Israel	--	--	508.957	6.349	484.303	5.652
Italy	523.826	4.601	497.248	3.551	497.900	4.752
Japan	553.183	1.765	554.475	1.754	549.653	2.227
Korea, Rep. of	575.571	2.119	545.778	2.045	548.642	2.583
Latvia (LSS)	486.383	4.905	476.156	3.332	502.693	4.837
Lithuania	---	--	463.564	4.049	488.152	4.105
Netherlands	530.332	3.173	541.418	6.029	544.749	6.870
New Zealand	505.117	5.299	510.862	4.858	509.634	4.905
Romania	--	--	470.926	5.134	471.865	5.823
Russian Federation	--	--	522.581	4.486	529.220	6.395
Singapore	523.400	4.803	580.352	5.483	567.894	8.034
Slovak Republic	--	--	531.913	3.309	535.009	3.290
Slovenia	521.966	4.030	540.980	2.794	533.255	3.218
South Africa	--	--	262.941	11.092	242.640	7.850
Thailand	--	--	510.045	4.704	482.314	3.983
United States	541.863	3.258	512.587	5.560	514.915	4.553
International Avg.	513.734	0.888	518.137	0.889	521.211	0.897

Because of differences from 1995 to 1999 in the sampling of student populations, the results for Israel and Italy in exhibits of trend data differ from those containing just 1999 data. In TIMSS 1995, Israel tested only Hebrew-speaking students, while in 1999 the target population included both Hebrew and Arab speaking students. To provide meaningful trend analysis, TIMSS compared 1995 and 1999 using the Hebrew-speaking part of the population only. In Italy, the 1995 assessment sampled students from most but not all provinces, whereas in 1999 all provinces were included. The TIMSS 1999 trend data for Italy represents those provinces that participated in TIMSS 1995 only.

16.6 International Benchmarks of Achievement

In order to provide more information about student achievement, TIMSS identified four points on each of the mathematics and science scales for use as international benchmarks. The Top 10% benchmark was defined as the 90th percentile on the TIMSS scale, computed across all students in all participating countries, with countries weighted in proportion to the size of their eighth-grade population. This point on each scale (mathematics and science) is the point above which the top 10% of students in the 1999 TIMSS assessment scored. The upper quarter benchmark is the 75th percentile on the scale, above which the top 25% of students scored. The median benchmark is the 50th percentile, above which the top half of students scored. Finally, the lower quarter benchmark is the 25th percentile, the point reached by the top 75% of students.

The percentage of students in each country meeting or exceeding the marker levels were reported. In computations of the international benchmarks of achievement, each country was weighted to contribute as many students as there were students in the target population. In other words, each country's contribution to setting the international benchmarks was proportional to the estimated population enrolled in the eighth grade. Exhibit 16.4 shows the contribution of each country to the estimation of the international benchmarks.

Exhibit 16.4 Estimated Enrollment at the Eighth Grade Within Country

Country	Sample Size	Estimated Enrollment
Australia	4032	260130
Belgium (Flemish)	5259	65539
Bulgaria	3272	88389
Canada	8770	371062
Chile	5907	208910
Chinese Taipei	5772	310429
Cyprus	3116	9786
Czech Republic	3453	119462
England	2960	552231
Finland	2920	59665
Hong Kong, SAR	5179	79097
Hungary	3183	111298
Indonesia	5848	1956221
Iran, Islamic Rep.	5301	1655741
Israel	4195	81486
Italy	3328	548711
Japan	4745	1416819
Jordan	5052	89171
Korea, Rep. of	6114	609483
Latvia (LSS)	2873	18122
Lithuania	2361	40452
Macedonia, Rep. of	4023	30280
Malaysia	5577	397762
Moldova	3711	59956
Morocco	5402	347675
Netherlands	2962	198144
New Zealand	3613	51553
Philippines	6601	1078093
Romania	3425	2596
Russian Federation	4332	2057413
Singapore	4966	41346
Slovak Republic	3497	72521
Slovenia	3109	23514
South Africa	8146	844706
Thailand	5732	727087
Tunisia	5051	139639
Turkey	7841	618058
United States	9072	3336295

If all countries had the same distribution of student achievement, approximately 10% of students within each country would be above the 90th percentile in the international distribution, regardless of the country's population size. That this is not the case, and that countries vary considerably, is evident from the fact that, 46% of students in Singapore reached the top 10% benchmark, compared to less than 1% in Tunisia, the Philippines, South Africa, and Morocco.

Because of the imputation technology used to derive the proficiency scores, the international benchmarks had to be computed once for each of the five plausible values, and the results averaged to arrive at the final figure. The standard errors presented in the exhibits are computed taking into account the sampling design as well as the variance due to imputation. The international benchmarks are presented in Exhibit 16.5 and 16.6 for mathematics and science, respectively.

Exhibit 16.5 **International Benchmarks of Mathematics Achievement for the Eighth Grade**

Proficiency Score	25th Percentile	50th Percentile	75th Percentile	90th Percentile
Plausible Value 1	396.86	479.20	554.49	615.15
Plausible Value 2	395.76	478.79	554.74	615.37
Plausible Value 3	395.62	478.56	554.83	616.23
Plausible Value 4	394.57	478.09	554.03	615.02
Plausible Value 5	396.30	479.10	554.56	615.76
Mean Plausible Value	395.82	478.75	554.53	615.51

Exhibit 16.6 **International Benchmarks of Science Achievement for the Eighth Grade**

Proficiency Score	25th Percentile	50th Percentile	75th Percentile	90th Percentile
Plausible Value 1	409.03	487.76	558.66	617.01
Plausible Value 2	409.87	487.61	557.60	615.88
Plausible Value 3	410.38	488.04	557.27	616.12
Plausible Value 4	410.05	487.54	557.47	615.82
Plausible Value 5	410.87	487.59	557.79	615.88
Mean Plausible Value	410.04	487.71	557.76	616.14

16.7 Gender Differences within Countries

TIMSS reported gender differences in overall student achievement in mathematics and science overall, as well as in content areas. Gender differences were presented in an exhibit showing mean achievement for males and females and the differences between them, with an accompanying graph indicating whether the difference was statistically significant. The significance test was adjusted for multiple comparisons based on the number of countries presented.

Because in most countries males and females attend the same schools, the samples of males and females cannot be treated as independent for the purpose of statistical tests. Accordingly, TIMSS used a jackknife procedure applicable to correlated samples for estimating the standard error of the male-female difference. This involves computing the differences between boys and girls once for each of the 75 replicate samples, and five more times, once for each plausible value, as described in Chapter 12.

16.8 Relative Performance by Content Areas

In addition to performance in mathematics and science overall, it was of interest to see how countries performed on the content areas relative to performance on the subject overall. Five content areas in mathematics and six in science were used in this analysis. Relative performance on the content areas was examined separately for the two subjects. TIMSS 1999 computed the average across content area scores for each country, and then displayed country performance in each content area as the difference between that average and the overall average. Confidence intervals were estimated for each difference.

In order to do this, TIMSS computed the vector of average proficiencies for each of the content areas on the test, and joined each column vector to form a matrix called R_{ks}, where a row contains the average proficiency score for country *k* on scale *s* for a specific subject. This R_{ks} matrix had also a "zeroth" row and column. The elements in r_{k0} contains the average of the elements on the *k*th row of the R_{ks} matrix. These are the country averages across the content areas. The elements in r_{0s} contains the average of the elements of the *s*th column of the R_{ks} matrix. These are the content area averages across all countries. The element r_{00} contains the overall average for the elements in vector r_{0j} or r_{k0}. Based on this information, the matrix I_{ks} was constructed in which the elements are computed as

$$i_{ks} = r_{ks} + r_{00} - r_{0s} - r_{k0}$$

Each of these elements can be considered as the interaction between the performance of country *k* on content area *s*. A value of zero for an element i_{ks} indicates a level of performance for country *k* on content area *s* that would be expected given its performance on other content areas and its performance relative to other countries on that content area. A negative value for an element i_{ks} indicates a performance for country *k* on content area *s* lower than would be expected on the basis of the country's overall performance. A positive value for an element i_{ks} indicates a better than expected performance for country *k* on the content areas. This procedure was applied to each of the 5 plausible values and the results were averaged.

To construct confidence intervals the standard error for each content area in each country first had to be estimated. These were then combined with a Bonferroni adjustment, based on the number of content areas. The imputation portion of the error was obtained from combining the results from the five calculations, one with each separate plausible value.

To compute the sampling portion of the standard error, the vector of average proficiency was computed for each of the country replicates for each content area on the test. For each country and each content area 75 replicates were created.[1] Each replicate was randomly reassigned to one of 75 sampling zones or replicates (h). These column vectors were then joined to form a new set of matrices each called R^h_{ks}, where a row contains the average proficiency for country k on content area s for a specific subject, for the hth international set of replicates. Each of these R^h_{ks} matrices has also a "zeroth" row and column. The elements in r^h_{k0} contain the average of the elements on the kth row of the R^h_{ks} matrix. These are the country averages across the content areas. The elements in r^h_{0s} contain the average of the elements of the sth column of the R^h_{ks} matrix. These are the content area averages across all countries. The element r^h_{00} contains the overall average for the elements in vector r^h_{0j} or r^h_{k0}. Based on this information the set of matrices $R^h_{ks'}$ were constructed, in which the elements were computed as

$$i^h_{ks} = r^h_{ks} + r^h_{00} - r^h_{0s} - r^h_{k0}$$

The JRR standard error is then given by the formula

$$jse_{r_{ks}} = \sqrt{\sum_h (i_{ks} - i^h_{ks})}^2$$

○○○

[1] In countries where the were less than 75 jackknife zones, 75 replicates were also created by assigning the overall mean to the as many replicates as were necessary to have 75.

The overall standard error was computed by combining the JRR and imputation variances. A relative performance was considered significantly different from the expected if the 95% confidence interval built around it did not include zero. The confidence interval for each of the i_{ks} elements was computed by adding to and subtracting from the i_{ks} element its corresponding standard error multiplied by the critical value for the number of comparisons.

The critical values were determined by adjusting the critical value for a two-tailed test, at the alpha 0.05 level of significance for multiple comparisons according the Dunn-Bonferroni procedure. The critical value for mathematics, with five content scales, was 2.5758, and for science with six content scales, was 2.6383.

16.9 Percent Correct for Individual Items

To portray student achievement as fully as possible, the TIMSS 1999 international reports present many examples of the items used in the TIMSS 1999 tests, together with the percentage of students in each country responding correctly to the item. This percentage was based on the total number of students tested on the items. Omitted and not-reached items were treated as incorrect. For multiple-choice items the percentage was the weighted percentage of students that answered the item correctly. For free-response items with more than one score level, it was the weighted percentage of students that achieved the highest score possible.

When the% correct for example items was computed, student responses were classified in the following way. For multiple-choice items, a response to item j was classified as correct (C_j) when the correct option was selected, incorrect (W_j) when the incorrect option or no option was selected, invalid (I_j) when two or more options were selected, not reached (R_j) when it was assumed that the student stopped working on the test before reaching the question, and not administered (A_j) when the question was not included in the student's booklet or had been mistranslated or misprinted. For free-response items, student responses to item *j* were classified as correct (C_j) when the maximum number of points was obtained, incorrect (W_j) when the wrong answer or an answer not worth all the points in the question was given, invalid (N_j) when the response was not legible or interpretable or was simply left blank, not reached (R_j) when it was determined that the student stopped working on the test before reaching the question, and not administered *(A_j)* when the question was not included in the student's booklet or had been mistranslated or misprinted. The% correct for an item *(P_j)* was computed as

$$P_j = \frac{c_j}{c_j + w_j + i_j + r_j + n_j}$$

where c_j, w_j, i_j, r_j and n_j are the weighted counts of the correct, wrong, invalid, not reached, and not interpretable responses to item *j,* respectively.

16.10 The Test-Curriculum Matching Analysis

TIMSS 1999 developed international tests of mathematics and science that reflect, as far as possible, the various curricula of the participating countries. The subject matter coverage of these tests was reviewed by the TIMSS 1999 Subject Matter Item Replacement Committee, which consisted of mathematics and science educators and practitioners from around the world, and the tests were approved for use by the National Research Coordinators (NRCs) of the participating countries. Although every effort was made in TIMSS 1999 to ensure the widest possible subject matter coverage, no test can measure all that is taught or learned in every participating country. The question therefore arises how well the items on the tests match the curricula of the participating countries. To address this issue, TIMSS 1999 asked each country to indicate which items on the tests, if any, were inappropriate to its curriculum. For each country, TIMSS 1999 then took the list of remaining items and computed the average percentage correct on those items for that country and all other countries. This allowed each country to select only those items on the tests that they would like included, and to compare the performance of their students on those items with that of the students in the other participating countries. However, in addition to comparing the performance of all countries on the set of items chosen by each country, the Test-Curriculum Matching Analysis (TCMA) also shows each country's performance on the items chosen by each of the other countries. In these analyses, each country was able to see not only the performance of all countries on the items appropriate for its curriculum, but also the performance of its students on items judged appropriate for the curriculum in other countries. The analytical method of the TCMA is described in Beaton and Gonzalez (1997).

The TCMA results show that the TIMSS 1999 tests provide a reasonable basis for comparing achievement across the participating countries. The analysis shows that omitting items considered by one country to be difficult for their students tends to improve the results for that country, but tends to improve the results for all other countries also, so that the overall pattern of relative performance is largely unaffected.

References

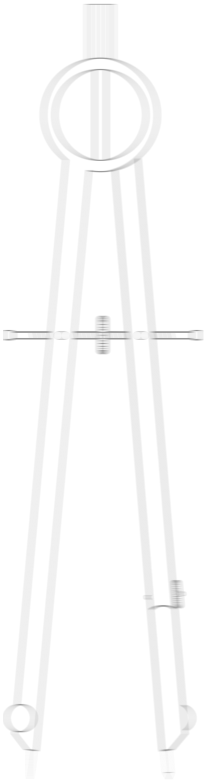
Beaton, A. E. & Gonzalez, E. J. (1997). "TIMSS Test-Curriculum Matching Analysis" in Martin, M.O. and Kelly, D.L. (Eds.), *TIMSS technical report, volume II: Implementation and Analysis.* Chestnut Hill, MA: Boston College.

Dunn, O.J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association, 56,* 52-64.

Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., Gregory, K.D., Smith, T.A., Chrostowski, S.J., Garden, R.A., & O'Connor, K.M. (2000). *TIMSS 1999 International Science Report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade.* Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., Gregory, K.D., Garden, R.A., O'Connor, K.M., Chrostowski, S.J., & Smith, T.A. (2000). *TIMSS 1999 International Mathematics Report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade.* Chestnut Hill, MA: Boston College.

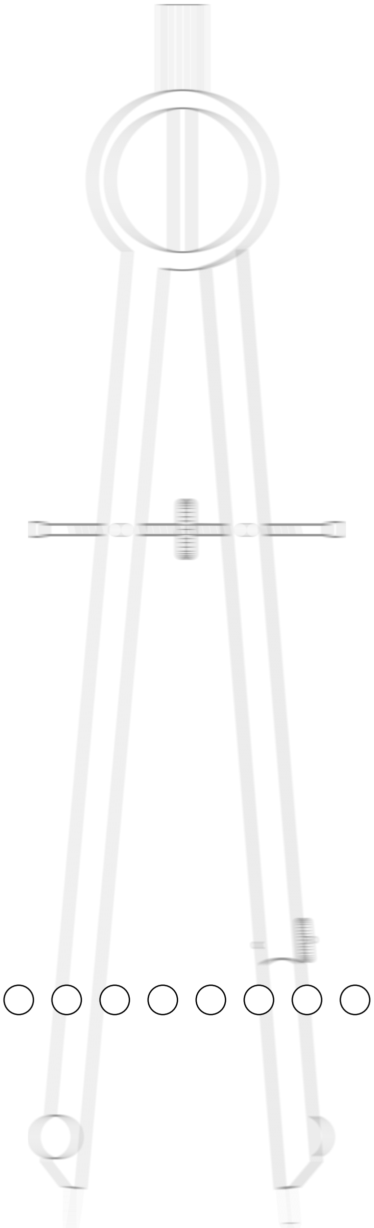
Winer, B.J., Brown, D.R., & Michels, K.M. (1991). *Statistical principles in experimental design.* New York: McGraw Hill.



17

Reporting Questionnaire Data

Teresa A. Smith



17 Reporting Questionnaire Data

Teresa A. Smith

17.1 Overview

This chapter documents the analysis and reporting procedures used for the background questionnaire data in producing the TIMSS 1999 international reports. In particular, it provides an overview of the consensus process used to develop the report outlines and prototype exhibits; discusses the development and computation of indices based on student, teacher, and school background variables; presents the approach used in reporting trends in background data; describes special considerations in reporting the student, teacher, school, and country questionnaire data; and explains how TIMSS 1999 handled issues of nonresponse in reporting these data.

17.2 Background Questionnaires

As described in chapter 4, TIMSS 1999 used four types of background questionnaires to gather information at various levels of the educational system:

1. Curriculum questionnaires that addressed issues of curriculum design and curricular emphasis in mathematics and science were completed by National Research Coordinators
2. A school questionnaire that provided information about school staffing and facilities, as well as curricular and instructional arrangements, was completed by school principals
3. Teacher questionnaires completed by mathematics and science teachers, provided information about their backgrounds, attitudes, and teaching activities and approaches
4. Students completed a student questionnaire providing information about their home backgrounds and attitudes, and their experiences in mathematics and science classes; there were two versions: a general science version intended for systems where science is taught as a single integrated subject, and a version intended for systems where science is taught as separate subjects (biology, chemistry, earth science, and physics)

17.3 TIMSS 1999 Reporting Approach

As in TIMSS 1995, the TIMSS 1999 results were reported separately by subject area, with the mathematics and science results appearing in separate volumes (Mullis et al., 2000; Martin et al., 2000). The TIMSS 1999 reports contain four chapters devoted to the questionnaire data, dealing with students' backgrounds and attitudes, the nature and coverage of the curriculum, teachers and instruction, and school contexts for learning. The 1999 reports included a number of innovations. First, summary indices based on some of the student, teacher, and school background data were presented to focus the reports more closely on issues related to good educational practice. Second, since TIMSS 1999 was designed to measure trends in student achievement and in the related educational contexts for learning and instruction, trends were presented in cases where comparable background data were obtained in both assessments. Third, the report was designed to give prominence to the background indices, with displays of secondary importance relegated to a resource reference section at the end of the reports.

17.3.1 Summary Indices from Background Data

In an effort to summarize the information obtained from the background questionnaires concisely and focus attention on educationally relevant support and practice, TIMSS sometimes combined information to form an index that was more global and reliable than the component questions (e.g., students' home educational resources and attitudes towards mathematics or science; teachers' emphasis on reasoning and problem-solving, and confidence in their preparation to teach mathematics or science; availability of school resources for mathematics or science instruction). According to the responses of students, their teachers or their schools, students were placed in a "high," "medium," or "low" category for the index, with the high level being set so that it corresponds to conditions or activities generally associated with higher academic achievement. For example, a three-level index of home educational resources was constructed from students' responses to three questions: number of books in the home, educational aids in the home (computer, study desk/table for own use, dictionary), and parents' education. Students were assigned to the high level if they reported having more than 100 books, having all three educational aids, and that at least one parent finished university. Students at the low level reported having

25 or fewer books in the home, not all three educational aids, and some secondary or less to be the highest level of education for either parent. Students with all other response combinations were assigned to the middle category.

The 17 indices computed for the TIMSS 1999 report are listed in Exhibit 17.1, which identifies the name of the index; the label used to identify it in the international report and database; the mathematics or science exhibit where the index data were reported; and the method used to compute the index.

Exhibit 17.1 Summary Indices from Background Data in the TIMSS-1999 International Report

Name of Index	Label	Exhibit[a]	Analysis Method
Index of Home Educational Resources	HER	4.1 (M) 4.1 (S)	Index based on students' responses to three questions about home educational resources: number of books in the home; educational aids in the home (computer, study desk/table for own use, dictionary); parents' education. High level indicates more than 100 books in the home; all three educational aids; and either parent's highest level of education is finished university. Low level indicates 25 or fewer books in the home; not all three educational aids; and both parents' highest level of education is some secondary or less or is not known. Medium level includes all other possible combinations of responses. Response categories were defined by each country to conform to their own educational system and may not be strictly comparable across countries.
Index of Out-of-School Study Time	OST	4.5 (M) 4.5 (S)	Index based on students' responses to three questions about out-of-school study time: time spent after school studying mathematics or doing mathematics homework; time spent after school studying science or doing science homework; time spent after school studying or doing homework in school subjects other than mathematics and science. Number of hours based on: no time = 0, less than 1 hour = 0.5, 1-2 hours = 1.5, 3-5 hours = 4, more than 5 hours = 7. High level indicates more than three hours studying all subjects combined. Medium level indicates more than one hour to three hours studying all subjects combined. Low level indicates one hour or less studying all subjects combined.
Index of Students' Self-Concept in Mathematics	SCM	4.8 (M)	Index based on students' responses to five statements about their mathematics ability: 1) I would like mathematics much more if it were not so difficult; 2) although I do my best, mathematics is more difficult for me than for many of my classmates; 3) nobody can be good in every subject, and I am just not talented in mathematics; 4) sometimes, when I do not understand a new topic in mathematics initially, I know that I will never really understand it; 5) mathematics is not one of my strengths. High level indicates student disagrees or strongly disagrees with all five statements. Low level indicates student agrees or strongly agrees with all five statements. Medium level includes all other possible combinations of responses.
Index of Students' Self-Concept in the Sciences*	SCS-G SCS-E SCS-B SCS-P SCS-C	4.8 (S)	Index based on students' responses to four statements about their science ability: 1) I would like science much more if it were not so difficult; 2) although I do my best, science is more difficult for me than for many of my classmates; 3) nobody can be good in every subject, and I am just not talented in science; 4) science is not one of my strengths. In countries where science is taught as separate subjects, students were asked about each subject area separately. High level indicates student disagrees or strongly disagrees with all four statements. Low level indicates student agrees or strongly agrees with all four statements. Medium level includes all other possible combinations of responses.

Name of Index	Label	Exhibit[a]	Analysis Method
Index of Positive Attitudes Towards Mathematics	PATM	4.10 (M)	Index based on students' responses to five statements about mathematics: 1) I like mathematics; 2) I enjoy learning mathematics; 3) mathematics is boring (reversed scale); 4) mathematics is important to everyone's life; 5) I would like a job that involved using mathematics. Average is computed across the five items based on a 4-point scale: 1 = strongly negative; 2 = negative; 3 = positive; 4 = strongly positive. High level indicates average is greater than 3. Medium level indicates average is greater than 2 and less than or equal to 3. Low level indicates average is less than or equal to 2.
Index of Positive Attitudes Towards the Sciences*	PATS-G PATS-E PATS-B PATS-P PATS-C	4.10 (S)	Index based on students' responses to five statements about science: 1) I like science; 2) I enjoy learning science; 3) science is boring (reversed scale); 4) science is important to everyone's life; 5) I would like a job that involved using science. Average is computed across the five items based on a 4-point scale: 1 = strongly negative; 2 = negative; 3 = positive; 4 = strongly positive. In countries where science is taught as separate subjects, students were asked about each subject area separately. High level indicates average is greater than 3. Medium level indicates average is greater than 2 and less than or equal to 3. Low level indicates average is less than or equal to 2.
Index of Confidence in Preparation to Teach Mathematics	CPTM	6.3 (M)	Index based on teachers' responses to 12 questions about how prepared they feel to teach different mathematics topics based on a 3-point scale: 1 = not well prepared; 2 = somewhat prepared; 3 = very well prepared. Average is computed across the 12 items for topics for which the teacher did not respond "do not teach". High level indicates average is greater than or equal to 2.75. Medium level indicates average is greater than or equal to 2.25 and less than 2.75. Low level indicates average is less than 2.25.
Index of Confidence in Preparation to Teach Science	CPTS	6.3 (S)	Index based on teachers' responses to 10 questions about how prepared they feel to teach different science topics (see reference exhibit R3.2) based on a 3-point scale: 1 = not well prepared; 2 = somewhat prepared; 3 = very well prepared. Average is computed across the 10 items for items for which the teacher did not respond "do not teach". High level indicates average is greater than or equal to 2.75. Medium level indicates average is greater than or equal to 2.25 and less than 2.75. Low level indicates average is less than 2.25.
Index of Teachers' Emphasis on Scientific Reasoning and Problem-Solving	ESRPS	6.12 (S)	Index based on teachers' responses to five questions about how often they ask students to: 1) explain the reasoning behind an idea; 2) represent and analyze relationships using tables, charts, graphs; 3) work on problems for which there is no immediately obvious method of solution; 4) write explanations about what was observed and why it happened; 5) put events or objects in order and give a reason for the organization. Average is computed across the five items based on a 4-point scale: 1 = never or almost never; 2 = some lessons; 3 = most lessons; 4 = every lesson. High level indicates average is greater than or equal to 3. Medium level indicates average is greater than or equal to 2.25 and less than 3. Low level indicates average is less than 2.25.
Index of Teachers' Emphasis on Mathematics Reasoning and Problem-Solving	EMRPS	6.13 (M)	Index based on teachers' responses to four questions about how often they ask students to: 1) explain the reasoning behind an idea; 2) represent and analyze relationships using tables, charts, or graphs; 3) work on problems for which there is no immediately obvious method of solution; 4) write equations to represent relationships. Average is computed across the four items based on a 4-point scale: 1 = never or almost never; 2 = some lessons; 3 = most lessons; 4 = every lesson. High level indicates average is greater than or equal to 3. Medium level indicates average is greater than or equal to 2.25 and less than 3. Low level indicates average is less than 2.25.

