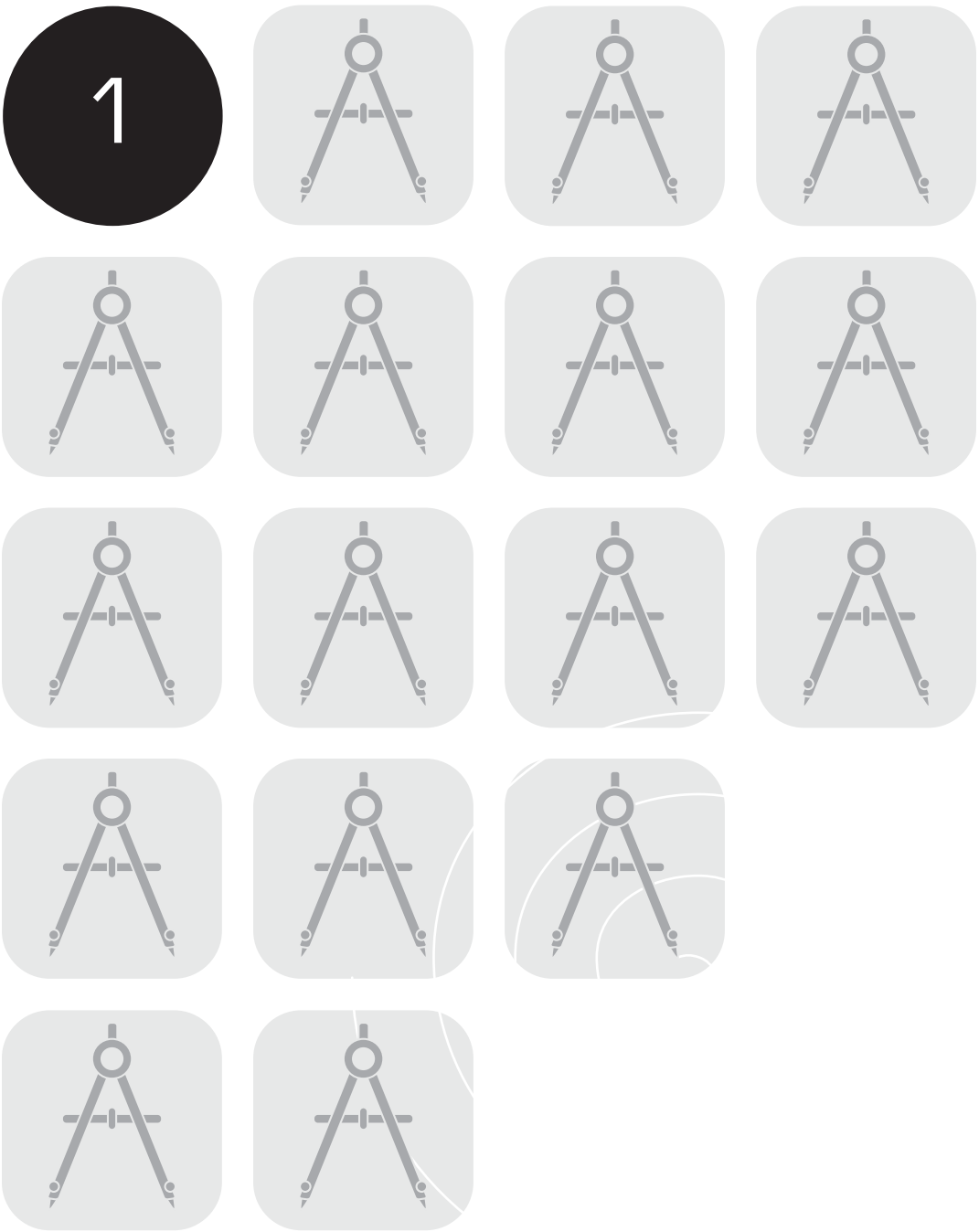
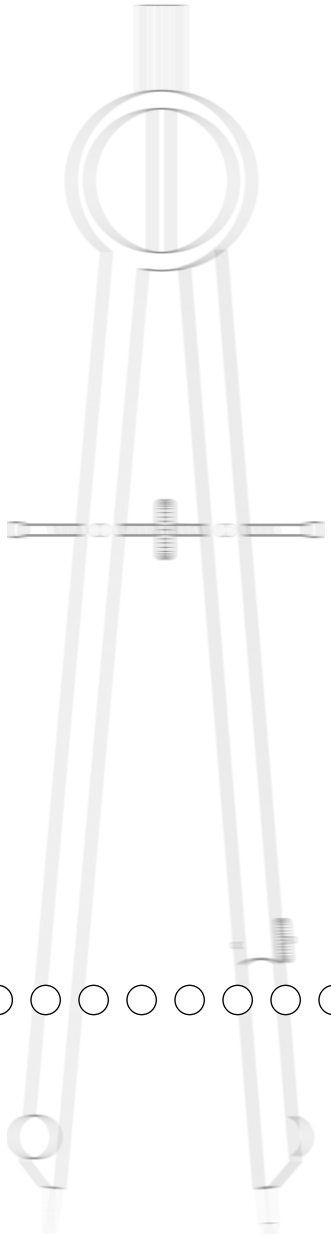


