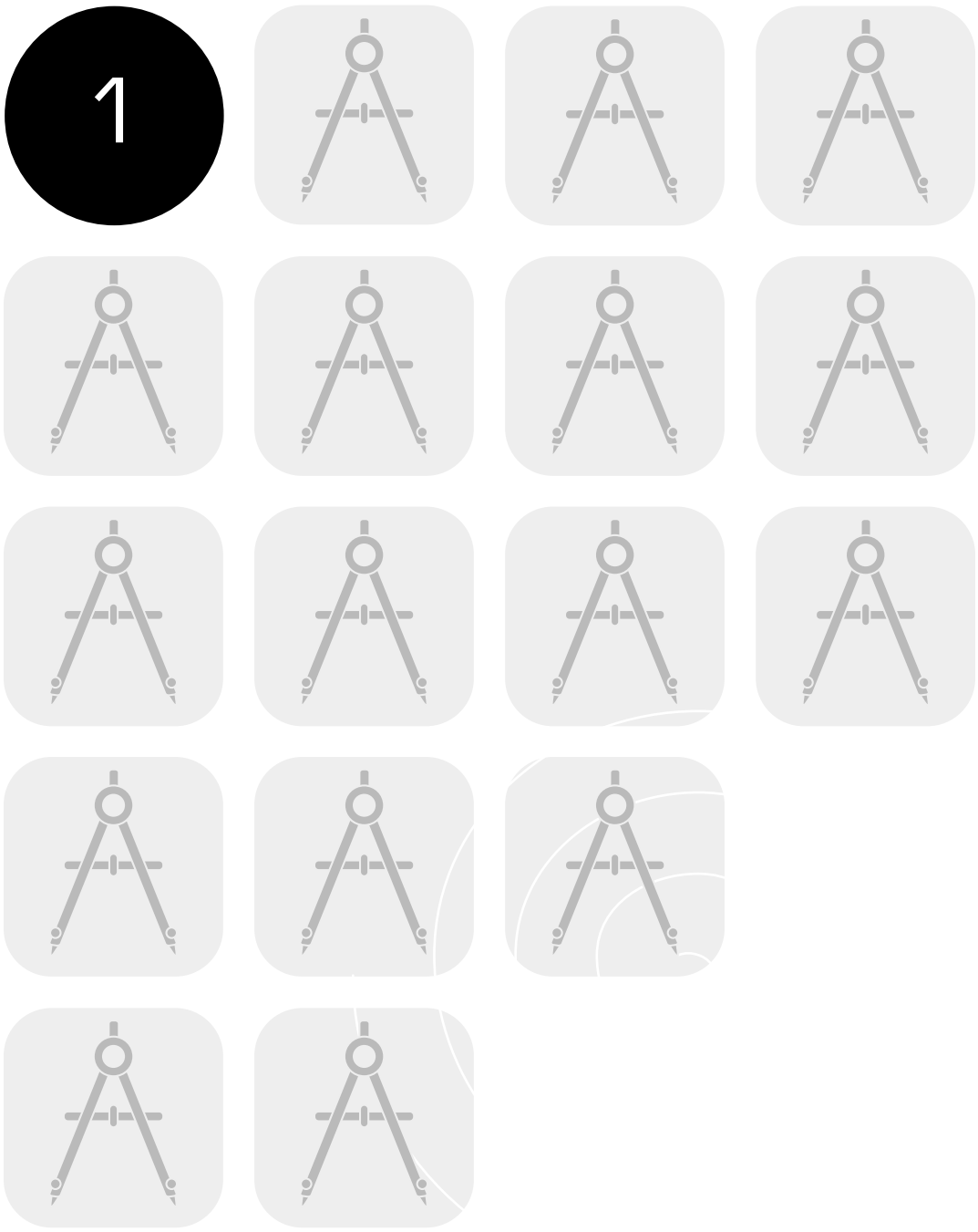
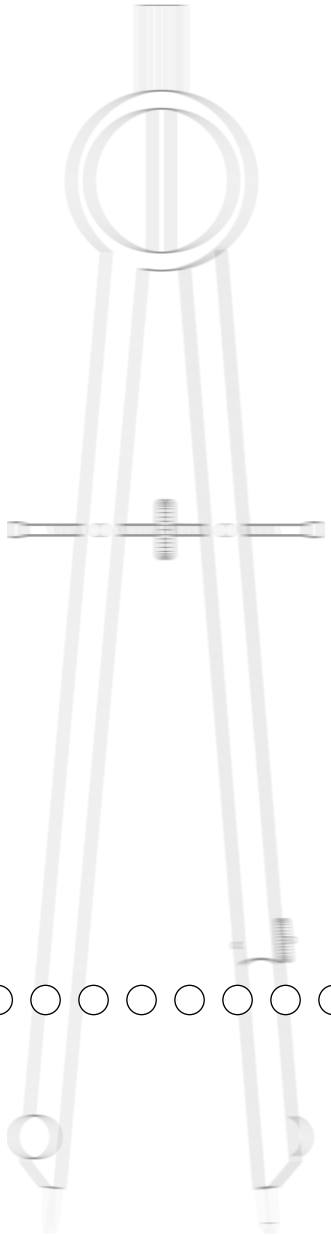
1



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 interna-

tional 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 mathe-

matics 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 rela-

tively 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 back-

ground 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.