Name of Index	Label	Exhibit[a]	Analysis Method
Index of Emphasis on Conducting Experiments in Science Classes[*]	ECES-G ECES-E ECES-B ECES-P ECES-C	6.14 (S)	Index based on teachers' reports on the percentage of time they spend demonstrating experiments; teachers' reports on the percentage of time students spend conducting experiments; students' reports on how often the teacher gives a demonstration of an experiment in science lessons; students' reports on how often they conduct an experiment or practical investigation in class. In countries where science is taught as separate subjects, students were asked about each subject area separately, and only teachers who teach a particular subject are included in the index shown for that subject. High level indicates teacher reported that at least 25% of class time is spent on the teacher demonstrating experiments or students conducting experiments, and the student reported that the teacher gives a demonstration of an experiment or the student conducts an experiment or practical investigation in class almost always or pretty often. Low level indicates the teacher reported that less than 10% of class time is spent on the teacher demonstrating experiments or students conducting experiments, and student reported that the teacher gives a demonstration of an Experiment and the student conducts an experiment or practical investigation in class once in a while or never. Medium level includes all other possible combinations of responses.
Index of Emphasis on Calculators in Mathematics Class	ECMC	6.16 (M)	Index based on students' reports of the frequency of using calculators in mathematics lessons and teachers' reports of students' use of calculators in mathematics class for five activities: checking answers; tests and exams; routine computation; solving complex problems; and exploring number concepts. High level indicates the student reported using calculators in mathematics lessons almost always or pretty often, and the teacher reported students use calculators at least once or twice a week for any of the tasks. Low level indicates the student reported using calculators once in a while or never, and the teacher reported students use calculators never or hardly ever for all of the tasks. Medium level includes all other possible combinations of responses.
Index of Teachers' Emphasis on Science Homework	ESH	6.18 (S)	Index based on teachers' responses to two questions about how often they usually assign science homework and how many minutes of science homework they usually assign students. High level indicates the assignment of more than 30 minutes of homework at least once or twice a week. Low level indicates the assignment of less than 30 minutes of homework less than once a week or never assigning homework. Medium level includes all other possible combinations of responses.
Index of Teachers' Emphasis on Mathematics Homework	EMH	6.21 (M)	Index based on teachers' responses to two questions about how often they usually assign mathematics homework and how many minutes of mathematics homework they usually assign students. High level indicates the assignment of more than 30 minutes of homework at least once or twice a week. Low level indicates the assignment of less than 30 minutes of homework less than once a week or never assigning homework. Medium level includes all other possible combinations of responses.
Index of Availability of School Resources for Mathematics Instruction	ASRMI	7.1 (M)	Index based on schools' average response to five questions about shortages that affect general capacity to provide instruction (instructional materials; budget for supplies; school buildings and grounds; heating/cooling and lighting systems; instructional space), and the average response to five questions about shortages that affect mathematics instruction (computers; computer software; calculators; library materials; audio-visual resources). High level indicates that both shortages, on average, affect instructional capacity none or a little. Medium level indicates that one shortage affects instructional capacity none or a little and the other shortage affects instructional capacity some or a lot. Low level indicates that both shortages affect instructional capacity some or a lot.
Index of Availability of School Resources for Science Instruction	ASRSI	7.1 (S)	Index based on schools' average response to five questions about shortages that affect general capacity to provide instruction (instructional materials; budget for supplies; school buildings and grounds; heating/cooling and lighting systems; instructional space), and the average response to six questions about shortages that affect science instruction (laboratory equipment and materials; computers; computer software; calculators; library materials; audio-visual resources). High level indicates that both shortages, on average, affect instructional capacity none or a little. Medium level indicates that one shortage affects instructional capacity none or a little and the other shortage affects instructional capacity some or a lot. Low level indicates that both shortages affect instructional capacity some or a lot.

Name of Index	Label	Exhibit[a]	Analysis Method
Index of Good School and Class Attendance	SCA	7.5 (M) 7.5 (S)	Index based on schools' responses to three questions about the seriousness of attendance problems in school: arriving late at school; absenteeism; skipping class. High level indicates that all three behaviors are reported to be not a problem. Low level indicates that two or more behaviors are reported to be a serious problem, or two behaviors are reported to be minor problems and the third a serious problem. Medium level includes all other possible combinations of responses.

a Exhibit number in the international report where data based on the index were presented. An (M) indicates mathematics report; (S) indicates science report.

* Separate indices were computed for general/integrated science (G), earth science (E), biology (B), physics (P), and chemistry (C)

The exhibit that displays each index shows the percentages of students at each level of the index, together with their mathematics or science achievement. In addition, the percentage at the high level was displayed graphically, with the countries ranked in order. For some of the sciences indices, the results were presented in separate panels for each science subject. The data for the component questions that made up the indices were usually presented in a section of the resource reference.

17.3.2 Reporting Trends in Background Data

Wherever possible and relevant, trend data were presented for the background indices as well as for other key variables from the background questionnaires. The exhibits containing trend data include all countries that participated in both the 1995 and 1999 assessments and that had internationally comparable data for the questions asked in both years.[1] In reporting trends for indices, the percentages of students in 1995 and 1999 at the high, medium, and low level of the index were presented, as were differences in the percentages from 1995 to 1999. Trend exhibits for some other key background variables presented percentages or average values for a number of reporting categories. In these exhibits, only the percentage of students in 1999 (or the average across students in 1999) and the corresponding difference between 1995 and 1999 were presented. This format was used most often in the science report, where the results for five science subject categories (general/integrated science, earth science, biology, physics, and chemistry) were presented in a single display.

○○○

[1] Although they were included in the trend exhibits based only on achievement data, Bulgaria and South Africa were excluded from trend exhibits due to problems with their 1995 background data.

All trend exhibits indicate the statistical significance of the difference between 1995 and 1999 in percentage of students or average across students. The significance tests reported in these exhibits are adjusted for multiple comparisons based on a Bonferroni procedure that holds to five percent the overall probability of erroneously declaring as significant any of the pair-wise differences across or within countries. Therefore, the requirement for statistical significance of each pair-wise difference is more stringent than that required for a simple comparison of two percentages without adjusting for multiple comparisons, and fewer statistically significant differences are identified.[2] In all exhibits based on background data, standard errors were provided for major statistics, and these may be used to construct unadjusted confidence intervals, if required.

17.3.3 Resource Reference

In the TIMSS 1999 reports, the most important background data displays are provided in the body of the text, with supporting exhibits included in a resource reference section for each chapter. The resource reference provides support for the main report chapters, containing detailed information about the component variables that went into computing the indices and on other variables of secondary interest, particularly some that were included in the 1995 report. In addition, trend data for component variables of an index were sometimes presented in the resource reference. For example, the index of home educational resources was supported by five exhibits presenting the component variables used to compute the index: number of books in the home; educational aids in the home; highest level of education of either parent; trends in educational aids in the home; and trends in number of books in the home. In addition, an exhibit in the resource reference section described country modifications in the definitions of educational levels for the parents' education.

17.4 Development of the International Reports

Like TIMSS in 1995, TIMSS 1999 was designed to investigate student learning of mathematics and science and the way in which aspects of the education systems, the schools, the teachers, and the students themselves relate to the learning opportunities and experiences of individual students. In trying to assess the influences on student learning put forth by the model as key determi-

[2] See Chapter 16 for a description of the Bonferroni correction.

nants of achievement – the system, schools, teachers, and students – the TIMSS International Study Center included in the initial report outlines as much information as possible about the following major areas:

- The curricular context of students' learning
- System-level characteristics
- School contexts
- Teacher qualifications and characteristics
- Instructional organization and activities
- Students' backgrounds and attitudes towards mathematics and science

Within each category aspects identified as key features of the educational process were included in the outlines as proposed subsections.

The goal of the international reports was to present as much descriptive information about the contexts for learning mathematics and science as possible without overburdening the reader. Indices based on variables from the TIMSS 1999 background questionnaires were proposed to summarize information. The TIMSS 1995 reports were reviewed to identify other key variables that should be included in 1999. Trend analyses were proposed for all indices and other key variables where comparable data were obtained in 1995 and 1999.

Analyses required to present indices and other descriptive data were planned and prototype exhibits prepared. This required a careful review of the questionnaires, detailed documentation of the variables and response categories, the development of general analysis plans (including the cutoffs for high, medium, and low levels of indices), and the specification of any country-specific modified analyses required to account for national adaptations. These plans were documented in analysis notes for each proposed exhibit.

The analysis plans, report outlines, and prototype exhibits were drafted by the International Study Center and underwent a lengthy review involving the National Research Coordinators and project staff. Consensus was then built among the constituents as to the reporting priorities for the first international reports including which indices and variables should be reported, how much information should be included, and which trend tables to present. The analysis plans, outlines, and prototype exhibits were again reviewed at the fifth meeting of the TIMSS 1999 National Research Coordinators in Kuala Lumpur, Malaysia, in October 1999, and then at the sixth meeting in Antalya, Turkey, in February 2000. Following each meeting, the material was revised and updated to reflect the ideas and suggestions of the coordinators. Some exhibits were deleted or added, and some of the analyses or presentational modes were modified.

After the data for all countries became available for analysis in the spring of 2000, the International Study Center, with support from the IEA Data Processing Center, conducted the analyses documented in the analysis notes. NRCs were given the opportunity to review the first draft tables in light of their national data in a mailout review in May, 2000, and to comment on the quality and consistency of their background data. Feedback from NRCs was incorporated into the draft exhibits prepared for international review at the final meeting of National Research Coordinators held in August, 2000, in Helsinki, Finland. As a result of this review, some tables and figures were modified and some deleted. For example, the cutoffs for high, medium, and low levels of some indices were changed, and for some categorical variables, categories were modified to reflect the distribution of student responses. Further refinements were made following that meeting and final drafts were sent to NRCs in September, 2000. Final revisions were made in October and November, and the two reports were published in December 2000 (Mullis et al., 2000; Martin et al., 2000).

17.5 Reporting Student Background Data

Reporting the data from the student questionnaire was fairly straightforward. Most of the tables in the international reports present weighted percentages of students in each country for each response category, together with the mean achievement (mathematics or science) of those students. International averages are also displayed for each category. In general, jackknife standard errors accompany the statistics reported.[3] In addition to the exhibits showing percentages of students overall, the international reports include some information separately by gender. For gender-based exhibits, the percentages of boys and girls in each category were displayed, and the statistical significance of the difference between genders was indicated.

[3] See Chapter 12 for a description of the jackknife methodology.

Reporting student attitudes, self-perceptions, and activities related to science was complicated by the fact that in some countries, science is taught as a general, integrated subject, while in others the fields of science - earth science, physics, chemistry, and biology - are taught as separate subjects. Countries could choose the appropriate version of the student questionnaire: the general science version or the version for countries with separate science subjects. The exhibits showing results for questions that differed in the two versions have separate sections that display the data for countries that administered each one.

In the exhibits based on questions asked about the separate sciences, data were presented in five panels corresponding to the types of science subjects included in the international version of the student questionnaires: general/integrated science and the four separate science subjects (earth science, life science, physics, and chemistry). Countries appear in the appropriate panels. In some countries, earth science or chemistry was not applicable for the eighth grade, and these countries were excluded from these panels. Also, in some countries combined courses such as physical science (physics/chemistry) or natural science (biology/earth science) were taught. In these cases, separate questions were still asked about separate science subjects (earth science, biology, physics, and chemistry), and the student data were reported in all panels. An exception was the Netherlands, where students were asked about earth science, biology, and physics/chemistry. The data for the physics/chemistry questions for this country were presented in the physics panel, and no data were presented in the chemistry panel.

In TIMSS 1999, 23 countries administered the general version of the student questionnaire, and 15 countries the separate science subject version. Table 17.2 lists the countries administering the general and separate science versions and indicates which science subjects were taught in each of the latter. In two countries, Chinese Taipei and Indonesia, the sciences were taught as separate subjects but students receive a single science course grade, and so the general version of the student questionnaire was administered. In both countries, student data were displayed in the general/integrated science panel.

Table 17.2 **Countries that Administered the General Science and Separate Science Subject Versions of the Student Questionnaire**

Country	General Version	Separate Science Version			
	General / Integrated Science	Earth Science	Biology	Physics	Chemistry
Australia	●				
Belgium (Flemish)		●	●	●	
Bulgaria		●	●	●	●
Canada	●				
Chile	●				
[a] Chinese Taipei	●				
Cyprus	●				
Czech Republic		●	●	●	●
England	●				
Finland		●	●	●	●
Hong Kong, SAR	●				
Hungary		●	●	●	●
[b] Indonesia	●				
Iran, Islamic Republic	●				
Israel	●				
Italy	●				
Japan	●				
Jordan	●				
Korea, Republic of	●				
Latvia			●	●	●
Lithuania			●	●	●
Macedonia, Republic of		●	●	●	●
Malaysia	●				
Moldova		●	●	●	●
Morocco		●	●	●	●
[c] Netherlands		●	●	●	●
New Zealand	●				
Philippines	●				
Romania		●	●	●	●
Russian Federation		●	●	●	●
Singapore	●				
Slovak Republic		●	●	●	●
Slovenia			●	●	●
South Africa	●				
Thailand	●				
Tunisia	●				
Turkey	●				
United States	●				

[a] Chinese Taipei: separate sciences are taught starting in grade 7, with biology in grade 7 and physics/chemistry in grade 8. Since the students in the target grade take only one science course (physics/chemistry), the general version of the questionnaire was administered and students were asked about 'natural science', which would pertain to the physics/chemistry course in grade 8"

[b] Indonesia: students are taught 'IPA science' by separate biology and physics teachers, but students receive a single composite grade. The general version of the questionnaire was used, and students were asked about 'IPA science'.

[c] Netherlands: students were asked questions about integrated physics/chemistry; data for questions pertaining to physics/chemistry were reported in the physics panel.

17.6 Reporting Teacher Background Data

In the eighth grade, different teachers generally teach mathematics and science. Accordingly, there was a questionnaire for mathematics teachers and another for science teachers, the two having some general questions in common but different subject-matter-related questions.The procedure was to sample a mathematics class from each participating school, administer the test to those students, and ask all their mathematics and science teachers to complete a teacher questionnaire. In countries with different teachers for each of the science subjects, this included all science teachers of the students in the sampled classes.[4] The teacher questionnaire was divided into two sections: Section A asked about teachers' general background and Section B asked class-specific questions about instructional practices. Where teachers taught more than one mathematics or science class to the sampled students, they were to complete only one Section A but a separate Section B for each class taught. Thus, the information about instruction was tied directly to the students tested and the specific mathematics and science classes in which they were taught.

Because the sampling for the teacher questionnaires was based on participating students, these responses do not necessarily represent all of the teachers of the target grade in each of the TIMSS countries. Rather, they represent teachers of the representative samples of students assessed. It is important to note that in the international reports, the student is always the unit of analysis, even when information from the teacher questionnaires is being reported. That is, the data presented are the percentages of *students* whose teachers reported various characteristics or instructional strategies. Using the student as the unit of analysis makes it possible to describe the instruction received by representative samples of students. Although this approach may provide a different perspective from that obtained by simply collecting information from teachers, it is consistent with the TIMSS goals of illuminating students' educational contexts and performance.

Data collected from mathematics teachers were presented in the international mathematics report, and those collected from science teachers in the science report. As in reporting the student background data, most exhibits based on teacher responses dis-

○○○

[4] In Slovenia and the Slovak Republic, background questionnaires were administered to only one of the separate science subject area teachers for the sampled mathematics classes. As a result, science teacher background data are not available for more than half of the relevant science teachers, and Slovenia and the Slovak Republic are not included in the exhibits based on science teacher data.

played percentages of students in different categories for each country and on average internationally. Where possible and relevant, the average achievement of students was reported for each category in an exhibit to show the relationship with achievement. Trends in the percentages of students were also displayed where appropriate. For indices computed from teacher data, percentages of students and average achievement are displayed at the high, medium, and low level for the index.

The data obtained from the science teachers were displayed in two ways. Some of the general information data were presented together for all science teachers in each country. The data for information specific to the science subject, such as preparation to teach the sciences, instructional time in the sciences, and emphasis on experiments, were presented both for the general/integrated science and for the separate science subject area teachers. The tracking information provided by schools that identified teachers by the type of course taught to the sampled students - mathematics, physics, biology, chemistry, earth science, or integrated science - was used to organize the panels for exhibits showing data for the separate sciences.

In general, the countries displayed in the separate science panels correspond to those in Exhibit 17.2. Exceptions include Chinese Taipei and Indonesia, which were shown in the separate science panels in the exhibits based on science teacher data but in the general/integrated panels in the exhibits based on student data. Although the students were asked the general science questions, the teachers in Chinese Taipei were identified as physics/chemistry teachers and were reported in the physics panel; the teachers in Indonesia were identified as biology or physics teachers, and were reported in the corresponding panels. Furthermore, in a few other countries, some combined science subjects were taught by the same teachers. In Finland, Morocco, and the Netherlands, some teachers were identified as physics/chemistry teachers; in Finland and Morocco, some were identified as biology/earth science teachers. The data for teachers who teach more than one subject were reported in only one panel to avoid duplicating the information; biology/earth science was reported in the biology panel and physics/chemistry in the physics panel.

Another consequence of the TIMSS design was that since students were usually taught mathematics and science by different teachers and often were taught one subject by more than one teacher, they had to be linked to more than one teacher for reporting purposes. When a student was taught a subject by more than one teacher, the student's sampling weight used in reporting results for the subject was distributed among those teachers. The student's contribution to student population estimates thus remained constant regardless of the number of teachers.This was consistent with the policy of reporting attributes of teachers and their classrooms in terms of the percentages of students taught by teachers with these attributes. Exceptions were where student-level variables were based on composite responses of all of the students' teachers in a given subject. Analyses of this type involved computing the sum or determining the highest value reported across all of a student's teachers. The composite values obtained were then used to produce the reported student-weighted statistics (e.g., total instructional time in the subjects and the degree of content coverage in mathematics or science).

17.7 Reporting School Background Data

The principals of the selected schools in TIMSS completed questionnaires on the school contexts in which the learning and teaching of mathematics and science occur. Although schools constituted the first stage of sampling, the TIMSS school sample was designed to optimize the student sample, not to provide an optimal sample of schools.[5] Therefore, like the teacher data, the school-level data were reported using the student as the unit of analysis to describe the school contexts for the representative samples of students. In general, the exhibits based on the school data present percentages of students in schools with different characteristics for each country and for the international average. In a few instances, average numerical values for open-ended questions were computed across students (e.g., instructional time, and hours the principal spends on different activities).

17.8 Reporting Curriculum Questionnaire Data

One chapter in each of the 1999 international mathematics and science reports was devoted to data from the curriculum questionnaire. This chapter included summary information about the structure and organization of the mathematics and science curriculum: the level of centralization (i.e., national, regional, local); when the curriculum was introduced and its current status; meth-

[5] See Chapter 2 for a description of the TIMSS sample design.

ods used to support and monitor curriculum implementation; use of public examinations and system wide assessments; percentage of instructional time specified for mathematics and science; differentiation of instruction for students with different abilities or interests; emphasis placed on different approaches and processes; and subjects offered at the eighth grade (science only). For TIMSS countries without a national curriculum (i.e., Australia, Canada, and the United States), composite information that reflected the curriculum across the states or provinces was provided in answer to most questions.

A major function of the curriculum questionnaires was to collect information about which topics in mathematics and science were intended to have been taught by the end of the eighth grade. Responses were summarized to give the percentage of the topics in each content area that were intended to be taught to all or almost all of the eighth-grade students in each country. Detailed information on the percentage of students intended to be taught each individual mathematics or science topic was reported in the accompanying reference section. Most of these topics were addressed by items on the TIMSS achievement tests. (In the teacher questionnaires, these topics were also presented to the mathematics and science teachers, who were asked to what extent they had been covered in class during the year or in previous years.) The curriculum chapters in the international reports present both teachers' reports of the topics actually taught (i.e., the implemented curriculum) and National Research Coordinators' reports of topics intended to be taught (i.e., the intended curriculum), providing complementary perspectives on the coverage of the mathematics and science curriculum in each country.

17.9 Reporting Response Rates for Background Questionnaire Data

While it is desirable that all questions included in a data collection instrument be answered by all intended respondents, a certain percentage of non-response is inevitable. Not only do some questions remain unanswered; sometimes entire questionnaires are not completed or not returned. In TIMSS 1999, since teachers, students, or principals sometimes did not complete the questionnaire assigned to them or some questions within it, certain variables had less than a 100% response rate.

The handling of non-responses varied depending on how the data were to be reported. For background variables that were reported directly, the non-response rates indicate the percentage of students for whom no response was available for a given question. In general, derived variables based on more than one background question were coded as missing if data for any of the required background variables were missing. An exception were indices. Cases were coded as missing for an index variable only if there was no response for more the one-third of the questions used to compute the index; index values would be computed if there were valid data for at least two-thirds of the required variables.

The tables in the TIMSS international reports contain special notations on response rates for the background variables. Although in general the response rates for the student and school background variables were high, some variables and some countries exhibited less than acceptable rates. The non-response rates were somewhat higher for the teacher background data, particularly in cases where teachers were required to complete more than one questionnaire. Since the student is the unit of analysis, the non-response rates given in the international reports always reflect the percentage of students for whom the required responses from students, teachers, or schools were not available. The following special notations were used to convey information about response rates in tables in the international reports.[6]

- For a country where student, teacher or school responses were available for 70% to 84% of the students, an “r” appears next to the data for that country.
- When student, teacher or school responses were available for 50% to 69% of the students, an “s” appears next to the data for that country.

○○○

[6] Since the information from the country questionnaires was obtained at the national level, no non-response flags were necessary in exhibits based on these data.

- When student, teacher or school responses were available for fewer than 50% of the students, an “x” replaces the data.
- When the percentage of students in a particular category fell below 2%, achievement data were not reported in that category. The data were replaced by a tilde (~).
- When data were unavailable for all respondents in a country, dashes (–) were used in place of data in all of the affected columns.[7]

For the trend exhibits, which displayed data for both 1995 and 1999, the non-response notation was determined by the lower of the two response rates. Since response rates for some variables were lower in 1995, this sometimes led to the data for a country being replaced with xx’s or dashes in the trend exhibit, even though response rates for their 1999 data were acceptable.

17.10 Summary

This chapter presented how TIMSS reported and analyzed the background data from students, teachers, schools and NRCs. It documented how summary indices were created, trend data was reported, and the consensus approach used in developing the international reports.

[7] A dash usually indicates that a background question was not administered in a country, but could also reflect translation problems or the administration of a question that was judged to be not internationally comparable. In the exhibits based on the separate science subjects, dashes for specific countries reflect the specific science subjects not included in each country.

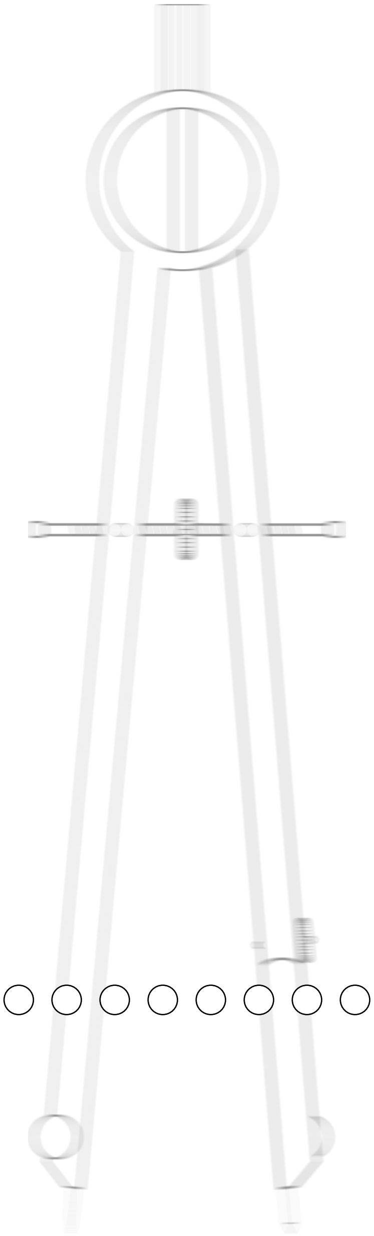
References

Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., Gregory, K.D., Smith, T.A., Chrostowski, S.J., Garden, R.A., & O'Connor, K.M. (2000). *TIMSS 1999 International Science Report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade.* Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., Gregory, K.D., Garden, R.A., O'Connor, K.M., Chrostowski, S.J., & Smith, T.A. (2000). *TIMSS 1999 International Mathematics Report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade.* Chestnut Hill, MA: Boston College.

A

Appendix A: Acknowledgements



A Acknowledgements

TIMSS 1999 was truly a collaborative effort among hundreds of individuals around the world. Staff from the national research centers in each participating country, the International Association for the Evaluation for Educational Achievement (IEA), the International Study Center (ISC) at Boston College, advisors, and funding agencies worked closely to develop and implement TIMSS 1999. The project would not have been possible without the tireless efforts of all involved. Below, the individuals and organizations are acknowledged for their contributions. Given that implementing TIMSS 1999 has spanned approximately four years and involved so many people and organizations, this list may not pay heed to all who contributed throughout the life of the project. Any omission is inadvertent. TIMSS 1999 also acknowledges the students, teachers, and school principals who contributed their time and effort to the study. This report would not be possible without them.

Funding Agencies

Funding for the international coordination of TIMSS 1999 was provided by the National Center for Education Statistics of the U.S. Department of Education, the U.S. National Science Foundation, the World Bank, and participating countries. Valena Plisko, Eugene Owen, and Patrick Gonzales of the National Center for Education Statistics; Larry Suter, Elizabeth VanderPutten, and Janice Earle of the National Science Foundation; and Marlaine Lockheed of the World Bank each played a crucial role in making TIMSS 1999 possible and for ensuring the quality of the study. Each participating country was responsible for finding national project costs and implementing TIMSS 1999 in accordance with the international procedures.

Management and Operations

TIMSS 1999 was conducted under the auspices of the IEA. TIMSS 1999 was co-directed by Michael O. Martin and Ina V.S. Mullis, and managed centrally by the staff of the International Study Center at Boston College, Lynch School of Education. Although the study was directed by the International Study Center and its staff members implemented various parts of TIMSS 1999, important activities also were carried out in centers around the world. In the IEA Secretariat, Hans Wagemaker, Executive Director, was responsible for overseeing fundraising and country participation. The IEA Secretariat also coordinated translation verification and recruiting of quality control monitors. The data were processed centrally by the IEA Data Processing Center in Hamburg. Statistics Canada was responsible for collecting and evaluating the sampling documentation from each country and for calculating the sampling weights. Educational Testing Service in Princeton, New Jersey conducted the scaling of the achievement data.