TIMSS 1999 Benchmarking: an Overview

Michael O. Martin
Ina V.S. Mullis



1 TIMSS 1999 Benchmarking: 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. The Third International Mathematics and Science Study (TIMSS), conducted in 1994-1995, was the largest and most complex IEA study, 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. The results of the TIMSS 1999 mathematics assessment are presented in Mullis, Martin, Gonzalez, Gregory, Garden, O'Connor, Chrostowski, and Smith (2000) and the science assessment in Martin, Mullis, Gonzalez, Gregory, Smith, Chrostowski, Garden, and O'Connor (2000). Technical aspects of the project are documented in Martin, Gregory, and Stemler (2000).

To provide U.S. states and school districts with an opportunity to benchmark the performance of their students against that of students in the high-performing TIMSS countries, the International Study Center at Boston College, with the support of the National Center for Education Statistics and the National Science Foundation, established the TIMSS 1999 Benchmarking Study. Through this project, the TIMSS mathematics and science achievement tests and questionnaires were administered to representative samples of students in participating states and school districts in the spring of 1999, at the same time the tests and questionnaires were administered in the TIMSS countries. Participation in TIMSS Benchmarking was intended to help states and districts understand their comparative educational standing, assess the rigor and effectiveness of their own mathematics and science programs in an international context, and improve the teaching and learning of mathematics and science. Mathematics results for the Benchmarking participants are presented in Mullis, Martin,

Gonzalez, O'Connor, Chrostowski, Gregory, Garden, and Smith (2001), and science results in Martin, Mullis, Gonzalez, O'Connor, Chrostowski, Gregory, Smith, and Garden (2001). The purpose of this present volume is to describe the technical procedures underlying the Benchmarking reports.

1.2 Participants in TIMSS Benchmarking

Thirteen states availed of the opportunity to participate in the Benchmarking Study. Eight public school districts and six consortia also participated, for a total of fourteen districts and consortia. They are listed in Exhibit 1 of the Introduction, together with the 38 countries that took part in TIMSS 1999.

1.3 The Student Population

TIMSS 1999 had as its target population students enrolled in the upper of the two adjacent grades that contained the largest proportion of 13-year-old students at the time of testing, which was the eighth grade in most countries, including the United States. The eighth grade was the target population for all of the Benchmarking participants.

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. Data collection among Benchmarking participants took place at the time of the U.S. national TIMSS data collection.

1.5 The TIMSS 1999 Assessment Framework

An essential attribute of the TIMSS 1999 Benchmarking study was that students in the Benchmarking jurisdictions were presented with the same mathematics and science assessment as students participating in the international study.

The designers of TIMSS 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).

Exhibit 1.1 TIMSS 1999 Countries and Benchmarking Participants

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

States
Connecticut
Idaho
Illinois
Indiana
Maryland
Massachusetts
Michigan
Missouri
North Carolina
Oregon
Pennsylvania
South Carolina
Texas

Districts and Consortia
Academy School Dist. #20, CO
Chicago Public Schools, IL
Delaware Science Coalition, DE
First in the World Consort., IL
Fremont/Lincoln/WestSide PS, NE
Guilford County, NC
Jersey City Public Schools, NJ
Miami-Dade County PS, FL
Michigan Invitational Group, MI
Montgomery County, MD
Naperville Sch. Dist. #203, IL
Project SMART Consortium, OH
Rochester City Sch. Dist., NY
SW Math/Sci. Collaborative, PA

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.

The organization and coverage of the intended curriculum were investigated in TIMSS 1999 through curriculum questionnaires that were completed by National Research Coordinators (NRCs) and their curriculum advisors. 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 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.[1] 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.

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.

○○○

1. 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		

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. The assessment consisted of eight booklets, each requiring 90 minutes of response time. Each participating student was assigned one booklet only. The mathematics and science items were assembled into 26

groups or clusters, which were assigned to the student booklets in accordance with the design (seven clusters per booklet) so that representative samples of students responded to each item cluster. In all, the design provided 396 testing minutes, 198 for mathematics and 198 for science.