IEA Secretariat

Hans Wagemaker, Executive Director
Barbara Malak, Manager Membership Relations
Leendert Dijkhuizen, Fiscal Officer

International Study Center at Boston College

Responsible for TIMSS and PIRLS

Michael O. Martin, Co-Director
Ina V.S. Mullis, Co-Director
Eugenio J. Gonzalez, Director of Operations and Data Analysis
Kelvin D. Gregory, TIMSS Project Coordinator
Teresa A. Smith, TIMSS Science Coordinator
Robert Garden, TIMSS Mathematics Coordinator
Kathleen O'Connor, TIMSS Benchmarking Coordinator
Dana L. Kelly, PIRLS Project Coordinator
Steven Chrostowski, Research Associate
Ce Shen, Research Associate
Julie Miles, Research Associate
Steven Stemler, Research Associate
Ann Kennedy, Research Associate
Joseph Galia, Statistician/Programmer
Lana Seliger, Statistician/Programmer
Andrea Pastelis, Database Manager
Kieran Brosnan, Technology Support Specialist
Christine Conley, Publications Design Manager
Jose Nieto, Publications Production Manager
Mario Pita, Data Graphics Specialist
Christina Lopez, Data Graphics Specialist
Betty Hugh, Data Graphics Specialist
Isaac Li, Data Graphics Assistant
Kathleen Packard, Manager, Finance
Susan Comeau, Manager, Office Administration
Ann Tan, Manager, Conference Administration
Monica Guidi, Administrative Coordinator
Laura Misas, Administrative Coordinator

Statistics Canada

Pierre Foy, Senior Methodologist
Marc Joncas, Senior Methodologist
Andrea Farkas, Junior Methodologist
Salina Park, Cooperative Exchange Student

IEA Data Processing Center

Dirk Hastedt, Senior Researcher
Heiko Sibberns, Senior Researcher
Knut Schwippert, Senior Researcher
Caroline Dupeyrat, Researcher
Oliver Neuschmidt, Researcher
Stephan Petzchen, Research Assistant
Anneke Niemeyer, Research Assistant
Juliane Pickel, Research Assistant

Educational Testing Services

Kentaro Yamamoto, Principal Research Scientist
Ed Kulick, Manager, Research Data Analysis

Project Management Team

Michael Martin, International Study Center
Ina Mullis, International Study Center
Eugenio Gonzalez, International Study Center
Hans Wagemaker, IEA Secretariat
Dirk Hastedt, IEA Data Processing Center
Pierre Foy, Statistics Canada
Kentaro Yamamoto, Educational Testing Services
Eugene Johnson, American Institutes for Research

Sampling Referee

Keith Rust, Westat, Inc.

National Research Coordinators

The TIMSS 1999 National Research Coordinators and their staff had the enormous task of implementing the TIMSS 1999 design. This required obtaining funding for the project; participating in the development of the instruments and procedures; conducting field tests; participating in and conducting training sessions; translating the instruments and procedural manuals into the local language; selecting the sample of schools and students; working with the schools to arrange for the testing; arranging for data collection, coding, and data entry; preparing the data files for submission to the IEA Data Processing Center; contributing to the development of the international reports; and preparing national reports. The way in which the national centers operated and the resources that were available varied considerably across the TIMSS 1999 countries. In some countries, the tasks were conducted centrally, while in others, various components were subcontracted to other organizations. In some countries, resources were more than adequate, while in some cases, the national centers were operating with limited resources. Of course, across the life of the project, some NRCs have changed. This list attempts to include all past NRCs who served for a significant period of time as well as all the present NRCs. All of the TIMSS 1999 National Research Coordinators and their staff members are to be commended for their professionalism and their dedication in conducting all aspects of TIMSS.

Australia
Susan Zammit
Australian Council for Educ. Res.(ACER)
19 Prospect Hill Rd.
Private Bag 55
Camberwell, Victoria 3124

Belgium (Flemish)
Christiane Brusselmans-Dehairs
Jean-Pierre Verhaeghe
Vakgroep Onderwijskunde Universiteit Gent
Henri Dunantlaan 2
B 9000 Gent

Ann Van Den Broeck
Dekenstraat 2
AFD.Didaktiek
3000 Leuven

Jan Van Damme
Afd. Didactiek
Vesaliusstraat 2
B-3000 Leuven

Bulgaria
Kiril Bankov
Faculty of Mathematics and Informatics
University of Sofia
1164 Sophia

Canada
Alan Taylor
Applied Research and Evaluation Services (ARES)
University of British Columbia
6058 Pearl Avenue,
Burnaby, BC V5H 3P9

Richard Jones
Education Quality & Accountability Office(EQAO)
2 Carlton St., Suite 1200
Toronto, ON M5B2M9

Jean-Louis Lebel
Direction de la sanction des etudes
1035 rue De La Chevrotiere
26 etage
Quebec GIR 5A5

Michael Marshall
University of British Columbia
Faculty of Education, Rm 6
2125 Main Mall
Vancouver, BC V6T1Z4

Chile
Maria Ines Alvarez
Unidad de Curriculum y Evaluacion
Ministerio de Educacion
Alameda 1146
Sector B, Piso 8

Chinese Taipei
Jau-D Chen
Dean of General Affairs
National Taiwan Normal University
162, E. Hoping Rd. Sec. 1
Taipei, Taiwan 117

Cyprus
Constantinos Papanastasiou
Department of Education
University of Cyprus
P.O. Box 20537
Nicosia CY-1678

Czech Republic
Jana Paleckova
Institute for Information of Education (UIV)
Senovazne nam.26
111 21 Praha 1

England
Graham Ruddock
National Foundation for Educational Research (NFER)
The Mere, Upton Park
Slough, Berkshire
SL1 2DQ

Finland
Pekka Kupari
University of Jyvaskyla
Institute for the Educational Research
P. O. Box 35
SF – 40351 Jyvaskyla

Hong Kong, SAR
Frederick Leung
The University of Hong Kong – Department of Curriculum
Faculty of Education, Rm. 219
Pokfulam Road

Hungary
Péter Vari
National Institute of Public Education
Centre for Evaluation Studies
Dorottya u.8, Pf 701/420
1051 Budapest

Indonesia
Jahja Umar
Examiniation Development Center
Jalan Gunung Sahari Raya – 4
Jakarta Pusat
Jakarta

Iran, Islamic Republic
Ali Reza Kiamanesh
Ministry of Education
196, Institute for Education Research
Keshavaraz Blvd.
Tehran, 14166

Israel
Ruth Zuzovsky
Tel Aviv University
School of Education
Center for Science and Technology Education
Ramat Aviv 69978

Italy
Anna Maria Caputo
Ministerio della Pubblica Istruzione
Centro Europeo Dell 'Educazione (CEDE)
5- Villa Falconieri
Frascati (Roma)
00044

Japan
Yuji Saruta
Hanako Senuma
National Institute for Educational Research (NIER)
6-5-22 Shimomeguro
Meguro-ku, Tokyo
153-8681

Jordon
Tayseer Al-Nhar
National Center for Human Resources Development
P. O. Box 560
Amman, Jordan 11941

Korea, Republic of
Sungsook Kim
Chung Park
Korea Institute of Curriculum & Evaluation (KICE)
KICE
25-1 Samchung-dong
GhongRo-Gu, Seoul
110-230

Latvia
Andrejs Geske
University of Latvia
IEA National Research Center
Jurmalas Gatve 74/76, Rm. 204A
Riga
LV-1083

Lithuania

Algirdas Zabulionis
National Examinations Center
Ministry of Education and Science
M. Katkaus 44
Vilnius
LT2051

Macedonia, Republic of

Anica Aleksova
Ministry of Education and Science
Bureau for Development of Education
Ruder Boskovic St. bb.
1 000 Skopje

Malaysia

Azmi Zakaria
Ministry of Education
Level 2,3 &5 Block J South
Pusat Bandar Damansara, Kuala Lumpur
50604

Moldova, Republic Of

Ilie Nasu
Ministry of Education and Science
University "A. Russo"
Str. Puschin 38
Balti 3100

Lidia Costiuc
1 Piata Mazzi Adunazi Nationale
Chisinau

Morocco

Mohamed Fatihi
Direction de l'Evaluation du Systeme Educatif
INNOVATIONS PEDAGOGIQUES
32 Boulevard Ibn Toumert
Place Bob Rouah, Rabat

Netherlands

Klaas Bos
University of Twente
Centre for Applied Research in Education (OCTO)
P.O. Box 217
7500 AE Enschede

New Zealand

Megan Chamberlain
Ministry of Education
CER Unit-Research Division
45-47 Pipitea Street
Thorndon, Wellington

Philippines

Ester Ogena
DOST-Science Education Institute
3F PTRI Blg
Bicutan, Taguig
Metro Manila 1604

Vivien Talisayon
Institute Of Science & Mathematics Education Development
University of the Philippines UPISMED
Diliman, Quezon City

Romania

Gabriela Noveanu
Institute for Educational Sciences
Evaluation and Forecasting Division
Str. Stirbei Voda 37
Bucharest Ro-70732

Russian Federation

Galina Kovalyova
Center for Evaluating the Quality of Education
Institute of General Secondary Education
ul. Pogodinskaya 8
Moscow
119905

Singapore

Cheow Cher Wong
Research and Evaluation Branch
Ministry of Education
1 North Buona Vista Dr /MOE Building
Singapore, Singapore
138675

Slovak Republic

Olga Zelmanova
Maria Berova
SPU-National Institute for Education
Pluhova 8, P. O. Box 26
Brastislava
830 00

Slovenia

Barbara Japelj
Educational Research Institute Ljubljana
Gerbiceva 62
Ljubljana
1000

South Africa

Sarah Howie
Human Sciences Research Council
134 Pretorius Street
Private Bag X41
Pretoria
0001

Thailand

Precharn Dechsri
Institute For the Promotion of Teaching Science & Technology (IPST)
924 Sukhumvit Rd. Ekamai
Bangkok
10100

Tunisia

Ktari Mohsen
Ministere de l'Education
Boulevard Bab-Bnet
Tunis

Turkey

Yurdanur Atlioglu
Educational Research and Development Directorate
Gazi Mustafa Kemal Bulvani
No 109/5-6-7
Maltepe, Ankara
06570

USA

Patrick Gonzales
National Center for Education Statistics
U.S. Department of Education
1990 K St., NW Rm 9071
Washington, DC 20006

TIMSS 1999 Advisory Committees

The International Study Center at Boston College was supported in its work by advisory committees. The Subject Matter Item Replacement Committee was instrumental in developing of TIMSS 1999 tests, and the Questionnaire Item Review Committee revised the TIMSS questionnaires. The Scale Anchoring Panel developed the descriptions of the international benchmarks in mathematics and science.

Subject Matter Item Replacement Committee

Mathematics

Antoine Bodin, France
Anna-Maria Caputo, Italy
Nobert Delagrange, Belgium (Flemish)
Jan de Lange, Netherlands
Hee-Chan Lew, Republic of Korea
Mary Lindquist, United States
David Robitaille, Canada

Science

Hans Ernst Fischer, Germany
Galina Kovalyova, Russian Federation
Svein Lie, Norway
Masao Miyake, Japan
Graham Orpwood, Canada
Jana Strakova, Czech Republic
Carolyn Swain, England

Subject Area Coordinators

Bob Garden, New Zealand (Mathematics)
Teresa Smith, United States (Science)

Special Consultants

Chancey Jones, Mathematics
Christine O'Sullivan, Science

Questionnaire Item Review Committee

Im Hyung, Republic of Korea
Barbara Japelj, Slovenia
Trevor Williams, United States
Graham Ruddock, England
Klaas Bos, Netherlands

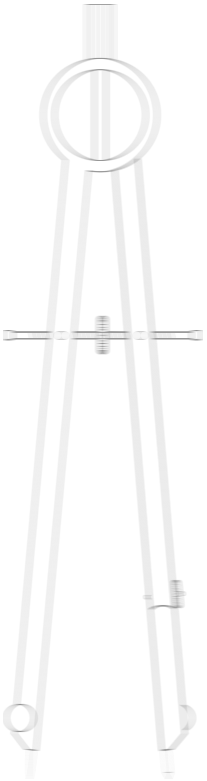
Scale Anchoring Committees

Mathematics

Anica Aleksova, Republic of Macedonia
Lillie Albert, United States
Kiril Bankov, Bulgaria
Jau-D Chen, Chinese Taipei
John Dossey, United States
Barbara Japelj, Slovenia
Mary Lindquist, United States
David Robitaille, Canada
Graham Ruddock, United Kingdom
Hanako Senuma, Japan
Pauline Vos, The Netherlands

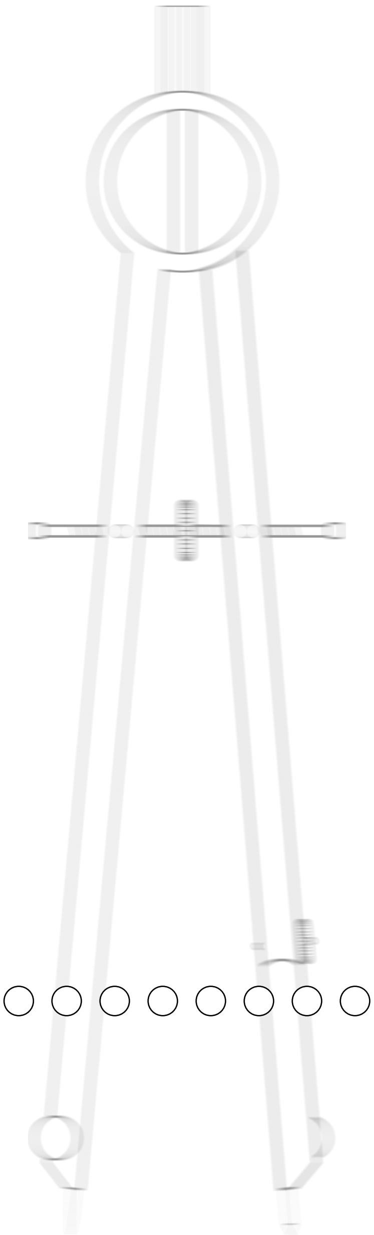
Science

Audrey Champagne, United States
Galina Kovalyova, Russian Federation
Jan Lokan, Australia
Jana Paleckova, Czech Republic
Senta Raizen, United States
Vivien Talisayon, Philippines
Hong Kim Tan, Singapore



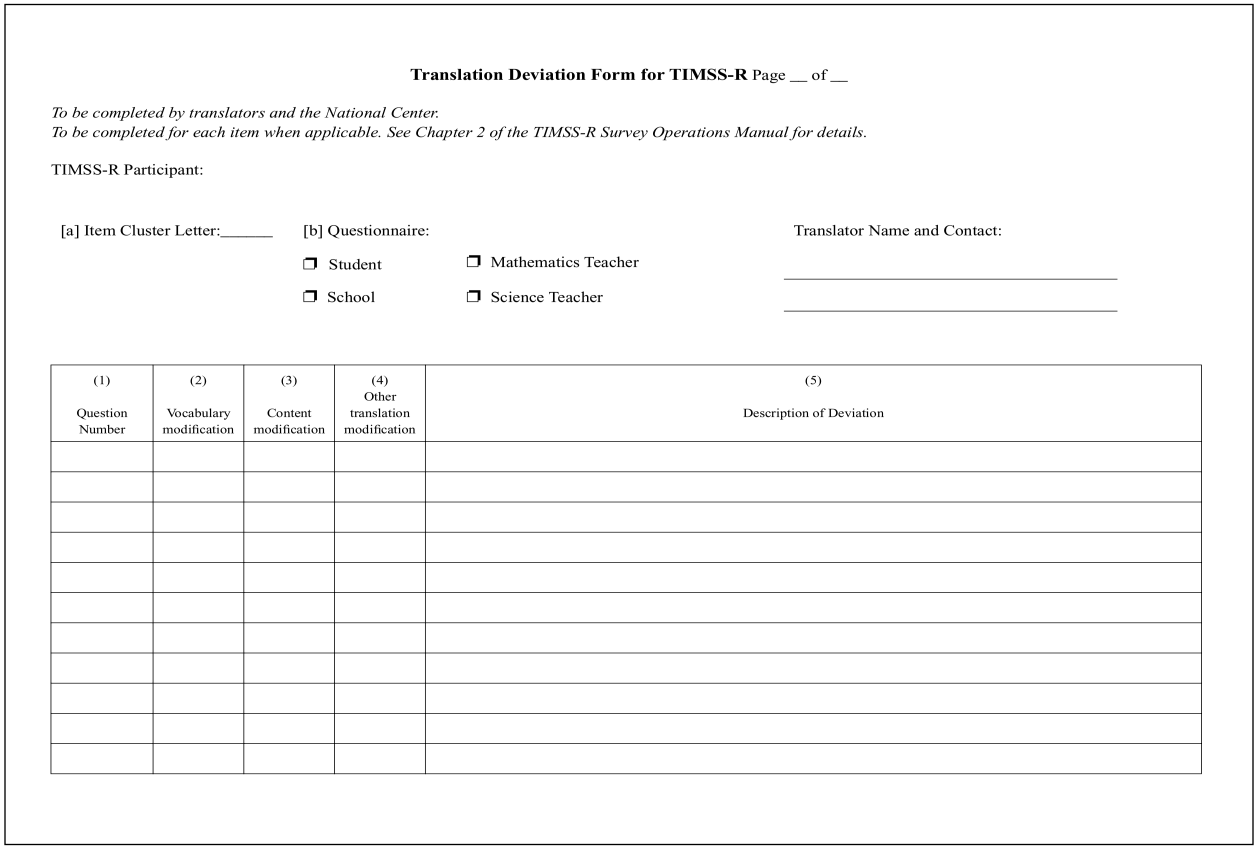
B

Appendix B: Translation Deviation Form and List of Instruments Submitteed



B Translation Deviation Form and List of Instruments Submitted

Exhibit B.1 **Translation Deviation Form for TIMSS-R**



Translation Deviation Form for TIMSS-R Page __ of __

To be completed by translators and the National Center.
To be completed for each item when applicable. See Chapter 2 of the TIMSS-R Survey Operations Manual for details.

TIMSS-R Participant:

[a] Item Cluster Letter:______

[b] Questionnaire:

❒ Student
❒ School
❒ Mathematics Teacher
❒ Science Teacher

Translator Name and Contact:

____

____

(1) Question Number	(2) Vocabulary modification	(3) Content modification	(4) Other translation modification	(5) Description of Deviation

Exhibit B.2 List of Instruments Submitted

Country	Language(s)	Clusters A-Z	Booklets 1-8	Student Questionnaire	Mathematics Teacher Questionnaire	Science Teacher Questionnaire	School Questionnaire
Australia	English	●	–	●	●	●	●
Belgium (Flemish)	Flemish	●	●	●	●	●	●
Bulgaria	Bulgarian	●	●	●	●	●	●
Canada	English	–	–	–	–	–	–
Canada	French	●	●	–	–	–	–
Chile	Spanish	●	●	●	●	●	●
Chinese Taipei	Chinese	●	●	●	●	●	●
Cyprus	Greek	●	●	●	●	●	●
Czech Republic	Czech	●	–	●	●	●	●
England	English	●	–	●	●	●	●
Finland	Finnish	●	●	●	●	●	●
Finland	Swedish	●	●	●	●	●	●
Hong Kong, SAR	Chinese	●	–	●	●	●	●
Hong Kong, SAR	English[a)]	–	–	●	●	●	●
Hungary	Hungarian	●	●	●	●	●	●
Indonesia	Indonesian	●	●	●	●	●	●
Iran, Islamic Rep.	Iranian	●	●	●	●	●	●
Israel	Hebrew	●	●	●	●	●	●
Israel	Arabic	●	●	●	●	●	●
Italy	Italian	●	●	●	●	●	●
Japan	Japanese	●	●	●	●	●	●
Jordan	Arabic	●	●	●	●	●	●
Korea, Rep. of	Korean	●	●	●	●	●	●
Latvia	Latvian	–	●	●	●	●	●
Lithuania	Lithuanian	●	●	–	–	–	–
Macedonia, Rep. of	Macedonian	●	–	●	●	●	●
Macedonia, Rep. of	Albanian	●	–	●	●	●	–[b]
Malaysia	Malay	●	●	●	●	●	●
Moldova	Moldavian	●	●	●	●	●	●

a. Hong Kong used the original English version of student test as developed by ISC

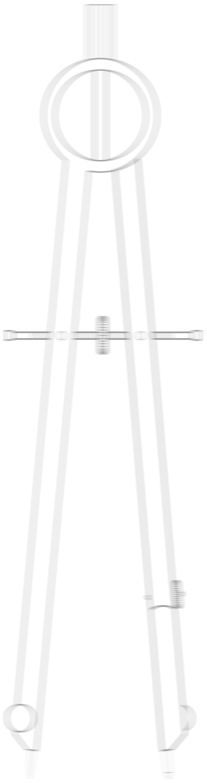
b. Macedonian version of School Questionnaire was used for all schools

Exhibit B.2 List of Instruments Submitted (Continued)

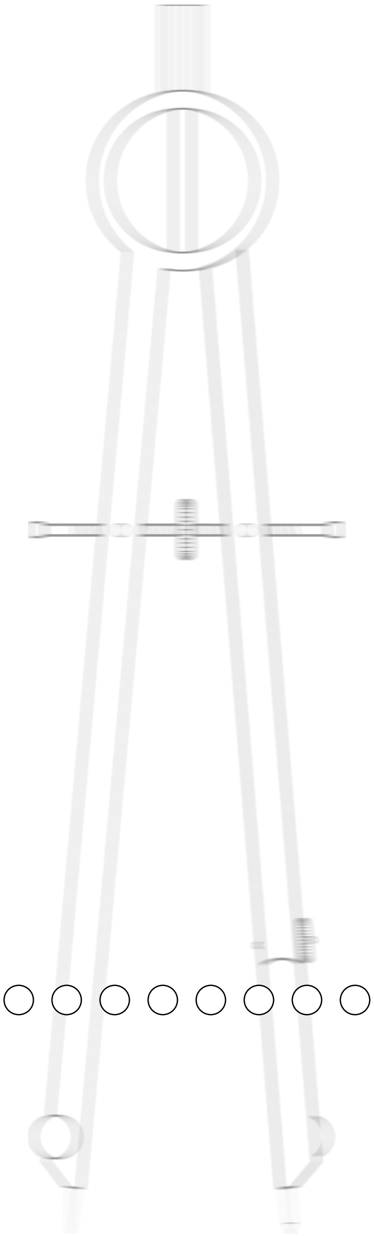
Country	Language(s)	Clusters A-Z	Booklets 1-8	Student Questionnaire	Mathematics Teacher Questionnaire	Science Teacher Questionnaire	School Questionnaire
Moldova	Russian	●	●	●	●	●	●
Morocco	Arabic	–	●	●	●	●	●
Netherlands	Netherlands	●	●	●	●	●	●
New Zealand	English	●	–	●	●	●	●
Philippines	English	●	●	●	●	●	●
Philippines	Filipino	●	●	●	●	●	●
Romania	Romanian	●	●	●	●	●	●
Russian Federation	Russian	–	●[a]	●	●	●	●
Singapore	English	●	●	●	●	●	●
Slovak Republic	Slovak	–	●	●	●	●	●
Slovenia	Slovenian	●	●	●	●	●	●
South Africa	English	–	●	●	●	●	●
South Africa	Afrikaans	–	●	●	●	●	●
Thailand	Thai	●	●	●	●	●	●
Tunisia	Arabic	●	●	●	–	–	●
Tunisia	French	–	–	–	●	●[b]	–
Turkey	Turkish	–	●	●	●	●	●
United States	English	–	●	●	●	●	●

a. Except clusters A–H

b. Students were tested only in Arabic and only Arabic version of Student and School Questionnaires were used



Appendix C: Sample Implementation



C Sample Implementation

C.1 Introduction

For each country participating in TIMSS-R, this appendix describes the target population definition where necessary, coverage and exclusions, use of stratification variables, and any deviations from the general TIMSS-R design.

C.2 Australia

C.2.1 Target Population

In Australia, the target grades varied by State and Territory. The target grade was the 8th grade in New South Wales, Victoria, Tasmania and the Australian Capital Territory. The target grade was the 9th grade in Queensland, South Australia, Western Australia and the Northern Territory. This variation is due to different age entrance rules applied in the Australian States and Territories.

C.2.2 Coverage and Exclusions

School-level exclusions consisted of very small schools, special schools (distance-education schools, hospital schools, schools for learning difficulties) and catholic and independent schools in the Northern Territory.

C.2.3 Sample Design

- Explicit stratification by States and Territories and school type (government, catholic and independent), for a total of 24 strata.
- No implicit stratification.
- Because there were many explicit strata, explicit strata within States and Territories were treated as implicit strata for variance estimation.
- Australia used a modified school sampling method. The method is acceptable, but an alternate method of identifying replacement schools was used in the strata marked with ($^{\circ}$) in table C1.
- Large school sample size in the larger States to produce reliable state-level estimates.

Exhibit C.1: Allocation of School Sample in Australia

Explicit Stratum		Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
				Sampled	1st Replacement	2nd Replacement	
Australian Capital Territory	Catholic	1	0	1	0	0	0
	Government	2	0	1	0	0	1
	Independent	1	0	1	0	0	0
New South Wales	Catholic[○]	10	0	8	1	0	1
	Government[○]	33	0	27	3	0	3
	Independent[○]	5	0	4	0	0	1
Victoria	Catholic[○]	8	0	5	2	0	1
	Government[○]	23	1	19	3	0	0
	Independent[○]	5	0	5	0	0	0
Queensland	Catholic	5	0	3	2	0	0
	Government[○]	20	0	18	0	0	2
	Independent	5	0	4	1	0	0
South Australia	Catholic	5	0	5	0	0	0
	Government[○]	19	0	17	1	0	1
	Independent	4	0	4	0	0	0
Western Australia	Catholic	3	0	1	1	0	1
	Government[○]	10	0	9	1	0	0
	Independent	2	0	2	0	0	0
Tasmania	Catholic	3	0	3	0	0	0
	Government[○]	15	0	11	3	0	1
	Independent	2	0	2	0	0	0
Northern Territory	Catholic	0	0	0	0	0	0
	Government	2	0	2	0	0	0
	Independent	1	1	0	0	0	0
Total		**184**	**2**	**152**	**18**	**0**	**12**

C.3 Belgium (Flemish)

C.3.1 Coverage and Exclusions

School-level exclusions consisted of very small schools (MOS<10).

C.3.2 Sample Design

- Explicit stratification by school size (very large schools and large schools), for a total of 2 explicit strata.
- Implicit stratification by school type (state, local board and catholic) and school program (schools with or without the technical program), for a total of 6 strata.
- Two classrooms per school in the general program (when available).
- Belgium sub-sampled 15 schools among the 80 sampled schools with the technical program, to select one classroom from the technical program.

Exhibit C.2 Allocation of School Sample in Belgium (Flemish)

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Large schools	149	0	105	21	8	15
Very large schools	1	0	1	0	0	0
Total	**150**	**0**	**106**	**21**	**8**	**15**
Vocational component	*15*	*1*	*12*	*0*	*0*	*2*

C.4 Bulgaria

C.4.1 Target Population

Bulgaria selected the same target grade as they had in TIMSS in 1995, i.e., the 8th grade. However, because of changes in age entrance policies, the 1999 target population is older than their 1995 target population.

C.4.2 Coverage and Exclusions

School-level exclusions consisted of specials schools for the physically and mentally disabled, schools for students with criminal behavior and very small schools (MOS<9).

C.4.3 Sample Design

- Explicit stratification by school size (large schools and small schools), for a total of 2 explicit strata.
- No implicit stratification.
- Schools in the "Small schools" stratum selected with equal probabilities.

Exhibit C.3: Allocation of School Sample in Bulgaria

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Large schools	150	0	144	0	0	6
Small schools	22	3	19	0	0	0
Total	**172**	**3**	**163**	**0**	**0**	**6**

C.5 Canada

C.5.1 Coverage and Exclusions

School-level exclusions consisted of offshore schools, schools where students are taught in aboriginal languages, very small schools, schools in Prince Edward Island, French schools in New Brunswick and schools in the Territories.

C.5.2 Sample Design

- Explicit stratification by province, language (French and English in New Brunswick, Québec and Ontario), school size (very large schools and large schools in Newfoundland, large schools and small schools in Saskatchewan) and school type (government and independent in Québec), for a total of 16 explicit strata.

- Implicit stratification by region (in Ontario English), language (French and English in Nova Scotia) and school type (public and independent in British Columbia), for a total of 26 implicit strata.
- Schools in the “Newfoundland - Very large schools”, “Ontario French” & “Saskatchewan - Small schools” strata selected with equal probabilities.
- Large school sample size in Ontario, Newfoundland, Québec, Alberta and British Columbia to produce reliable provincial estimates.

Exhibit C.4 Allocation of School Sample in Canada

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Newfoundland-Very large schools	2	0	2	0	0	0
Newfoundland-Large schools	38	1	37	0	0	0
Nova Scotia	5	0	3	0	0	2
New Brunswick-English	2	0	2	0	0	0
New Brunswick-French	2	2	0	0	0	0
Québec-Government-English	4	0	4	0	0	0
Québec-Government-French	37	0	30	3	2	2
Québec-Independent-English	2	0	2	0	0	0
Québec-Independent-French	7	1	6	0	0	0
Ontario-English	120	3	112	1	0	4
Ontario-French	80	3	73	1	0	3
Manitoba	6	0	5	0	0	1
Saskatchewan-Large schools	4	0	4	0	0	0
Saskatchewan-Small schools	2	0	2	0	0	0
Alberta	55	1	52	2	0	0
British Columbia	44	1	42	0	0	1
Total	**410**	**12**	**376**	**7**	**2**	**13**

C.6 Chile

C.6.1 Target Population

The target grade selected for the national desired target population was the 8th grade. Students in the 7th grade were tested for national purposes.

C.6.2 Coverage and Exclusions

School-level exclusions consisted of geographically inaccessible schools and very small schools (MOS<15).

C.6.3 Sample Design

- No explicit stratification.
- Implicit stratification by school type (public and private) and urbanization (rural and urban), for a total of 4 implicit strata.
- Large school sample size because of expected large intraclass correlation.

Exhibit C.5 Allocation of School Sample in Chile

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Chile	186	0	182	4	0	0
Total	**186**	**0**	**182**	**4**	**0**	**0**

C.7 Chinese Taipei

C.7.1 Coverage and Exclusions

School-level exclusions consisted of schools on isolated islands (Kinnen, Matsu, Penghu and two islands in Taituag county) and very small schools (MOS < 20).

C.7.2 Sample Design

- No explicit stratification.
- Implicit stratification by region (North, East, South & Middle), for a total of 4 implicit strata.

Exhibit C.6: Allocation of School Sample in Chinese Taipei

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Chinese Taipei	150	0	150	0	0	0
Total	**150**	**0**	**150**	**0**	**0**	**0**

C.8 Cyprus

C.8.1 Coverage and Exclusions

All schools are included.

C.8.2 Sample Design

- All national schools included in the sample.
- Two classrooms sampled per school.

Exhibit C.7: Allocation of School Sample in Cyprus

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Nicosia	23	0	23	0	0	0
Lemesos	16	0	16	0	0	0
Larnaka	14	0	14	0	0	0
Pafos	8	0	8	0	0	0
Total	**61**	**0**	**61**	**0**	**0**	**0**

C.9 Czech Republic

C.9.1 Coverage and Exclusions

School-level exclusions consisted of schools for the disabled, very small schools (MOS<10) and Polish language schools.

C.9.2 Sample Design

- Explicit stratification by school level (Basic schools and Gymnasium), for a total of 2 explicit strata.
- Implicit stratification by urbanization (5 levels), for a total of 10 implicit strata.
- Large school sample size in the "Gymnasiums" stratum to produce reliable estimates by school level.

Exhibit C.8: Allocation of School Sample in Czech Republic

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Basic schools	90	2	82	6	0	0
Gymnasiums	60	6	54	0	0	0
Total	**150**	**8**	**136**	**6**	**0**	**0**

C.10 England

C.10.1 Coverage and Exclusions

School-level exclusions consisted of special-needs schools and very small schools (MOS<13).

C.10.2 Sample Design

- No explicit stratification.
- Implicit stratification by school type (independent, grant and other) and school performance (5 levels), for a total of 11 implicit strata.
- In schools where mathematics instruction was streamed, home rooms were sampled rather than mathematics classes.

Exhibit C.9: Allocation of School Sample in England

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
England	150	0	76	34	18	22
Total	**150**	**0**	**76**	**34**	**18**	**22**

C.11 Finland

C.11.1 Coverage and Exclusions

School-level exclusions consisted of schools from the autonomous province of Ahvenanmaa (Âland), special schools & Rudolph Steiner schools, foreign language schools and very small schools (MOS<10).

C.11.2 Sample Design

- Explicit stratification by region (Uusimaa, Southern Finland, Eastern Finland, Mid-Finland and Northern Finland), for a total of 5 explicit strata.
- Implicit stratification by urbanization (urban, semi-urban and rural), for a total of 15 implicit strata.
- Equal sample allocation by explicit strata and large school sample size to produce reliable regional estimates.

Exhibit C.10: Allocation of School Sample in Finland

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Uusimaa	32	0	31	1	0	0
Southern Finland	32	0	31	1	0	0
Eastern Finland	32	0	29	2	0	1
Mid-Finland	32	0	32	0	0	0
Northern Finland	32	0	32	0	0	0
Total	**160**	**0**	**155**	**4**	**0**	**1**

C.12 Hong Kong, SAR

C.12.1 Coverage and Exclusions

School-level exclusions consisted of special-needs schools.

C.12.2 Sample Design

- No explicit stratification.
- Implicit stratification by funding (aided, government and private) and gender (co-ed, girls and boys), for a total of 9 implicit strata.
- Large school sample size because of expected large intraclass correlation.

Exhibit C.11: Allocation of School Sample in Hong Kong, SAR

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Hong Kong, SAR	180	0	135	0	2	43
Total	**180**	**0**	**135**	**0**	**2**	**43**

C.13 Hungary

C.13.1 Coverage and Exclusions

School-level exclusions consisted of specials schools for the disabled and very small schools (MOS<10).

C.13.2 Sample Design

- No explicit stratification.
- Implicit stratification by region (20) and urbanization (large towns, small towns and villages), for a total of 58 implicit strata.
- Hungary used an alternate, and acceptable, school sampling method.

Exhibit C.12: Allocation of School Sample in Hungary

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Hungary	150	0	147	0	0	3
Total	**150**	**0**	**147**	**0**	**0**	**3**

C.14 Indonesia

C.14.1 Coverage and Exclusions

No school-level exclusions.

C.14.2 Sample Design

- Explicit stratification by school type (public and private), for a total of 2 explicit strata.
- Implicit stratification by performance (5 levels), for a total of 10 implicit strata.

Exhibit C.13: Allocation of School Sample in Indonesia

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Public schools	100	0	89	8	3	0
Private schools	50	0	43	4	3	0
Total	**150**	**0**	**132**	**12**	**6**	**0**

C.15 Iran, Islamic Rep.

C.15.1 Coverage and Exclusions

School-level exclusions consisted of specials schools for the disabled.

C.15.2 Sample Design

- Explicit stratification by school size (small schools and large schools), for a total of 2 explicit strata.
- No implicit stratification.
- Large school sample size because of expected large intraclass correlation.

Exhibit C.14: Allocation of School Sample in Islamic Republic of Iran

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Large schools	117	0	113	4	0	0
Small schools	53	0	51	2	0	0
Total	**170**	**0**	**164**	**6**	**0**	**0**

C.16 Israel

C.16.1 Coverage and Exclusions

School-level exclusions consisted of special education schools, very orthodox religious schools and Jordanian schools.

C.16.2 Sample Design

- No explicit stratification.
- Implicit stratification by language (Hebrew and Non-Hebrew), school type (religious and secular) and school level (elementary and junior high), for a total of 6 implicit strata.

Exhibit C.15: Allocation of School Sample in Israel

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Israel	150	11	137	2	0	0
Total	**150**	**11**	**137**	**2**	**0**	**0**

C.17 Italy

C.17.1 Coverage and Exclusions

School-level exclusions consisted of non-government middle schools (catholic, independent, municipal, etc.).

C.17.2 Sample Design

- No explicit stratification.
- Implicit stratification by region and type of municipality (capital towns and other small towns), for a total of 38 implicit strata.
- Large school sample size because of expected large intraclass correlation.

Exhibit C.16: Allocation of School Sample in Italy

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Italy	180	0	170	9	1	0
Total	**180**	**0**	**170**	**9**	**1**	**0**

C.18 Japan

C.18.1 Coverage and Exclusions

School-level exclusions consisted of specials schools for the physically and mentally disabled, schools with atypical systems and very small schools (MOS<18).

C.18.2 Sample Design

- Explicit stratification by school type (national/private and public) and urbanization (big city area, city area and not city area), for a total of 4 explicit strata.
- No implicit stratification.

Exhibit C.17: Allocation of School Sample in Japan

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Public schools - Big city area	24	0	19	0	0	5
Public schools - City area	82	0	82	0	0	0
Public schools - Not city area	35	0	34	0	0	1
National & Private schools	9	0	5	0	0	4
Total	**150**	**0**	**140**	**0**	**0**	**10**

C.19 Jordan

C.19.1 Coverage and Exclusions

School-level exclusions consisted of very small schools (MOS<15).

C.19.2 Sample Design

- Explicit stratification by school size (small rural schools and large schools), for a total of 2 explicit strata.
- Implicit stratification by education authority (public, private and UNRWA) and urbanization (rural and urban), for a total of 6 implicit strata.
- Schools in the “Small rural schools” stratum selected with equal probabilities.

Exhibit C.18: Allocation of School Sample in Jordan

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Large schools	142	2	139	1	0	0
Small rural schools	8	1	7	0	0	0
Total	**150**	**3**	**146**	**1**	**0**	**0**

C.20 Korea, Rep. of

C.20.1 Target Population

Because Korea performed the TIMSS-R assessment 4 months later in the school year than they did in TIMSS, their TIMSS-R target population is older when compared to their TIMSS target population.

C.20.2 Coverage and Exclusions

School-level exclusions consisted of schools located in remote places, islands and border areas, physical education middle schools and very small schools (MOS<18).

C.20.3 Sample Design

- Explicit stratification by province (16), for a total of 16 explicit strata.
- Implicit stratification by urbanization (metro, urban and rural) and gender (boys, girls and co-ed), for a total of 75 implicit strata.
- Because there were many explicit strata, they were treated as implicit strata for variance estimation.

Exhibit C.19: Allocation of School Sample in Republic of Korea

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Seoul	32	0	32	0	0	0
Pusan	13	0	13	0	0	0
Taegu	8	0	8	0	0	0
Inchon	9	0	9	0	0	0
Kwangju	5	0	5	0	0	0
Taejon	5	0	5	0	0	0
Ulsan	4	0	4	0	0	0
Kyunggi-do	26	0	26	0	0	0
Kangwon-do	4	0	4	0	0	0
Chungchongbuk-do	5	0	5	0	0	0
Chungchongnam-do	6	0	6	0	0	0
Chollabuk-do	7	0	7	0	0	0
Chollanam-do	6	0	6	0	0	0
Kyongsangbuk-do	8	0	8	0	0	0
Kyongsangnam-do	10	0	10	0	0	0
Cheju-do	2	0	2	0	0	0
Total	**150**	**0**	**150**	**0**	**0**	**0**

C.21 Latvia

C.21.1 Coverage and Exclusions

Coverage in Latvia was restricted to students whose language of instruction is Latvian. School-level exclusions consisted of specials schools for the physically and mentally disabled and very small schools (MOS<8).

C.21.2 Sample Design

- Explicit stratification by school size (very large schools, large schools and small rural schools), for a total of 3 explicit strata.
- Implicit stratification by urbanization (rural and urban) and region (5), for a total of 16 implicit strata.
- Schools in the “Very large schools” & “Small rural schools” strata selected with equal probabilities.

Exhibit C.20: Allocation of School Sample in Latvia

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Very large schools	21	0	21	0	0	0
Large schools	104	0	100	2	0	2
Small rural schools	25	2	22	0	0	1
Total	**150**	**2**	**143**	**2**	**0**	**3**

C.22 Lithuania

C.22.1 Target Population

Lithuania tested the 9th grade at the beginning of the school year. Because of this factor, combined with changes in age entrance policies, their TIMSS 1999 target population is now older when compared to their TIMSS 1995 target population.

C.22.2 Coverage and Exclusions

Coverage in Lithuania was restricted to students whose language of instruction is Lithuanian. School-level exclusions consisted of specials schools and very small schools (MOS<7).

C.22.3 Sample Design

- Explicit stratification by school size (large schools and small schools), for a total of 2 explicit strata.
- Implicit stratification by school level (basic and secondary), for a total of 4 implicit strata.
- Schools in the “Small rural schools” stratum selected with equal probabilities.

Exhibit C.21: Allocation of School Sample in Lithuania

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Large schools	133	0	133	0	0	0
Small schools	17	0	17	0	0	0
Total	**150**	**0**	**150**	**0**	**0**	**0**

C.23 Republic of Macedonia

C.23.1 Target Population

The Republic of Macedonia selected the 8th grade as their target population. Their target population is somewhat older than most other TIMSS 1999 participating countries.

C.23.2 Coverage and Exclusions

School-level exclusions consisted of specials schools and very small schools (MOS<14).

C.23.3 Sample Design

- Explicit stratification by school size (very large schools and large schools), for a total of 2 explicit strata.
- Implicit stratification by language (Albanian and Macedonian), for a total of 2 implicit strata.
- Schools offering both languages were split into components to fit the implicit stratification by language. Thus 5 schools were sampled twice, once from each language group.
- Schools in the "Very large schools" stratum selected with equal probabilities.

Exhibit C.22: Allocation of School Sample in Republic of Macedonia

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Large schools	129	0	128	0	0	1
Very large schools	21	0	21	0	0	0
Total	**150**	**0**	**149**	**0**	**0**	**1**

C.24 Malaysia

C.24.1 Coverage and Exclusions

School-level exclusions consisted of private secondary schools, private Chinese secondary schools, international secondary schools, specials secondary schools for the physically and mentally disabled and very small schools (MOS<18).

C.24.2 Sample Design

- No explicit stratification.
- Implicit stratification by region (14) and urbanization (rural and urban), for a total of 28 implicit strata.

Exhibit C.23: Allocation of School Sample in Malaysia

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Malaysia	150	0	148	1	1	0
Total	**150**	**0**	**148**	**1**	**1**	**0**

C.25 Moldova

C.25.1 Coverage and Exclusions

School-level exclusions consisted of specials schools for the physically and mentally disabled, schools with neither Russian or Romanian as language of instruction and very small schools (MOS<13).

C.25.2 Sample Design

- No explicit stratification.
- Implicit stratification by urbanization (rural and urban), language (National, Russian and mixed) and region (central, north and south), for a total of 17 implicit strata.

Exhibit C.24: Allocation of School Sample in Moldova

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Moldova	150	0	145	5	0	0
Total	**150**	**0**	**145**	**5**	**0**	**0**

C.26 Morocco

C.26.1 Coverage and Exclusions

School-level exclusions consisted of specials education institutions (blind, disabled & jail centers), schools of University and Cultural French Mission and very small schools (MOS<9).

C.26.2 Sample Design

- No explicit stratification.
- Implicit stratification by region (14) and urbanization (rural and urban), for a total of 28 implicit strata.
- Two classrooms per school, sampled with equal probability.
- A sub-sample of 17 students per classroom.
- Large school sample size because of expected large intraclass correlation.

Exhibit C.25: Allocation of School Sample in Morocco

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Morocco	174	0	172	1	0	1
Total	**174**	**0**	**172**	**1**	**0**	**1**

C.27 Netherlands

C.27.1 Coverage and Exclusions

School-level exclusions consisted of schools with renewing program (vrijescholen) and schools with English stream.

C.27.2 Sample Design

- Explicit stratification by school size (very large schools and large schools), for a total of 2 explicit strata.
- Implicit stratification by school program (VBO, MAVO, VBO/AVO, MAVO/HAVO/VWO, HAVO/VWO, VBO/AVO/VWO), for a total of 7 implicit strata.
- The sample consists of 150 administrative schools. For many of these schools, an additional sampling stage occurred to select a physical school within administrative schools using PPS.
- Schools in the "Very large schools" stratum selected with equal probabilities.

Exhibit C.26: Allocation of School Sample in the Netherlands

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Large schools	134	2	77	29	6	20
Very large schools	16	0	9	4	1	2
Total	**150**	**2**	**86**	**33**	**7**	**22**

C.28 New Zealand

C.28.1 Coverage and Exclusions

School-level exclusions consisted of correspondence schools, specials schools, Rudolph Steiner & Full Immersion Maori language schools and very small schools (MOS<13).

C.28.2 Sample Design

- Explicit stratification by school size (very large schools and large schools), for a total of 2 explicit strata.
- Implicit stratification by school type (state and private), gender (boys, girls and co-ed), SES (low, middle and high) and urbanization (rural and urban), for a total of 10 implicit strata.
- Schools in the "Very large schools" stratum selected with equal probabilities.

Exhibit C.27: Allocation of School Sample in New Zealand

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Very large schools	16	0	14	0	0	2
Large schools	140	0	131	6	1	2
Total	**156**	**0**	**145**	**6**	**1**	**4**

C.29 Philippines

C.29.1 Coverage and Exclusions

School-level exclusions consisted of all schools from the Autonomous Region of Muslim Mindanao and very small schools (MOS<49).

C.29.2 Sample Design

- No explicit stratification.
- Implicit stratification by region (15) and school type (public and private), for a total of 30 implicit strata.

Exhibit C.28: Allocation of School Sample in the Philippines

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Philippines	150	0	148	2	0	0
Total	**150**	**0**	**148**	**2**	**0**	**0**

C.30 Romania

C.30.1 Target Population

Romania selected the same target grade as they had in TIMSS 1995, i.e., the 8th grade. Their target population is older, when compared to most other TIMSS 1999 participating countries, but of the same age as in TIMSS 1995.

C.30.2 Coverage and Exclusions

School-level exclusions consisted of specials schools for the physically and mentally disabled, very small schools (MOS<8) and other schools with different characteristics.

C.30.3 Sample Design

- Explicit stratification by school size (small rural schools and large schools), for a total of 2 explicit strata.
- Implicit stratification by urbanization (rural and urban), for a total of 3 implicit strata.
- Schools in the “Small rural schools” stratum selected with equal probabilities.

Exhibit C.29: Allocation of School Sample in Romania

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Large schools	125	0	122	0	0	3
Small rural schools	25	0	25	0	0	0
Total	**150**	**0**	**147**	**0**	**0**	**3**

C.31 Russian Federation

C.31.1 Coverage and Exclusions

School-level exclusions consisted of specials schools for the physically and mentally disabled and special schools with Non-Russian teaching language.

C.31.2 Sample Design

- Preliminary sampling of 45 regions from a list of 89 regions; 19 regions were large enough to be sampled with certainty, they are marked with ($^{\bigcirc}$) in table C30.
- No explicit stratification, the explicit strata shown in table C30 are the 45 sampled regions.
- Implicit stratification by school size (small schools and large schools) and by urbanization (village, settlement, small town, middle town, large town and metropolis) for large schools only.
- Four schools sampled per region; more schools sampled in some certainty regions.
- Schools in the “Small schools” implicit strata sampled with equal probabilities within the selected regions.
- Large school sample size because of preliminary sampling stage.

Exhibit C.30: Allocation of School Sample in the Russian Federation

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
1. Bashkortostan ❍	4	0	4	0	0	0
2. Kabardino-Balkaria	4	0	4	0	0	0
3. Kalmykia	4	0	4	0	0	0
4. Marii Al	4	0	4	0	0	0
5. Tataria	4	0	4	0	0	0
6. Udmuttia	4	0	4	0	0	0
7. Krasnodar Kr. ❍	6	0	6	0	0	0
8. Altay Kr. ❍	4	0	4	0	0	0
9. Krasnoyarsk Kr. ❍	4	0	4	0	0	0
10. Primor Kr.	4	0	4	0	0	0
11. Stavropol Kr. ❍	4	0	4	0	0	0
12. Habarovsk Kr.	4	0	4	0	0	0
13. Belgorod Obl.	4	0	4	0	0	0
14. Vladimir Obl.	4	0	4	0	0	0
15. Volgograd Obl. ❍	4	0	3	0	1	0
16. Vologda Obl.	4	0	4	0	0	0
17. Ust Orda Ok. & Irkutsk Obl. ❍	4	0	4	0	0	0
18. Kemerovo Obl. ❍	4	0	4	0	0	0
19. Kirov Obl.	4	0	4	0	0	0
20. Leningrad Obl.	4	0	4	0	0	0
21. Moscow Obl. ❍	6	0	6	0	0	0
22. Murmansk Obl.	4	0	4	0	0	0
23. N. Novgorod Obl. ❍	4	0	4	0	0	0
24. Novgorod Obl.	4	0	4	0	0	0
25. Omsk Obl. ❍	4	0	4	0	0	0
26. Novosibirsk Obl.	4	0	4	0	0	0
27. Orenburg Obl.	4	0	4	0	0	0
28. Orel Obl.	4	0	4	0	0	0
29. Komi Perm Ok. & Perm Obl. ❍	4	0	3	1	0	0
30. Rostov Obl. ❍	4	0	4	0	0	0

Exhibit C.30: Allocation of School Sample in the Russian Federation (Continued)

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
31. Rasan Obl.	4	0	4	0	0	0
32. Samara Obl. ○	4	0	4	0	0	0
33. Saratov Obl. ○	4	0	4	0	0	0
34. Sahalin Obl.	4	0	3	0	0	1
35. Sverdlovsk Obl. ○	6	0	6	0	0	0
36. Smolensk Obl.	4	0	4	0	0	0
37. Tambov Obl.	4	0	4	0	0	0
38. Tver Obl.	4	0	4	0	0	0
39. Tomsk Obl.	4	0	3	0	1	0
40. Ulianovsk Obl.	4	0	4	0	0	0
41. Chelyabinsk Obl. ○	4	0	4	0	0	0
42. Chita Obl.	4	0	4	0	0	0
43. Moscow ○	8	0	8	0	0	0
44. Sankt Petersburg ○	4	0	4	0	0	0
45. Khanty Mansi Ok.	4	0	4	0	0	0
Total	**190**	**0**	**186**	**1**	**2**	**1**

C.32 Singapore

C.32.1 Coverage and Exclusions

There are no school-level exclusions.

C.32.2 Sample Design

- All national schools are in their sample.

Exhibit C.31: Allocation of School Sample in Singapore

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Singapore	145	0	145	0	0	0
Total	**145**	**0**	**145**	**0**	**0**	**0**

C.33 Slovak Republic

C.33.1 Coverage and Exclusions

School-level exclusions consisted of special-needs schools, schools with non-native language speakers and very small schools (MOS<13).

C.33.2 Sample Design

- Explicit stratification by school level (basic school and gymnasium) and school size (very large gymnasiums and large gymnasiums), for a total of 3 explicit strata.
- Implicit stratification by region and school type (private and other), for a total of 11 implicit strata.
- Schools in the “Very large gymnasiums” stratum selected with equal probabilities.
- Large school sample size in the two gymnasiums strata to produce estimates by school level.

Exhibit C.32: Allocation of School Sample in Slovak Republic

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Very large gymnasiums	2	0	2	0	0	0
Large gymnasiums	28	0	27	1	0	0
Basic schools	120	0	114	1	0	5
Total	**150**	**0**	**143**	**2**	**0**	**5**

C.34 Slovenia

C.34.1 Target Population

Slovenia selected the same target grade as they had in TIMSS 1995, i.e., the 8th grade. Their target population is older, when compared to most other TIMSS 1999 participating countries, but of the same age as in TIMSS 1995.

C.34.2 Coverage and Exclusions

School-level exclusions consisted of specials schools for the physically and mentally disabled, schools where the language of instruction is Italian or Hungarian and very small schools (MOS<11).