1.8 Background Questionnaires

TIMSS in 1999 administered a broad array of questionnaires both in participating countries and Benchmarking jurisdictions to collect data on the educational context for student achievement. *Benchmark Coordinators* and *National Research Coordinators* from participating countries, 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. The mathematics and science *teachers* of sampled students responded to questions about teaching emphasis on topics in the curriculum frameworks, instructional practices, professional training and education, and their views on mathematics and science. The heads of *schools* responded to questions about school staffing and resources, mathematics and science course offerings, and teacher support.

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. The translation process and its verification represented an enormous effort for the national centers and for the international management team. Even though the United States and the Benchmarking participants tested in English, it was nonetheless necessary to make minor cultural adaptations to reflect U.S. language usage.

1.10 Sampling

To meet the TIMSS' sampling standards, the Benchmarking sample design had to result in probability samples that gave accurately weighted estimates of population parameters in each Benchmarking jurisdiction, and for which estimates of sampling variance could be computed. Sampling for the Benchmarking study was conducted by Westat, following the sampling design for the U.S. national TIMSS sample as much as possible, but with adaptations to suit the circumstances of individual Benchmarking participants.

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, which may be stratified; the second stage consisted of a single classroom selected at random from the target grade in sampled schools. Large countries like the United States added an extra preliminary stage in which school districts were sampled first, and then schools within districts.

Although in the second sampling stage the sampling units were intact mathematics classrooms, the ultimate sampling units 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 science class or classes. In some systems, however, it may have been the case that some of the students in the selected mathematics class were not enrolled in a science class, and possibly some students in the science class were not enrolled in any mathematics class.

TIMSS 1999 Benchmarking study participants included thirteen states, eight public school districts, and six self-defined school consortia. Samples were selected according to a two-stage stratified systematic sample design. Schools were selected independently within the sampling strata, then classes were selected within schools. The student sample consisted of all eligible students within the selected classes.

Sampling strata were defined by public/private status, where regular public, Bureau of Indian Affairs, Department of Defense, and state schools were “public”; Catholic, non-Catholic religious, and non-religious private schools were “private”. The public school target sample size was 50 for states and 25 for districts and consortia. If schools from a participating Benchmarking jurisdiction were selected as part of the U.S. sample for the TIMSS 1999 international study (U.S. national sample), those schools were also included in the TIMSS 1999 Benchmarking study sample. Target stratum sample sizes were assigned so that the distribution of the Benchmarking study sample would be proportional to strata eighth grade enrollments.

1.11 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. As the data collection contractor for the U.S. national TIMSS, Westat was fully acquainted with the TIMSS procedures, and applied them in each of the Benchmarking jurisdictions in the same way as in the national data collection.

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 recruited and trained a team of 71 international quality control monitors to observe the data collection in each country. Quality control monitors visited a sample of approximately 15 schools in each of the 38 TIMSS countries, where they observed testing sessions and interviewed school coordinators. In all, a total of 550 testing sessions were observed. Reports from monitors indicated a high degree of compliance with prescribed procedures.

As a parallel quality control effort for the Benchmarking project, the International Study Center recruited and trained a team of 18 quality control observers, and sent them to observe the data collection activities of the Westat test administrators in a sample of about 10 percent of the schools in the study (98 schools in all). In line with the experience internationally, the observers reported that the data collection was conducted successfully according to the prescribed procedures, and that no serious problems were encountered.

1.12 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. Analyses of responses based on the second digit should provide insight into ways to help students better understand mathematics concepts and problem-solving approaches.

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. In the United States, the scoring was conducted by National Computer Systems (NCS) under contract to Westat. To ensure that student responses from the Benchmarking jurisdictions were scored in the same way as those from the U.S. national sample, NCS had both sets of data scored at the same time and by the same scoring staff.

1.13 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. In the United States, the creation of the data files for both the Benchmarking jurisdictions and the U.S. national TIMSS effort was the responsibility of Westat, working closely with NCS. After the data files were checked carefully by Westat, they were sent to the IEA Data Processing Center, where they underwent further validity checks before being forwarded to the International Study Center at Boston College.

1.14 IRT Scaling and Data Analysis

The reporting of the TIMSS achievement data was based primarily on item response theory (IRT) scaling methods. The achievement 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 and science 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) and six science content areas (earth science, life science, physics, chemistry, environmental and resource issues, and scientific inquiry and the nature of science), as well as for mathematics and science 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. The TIMSS scale was set to have an average over those countries that participated in TIMSS in 1995 of 500 and a standard deviation of 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. Students tested in the Benchmarking jurisdictions were assigned scores on this scale using the TIMSS IRT procedures.