C.34.3 Sample Design

- Explicit stratification by school size (very large schools and large schools), for a total of 2 explicit strata.
- Implicit stratification by urbanization (5 levels), for a total of 6 implicit strata.
- Because Slovenia used the same sampled schools for TIMSS 1999 & the IEA Civics in Education Study, special accommodation was made for schools with only one classroom, whereby the sampled schools and their replacement schools were alternately shared between the two studies.
- Schools in the "Very large schools" stratum selected with equal probabilities.

Exhibit C.33: Allocation of School Sample in Slovenia

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Large schools	148	0	145	2	0	1
Very large schools	2	0	2	0	0	0
Total	**150**	**0**	**147**	**2**	**0**	**1**

C.35 South Africa

C.35.1 Coverage and Exclusions

School-level exclusions consisted of specials schools and very small schools (MOS<28).

C.35.2 Sample Design

- Explicit stratification by province (9) and language (English and other in Gauteng province), for a total of 10 explicit strata.
- Implicit stratification by language (English, Afrikaans and other) and school funding (state, state-aided and private), for a total of 61 implicit strata.
- Equal sample allocation and large sample size to produce reliable provincial estimates.

Exhibit C.34: Allocation of School Sample in South Africa

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Eastern Cape	25	0	25	0	0	0
Free State	25	1	19	2	0	3
Gauteng - English	22	1	13	2	1	5
Gauteng - Other	3	1	2	0	0	0
Kwazulu Natal	25	0	23	2	0	0
Mpumalanga	25	1	20	1	0	3
North West	25	0	15	1	0	9
Northern Cape	25	1	22	0	0	2
Northern Province	25	0	21	1	0	3
Western Cape	25	1	23	1	0	0
Total	**225**	**6**	**183**	**10**	**1**	**25**

C.36 Thailand

C.36.1 Coverage and Exclusions

School-level exclusions consisted of a variety of special schools and very small schools (MOS<15/20)

C.36.2 Sample Design

- Explicit stratification by school type (secondary, primary and private) and school size (small schools and large schools), for a total of 4 explicit strata.
- Implicit stratification by region (13), for a total of 50 implicit strata.

Exhibit C.35: Allocation of School Sample in Thailand

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Secondary (DGE)	107	0	104	2	1	0
National Primary (ONPEC) Large schools	26	0	25	1	0	0
National Primary (ONPEC) Small schools	7	0	4	0	3	0
Private Education (OPEC)	10	0	10	0	0	0
Total	**150**	**0**	**143**	**3**	**4**	**0**

C.37 Tunisia

C.37.1 Coverage and Exclusions

School-level exclusions consisted of special schools for the blind.

C.37.2 Sample Design

- No explicit stratification.
- Implicit stratification by region (Interior and Coast), for a total of 2 implicit strata.

Exhibit C.36: Allocation of School Sample in Tunisia

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Tunisia	150	1	126	17	6	0
Total	**150**	**1**	**126**	**17**	**6**	**0**

C.38 Turkey

C.38.1 Target Population

Turkey selected the 8th grade for the state schools and the 7th grade for the Anatolian high schools.

C.38.2 Coverage and Exclusions

School-level exclusions consisted of specials schools for the physically and mentally disabled, schools with bussing system and very small schools (MOS<20).

C.38.3 Sample Design

- Preliminary sampling of 40 provinces from a list of 80 provinces; 13 provinces were large enough to be sampled with certainty, they are marked with ($^{○}$) in table C37.
- No explicit stratification, the explicit strata shown in table C37 are the 45 sampled provinces.
- Implicit stratification by county within sampled provinces.
- Four schools sampled per province; more schools sampled in some certainty provinces.
- Large school sample size because of preliminary sampling stage.

Exhibit C.37: Allocation of School Sample in Turkey

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
1. Adana ○	6	0	6	0	0	0
2. Afyon	4	0	4	0	0	0
3. Ankara ○	14	0	14	0	0	0
4. Antalya ○	4	0	4	0	0	0
5. Ardahan	4	0	4	0	0	0
6. Artvin	4	0	4	0	0	0
7. Balikesir	4	0	4	0	0	0
8. Bingol	4	0	4	0	0	0
9. Bursa ○	6	0	6	0	0	0
10. Denizli	4	0	4	0	0	0
11. Diyarbakir	4	0	4	0	0	0
12. Elazig	4	0	4	0	0	0
13. Erzurum	4	0	4	0	0	0
14. Eskisehir	4	0	3	1	0	0
15. Gaziantep ○	4	0	4	0	0	0
16. Hatay ○	4	0	4	0	0	0
17. Isparta	4	0	4	0	0	0
18. Istambul ○	28	0	28	0	0	0
19. Izmir ○	10	0	10	0	0	0
20. Içel ○	4	0	4	0	0	0
21. K. Maras	4	0	4	0	0	0
22. Kayseri ○	4	0	4	0	0	0
23. Kirikkale	4	0	4	0	0	0
24. Kirklareli	4	0	4	0	0	0
25. Kocaeli ○	4	0	4	0	0	0
26. Konya ○	4	0	4	0	0	0
27. Malatya	4	0	4	0	0	0
28. Manisa	4	0	4	0	0	0
29. Mugla	4	0	4	0	0	0
30. Nigde	4	0	4	0	0	0

Exhibit C.37: Allocation of School Sample in Turkey (Continued)

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
31. Osmaniye	4	0	4	0	0	0
32. Rize	4	0	4	0	0	0
33. Samsun ○	4	0	4	0	0	0
34. Sanliurfa	4	0	3	1	0	0
35. Sinop	4	0	4	0	0	0
36. Tekirdag	4	0	4	0	0	0
37. Trabzon	4	0	4	0	0	0
38. Van	4	0	4	0	0	0
39. Zonguldak	4	0	4	0	0	0
40. Çanakkale	4	0	4	0	0	0
Total	**204**	**0**	**202**	**2**	**0**	**0**

C.39 United States of America

C.39.1 Coverage and Exclusions

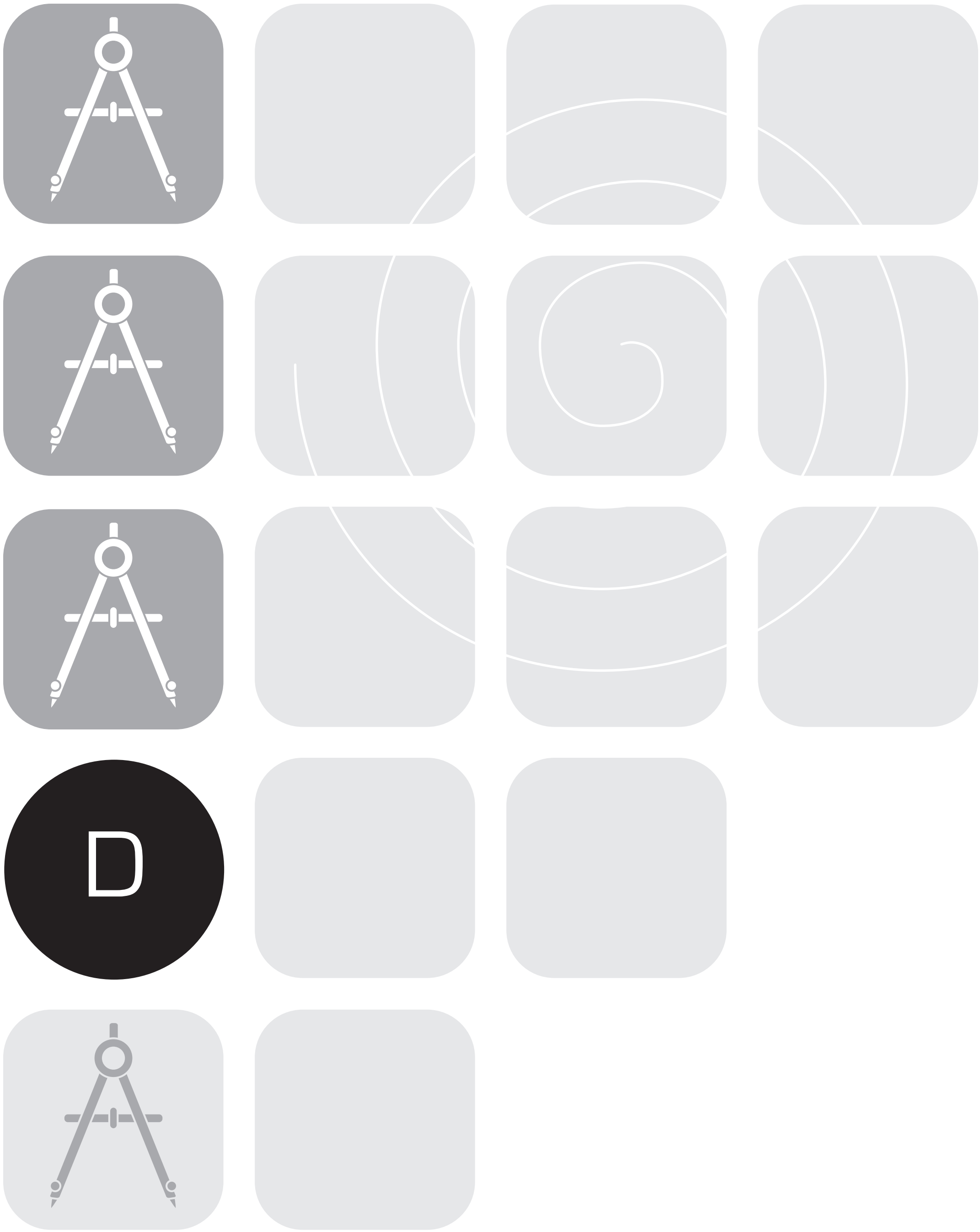
School-level exclusions consisted of schools in the Territories.

C.39.2 Sample Design

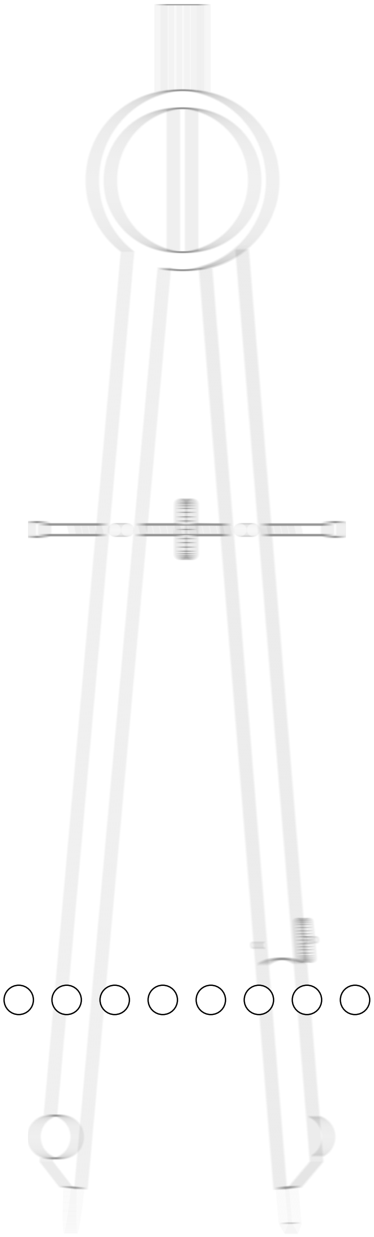
- Preliminary sampling of 52 primary sampling units (PSUs) from a list of 1 027 PSUs; 10 PSUs were large enough to be sampled with certainty.
- Special explicit stratification applied to the USA design, by school type and PSU size. This stratification is used for the computation of school participation adjustments and is presented in table C38.
- Implicit stratification by religious denomination and PSU within the private schools and by PSU and minority status within the public schools.
- Large school sample size because of preliminary sampling stage.

Exhibits C.38: Allocation of School Sample in the United States

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Private - Certainty PSUs (10)	18	0	12	2	0	4
Private - Large PSUs (6)	7	0	5	1	1	0
Private - Small PSUs (36)	25	1	18	4	1	1
Public - Certainty PSUs (10)	59	1	45	4	1	8
Public - Large PSUs (6)	23	0	18	2	0	3
Public - Small PSUs - Metro (18)	79	1	69	1	1	7
Public - Small PSUs - Non-Metro (18)	39	1	35	0	1	2
Total	**250**	**4**	**202**	**14**	**5**	**25**



Appendix D: Country Adaptations To Items and Item Scoring



D Country Adaptations To Items and Item Scoring

D.1 Items To Be Deleted In Countries

All Countries
M02 Mathematics (figure incorrectly drawn; no answer)

Bulgaria
H07 Mathematics (Barchart histogram of travel time – y-axis mislabeled)

Canada, French
H07 Mathematics (Barchart histogram of travel time – y-axis mislabeled)
I02 Mathematics (4/5 of books more than 2/3 – option C missing; it is the key)
T04 Mathematics (Height of stack from paper thickness – stem not correct)

Cyprus
F03 Science (Humans interpret senses – Wrong in 95, kept the same)
H01 Science (Not a function of the blood – Wrong in 95, kept the same)

Czech Republic
R01 Science (Bacteria/mold experiment – mistranslated stem)

Finland, Swedish
O04 Mathematics (Decimal rounded to nearest hundredth – mistranslated stem)

Hong Kong
E12 Science (Stone in underground caves – poor discrimination)
J18 Mathematics (Distance between towns from map reduced; scale incorrect)
Z02 Science (Diagram of rain from sea – scoring reliability less than 70%)

Hungary
C12 Science (Substance Not a fossil fuel – “fossil fuels” in stem is mistranslated)

Indonesia
I12 Science (Interdependence among organisms – five options instead of four)
J05 Science (Gravity acting on rocket – negative discrimination)

Iran
B07 Mathematics (Graph showing greatest increase – stem mistranslated in 1995 and 1999)

Israel, Hebrew

I07 Mathematics (Area of paved walkway around pool – misprint in key C)
O11 Science (Chemical change involving elements – mistranslated key)

Israel, Arabic

I17 Science (Animal on Earth longest time – translation error)

Italy

L11 Mathematics (Graph of humidity in room – vertical scale on map incorrect)

Japan

E12 Science (Stone in underground caves – poor discrimination)

Jordan

I17 Science (Animal on Earth longest time – translation error)
N09 Science (Balancing 10 and 5 liter buckets – printing problem, figure not clear)

Korea

A11 Science (Overgrazing by livestock – misprint)

Latvia

E12 Science (Stone in underground cave – poor discrimination)
L03 Science (Physical characteristic of prey – mistranslated in stem)

Macedonia (Both Languages)

I11 Science (Observation of bubbling electrodes – negative discrimination)

Macedonia, Albanian

P09 Mathematics (Expression equivalent to n x n x n – stem misprinted; underscores between n's instead of multiplication signs)

Macedonia, Macedonian

P09 Mathematics (Expression equivalent to n x n x n – stem misprinted; nothing between N’s)

Moldova (Both Languages)

G12 Science (Nonrenewable natural resource – negative discrimination)
R01 Science (Bacteria/Mold experiment – mistranslated stem)

Moldova, Russia

B07 Mathematics (Graph showing greatest increase – stem mistranslated)
E09 Science (Time/temperature table – mistranslated response options)
H07 Mathematics (Barchart histogram of travel time – y-axis mislabeled)

Morocco

P10 Mathematics (Diagram of similar triangles – error in labeled angle size)
P16 Mathematics (Day/Time in table at shown temperature – printing error; thermometer incorrectly shaded)

Philippines, Filipino

L03 Science (Physical characteristics of prey – mistranslation in stem)
P16 Mathematics (Day/Time in table at shown temperature – printing error; thermometer not shaded)

Romania, Hungarian

C12 Science (Substance Not a fossil fuel – "fossil fuels" in stem is mistranslated)

Romania, Romanian

E09 Science (Time/temperature table – Duplicate key resulting in double gridding)

Slovak Republic

F03 Science (Humans interpret senses – mistranslated stem)

South Africa, Afrikaans

B02 Science (Energy released from car engine – error in key, B)
H07 Mathematics (Barchart histogram of travel time – y-axis mislabeled)
I10 Science (Gas need for rust to form – poor discrimination)
I14 Science (Transmission of sound on the moon – five options instead of four)

Thailand

N17 Mathematics (Amount of paint left – mistranslated stem)
P16 Mathematics (Day/time in table at shown temperature – printing error; thermometer not shaded)

Tunisia

E12 Science (Stone in underground cave – poor discrimination)
O01 Mathematics (Graph of cooling water – printed twice in booklet)
P16 Mathematics (Day/time in table at shown temperature – printing error; thermometer not shaded)
U01 Mathematics (Printing error in table – missing "2" in first cooling time)

Turkey

I17 Science (Animal on Earth longest time – negative discrimination)
K11 Science (Thermometer scale for boiling water – negative discrimination)
R01 Science (Bacteria/mold experiment – negative discrimination)

D.2 Items Needing Options Changed

Korea

C02 Mathematics (Pie graph of crop distribution – switch options C & D)

Moldova (Both languages)

I10 Science (Gas need for rust to form – switch options A and B)

D.3 Items Needing Free-Response Category Recoding

Science

P06 (Digestion in stomach)

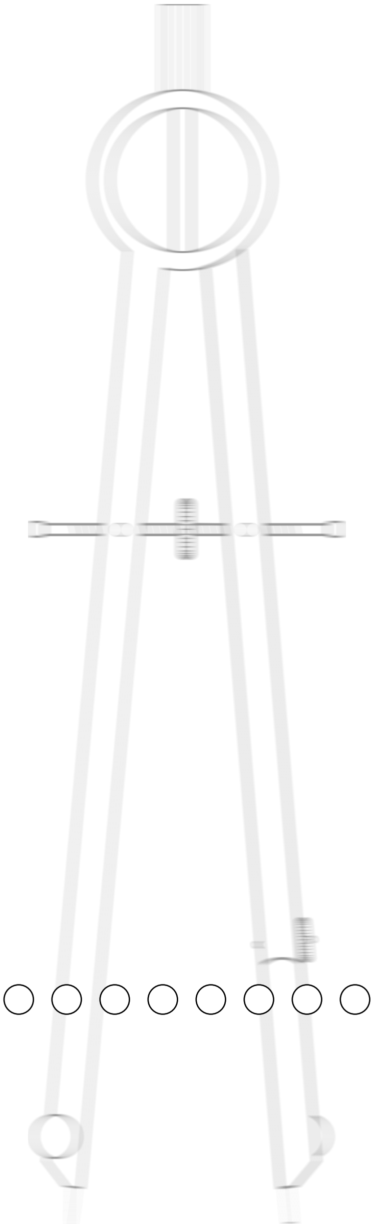
- Recode 10 into 11
- Combine 20, 21, and 22 and recode into 10
- Recode 29 into 19

W02 (Two reasons for water shortage) – For W02D only!

- Recode 10 into 71
- Recode 20 into 10

E

Appendix E: Parameters for IRT Analyses of TIMSS Achievement Data



E

Parameters for IRT Analyses of TIMSS Achievement Data

Exhibit E.1 IRT Parameters for Re-analysis of TIMSS 1995 Eighth-Grade Mathematics

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.
M012001	1.588	0.035	0.098	0.013	0.187	0.007
M012002	0.603	0.020	-0.984	0.081	0.195	0.029
M012003	0.747	0.017	-0.421	0.029	0.062	0.012
M012004	1.077	0.031	0.430	0.020	0.243	0.008
M012005	0.812	0.031	0.342	0.037	0.322	0.012
M012006	0.662	0.028	-0.573	0.086	0.380	0.025
M012007	0.807	0.033	-0.176	0.051	0.196	0.020
M012008	0.431	0.024	-0.363	0.121	0.301	0.028
M012009	0.943	0.049	0.448	0.042	0.343	0.014
M012010	1.395	0.038	0.150	0.016	0.072	0.008
M012011	0.978	0.035	-0.195	0.037	0.177	0.016
M012012	1.277	0.043	-0.124	0.027	0.227	0.013
M012013	1.179	0.050	0.176	0.032	0.234	0.014
M012014	0.978	0.045	-0.632	0.063	0.304	0.026
M012015	0.939	0.039	-0.137	0.043	0.172	0.018
M012016	1.390	0.093	0.940	0.033	0.428	0.010
M012017	0.781	0.038	0.253	0.048	0.164	0.018
M012018	0.549	0.029	-0.818	0.125	0.204	0.039
M012019	0.881	0.045	0.088	0.053	0.273	0.020
M012020	1.096	0.059	0.141	0.046	0.396	0.017
M012021	1.232	0.039	-0.431	0.027	0.099	0.014
M012022	0.729	0.058	1.125	0.054	0.273	0.016
M012023	0.542	0.026	-1.914	0.149	0.226	0.049
M012024	0.711	0.041	-0.434	0.097	0.310	0.032
M012025	0.655	0.030	-0.669	0.082	0.159	0.030
M012026	1.135	0.052	0.586	0.027	0.181	0.011
M012027	1.115	0.053	0.289	0.035	0.259	0.015
M012028	0.956	0.042	-0.258	0.049	0.232	0.021
M012029	0.953	0.040	0.129	0.036	0.154	0.016
M012030	1.384	0.057	0.659	0.021	0.157	0.009
M012031	1.382	0.071	1.112	0.024	0.163	0.007
M012032	0.421	0.027	-0.016	0.126	0.197	0.032
M012033	1.042	0.049	-0.011	0.045	0.301	0.018
M012034	0.734	0.033	0.104	0.048	0.125	0.019
M012035	1.512	0.055	0.402	0.019	0.135	0.009
M012036	0.634	0.037	0.266	0.072	0.192	0.024
M012037	0.668	0.050	0.666	0.071	0.287	0.022
M012038	0.913	0.045	-0.564	0.069	0.321	0.027
M012039	1.188	0.059	0.566	0.030	0.258	0.012
M012040	1.227	0.055	-0.076	0.038	0.323	0.016

Exhibit E.1 IRT Parameters for Re-analysis of TIMSS 1995 Eighth-Grade Mathematics (continued)

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.
M012041	1.041	0.040	-0.034	0.034	0.158	0.016
M012042	1.185	0.056	0.742	0.026	0.185	0.010
M012043	0.791	0.043	-0.120	0.071	0.292	0.025
M012044	1.135	0.047	-0.503	0.045	0.262	0.021
M012045	0.738	0.030	-1.560	0.093	0.181	0.040
M012046	1.227	0.056	0.603	0.026	0.197	0.011
M012047	1.180	0.051	0.128	0.034	0.260	0.015
M012048	1.032	0.046	-0.474	0.053	0.274	0.023
M012049	1.672	0.131	1.121	0.039	0.242	0.012
M012050	1.478	0.108	0.301	0.042	0.261	0.019
M012051	1.044	0.120	1.236	0.064	0.271	0.017
M012052	0.650	0.031	0.329	0.039	0.000	0.000
M012053	0.827	0.055	-0.684	0.093	0.194	0.037
M012054	0.430	0.028	-1.784	0.116	0.000	0.000
M012055	0.528	0.058	-0.024	0.185	0.273	0.049
M012056	0.980	0.102	1.003	0.061	0.247	0.020
M012057	1.076	0.070	-0.198	0.060	0.194	0.027
M012058	1.771	0.112	0.534	0.027	0.129	0.013
M012059	0.789	0.086	0.870	0.079	0.246	0.026
M012060	0.563	0.030	0.421	0.045	0.000	0.000
M012061	0.608	0.033	-1.765	0.087	0.000	0.000
M012062	1.258	0.109	0.777	0.045	0.236	0.018
M012063	0.611	0.062	-0.047	0.151	0.266	0.045
M012064	0.844	0.062	0.263	0.065	0.153	0.025
M012065	0.654	0.047	-0.421	0.108	0.171	0.037
M012066	1.014	0.129	1.353	0.073	0.301	0.018
M012067	0.555	0.041	-0.676	0.136	0.174	0.042
M012068	0.745	0.033	-0.586	0.039	0.000	0.000
M012069	0.682	0.074	0.013	0.146	0.354	0.042
M012070	1.200	0.126	1.114	0.053	0.259	0.016
M012071	1.482	0.057	0.390	0.021	0.000	0.000
M012072	0.886	0.086	0.942	0.061	0.188	0.021
M012073	1.566	0.133	0.612	0.041	0.312	0.017
M012074	0.964	0.081	1.033	0.051	0.128	0.015
M012075	1.313	0.082	0.303	0.037	0.143	0.017
M012076	0.575	0.056	0.158	0.132	0.209	0.040
M012077	0.669	0.048	-1.519	0.153	0.221	0.053
M012078	0.479	0.040	-2.291	0.230	0.221	0.056
M012079	1.542	0.143	1.673	0.068	0.276	0.011
M012080	0.966	0.059	-0.638	0.071	0.161	0.032
M012081	0.614	0.047	-2.218	0.188	0.225	0.057
M012082	1.031	0.115	1.527	0.069	0.140	0.013
M012083	0.961	0.078	0.906	0.050	0.131	0.017
M012084	0.895	0.039	0.483	0.031	0.000	0.000
M012085	1.288	0.084	0.342	0.039	0.137	0.018
M012086	0.814	0.055	-1.442	0.124	0.219	0.050
M012087	0.714	0.087	0.399	0.129	0.373	0.037

Exhibit E.1 IRT Parameters for Re-analysis of TIMSS 1995 Eighth-Grade Mathematics (continued 2)

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.
M012088	0.826	0.058	-0.726	0.105	0.224	0.042
M012089	1.396	0.082	0.551	0.030	0.073	0.012
M012090	0.702	0.065	0.507	0.085	0.176	0.029
M012091	0.989	0.043	0.612	0.030	0.000	0.000
M012092	0.634	0.048	-0.695	0.135	0.198	0.045
M012093	0.476	0.028	0.225	0.051	0.000	0.000
M012095	0.853	0.058	-1.210	0.116	0.229	0.048
M012096	0.752	0.086	0.292	0.127	0.395	0.036
M012097	0.921	0.038	0.145	0.029	0.000	0.000
M012098	1.331	0.107	0.073	0.060	0.366	0.025
M012099	0.305	0.033	-0.452	0.268	0.233	0.048
M012100	0.665	0.094	1.250	0.107	0.285	0.028
M012101	1.447	0.135	1.391	0.057	0.307	0.012
M012102	0.976	0.076	0.289	0.064	0.214	0.025
M012103	1.414	0.054	0.022	0.022	0.000	0.000
M012104	0.855	0.053	0.003	0.060	0.116	0.024
M012105	1.557	0.105	1.007	0.031	0.083	0.009
M012106	1.340	0.102	0.592	0.041	0.213	0.017
M012107	1.132	0.103	0.835	0.051	0.244	0.018
M012108	1.858	0.110	0.402	0.026	0.120	0.012
M012109	0.763	0.034	-0.838	0.044	0.000	0.000
M012110	0.916	0.076	-0.167	0.094	0.308	0.035
M012111	0.976	0.078	-0.067	0.081	0.294	0.032
M012112	1.166	0.045	0.110	0.024	0.000	0.000
M012113	1.312	0.104	0.347	0.050	0.274	0.021
M012114	1.497	0.136	1.259	0.049	0.281	0.013
M012115	1.245	0.095	0.328	0.050	0.243	0.022
M012116	0.504	0.053	0.404	0.145	0.195	0.040
M012117	0.775	0.063	-0.449	0.117	0.256	0.042
M012118	0.861	0.058	-0.401	0.081	0.175	0.033
M012119	0.980	0.083	-0.533	0.107	0.366	0.040
M012120	1.179	0.093	0.096	0.062	0.297	0.026
M012121	1.018	0.044	0.596	0.030	0.000	0.000
M012122	0.391	0.037	-2.464	0.284	0.219	0.056
M012123	0.932	0.091	0.856	0.062	0.216	0.022
M012124	0.917	0.087	0.690	0.067	0.250	0.024
M012125	0.969	0.081	0.973	0.051	0.130	0.016
M012126	0.556	0.041	-1.827	0.175	0.205	0.052
M012127	0.841	0.072	-0.088	0.098	0.281	0.036
M012128	0.869	0.067	0.732	0.055	0.124	0.019
M012129	1.037	0.079	0.097	0.065	0.232	0.027
M012130	1.880	0.117	0.419	0.027	0.136	0.013
M012131	1.112	0.087	0.390	0.054	0.208	0.022
M012132	0.704	0.033	0.255	0.037	0.000	0.000
M012133	0.622	0.043	-0.991	0.128	0.180	0.043
M012134	1.365	0.103	0.615	0.040	0.205	0.017
M012135	0.889	0.082	0.650	0.068	0.242	0.024

Exhibit E.1 IRT Parameters for Re-analysis of TIMSS 1995 Eighth-Grade Mathematics (continued 3)

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.	step parameter d_{ji}	S.E.
M012136	1.353	0.097	0.625	0.037	0.160	0.016		
M012137	0.934	0.071	0.328	0.061	0.183	0.024		
M012138	1.072	0.093	0.704	0.053	0.220	0.020		
M012139	0.804	0.054	-1.576	0.125	0.208	0.049		
M012140	1.066	0.045	0.623	0.029	0.000	0.000		
M012141	1.383	0.055	0.405	0.022	0.000	0.000		
M012142A	0.502	0.030	-1.689	0.098	0.000	0.000		
M012142B	0.882	0.044	1.110	0.045	0.000	0.000		
M012143A	0.872	0.040	-0.461	0.038	0.000	0.000		
M012143B	1.426	0.064	0.758	0.025	0.000	0.000		
M012143C	1.108	0.056	0.942	0.035	0.000	0.000		
M012144D	0.436	0.007	0.644	0.018	0.000	0.000		
						d0		
						d1	-3.183	0.101
						d2	0.074	0.146
						d3	3.108	0.114
M012145A	1.183	0.040	1.022	0.024	0.000	0.000		
M012145B	0.716	0.042	2.316	0.098	0.000	0.000		
M012146D	0.459	0.011	0.752	0.026	0.000	0.000		
						d0		
						d1	-0.910	0.051
						d2	0.910	0.057
M012147A	0.569	0.013	0.525	0.020	0.000	0.000		
						d0		
						d1	-1.277	0.053
						d2	1.277	0.056
M012147B			0.869	0.028	1.238	0.026	0.000	0.000
						d0		
						d1	0.095	0.029
						d2	-0.095	0.043
M012148	0.887	0.027	-0.088	0.022	0.000	0.000		
M012149	0.519	0.011	0.812	0.019	0.000	0.000		
						d0		
						d1	-0.227	0.041
						d2	-1.043	0.073
						d3	1.270	0.074
M012150	1.091	0.052	0.567	0.029	0.096	0.012		
M012151	0.987	0.030	0.423	0.020	0.000	0.000		

Exhibit E.2 IRT Parameters for Re-analysis of TIMSS 1995 Eighth-Grade Science

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.
S012001	0.453	0.016	-0.529	0.078	0.209	0.021
S012002	0.415	0.015	-0.317	0.080	0.247	0.019
S012003	0.771	0.022	-0.926	0.054	0.201	0.023
S012004	0.497	0.021	-0.188	0.085	0.333	0.021
S012005	0.584	0.022	0.060	0.058	0.192	0.019
S012006	0.656	0.024	0.072	0.050	0.212	0.017
S012007	0.567	0.025	-1.864	0.137	0.295	0.044
S012008	0.800	0.073	1.370	0.056	0.431	0.013
S012009	0.907	0.054	1.511	0.037	0.143	0.008
S012010	0.714	0.028	-1.928	0.107	0.206	0.046
S012011	0.567	0.060	1.493	0.077	0.344	0.019
S012012	0.730	0.049	-0.503	0.126	0.557	0.029
S012013	0.603	0.037	1.001	0.053	0.122	0.017
S012014	0.659	0.038	-0.509	0.109	0.281	0.036
S012015	0.504	0.031	-0.545	0.140	0.327	0.036
S012016	0.426	0.025	-0.995	0.155	0.339	0.035
S012017	1.057	0.060	0.837	0.034	0.259	0.012
S012018	0.938	0.089	1.339	0.054	0.433	0.012
S012019	0.561	0.035	0.871	0.060	0.126	0.019
S012020	0.738	0.034	-0.724	0.082	0.223	0.032
S012021	0.806	0.045	0.973	0.038	0.150	0.013
S012022	0.618	0.049	0.629	0.085	0.316	0.024
S012023	0.842	0.048	-0.014	0.067	0.340	0.023
S012024	0.758	0.040	-0.489	0.085	0.298	0.030
S012025	0.818	0.090	1.768	0.073	0.337	0.012
S012026	0.442	0.025	-1.302	0.155	0.303	0.038
S012027	0.583	0.027	-1.291	0.115	0.179	0.041
S012028	0.601	0.034	0.227	0.074	0.168	0.025
S012029	0.537	0.047	0.528	0.115	0.308	0.030
S012030	0.550	0.055	0.940	0.102	0.336	0.027
S012031	0.729	0.041	-0.155	0.080	0.288	0.027
S012032	0.988	0.038	-0.257	0.040	0.162	0.018
S012033	0.295	0.020	-0.351	0.159	0.276	0.028
S012034	0.719	0.038	-0.390	0.083	0.252	0.029
S012035	0.633	0.032	-1.269	0.124	0.246	0.044
S012036	0.813	0.037	-0.446	0.064	0.216	0.026
S012037	0.547	0.027	-2.017	0.151	0.233	0.050
S012038	0.763	0.051	0.405	0.068	0.337	0.021
S012039	0.542	0.032	-0.717	0.138	0.308	0.039
S012040	1.266	0.069	0.681	0.031	0.322	0.011
S012041	0.529	0.053	1.063	0.096	0.293	0.026
S012042	0.742	0.068	1.163	0.063	0.371	0.017
S012043	0.586	0.038	-0.361	0.126	0.289	0.037
S012044	0.427	0.023	-1.469	0.154	0.251	0.039
S012045	0.959	0.053	-0.490	0.075	0.430	0.026
S012046	0.530	0.041	0.687	0.092	0.221	0.027
S012047	0.738	0.061	1.901	0.067	0.126	0.010

Exhibit E.2 IRT Parameters for Re-analysis of TIMSS 1995 Eighth-Grade Science (continued)

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.
S012048	0.795	0.042	0.422	0.049	0.199	0.018
S012049	0.414	0.039	-1.065	0.246	0.249	0.057
S012050	0.790	0.078	0.742	0.077	0.219	0.026
S012051	1.121	0.101	1.048	0.048	0.178	0.016
S012052	0.919	0.074	0.097	0.079	0.241	0.031
S012053	0.436	0.053	0.513	0.192	0.231	0.046
S012054	0.662	0.054	0.431	0.083	0.138	0.028
S012055	0.450	0.039	-1.990	0.250	0.254	0.059
S012056	0.573	0.086	1.441	0.122	0.234	0.031
S012057	0.636	0.032	0.719	0.046	0.000	0.000
S012058	1.016	0.069	0.286	0.052	0.149	0.022
S012059	0.758	0.101	1.343	0.090	0.246	0.023
S012060	0.685	0.062	-0.083	0.121	0.256	0.040
S012061	0.609	0.034	1.145	0.062	0.000	0.000
S012062	1.170	0.111	0.948	0.051	0.242	0.017
S012063	0.587	0.046	-0.362	0.124	0.174	0.040
S012064	1.076	0.113	1.523	0.064	0.122	0.012
S012065	0.639	0.066	0.666	0.099	0.194	0.032
S012066	0.833	0.074	0.679	0.067	0.200	0.023
S012067	0.615	0.033	-1.460	0.072	0.000	0.000
S012068	0.261	0.025	1.587	0.169	0.000	0.000
S012069	0.817	0.076	0.353	0.086	0.259	0.030
S012070	0.640	0.061	0.389	0.106	0.208	0.034
S012071	0.835	0.057	-0.745	0.099	0.214	0.040
S012072	0.756	0.058	-1.067	0.140	0.279	0.051
S012073	0.885	0.082	0.280	0.086	0.306	0.030
S012074	0.888	0.089	1.080	0.062	0.177	0.019
S012075	0.737	0.078	0.672	0.092	0.277	0.029
S012076	0.487	0.054	0.388	0.159	0.222	0.042
S012078	0.597	0.057	0.557	0.102	0.160	0.032
S012079	0.784	0.074	0.519	0.082	0.239	0.028
S012080	0.800	0.066	-0.311	0.110	0.261	0.040
S012081	0.700	0.034	0.744	0.044	0.000	0.000
S012082	0.415	0.042	-0.324	0.219	0.233	0.051
S012083	0.617	0.064	0.483	0.115	0.221	0.035
S012084	0.574	0.044	-0.606	0.138	0.198	0.043
S012085	0.558	0.068	0.911	0.119	0.217	0.034
S012086	0.793	0.035	-0.646	0.039	0.000	0.000
S012087	0.692	0.032	-0.117	0.037	0.000	0.000
S012088	0.752	0.067	0.540	0.077	0.173	0.027
S012089	0.401	0.027	-1.404	0.103	0.000	0.000
S012090	0.800	0.129	1.571	0.105	0.328	0.022
S012091	0.611	0.086	1.414	0.111	0.226	0.028
S012092	0.664	0.063	0.075	0.125	0.239	0.041
S012093	0.331	0.041	0.820	0.216	0.186	0.042
S012094	1.010	0.160	2.089	0.135	0.268	0.013
S012095	0.755	0.070	0.042	0.114	0.284	0.038
S012096	0.891	0.049	-1.894	0.075	0.000	0.000

Exhibit E.2 IRT Parameters for Re-analysis of TIMSS 1995 Eighth-Grade Science (continued 2)

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.	step parameter d_{ji}	S.E.
S012097	0.466	0.040	-0.764	0.189	0.219	0.050		
S012098	0.735	0.064	0.540	0.080	0.168	0.028		
S012099	0.585	0.030	-0.096	0.043	0.000	0.000		
S012100	0.564	0.028	-0.197	0.044	0.000	0.000		
S012101	1.035	0.131	1.422	0.070	0.256	0.017		
S012102	0.961	0.090	1.395	0.059	0.104	0.013		
S012103	0.447	0.045	0.056	0.179	0.211	0.045		
S012104	0.745	0.033	-0.163	0.035	0.000	0.000		
S012105	0.866	0.101	1.217	0.071	0.229	0.021		
S012106	0.633	0.030	-0.366	0.042	0.000	0.000		
S012107	0.354	0.025	-0.317	0.070	0.000	0.000		
S012108	0.898	0.062	-1.139	0.108	0.242	0.046		
S012109	0.695	0.039	1.349	0.065	0.000	0.000		
S012110	0.747	0.038	-1.341	0.061	0.000	0.000		
S012111	0.688	0.069	0.318	0.109	0.221	0.037		
S012112	0.511	0.029	-0.307	0.050	0.000	0.000		
S012113	0.506	0.018	0.086	0.030	0.000	0.000		
						d0		
						d1	-0.413	0.062
						d2	0.413	0.062
S012114	0.355	0.050	0.922	0.226	0.249	0.044		
S012115	0.949	0.112	1.114	0.069	0.269	0.021		
S012116	0.500	0.029	0.205	0.049	0.000	0.000		
S012117	0.631	0.054	-0.055	0.119	0.197	0.039		
S012118	0.486	0.079	1.250	0.159	0.241	0.041		
S012119	0.874	0.119	1.546	0.088	0.207	0.019		
S012120	1.000	0.149	1.806	0.101	0.187	0.015		
S012121	0.508	0.030	-0.778	0.065	0.000	0.000		
S012122	0.270	0.010	0.889	0.058	0.000	0.000		
						d0		
						d1	-3.037	0.155
						d2	3.037	0.164
S012123	0.453	0.054	-0.308	0.252	0.321	0.057		
S012124	0.893	0.132	1.514	0.091	0.280	0.019		
S012125	0.622	0.027	0.987	0.036	0.000	0.000		
						d0		
						d1	0.666	0.043
						d2	-0.666	0.063
S012126	0.871	0.037	0.021	0.030	0.000	0.000	0.51	
S012127	0.913	0.038	0.046	0.029	0.000	0.000	0.72	
S012128A			0.422	0.021	-2.121	0.102	0.000	0.000
S012128B			0.298	0.018	0.665	0.065	0.000	0.000
S012129	0.399	0.011	0.375	0.027	0.000	0.000		
						d0		
						d1	-0.781	0.057
						d2	0.781	0.060

Exhibit E.2 IRT Parameters for Re-analysis of TIMSS 1995 Eighth-Grade Science (continued 3)

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.	step parameter d_{ji}	S.E.
S012130	0.504	0.016	1.364	0.039	0.000	0.000		
						d0		
						d1	0.014	0.041
						d2	-0.014	0.059
S012131A	0.383	0.020	-1.267	0.078	0.000	0.000		
S012131B	0.530	0.023	0.902	0.045	0.000	0.000		
S012132	0.893	0.046	2.286	0.080	0.000	0.000		
S012133	0.253	0.013	1.885	0.092	0.000	0.000		
						d0		
						d1	0.375	0.074
						d2	-0.375	0.107
S012134A	0.455	0.028	-0.721	0.068	0.000	0.000		
S012134D	0.624	0.027	0.742	0.033	0.000	0.000		
						d0		
						d1	0.289	0.048
						d2	-0.289	0.058
S012135D	0.369	0.020	-0.637	0.057	0.000	0.000		
						d0		
						d1	0.849	0.098
						d2	-0.849	0.076

Exhibit E.3 **IRT Parameters for TIMSS Joint 1995-1999 Eighth Grade Mathematics**

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.
M012001	1.557	0.022	-0.079	0.009	0.159	0.005
M012002	0.619	0.010	-1.253	0.038	0.068	0.015
M012003	0.882	0.012	-0.377	0.015	0.048	0.006
M012004	1.281	0.026	0.481	0.012	0.303	0.005
M012005	0.668	0.016	-0.211	0.037	0.233	0.013
M012006	0.647	0.016	-0.894	0.055	0.269	0.019
M012007	0.824	0.022	-0.228	0.034	0.201	0.014
M012008	0.462	0.015	-0.919	0.079	0.176	0.023
M012009	0.879	0.030	0.221	0.033	0.330	0.012
M012010	1.434	0.025	0.091	0.010	0.055	0.005
M012011	1.074	0.024	-0.291	0.021	0.157	0.010
M012012	1.480	0.031	-0.433	0.016	0.202	0.009
M012013	1.181	0.030	-0.049	0.021	0.184	0.010
M012014	0.930	0.027	-0.878	0.043	0.215	0.019
M012015	0.942	0.023	-0.475	0.028	0.117	0.013
M012016	1.648	0.070	0.803	0.019	0.444	0.006
M012017	0.790	0.023	-0.016	0.031	0.111	0.013
M012018	0.644	0.024	-0.628	0.070	0.237	0.024
M012019	0.771	0.022	-0.414	0.040	0.139	0.016
M012020	1.328	0.043	-0.127	0.027	0.401	0.011
M012021	1.419	0.031	-0.379	0.017	0.142	0.009
M012022	0.588	0.031	0.602	0.057	0.237	0.018
M012023	0.650	0.022	-1.724	0.100	0.206	0.039
M012024	0.831	0.030	-0.286	0.049	0.334	0.017
M012025	0.699	0.019	-0.861	0.047	0.095	0.019
M012026	1.252	0.033	0.260	0.018	0.192	0.008
M012027	1.317	0.037	0.136	0.020	0.254	0.009
M012028	1.117	0.028	-0.423	0.026	0.198	0.012
M012029	1.009	0.026	-0.093	0.024	0.164	0.011
M012030	1.450	0.036	0.424	0.013	0.146	0.006
M012031	1.440	0.043	0.853	0.014	0.153	0.005
M012032	0.413	0.016	-0.398	0.086	0.126	0.023
M012033	1.119	0.032	-0.105	0.027	0.282	0.012
M012034	0.730	0.023	0.098	0.034	0.142	0.013
M012035	1.696	0.041	0.294	0.012	0.171	0.007
M012036	0.724	0.028	0.331	0.039	0.216	0.014
M012037	0.676	0.029	0.384	0.046	0.221	0.016
M012038	0.982	0.031	-0.538	0.041	0.329	0.017
M012039	0.988	0.028	0.088	0.025	0.168	0.011
M012040	1.207	0.032	-0.376	0.026	0.257	0.013
M012041	1.111	0.028	-0.127	0.022	0.163	0.011
M012042	1.148	0.031	0.132	0.021	0.185	0.010
M012043	0.833	0.027	-0.348	0.043	0.214	0.017
M012044	1.196	0.031	-0.535	0.027	0.223	0.013
M012045	0.822	0.022	-1.531	0.053	0.123	0.024
M012046	1.452	0.037	0.242	0.015	0.187	0.008
M012047	1.278	0.036	-0.013	0.023	0.277	0.011

Exhibit E.3 IRT Parameters for TIMSS Joint 1995-1999 Eighth Grade Mathematics (continued)

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.
M012048	1.063	0.031	-0.550	0.035	0.295	0.016
M012049	1.814	0.139	0.879	0.032	0.255	0.012
M012050	1.719	0.109	0.110	0.035	0.258	0.019
M012051	1.112	0.113	1.030	0.053	0.270	0.017
M012052	0.725	0.032	0.208	0.032	0.000	0.000
M012053	0.981	0.062	-0.616	0.075	0.214	0.033
M012054	0.578	0.029	-1.343	0.069	0.000	0.000
M012055	0.519	0.048	-0.235	0.160	0.218	0.045
M012056	0.936	0.086	0.844	0.056	0.200	0.021
M012057	1.138	0.067	-0.382	0.055	0.203	0.026
M012058	1.639	0.091	0.302	0.028	0.116	0.014
M012059	0.678	0.059	0.317	0.090	0.167	0.032
M012060	0.684	0.031	-0.043	0.035	0.000	0.000
M012061	0.732	0.035	-1.471	0.065	0.000	0.000
M012062	1.056	0.085	0.467	0.056	0.233	0.023
M012063	0.589	0.043	-0.722	0.132	0.168	0.042
M012064	0.939	0.062	-0.004	0.062	0.167	0.026
M012065	0.736	0.052	-0.351	0.096	0.186	0.036
M012066	0.727	0.087	1.098	0.082	0.226	0.026
M012067	0.587	0.043	-0.765	0.139	0.183	0.045
M012068	0.797	0.033	-0.682	0.038	0.000	0.000
M012069	0.778	0.068	-0.058	0.108	0.306	0.036
M012070	1.143	0.103	0.749	0.052	0.273	0.020
M012071	1.488	0.053	0.217	0.019	0.000	0.000
M012072	0.946	0.070	0.598	0.051	0.151	0.020
M012073	1.539	0.105	0.261	0.040	0.270	0.019
M012074	1.020	0.071	0.762	0.042	0.123	0.016
M012075	1.408	0.077	0.014	0.036	0.144	0.018
M012076	0.590	0.049	-0.054	0.120	0.184	0.038
M012077	0.844	0.052	-1.173	0.101	0.178	0.040
M012078	0.625	0.039	-1.732	0.135	0.164	0.043
M012079	1.984	0.124	1.355	0.039	0.290	0.010
M012080	0.978	0.055	-0.620	0.065	0.152	0.028
M012081	0.718	0.046	-2.168	0.141	0.180	0.048
M012082	1.179	0.107	1.225	0.044	0.138	0.013
M012083	1.071	0.069	0.567	0.040	0.120	0.017
M012084	0.904	0.036	-0.045	0.028	0.000	0.000
M012085	1.343	0.083	0.101	0.042	0.195	0.021
M012086	0.821	0.053	-1.580	0.124	0.208	0.049
M012087	0.618	0.051	-0.381	0.134	0.220	0.043
M012088	0.794	0.049	-0.748	0.089	0.168	0.035
M012089	1.265	0.071	0.486	0.031	0.087	0.013
M012090	0.782	0.057	0.208	0.070	0.154	0.027
M012091	1.013	0.040	0.361	0.026	0.000	0.000
M012092	0.660	0.054	-0.641	0.145	0.257	0.047
M012093	0.462	0.026	-0.187	0.050	0.000	0.000
M012095	1.187	0.082	-0.735	0.079	0.316	0.037
M012096	0.775	0.074	0.043	0.119	0.354	0.038

Exhibit E.3 IRT Parameters for TIMSS Joint 1995-1999 Eighth Grade Mathematics (continued 2)

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.
M012097	0.815	0.034	-0.303	0.033	0.000	0.000
M012098	1.331	0.093	-0.098	0.057	0.316	0.027
M012099	0.383	0.037	-0.511	0.226	0.204	0.051
M012100	1.044	0.121	1.052	0.064	0.347	0.020
M012101	1.460	0.141	1.009	0.043	0.287	0.015
M012102	1.066	0.079	0.298	0.058	0.240	0.024
M012103	1.495	0.055	-0.087	0.020	0.000	0.000
M012104	0.923	0.058	-0.013	0.058	0.146	0.024
M012105	1.631	0.101	0.828	0.027	0.084	0.010
M012106	1.047	0.069	0.294	0.049	0.161	0.021
M012107	0.938	0.083	0.672	0.060	0.221	0.022
M012108	1.706	0.093	0.257	0.027	0.112	0.013
M012109	0.927	0.038	-0.495	0.032	0.000	0.000
M012110	1.130	0.074	-0.615	0.071	0.258	0.033
M012111	1.034	0.075	-0.428	0.081	0.305	0.034
M012112	1.278	0.048	0.036	0.022	0.000	0.000
M012113	1.191	0.081	0.028	0.053	0.256	0.024
M012114	1.802	0.142	0.842	0.031	0.230	0.012
M012115	1.444	0.098	0.046	0.046	0.273	0.022
M012116	0.603	0.052	0.291	0.100	0.160	0.032
M012117	0.947	0.064	-0.519	0.081	0.234	0.034
M012118	0.987	0.054	-0.485	0.056	0.125	0.025
M012119	1.129	0.088	-0.344	0.081	0.399	0.032
M012120	1.325	0.083	-0.340	0.053	0.264	0.026
M012121	1.143	0.044	0.282	0.024	0.000	0.000
M012122	0.514	0.036	-1.820	0.167	0.173	0.046
M012123	0.807	0.064	0.299	0.072	0.172	0.028
M012124	1.062	0.074	0.238	0.053	0.209	0.022
M012125	0.927	0.070	0.706	0.049	0.120	0.018
M012126	0.738	0.044	-1.452	0.107	0.156	0.040
M012127	0.960	0.073	-0.072	0.077	0.298	0.030
M012128	0.958	0.061	0.474	0.044	0.104	0.017
M012129	1.219	0.073	-0.401	0.053	0.199	0.027
M012130	2.082	0.119	0.393	0.023	0.129	0.012
M012131	1.256	0.082	0.069	0.047	0.207	0.022
M012132	0.775	0.033	0.032	0.031	0.000	0.000
M012133	0.759	0.052	-0.799	0.107	0.201	0.040
M012134	1.453	0.101	0.461	0.037	0.222	0.017
M012135	0.815	0.065	0.518	0.064	0.171	0.024
M012136	1.372	0.087	0.451	0.035	0.156	0.016
M012137	0.987	0.064	0.056	0.055	0.158	0.024
M012138	1.311	0.097	0.478	0.043	0.249	0.018
M012139	0.863	0.056	-1.478	0.112	0.202	0.046
M012140	1.272	0.048	0.487	0.023	0.000	0.000
M012141	1.554	0.056	0.260	0.019	0.000	0.000
M012142A	0.732	0.034	-1.279	0.056	0.000	0.000

Exhibit E.3 IRT Parameters for TIMSS Joint 1995-1999 Eighth Grade Mathematics (continued 3)

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.	step parameter d_{ji}	S.E.
M012142B	0.883	0.039	0.783	0.034	0.000	0.000		
M012143A	1.014	0.042	-0.610	0.033	0.000	0.000		
M012143B	1.420	0.056	0.422	0.021	0.000	0.000		
M012143C	1.133	0.050	0.732	0.028	0.000	0.000		
M012144D	0.488	0.007	0.415	0.014	0.000	0.000		
						d0		
						d1	-3.009	0.092
						d2	0.281	0.124
						d3	2.728	0.089
M012145A	1.185	0.037	0.922	0.021	0.000	0.000		
M012145B	0.742	0.039	2.135	0.082	0.000	0.000		
M012146D	0.466	0.011	0.662	0.023	0.000	0.000		
						d0		
						d1	-0.887	0.047
						d2	0.887	0.052
M012147A	0.497	0.010	0.425	0.020	0.000	0.000		
						d0		
						d1	-1.491	0.056
						d2	1.491	0.058
M012147B	0.828	0.021	0.917	0.018	0.000	0.000		
						d0		
						d1	-0.120	0.028
						d2	0.120	0.035
M012148	0.941	0.027	-0.162	0.020	0.000	0.000		
M012149	0.546	0.011	0.659	0.016	0.000	0.000		
						d0		
						d1	-0.179	0.038
						d2	-1.053	0.065
						d3	1.232	0.065
M012150	1.233	0.054	0.408	0.027	0.128	0.012		
M012151	1.021	0.029	0.245	0.018	0.000	0.000		
M022002	1.800	0.117	1.035	0.029	0.142	0.009		
M022004	1.766	0.137	0.479	0.037	0.326	0.016		
M022005	1.243	0.132	1.076	0.052	0.303	0.016		
M022008	0.653	0.032	0.643	0.042	0.000	0.000		
M022010	0.932	0.055	-0.401	0.063	0.149	0.028		
M022012	0.624	0.029	-0.719	0.047	0.000	0.000		
M022016	0.987	0.100	0.964	0.058	0.233	0.020		
M022021	1.675	0.110	0.446	0.032	0.194	0.015		
M022022	1.928	0.131	0.607	0.028	0.213	0.013		
M022026	0.770	0.032	0.050	0.031	0.000	0.000		
M022030	1.040	0.041	-0.931	0.035	0.000	0.000		
M022031	1.166	0.088	0.737	0.041	0.186	0.016		
M022033	0.522	0.041	-0.405	0.137	0.164	0.040		
M022037	0.766	0.053	-0.179	0.080	0.185	0.029		
M022038	0.654	0.045	-0.266	0.091	0.154	0.031		
M022041	0.581	0.057	-0.296	0.168	0.291	0.047		

Exhibit E.3 IRT Parameters for TIMSS Joint 1995-1999 Eighth Grade Mathematics (continued 4)

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.	step parameter d_{ji}	S.E.
M022042	2.144	0.117	0.244	0.023	0.125	0.013		
M022043	0.639	0.044	-0.939	0.123	0.168	0.041		
M022046	0.735	0.032	-0.728	0.042	0.000	0.000		
M022049	0.650	0.062	-0.171	0.139	0.297	0.042		
M022050	1.078	0.087	0.852	0.044	0.159	0.016		
M022055	1.380	0.051	0.301	0.021	0.000	0.000		
M022057	0.397	0.033	-0.911	0.197	0.164	0.046		
M022062	1.044	0.066	0.472	0.042	0.119	0.017		
M022066	1.314	0.067	-0.137	0.036	0.102	0.017		
M022070	0.699	0.043	-1.135	0.105	0.152	0.037		
M022165	1.201	0.085	0.168	0.052	0.285	0.022		
M022166	1.197	0.074	0.245	0.042	0.166	0.019		
M022168	0.758	0.076	0.807	0.075	0.216	0.025		
M022169	0.913	0.060	-0.454	0.077	0.206	0.032		
M022172	1.241	0.085	0.108	0.050	0.254	0.022		
M022173	1.385	0.096	0.409	0.040	0.220	0.018		
M022176	0.897	0.073	-0.384	0.102	0.360	0.036		
M022178	1.220	0.046	0.260	0.023	0.000	0.000		
M022181	1.127	0.072	-0.958	0.078	0.241	0.036		
M022185	0.787	0.067	0.136	0.089	0.246	0.032		
M022188	0.964	0.090	0.804	0.059	0.260	0.020		
M022189	0.874	0.053	-0.827	0.080	0.172	0.033		
M022191	0.757	0.052	-0.503	0.095	0.187	0.036		
M022194	0.902	0.065	0.137	0.065	0.198	0.026		
M022196	1.313	0.069	-0.266	0.039	0.126	0.020		
M022198	1.185	0.084	0.512	0.043	0.191	0.018		
M022199	1.300	0.097	0.450	0.044	0.226	0.019		
M022202	0.800	0.035	0.631	0.035	0.000	0.000		
M022204	0.653	0.048	-1.044	0.141	0.224	0.047		
M022206	2.258	0.176	0.694	0.031	0.383	0.013		
M022207	1.034	0.054	0.016	0.047	0.235	0.020		
M022208	0.878	0.066	0.112	0.071	0.221	0.028		
M022210	1.515	0.088	0.456	0.029	0.123	0.013		
M022213	0.912	0.064	0.261	0.057	0.145	0.024		
M022219	1.038	0.042	0.593	0.028	0.000	0.000		
M022222	1.086	0.041	0.062	0.024	0.000	0.000		
M022227A	1.170	0.046	-0.453	0.027	0.000	0.000		
M022227B	1.485	0.056	0.392	0.020	0.000	0.000		
M022227C	1.360	0.055	0.703	0.024	0.000	0.000		
M022228	0.652	0.012	0.299	0.016	0.000	0.000		
						d0		
						d1	-1.647	0.053
						d2	1.647	0.054
M022231A	1.322	0.037	0.655	0.017	0.000	0.000		
M022231B	1.622	0.055	1.211	0.021	0.000	0.000		

Exhibit E.3 IRT Parameters for TIMSS Joint 1995-1999 Eighth Grade Mathematics (continued 5)

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.	step parameter d_{ji}	S.E.
M022232	0.557	0.012	1.250	0.025	0.000	0.000		
						d0		
						d1	-2.077	0.070
						d2	2.077	0.076
M022234A	0.815	0.016	0.445	0.014	0.000	0.000		
						d0		
						d1	-0.652	0.032
						d2	0.652	0.034
M022234B	0.863	0.017	0.791	0.014	0.000	0.000		
						d0		
						d1	-1.342	0.048
						d2	1.342	0.050
M022237	0.929	0.026	-0.033	0.020	0.000	0.000		
M022241	1.275	0.052	0.295	0.025	0.118	0.012		
M022243	1.014	0.030	0.168	0.019	0.000	0.000		
M022244	1.194	0.032	-0.010	0.017	0.000	0.000		
M022245	0.707	0.050	-0.899	0.121	0.214	0.043		
M022246	0.722	0.053	0.070	0.080	0.169	0.028		
M022249	1.012	0.071	0.287	0.055	0.203	0.022		
M022251	0.929	0.111	1.368	0.069	0.208	0.017		
M022252	0.989	0.071	-0.237	0.075	0.271	0.031		
M022253	1.067	0.041	-0.330	0.027	0.000	0.000		
M022256	0.755	0.018	0.381	0.015	0.000	0.000		
						d0		
						d1	0.093	0.027
						d2	-0.093	0.029
M022257	1.354	0.089	0.360	0.039	0.197	0.018		
M022258	0.645	0.047	-0.210	0.098	0.155	0.033		
M022260	0.932	0.068	-1.337	0.120	0.272	0.048		
M022261A	1.193	0.045	0.065	0.023	0.000	0.000		
M022261B	1.399	0.054	0.596	0.022	0.000	0.000		
M022261C	0.780	0.022	0.837	0.023	0.000	0.000		
						d0		
						d1	-1.893	0.094
						d2	1.893	0.097
M022262A	0.997	0.028	-0.770	0.025	0.000	0.000		
M022262B	0.988	0.028	-0.366	0.021	0.000	0.000		
M022262C	0.648	0.012	0.454	0.016	0.000	0.000		
						d0		
						d1	-1.520	0.052
						d2	1.520	0.053

Exhibit E.4 IRT Parameters for TIMSS Joint 1995 – 1999 Eighth-Grade Science

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.
S012001	0.494	0.012	-0.987	0.062	0.132	0.019
S012002	0.532	0.017	-0.201	0.057	0.306	0.015
S012003	0.878	0.017	-0.824	0.028	0.275	0.011
S012004	0.544	0.015	-0.552	0.054	0.226	0.016
S012006	0.867	0.018	0.092	0.021	0.245	0.008
S012007	0.843	0.033	-0.628	0.062	0.553	0.016
S012008	0.600	0.036	0.772	0.057	0.397	0.015
S012009	1.093	0.039	1.161	0.018	0.162	0.006
S012010	0.812	0.023	-1.757	0.064	0.206	0.024
S012011	0.584	0.034	1.075	0.046	0.288	0.014
S012012	1.041	0.044	-0.213	0.049	0.643	0.012
S012013	0.740	0.028	0.817	0.027	0.133	0.010
S012014	0.916	0.029	-0.624	0.042	0.302	0.017
S012015	0.712	0.025	-0.460	0.054	0.254	0.019
S012016	0.639	0.031	-0.487	0.089	0.437	0.023
S012018	0.658	0.041	0.685	0.058	0.376	0.016
S012019	0.558	0.023	0.594	0.041	0.118	0.014
S012020	0.918	0.028	-0.644	0.040	0.280	0.016
S012021	0.960	0.033	0.773	0.021	0.154	0.008
S012022	0.526	0.022	-0.198	0.068	0.170	0.021
S012023	0.926	0.030	-0.335	0.039	0.308	0.015
S012024	0.904	0.029	-0.462	0.041	0.305	0.016
S012025	0.679	0.055	1.547	0.053	0.331	0.012
S012026	0.535	0.025	-0.970	0.111	0.362	0.028
S012027	0.695	0.020	-1.183	0.053	0.113	0.020
S012028	0.682	0.023	0.093	0.037	0.145	0.014
S012029	0.776	0.044	0.676	0.046	0.427	0.013
S012030	0.520	0.028	0.323	0.070	0.230	0.020
S012031	1.011	0.038	0.006	0.037	0.415	0.013
S012032	1.089	0.027	-0.408	0.024	0.162	0.012
S012033	0.483	0.025	-0.142	0.096	0.275	0.025
S012035	0.772	0.026	-1.164	0.061	0.265	0.022
S012036	0.993	0.026	-0.415	0.028	0.194	0.013
S012037	0.577	0.022	-2.124	0.133	0.240	0.042
S012038	0.965	0.035	0.201	0.032	0.317	0.013
S012039	0.755	0.030	-0.384	0.058	0.372	0.018
S012040	1.601	0.053	0.482	0.016	0.320	0.008
S012041	0.430	0.022	0.176	0.084	0.216	0.021
S012043	0.726	0.026	-0.472	0.055	0.270	0.019
S012044	0.523	0.016	-1.592	0.079	0.110	0.025
S012045	1.335	0.047	-0.304	0.033	0.498	0.013
S012046	0.641	0.032	0.673	0.045	0.251	0.015
S012047	1.027	0.051	1.492	0.029	0.143	0.006
S012048	0.809	0.025	0.031	0.032	0.161	0.013
S012049	0.462	0.046	-1.169	0.256	0.291	0.059
S012050	0.832	0.071	0.571	0.065	0.211	0.024
S012053	0.458	0.045	0.330	0.132	0.172	0.035

Exhibit E.4 IRT Parameters for TIMSS Joint 1995 – 1999 Eighth-Grade Science (continued)

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.
S012055	0.627	0.044	-1.636	0.151	0.198	0.045
S012056	0.759	0.083	0.985	0.074	0.209	0.025
S012058	1.202	0.069	0.084	0.040	0.136	0.020
S012059	0.776	0.089	1.012	0.075	0.210	0.025
S012060	0.848	0.062	-0.082	0.075	0.218	0.030
S012061	0.731	0.037	0.696	0.039	0.000	0.000
S012062	1.112	0.100	0.600	0.054	0.294	0.022
S012063	0.708	0.044	-0.503	0.075	0.109	0.027
S012064	1.361	0.132	1.202	0.045	0.160	0.013
S012065	0.803	0.066	0.285	0.072	0.182	0.028
S012066	1.008	0.087	0.479	0.060	0.270	0.024
S012067	0.746	0.035	-1.131	0.055	0.000	0.000
S012068	0.467	0.030	1.087	0.074	0.000	0.000
S012069	1.141	0.087	0.184	0.057	0.294	0.025
S012070	0.834	0.067	0.195	0.074	0.246	0.028
S012071	0.821	0.057	-0.906	0.101	0.225	0.037
S012072	0.877	0.066	-1.162	0.118	0.297	0.042
S012073	0.884	0.061	-0.113	0.068	0.197	0.028
S012074	1.015	0.085	0.618	0.051	0.214	0.021
S012075	0.805	0.073	0.386	0.079	0.277	0.028
S012076	0.853	0.076	0.325	0.076	0.284	0.028
S012078	0.805	0.061	0.354	0.063	0.154	0.025
S012079	1.221	0.091	0.367	0.047	0.245	0.022
S012080	0.796	0.055	-0.494	0.087	0.200	0.033
S012081	0.893	0.040	0.559	0.030	0.000	0.000
S012082	0.448	0.046	-0.306	0.189	0.231	0.046
S012083	0.716	0.058	0.099	0.085	0.195	0.031
S012084	0.684	0.049	-0.581	0.101	0.186	0.034
S012085	0.694	0.075	0.816	0.084	0.223	0.028
S012086	0.734	0.034	-0.826	0.047	0.000	0.000
S012087	0.738	0.034	-0.153	0.034	0.000	0.000
S012088	0.824	0.065	0.308	0.067	0.192	0.026
S012089	0.515	0.028	-1.137	0.072	0.000	0.000
S012091	0.752	0.086	0.995	0.079	0.226	0.026
S012093	0.436	0.040	0.024	0.137	0.139	0.036
S012095	1.097	0.083	-0.064	0.069	0.308	0.030
S012096	0.804	0.042	-2.153	0.091	0.000	0.000
S012097	0.498	0.042	-0.927	0.168	0.202	0.045
S012098	0.679	0.051	0.062	0.081	0.133	0.029
S012099	0.623	0.032	-0.074	0.038	0.000	0.000
S012100	0.625	0.032	-0.276	0.041	0.000	0.000
S012101	0.962	0.105	0.985	0.062	0.249	0.022
S012102	1.088	0.091	1.090	0.045	0.107	0.014
S012103	0.479	0.043	-0.079	0.134	0.167	0.037
S012104	0.951	0.040	-0.119	0.028	0.000	0.000
S012105	0.782	0.084	0.847	0.075	0.224	0.026
S012106	0.898	0.039	-0.380	0.033	0.000	0.000

Exhibit E.4 IRT Parameters for TIMSS Joint 1995 – 1999 Eighth-Grade Science (continued 2)

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.	step parameter d_{ji}	S.E.
S012107	0.557	0.031	-0.306	0.047	0.000	0.000		
S012108	1.045	0.078	-0.784	0.093	0.367	0.037		
S012109	0.799	0.042	1.104	0.048	0.000	0.000		
S012110	0.887	0.041	-1.067	0.048	0.000	0.000		
S012111	0.753	0.062	0.100	0.085	0.193	0.032		
S012112	0.510	0.029	-0.321	0.049	0.000	0.000		
S012113	0.546	0.019	-0.072	0.026	0.000	0.000		
						d0		
						d1	-0.327	0.055
						d2	0.327	0.052
S012115	1.079	0.114	0.870	0.057	0.291	0.021		
S012116	0.591	0.031	0.074	0.039	0.000	0.000		
S012117	0.681	0.055	-0.103	0.099	0.202	0.034		
S012118	0.462	0.070	0.974	0.164	0.251	0.042		
S012119	0.964	0.120	1.313	0.071	0.219	0.019		
S012120	1.502	0.133	1.438	0.053	0.182	0.010		
S012121	0.684	0.034	-0.644	0.047	0.000	0.000		
S012122	0.328	0.011	0.720	0.043	0.000	0.000		
						d0		
						d1	-2.391	0.117
						d2	2.391	0.123
S012123	0.440	0.040	-1.168	0.212	0.224	0.050		
S012124	1.058	0.122	1.066	0.061	0.290	0.019		
S012125	0.759	0.030	0.744	0.026	0.000	0.000		
						d0		
						d1	0.534	0.033
						d2	-0.534	0.045
S012126	1.020	0.042	0.037	0.025	0.000	0.000		
S012127	0.967	0.041	0.037	0.026	0.000	0.000		
S012128A			0.462	0.021	-2.184	0.096	0.000	0.000
S012128B			0.384	0.020	0.316	0.043	0.000	0.000
S012129	0.474	0.014	0.298	0.022	0.000	0.000		
						d0		
						d1	-0.377	0.045
						d2	0.377	0.046
S012130	0.572	0.018	0.988	0.027	0.000	0.000		
						d0		
						d1	0.075	0.033
						d2	-0.075	0.044
S012131A	0.489	0.023	-1.176	0.064	0.000	0.000		
S012131B	0.661	0.026	0.493	0.028	0.000	0.000		
S012132	0.984	0.044	1.828	0.056	0.000	0.000		
S012133	0.279	0.014	1.715	0.079	0.000	0.000		
						d0		
						d1	0.375	0.064
						d2	-0.375	0.091
S012134A	0.471	0.029	-0.794	0.071	0.000	0.000		

Exhibit E.4 IRT Parameters for TIMSS Joint 1995 – 1999 Eighth-Grade Science (continued 3)

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.	step parameter d_{ji}	S.E.
S012134D	0.657	0.028	0.515	0.027	0.000	0.000		
						d0		
						d1	0.211	0.044
						d2	-0.211	0.049
S022002	0.898	0.063	0.214	0.059	0.172	0.024		
S022007	0.557	0.047	-0.452	0.138	0.192	0.041		
S022009	0.811	0.067	-1.280	0.147	0.370	0.046		
S022012	1.239	0.095	0.779	0.038	0.163	0.015		
S022014	0.420	0.047	-0.260	0.212	0.220	0.050		
S022017	0.946	0.042	0.560	0.029	0.000	0.000		
S022019	0.874	0.068	-0.201	0.088	0.283	0.034		
S022022	0.677	0.032	0.011	0.035	0.000	0.000		
S022030	0.785	0.054	-0.600	0.092	0.207	0.035		
S022035	0.313	0.024	-0.087	0.071	0.000	0.000		
S022040	0.619	0.043	-0.461	0.096	0.148	0.031		
S022043	0.745	0.040	1.073	0.050	0.000	0.000		
S022048	0.891	0.044	0.862	0.036	0.000	0.000		
S022049	0.848	0.038	0.385	0.029	0.000	0.000		
S022054	1.113	0.084	0.268	0.053	0.245	0.024		
S022058	0.757	0.081	0.152	0.117	0.393	0.035		
S022064	0.381	0.065	1.791	0.205	0.194	0.035		
S022069	0.682	0.024	0.248	0.025	0.000	0.000		
S022073	0.943	0.065	-0.759	0.088	0.282	0.035		
S022074	1.077	0.081	0.347	0.052	0.218	0.023		
S022078	1.117	0.045	-0.051	0.024	0.000	0.000		
S022081	0.774	0.034	-0.806	0.042	0.000	0.000		
S022082	0.963	0.091	1.141	0.054	0.135	0.015		
S022090	0.614	0.023	-0.200	0.028	0.000	0.000		
						d0		
						d1	-0.168	0.054
						d2	0.168	0.048
S022094	0.728	0.083	1.238	0.078	0.164	0.021		
S022099	0.657	0.083	0.771	0.106	0.325	0.030		
S022106	0.989	0.109	1.438	0.067	0.133	0.014		
S022115	0.957	0.069	0.041	0.065	0.233	0.027		
S022117	0.617	0.056	0.435	0.090	0.171	0.029		
S022123	0.659	0.069	0.468	0.103	0.260	0.033		
S022126	0.527	0.067	0.779	0.128	0.236	0.036		
S022131	0.716	0.053	-0.736	0.112	0.225	0.039		
S022132	0.888	0.096	1.059	0.065	0.210	0.020		
S022137	0.939	0.084	0.821	0.054	0.193	0.020		
S022140	0.959	0.040	-0.171	0.027	0.000	0.000		
S022141	1.007	0.045	0.812	0.031	0.000	0.000		
S022145	0.608	0.046	-0.099	0.096	0.153	0.032		
S022150	0.883	0.078	0.529	0.065	0.222	0.025		
S022152	1.035	0.042	-0.169	0.027	0.000	0.000		
S022154	0.526	0.028	-0.615	0.056	0.000	0.000		

Exhibit E.4 IRT Parameters for TIMSS Joint 1995 – 1999 Eighth-Grade Science (continued 4)

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.	step parameter d_{ji}	S.E.
S022157	0.783	0.071	0.506	0.072	0.203	0.027		
S022158	0.866	0.038	0.191	0.028	0.000	0.000		
S022160	0.505	0.029	0.371	0.047	0.000	0.000		
S022161	0.623	0.032	0.307	0.037	0.000	0.000		
S022165D	0.565	0.021	-0.254	0.028	0.000	0.000		
						d0		
						d1	0.039	0.053
						d2	-0.039	0.047
S022172A	0.682	0.024	-1.268	0.044	0.000	0.000		
S022172B	0.507	0.021	-1.530	0.067	0.000	0.000		
S022174	0.661	0.031	0.137	0.035	0.000	0.000		
S022178	0.969	0.073	0.418	0.054	0.187	0.023		
S022181	0.797	0.093	0.909	0.078	0.267	0.026		
S022183	1.420	0.122	0.948	0.038	0.208	0.014		
S022187	0.650	0.066	0.891	0.081	0.165	0.026		
S022188	0.654	0.101	1.009	0.123	0.386	0.031		
S022191	0.596	0.019	-0.873	0.036	0.000	0.000		
						d0		
						d1	-0.506	0.066
						d2	0.506	0.052
S022194	0.866	0.072	0.700	0.056	0.163	0.020		
S022198	0.972	0.123	1.289	0.071	0.252	0.019		
S022202	0.868	0.108	1.111	0.074	0.288	0.022		
S022206	0.717	0.108	1.315	0.102	0.295	0.026		
S022208	1.131	0.111	0.908	0.051	0.250	0.019		
S022213	0.723	0.040	1.238	0.058	0.000	0.000		
S022217A	0.825	0.038	0.405	0.030	0.000	0.000		
S022217D	0.614	0.024	0.632	0.028	0.000	0.000		
						d0		
						d1	-0.076	0.046
						d2	0.076	0.052
S022258	0.651	0.032	0.225	0.036	0.000	0.000		
S022275	1.452	0.127	1.243	0.043	0.171	0.012		
S022278	1.056	0.063	-0.253	0.054	0.171	0.025		
S022279	0.649	0.032	-0.145	0.037	0.000	0.000		
S022280	1.350	0.107	0.365	0.047	0.320	0.021		
S022281	0.478	0.031	1.042	0.072	0.000	0.000		
S022282	1.089	0.048	1.731	0.050	0.000	0.000		
S022283	0.732	0.034	-1.003	0.053	0.000	0.000		
S022284	0.876	0.040	0.554	0.031	0.000	0.000		
S022289	0.788	0.020	0.534	0.017	0.000	0.000		
						d0		
						d1	0.701	0.023
						d2	-0.701	0.029
S022290	1.228	0.085	0.151	0.048	0.231	0.023		
S022292	0.779	0.036	0.132	0.031	0.000	0.000		
S022293	0.930	0.087	0.827	0.057	0.207	0.021		

Exhibit E.4 IRT Parameters for TIMSS Joint 1995 – 1999 Eighth-Grade Science (continued 5)

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.	step parameter d_{ji}	S.E.
S022294	0.875	0.068	0.020	0.077	0.257	0.030		
S022295	0.862	0.058	-0.325	0.074	0.193	0.030		
SX12005	0.693	0.018	0.145	0.031	0.307	0.010		
SX12017	1.316	0.042	0.521	0.017	0.229	0.008		
SX12034	0.822	0.023	-0.621	0.036	0.156	0.015		
SX12042	0.876	0.041	0.661	0.034	0.333	0.012		
SX12051	1.097	0.087	0.759	0.043	0.169	0.018		
SX12052	1.040	0.073	0.058	0.058	0.223	0.026		
SX12054	0.740	0.059	0.381	0.069	0.153	0.026		
SX12057	0.793	0.036	0.312	0.030	0.000	0.000		
SX12077	0.626	0.031	-1.169	0.062	0.000	0.000		
SX12090	1.131	0.133	1.098	0.060	0.318	0.019		
SX12092	0.790	0.052	-0.292	0.074	0.141	0.029		
SX12094	1.076	0.171	1.624	0.093	0.290	0.016		
SX12114	0.413	0.048	0.316	0.176	0.203	0.042		
SX12135D	0.406	0.023	-0.744	0.057	0.000	0.000		
						d0		
						d1	0.712	0.090
						d2	-0.712	0.065
SX22041	0.753	0.052	-1.191	0.114	0.202	0.037		
SX22042	0.912	0.061	0.062	0.059	0.168	0.025		
SX22086	1.040	0.042	0.035	0.025	0.000	0.000		
SX22088A	0.721	0.024	-0.847	0.034	0.000	0.000		
SX22088B	0.485	0.020	-0.195	0.035	0.000	0.000		
SX22118	1.413	0.112	0.437	0.042	0.287	0.019		
SX22222	1.325	0.088	0.378	0.037	0.186	0.019		
SX22225	0.968	0.106	1.436	0.069	0.123	0.014		
SX22235	0.713	0.079	0.421	0.107	0.327	0.033		
SX22238	0.777	0.083	0.696	0.081	0.269	0.028		
SX22240	1.740	0.139	1.370	0.048	0.258	0.010		
SX22244	0.999	0.046	0.892	0.034	0.000	0.000		
SX22245	0.455	0.072	1.482	0.157	0.220	0.034		
SX22249D	0.651	0.023	-0.232	0.028	0.000	0.000		
SX22254	1.616	0.138	0.906	0.034	0.225	0.014		
SX22264	0.799	0.118	1.453	0.100	0.267	0.021		
SX22268	0.472	0.021	0.261	0.035	0.000	0.000		
SX22276	0.781	0.063	0.215	0.077	0.216	0.029		
SX22277D	0.542	0.020	-0.228	0.029	0.000	0.000		
						d0		
						d1	-0.102	0.055
						d2	0.102	0.050
SX22286	0.691	0.029	1.205	0.042	0.000	0.000		
SX22288	0.627	0.016	1.047	0.025	0.000	0.000		
						d0		
						d1	-0.479	0.036
						d2	0.479	0.045

Exhibit E.5 **IRT Parameters for TIMSS 1999 Eighth-Grade Mathematics - Algebra**

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.	step parameter d_{ji}	S.E.
1M012002	0.605	0.012	-1.171	0.053	0.133	0.020		
1M012012	1.392	0.031	-0.424	0.019	0.225	0.010		
1M012017	0.803	0.024	0.007	0.032	0.127	0.013		
1M012020	1.229	0.041	-0.139	0.030	0.407	0.012		
1M012022	0.516	0.029	0.585	0.073	0.223	0.021		
1M012029	0.970	0.026	-0.126	0.027	0.158	0.012		
1M012035	1.505	0.039	0.291	0.014	0.178	0.007		
1M012040	1.122	0.032	-0.432	0.030	0.253	0.014		
1M012042	1.398	0.037	0.140	0.017	0.203	0.008		
1M012046	1.686	0.044	0.254	0.013	0.205	0.007		
1M012048	0.956	0.030	-0.656	0.044	0.273	0.019		
1M022002	1.121	0.101	1.151	0.049	0.135	0.013		
1M022008	0.624	0.032	0.644	0.045	0.000	0.000		
1M022041	0.549	0.045	-0.594	0.158	0.192	0.048		
1M022042	3.838	0.223	0.205	0.015	0.105	0.010		
1M022050	1.413	0.109	0.806	0.035	0.177	0.014		
1M022078	0.773	0.072	0.406	0.087	0.258	0.030		
1M022083	1.070	0.080	0.797	0.042	0.132	0.016		
1M022089	0.787	0.033	-0.020	0.031	0.000	0.000		
1M022118	1.307	0.048	-0.339	0.023	0.000	0.000		
1M022165	1.723	0.122	0.241	0.037	0.311	0.018		
1M022166	1.785	0.106	0.263	0.028	0.174	0.015		
1M022176	0.841	0.070	-0.450	0.114	0.335	0.040		
1M022185	0.906	0.069	0.103	0.071	0.228	0.028		
1M022196	1.377	0.071	-0.252	0.037	0.118	0.019		
1M022210	1.540	0.086	0.420	0.027	0.098	0.012		
1M022228	0.585	0.011	0.284	0.017	0.000	0.000		
						d0		
						d1	-1.887	0.059
						d2	1.887	0.060
1M022251	0.711	0.088	1.447	0.089	0.180	0.021		
1M022253	0.951	0.038	-0.381	0.030	0.000	0.000		
1M022261A	1.790	0.065	0.061	0.017	0.000	0.000		
1M022261B	3.304	0.137	0.489	0.012	0.000	0.000		
1M022261C	1.719	0.053	0.687	0.013	0.000	0.000		
						d0		
						d1	-0.701	0.044
						d2	0.701	0.045
1M022262A	2.149	0.061	-0.566	0.013	0.000	0.000		
1M022262B	2.028	0.055	-0.268	0.012	0.000	0.000		
1M022262C	0.861	0.016	0.372	0.013	0.000	0.000		
						d0		
						d1	-1.073	0.039
						d2	1.073	0.040

Exhibit E.6 IRT Parameters for TIMSS 1999 Eighth-Grade Mathematics – Data Representation, Analysis, and Probability

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.	step parameter d_{ji}	S.E.
2M012006	0.546	0.010	-1.361	0.048	0.121	0.018		
2M012007	0.781	0.021	-0.388	0.041	0.150	0.017		
2M012014	0.925	0.023	-1.070	0.041	0.099	0.022		
2M012025	0.706	0.020	-0.884	0.054	0.102	0.024		
2M012032	0.424	0.014	-0.538	0.075	0.085	0.022		
2M012037	0.575	0.024	0.143	0.063	0.148	0.021		
2M012043	0.852	0.026	-0.488	0.044	0.167	0.020		
2M012047	0.902	0.026	-0.326	0.037	0.173	0.017		
2M022030	1.110	0.043	-0.911	0.031	0.000	0.000		
2M022101	0.632	0.037	-1.110	0.102	0.103	0.037		
2M022135	0.772	0.051	0.406	0.058	0.085	0.021		
2M022146	1.442	0.077	-0.272	0.038	0.123	0.020		
2M022181	1.064	0.059	-1.187	0.072	0.125	0.038		
2M022189	0.930	0.052	-0.880	0.073	0.115	0.035		
2M022208	0.690	0.046	-0.144	0.086	0.112	0.032		
2M022249	0.731	0.052	0.112	0.080	0.124	0.030		
2M022252	0.770	0.045	-0.631	0.082	0.108	0.034		
2M022256	0.739	0.017	0.332	0.016	0.000	0.000		
						d0		
						d1	0.114	0.027
						d2	-0.114	0.030
2M022257	0.804	0.060	0.248	0.072	0.153	0.027		
2M022258	0.635	0.041	-0.360	0.095	0.104	0.034		
2M022260	1.050	0.061	-1.459	0.079	0.120	0.040		

Exhibit E.7 IRT Parameters for TIMSS 1999 Eighth-Grade Mathematics – Fractions and Number Sense

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.
3M012001	1.569	0.022	-0.108	0.009	0.161	0.005
3M012004	1.208	0.024	0.439	0.013	0.290	0.005
3M012008	0.447	0.014	-0.869	0.076	0.206	0.021
3M012009	0.821	0.029	0.182	0.038	0.320	0.013
3M012010	1.496	0.027	0.113	0.011	0.078	0.005
3M012016	1.552	0.067	0.838	0.020	0.449	0.006
3M012018	0.563	0.018	-0.939	0.080	0.147	0.028
3M012021	1.393	0.030	-0.468	0.017	0.108	0.010
3M012024	0.783	0.029	-0.406	0.059	0.300	0.021
3M012027	1.209	0.034	0.072	0.022	0.237	0.010
3M012028	0.960	0.025	-0.570	0.034	0.158	0.016
3M012031	1.283	0.039	0.871	0.015	0.146	0.006
3M012033	0.994	0.030	-0.222	0.033	0.246	0.014
3M012036	0.639	0.025	0.217	0.051	0.177	0.018
3M012041	1.057	0.027	-0.192	0.025	0.144	0.012
3M012044	1.217	0.030	-0.645	0.026	0.174	0.014
3M012045	0.835	0.022	-1.525	0.053	0.126	0.027
3M022004	1.605	0.126	0.463	0.040	0.325	0.016
3M022010	0.943	0.057	-0.418	0.068	0.168	0.029
3M022012	0.662	0.030	-0.737	0.043	0.000	0.000
3M022026	0.671	0.029	0.027	0.035	0.000	0.000
3M022031	1.310	0.099	0.761	0.039	0.206	0.015
3M022038	0.669	0.049	-0.221	0.102	0.176	0.035
3M022043	0.621	0.041	-0.984	0.124	0.166	0.043
3M022046	0.689	0.030	-0.786	0.044	0.000	0.000
3M022057	0.395	0.034	-0.862	0.212	0.183	0.050
3M022066	1.078	0.059	-0.168	0.047	0.114	0.021
3M022070	0.668	0.043	-1.139	0.123	0.182	0.044
3M022073	0.647	0.051	-0.392	0.129	0.216	0.043
3M022093	1.228	0.096	0.431	0.051	0.283	0.020
3M022104	1.069	0.060	-0.491	0.060	0.163	0.028
3M022106	0.989	0.038	0.469	0.027	0.000	0.000
3M022110	0.348	0.023	0.613	0.071	0.000	0.000
3M022113	0.702	0.043	-1.202	0.113	0.173	0.043
3M022121	1.639	0.118	0.084	0.044	0.352	0.020
3M022126	1.137	0.123	1.176	0.057	0.308	0.016
3M022127	1.269	0.116	1.127	0.046	0.200	0.014
3M022128	1.008	0.090	1.053	0.052	0.183	0.017
3M022132	1.516	0.054	-0.101	0.020	0.000	0.000
3M022139	1.646	0.115	0.793	0.030	0.177	0.012
3M022144	0.692	0.063	0.406	0.097	0.222	0.032
3M022156	1.231	0.045	0.007	0.023	0.000	0.000
3M022169	0.846	0.057	-0.529	0.089	0.202	0.036
3M022172	1.075	0.073	-0.012	0.059	0.217	0.025
3M022173	1.292	0.090	0.385	0.043	0.216	0.018
3M022178	1.166	0.043	0.237	0.024	0.000	0.000
3M022191	0.813	0.065	-0.314	0.106	0.279	0.038
3M022194	0.918	0.071	0.206	0.069	0.236	0.026

Exhibit E.7 IRT Parameters for TIMSS 1999 Eighth-Grade Mathematics – Fractions and Number Sense (continued)

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.	step parameter d_{ji}	S.E.
3M022198	1.203	0.082	0.464	0.042	0.182	0.017		
3M022199	1.066	0.082	0.445	0.053	0.217	0.021		
3M022204	0.642	0.042	-1.174	0.129	0.186	0.045		
3M022206	1.851	0.165	0.703	0.038	0.380	0.014		
3M022207	0.930	0.050	-0.055	0.056	0.222	0.023		
3M022219	1.059	0.041	0.571	0.028	0.000	0.000		
3M022222	1.057	0.039	0.026	0.025	0.000	0.000		
3M022231A	1.366	0.038	0.638	0.017	0.000	0.000		
3M022231B	1.729	0.058	1.212	0.020	0.000	0.000		
3M022232	0.525	0.011	1.319	0.027	0.000	0.000		
						d0		
						d1	-2.195	0.075
						d2	2.195	0.081
3M022237	0.852	0.024	-0.040	0.021	0.000	0.000		
3M022241	1.128	0.046	0.264	0.028	0.097	0.012		
3M022245	0.679	0.044	-1.036	0.116	0.179	0.043		

Exhibit E.8 IRT Parameters for TIMSS 1999 Eighth-Grade Mathematics - Geometry

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.
4M012005	0.661	0.016	-0.379	0.043	0.181	0.016
4M012011	1.011	0.024	-0.351	0.026	0.162	0.013
4M012015	0.980	0.026	-0.505	0.032	0.133	0.016
4M012019	0.840	0.022	-0.515	0.037	0.094	0.017
4M012026	1.124	0.029	0.087	0.020	0.130	0.009
4M012039	0.901	0.025	-0.078	0.029	0.121	0.013
4M022016	0.524	0.057	0.963	0.115	0.147	0.033
4M022033	0.543	0.038	-0.482	0.128	0.133	0.041
4M022037	0.661	0.049	-0.245	0.110	0.177	0.039
4M022049	0.467	0.038	-0.648	0.181	0.183	0.050
4M022062	1.220	0.073	0.380	0.037	0.107	0.016
4M022085	1.216	0.087	0.686	0.042	0.170	0.016
4M022105	0.677	0.055	0.604	0.078	0.127	0.026
4M022108	0.765	0.048	-0.373	0.083	0.129	0.033
4M022116	0.929	0.065	0.469	0.053	0.139	0.020
4M022142	1.334	0.083	0.177	0.040	0.167	0.018
4M022154	0.912	0.056	-0.172	0.064	0.127	0.028
4M022160	1.135	0.093	0.929	0.047	0.195	0.015
4M022202	0.760	0.033	0.641	0.038	0.000	0.000
4M022213	1.120	0.073	0.267	0.047	0.165	0.020
4M022246	0.721	0.049	-0.048	0.083	0.131	0.031

Exhibit E.9 IRT Parameters for TIMSS-R 1998/1999 Mathematics - Measurement, Population 2

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.	step parameter d_{ji}	S.E.
5M012003	0.777	0.012	-0.428	0.022	0.056	0.009		
5M012013	1.003	0.028	-0.111	0.027	0.173	0.012		
5M012023	0.618	0.018	-1.892	0.088	0.155	0.037		
5M012030	1.268	0.035	0.472	0.016	0.161	0.007		
5M012034	0.647	0.019	-0.102	0.040	0.074	0.015		
5M012038	0.760	0.023	-1.012	0.061	0.156	0.027		
5M022005	0.932	0.105	1.150	0.067	0.280	0.019		
5M022021	1.673	0.109	0.474	0.032	0.189	0.015		
5M022022	1.198	0.082	0.564	0.040	0.174	0.016		
5M022055	1.349	0.049	0.295	0.022	0.000	0.000		
5M022079	1.104	0.062	-0.635	0.059	0.142	0.029		
5M022097	1.027	0.053	-0.512	0.054	0.104	0.024		
5M022124	0.610	0.048	-0.236	0.124	0.171	0.041		
5M022148	1.287	0.047	-0.408	0.024	0.000	0.000		
5M022159	1.363	0.112	1.275	0.040	0.109	0.010		
5M022168	1.062	0.090	0.752	0.051	0.228	0.018		
5M022188	0.668	0.071	0.780	0.090	0.207	0.029		
5M022227A	1.227	0.046	-0.457	0.026	0.000	0.000		
5M022227B	1.871	0.070	0.401	0.018	0.000	0.000		
5M022227C	1.642	0.065	0.704	0.021	0.000	0.000		
5M022234A	0.892	0.017	0.420	0.013	0.000	0.000		
						d0		
						d1	-0.537	0.030
						d2	0.537	0.031
5M022234B	1.057	0.021	0.760	0.013	0.000	0.000	0.67	
						d0		
						d1	-1.011	0.040
						d2	1.011	0.041
5M022243	0.969	0.028	0.132	0.020	0.000	0.000	0.78	
5M022244	1.153	0.031	-0.019	0.017	0.000	0.000	0.83	

Exhibit E.10 IRT Parameters for TIMSS 1999 Eighth-Grade Science - Chemistry

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.	step parameter d_{ji}	S.E.
1S012003	0.820	0.016	-1.151	0.040	0.153	0.020		
1S012016	0.456	0.015	-1.466	0.091	0.200	0.027		
1S012036	0.774	0.022	-0.718	0.048	0.118	0.022		
1S012040	0.862	0.032	0.386	0.035	0.269	0.013		
1S012048	0.700	0.023	-0.123	0.047	0.132	0.018		
1S022174	0.672	0.029	0.103	0.035	0.000	0.000		
1S022178	0.775	0.059	0.303	0.075	0.161	0.028		
1S022181	0.497	0.059	0.828	0.144	0.199	0.039		
1S022183	0.609	0.074	1.324	0.098	0.182	0.026		
1S022187	0.562	0.054	0.844	0.100	0.141	0.030		
1S022188	0.433	0.052	0.371	0.209	0.227	0.051		
1S022191	0.679	0.021	-0.934	0.029	0.000	0.000		
						d0		
						d1	-0.393	0.057
						d2	0.393	0.047
1S022194	0.642	0.061	0.786	0.088	0.154	0.028		
1S022198	0.785	0.104	1.500	0.091	0.259	0.020		
1S022202	0.509	0.063	0.993	0.136	0.201	0.037		
1S022206	0.394	0.054	1.137	0.189	0.176	0.044		
1S022208	0.838	0.075	0.787	0.066	0.193	0.022		
1S022213	0.644	0.033	1.328	0.063	0.000	0.000		
1S022217A	0.918	0.036	0.296	0.029	0.000	0.000		
1S022217D	0.557	0.020	0.587	0.031	0.000	0.000	10.50	
						d0		
						d1	-0.041	0.051
						d2	0.041	0.059

Exhibit E.11 IRT Parameters for TIMSS 1999 Eighth-Grade Science – Earth Science

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.	step parameter d_{ji}	S.E.
2S012006	0.761	0.017	-0.010	0.030	0.231	0.011		
2S012007	0.625	0.016	-1.800	0.072	0.162	0.031		
2S012011	0.393	0.017	0.794	0.068	0.185	0.016		
2S012013	0.715	0.027	0.796	0.031	0.128	0.011		
2S012021	0.703	0.026	0.772	0.031	0.118	0.011		
2S012027	0.655	0.019	-1.292	0.070	0.112	0.031		
2S012030	0.481	0.021	0.051	0.083	0.157	0.025		
2S012035	0.712	0.020	-1.521	0.069	0.124	0.033		
2S012041	0.342	0.013	-0.182	0.081	0.139	0.019		
2S012045	0.691	0.022	-1.323	0.077	0.189	0.034		
2S012046	0.541	0.027	0.488	0.071	0.196	0.022		
2S022073	0.833	0.048	-1.108	0.090	0.147	0.041		
2S022074	0.757	0.058	0.200	0.082	0.160	0.031		
2S022078	1.046	0.040	-0.149	0.025	0.000	0.000		
2S022081	0.768	0.033	-0.863	0.040	0.000	0.000		
2S022082	0.719	0.072	1.327	0.074	0.124	0.019		
2S022090	0.442	0.017	-0.419	0.036	0.000	0.000		
						d0		
						d1	-0.350	0.073
						d2	0.350	0.067
2S022275	0.817	0.084	1.463	0.070	0.135	0.016		
2S022283	0.648	0.032	-1.148	0.056	0.000	0.000		
2S022284	0.666	0.030	0.613	0.041	0.000	0.000		
2S022290	0.961	0.066	-0.022	0.066	0.186	0.029		
2S022294	0.578	0.043	-0.362	0.127	0.155	0.042		

Exhibit E.12 IRT Parameters for TIMSS 1999 Eighth-Grade Science – Life Science

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.	step parameter d_{ji}	S.E.
3S012001	0.519	0.012	-1.004	0.058	0.113	0.019		
3S012010	0.789	0.021	-1.900	0.064	0.149	0.028		
3S012014	0.983	0.030	-0.620	0.039	0.299	0.017		
3S012023	0.908	0.028	-0.457	0.038	0.262	0.016		
3S012024	0.781	0.027	-0.573	0.053	0.291	0.019		
3S012026	0.471	0.018	-1.339	0.108	0.279	0.028		
3S012028	0.652	0.023	0.101	0.043	0.153	0.015		
3S012031	0.914	0.036	0.002	0.043	0.420	0.014		
3S012033	0.379	0.016	-0.430	0.095	0.220	0.022		
3S012038	0.860	0.031	0.137	0.037	0.298	0.014		
3S012039	0.740	0.027	-0.589	0.060	0.301	0.021		
3S012043	0.640	0.023	-0.748	0.068	0.191	0.024		
3S012044	0.542	0.017	-1.540	0.087	0.128	0.029		
3S022094	0.821	0.093	1.350	0.074	0.194	0.019		
3S022099	0.325	0.041	0.737	0.214	0.238	0.040		
3S022106	0.954	0.099	1.462	0.065	0.129	0.014		
3S022115	0.727	0.055	-0.090	0.090	0.196	0.032		
3S022117	0.614	0.057	0.473	0.096	0.180	0.031		
3S022123	0.581	0.063	0.454	0.125	0.249	0.037		
3S022126	0.404	0.055	0.832	0.184	0.214	0.043		
3S022131	0.617	0.044	-0.936	0.128	0.181	0.041		
3S022132	0.948	0.094	1.028	0.059	0.208	0.019		
3S022137	0.952	0.082	0.836	0.054	0.196	0.020		
3S022140	0.883	0.037	-0.199	0.029	0.000	0.000		
3S022141	0.811	0.038	0.913	0.040	0.000	0.000		
3S022145	0.557	0.042	-0.164	0.106	0.140	0.033		
3S022150	0.817	0.070	0.500	0.070	0.209	0.026		
3S022152	0.957	0.039	-0.248	0.028	0.000	0.000		
3S022154	0.530	0.028	-0.666	0.054	0.000	0.000		
3S022157	0.653	0.054	0.395	0.082	0.153	0.028		
3S022158	0.864	0.036	0.185	0.028	0.000	0.000		
3S022160	0.523	0.029	0.363	0.046	0.000	0.000		
3S022161	0.568	0.029	0.319	0.041	0.000	0.000		
3S022165D	0.559	0.020	-0.271	0.028	0.000	0.000	6.77	
						d0		
						d1	0.054	0.054
						d2	-0.054	0.048
3S022172A	0.794	0.025	-1.146	0.036	0.000	0.000	0.93	
3S022172B	0.633	0.023	-1.284	0.048	0.000	0.000	2.48	
3S022258	0.580	0.029	0.232	0.040	0.000	0.000	1.07	
3S022289	0.706	0.018	0.564	0.019	0.000	0.000	1.91	
						d0		
						d1	0.764	0.026
						d2	-0.764	0.032
3S022293	0.854	0.075	0.800	0.061	0.188	0.022		
3S022295	0.836	0.058	-0.336	0.080	0.198	0.032		

Exhibit E.13 IRT Parameters for TIMSS 1999 Eighth-Grade Science - Physics

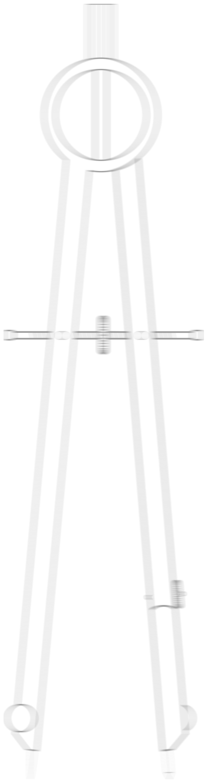
Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.
4S012002	0.383	0.009	-0.905	0.057	0.143	0.015
4S012004	0.468	0.011	-0.836	0.056	0.161	0.016
4S012008	0.443	0.024	0.652	0.080	0.351	0.018
4S012009	1.119	0.038	1.210	0.018	0.173	0.005
4S012012	0.756	0.031	-0.768	0.079	0.529	0.020
4S012015	0.686	0.024	-0.601	0.059	0.210	0.021
4S012018	0.428	0.022	0.276	0.087	0.260	0.021
4S012019	0.554	0.024	0.599	0.045	0.118	0.015
4S012020	0.914	0.028	-0.677	0.043	0.275	0.018
4S012022	0.516	0.024	-0.109	0.082	0.203	0.024
4S012025	0.408	0.042	1.814	0.092	0.284	0.021
4S012029	0.660	0.037	0.553	0.058	0.392	0.016
4S012032	1.007	0.025	-0.491	0.027	0.139	0.013
4S012037	0.605	0.020	-2.173	0.105	0.182	0.040
4S012047	0.992	0.051	1.659	0.032	0.160	0.006
4S022002	0.888	0.065	0.208	0.062	0.182	0.025
4S022007	0.513	0.038	-0.670	0.135	0.129	0.040
4S022009	0.808	0.056	-1.480	0.133	0.247	0.050
4S022012	1.351	0.100	0.828	0.036	0.173	0.014
4S022014	0.365	0.035	-0.544	0.216	0.150	0.049
4S022017	0.895	0.039	0.596	0.031	0.000	0.000
4S022019	0.745	0.063	-0.292	0.115	0.276	0.039
4S022022	0.739	0.033	-0.013	0.033	0.000	0.000
4S022030	0.820	0.054	-0.645	0.088	0.188	0.035
4S022035	0.320	0.024	-0.109	0.069	0.000	0.000
4S022040	0.660	0.047	-0.408	0.102	0.172	0.035
4S022043	0.688	0.036	1.144	0.054	0.000	0.000
4S022048	0.903	0.042	0.829	0.036	0.000	0.000
4S022049	0.764	0.035	0.381	0.032	0.000	0.000
4S022054	0.897	0.071	0.238	0.069	0.233	0.027
4S022058	0.514	0.054	-0.319	0.196	0.279	0.051
4S022064	0.404	0.069	1.804	0.191	0.205	0.037
4S022069	0.733	0.024	0.231	0.023	0.000	0.000
4S022278	0.923	0.059	-0.272	0.066	0.162	0.028
4S022279	0.662	0.030	-0.167	0.036	0.000	0.000
4S022280	1.088	0.079	0.186	0.056	0.246	0.024
4S022281	0.432	0.029	1.130	0.081	0.000	0.000
4S022282	0.917	0.040	1.992	0.061	0.000	0.000
4S022292	0.805	0.035	0.120	0.030	0.000	0.000

Exhibit E.14 IRT Parameters for TIMMS-R 1998/1999 Science - Environmental and resource issues, Population 2

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.	step parameter d_{ji}	S.E.
5S012005	0.488	0.017	-0.160	0.070	0.202	0.020		
5S012017	0.710	0.039	0.509	0.055	0.195	0.019		
5S012034	0.872	0.037	-0.544	0.060	0.205	0.026		
5S012042	0.702	0.042	0.379	0.068	0.258	0.022		
5S022086	0.680	0.030	-0.038	0.035	0.000	0.000		
5S022088A	1.097	0.030	-0.727	0.021	0.000	0.000		
5S022088B	0.646	0.021	-0.188	0.027	0.000	0.000		
5S022118	0.942	0.072	0.287	0.066	0.221	0.026		
5S022240	0.941	0.092	1.302	0.061	0.176	0.016		
5S022244	1.383	0.054	0.782	0.025	0.000	0.000		
5S022249D	0.708	0.022	-0.276	0.025	0.000	0.000		
5S022254	0.645	0.087	1.534	0.104	0.232	0.025		
5S022277D	0.789	0.025	-0.246	0.021	0.000	0.000		
						d0		
						d1	0.102	0.039
						d2	-0.102	0.036

Exhibit E.15 IRT Parameters for TIMMS-R 1998/1999 Science - Scientific inquiry and the nature of science, Population 2

Item	slope parameter a_j	S.E.	location parameter b_j	S.E.	guessing parameter c_j	S.E.	step parameter d_{ji}	S.E.
6S022041	0.971	0.063	-1.047	0.096	0.228	0.046		
6S022042	0.753	0.063	0.139	0.097	0.228	0.034		
6S022222	0.517	0.054	0.519	0.140	0.184	0.040		
6S022225	0.742	0.094	1.873	0.101	0.140	0.016		
6S022235	0.619	0.051	-0.075	0.124	0.183	0.041		
6S022238	0.586	0.055	0.411	0.118	0.179	0.037		
6S022245	0.608	0.077	1.252	0.106	0.229	0.029		
6S022264	0.660	0.083	1.367	0.097	0.221	0.025		
6S022268	0.317	0.017	0.272	0.052	0.000	0.000		
6S022276	0.669	0.054	0.052	0.106	0.179	0.037		
6S022286	0.519	0.022	1.423	0.058	0.000	0.000		
6S022288	1.123	0.026	0.877	0.015	0.000	0.000		
						d0		
						d1	0.002	0.022
						d2	-0.002	0.028



Typography	Body text was set in ITC New Baskerville, designed in by George W. Jones. Headings were set in Eurostile, by Aldo Novarese and A. Butti. Captions, footnotes and exhibit titles were set in Frutiger, designed by Adrian Frutiger.
Book Design & Illustrations	José R. Nieto
Layout & Production	Betty Hugh Nathan Pyritz
Production Assistants	Mario Pita Christina López
Cover Design	Christine Conley