IRT scales were also created for each of the five mathematics and six science content areas for the 1999 data. 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.

1.15 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. The U.S. National Center for Education Statistics was the TIMSS national center for the United States, with Patrick Gonzales serving as national research coordinator (NRC). Sampling and data collection activities were subcontracted to Westat.

TIMSS NRCs were responsible for a range of important activities, including: meeting with other NRCs and international project staff to review data collection instruments and procedures; conducting all national sampling activities; translating all of the tests, questionnaires, and administration manuals into the language of instruction of the country; 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 quality assurance site visits to schools during data collection; recruiting and training individuals to score the free-response questions in the achievement tests; recruiting and training data entry personnel for creating computerized data files, and conducting the data entry operation, using the software provided; and checking the accuracy and integrity of the data files before 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.[2]

○○○

2. A list of the TIMSS 1999 National Research Coordinators is provided in Appendix A.

All sampling and data collection activities for the Benchmarking project were conducted by Westat, under contract from the TIMSS International Study Center at Boston College. Scoring of constructed-response achievement items and data entry was carried out be National Computer Systems (NCS) under subcontract from Westat.

The TIMSS 1999 International Study Directors, Ina V.S. Mullis and Michael O. Martin, were responsible for the direction and coordination of both TIMSS 1999 internationally and the Benchmarking 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 studies. Several important TIMSS functions, including translation verification, sampling, data processing, and scaling, were conducted by centers around the world, under the direction of the TIMSS International Study Center. 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 both international and Benchmarking 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. Statistics Canada worked with Westat to ensure that all Benchmarking sampling activities were in compliance with established TIMSS procedures. Educational Testing Service, located in Princeton, New Jersey, was responsible for the psychometric scaling of the achievement data from both participating TIMSS countries and Benchmarking jurisdictions.

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.

1.16 Summary of the Report

In chapter 2, Robert Garden and Teresa Smith (subject matter coordinators in mathematics and science, respectively) describe the development of the TIMSS 1999 mathematics and science achievement tests, including the writing of 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. The TIMSS tests used in the Benchmarking study were identical to those used by the United States in the TIMSS 1999 international study.

Ina Mullis, Michael Martin, and Steven Stemler in chapter 3 provide an overview of the background questionnaires used in TIMSS 1999 and the Benchmarking study. 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, noting areas where the United States adapted the international versions to address issues of particular policy relevance.

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, as was the case for the Benchmarking jurisdictions and the United States nationally, adaptations had to be made to suit local language usage. In chapter 4, 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.

The selection of valid and efficient national samples of eighth-grade students in each country was crucial for the quality and success of TIMSS 1999. The international sampling design and sampling manual were developed at Statistics Canada by Pierre Foy and Marc Joncas, who also worked with participating countries in consultation with the TIMSS sampling referee to review national sampling plans, sampling data, sampling frames, and the quality of the national samples. In chapter 5, Pierre Foy and Marc Joncas describe the design and implementation of the international sampling for TIMSS 1999, paying particular attention to the coverage of the target population and to sampling precision requirements. They describe the use of stratification and multi-stage sampling, and illustrate the method used in sampling

schools in TIMSS. In addition, the authors describe the implementation of the sampling design in each of the TIMSS countries, including the grades tested, population coverage, exclusion rates, sample sizes, and participation rates for schools and students.

All sampling activities for the Benchmarking jurisdictions as well as for the U.S. national TIMSS sample were the responsibility of Westat. In chapter 6, Jean Fowler, Lou Rizzo, and Keith Rust describe the TIMSS 1999 Benchmarking sample design and how it relates to the international design set forth in the previous chapter. They present details of the stratification variables used, and describe school and student participation rates, and the procedure used to calculate sampling weights.

As a comparative sample survey of student achievement conducted simultaneously in 38 countries and 27 Benchmarking jurisdictions, TIMSS depended crucially on its data collection procedures to obtain high-quality data. In chapter 7, Eugenio Gonzalez and Dirk Hastedt describe the procedures developed for use in each country 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.

In the Benchmarking project, Westat was responsible for all aspects of data collection and preparation. In chapter 8, Dward Moore describes the field operations conducted by Westat, including within-school sampling activities and the administration of the achievement tests and questionnaires. He also outlines the data preparation tasks conducted by NCS under subcontract to Westat, including image processing and online scoring of free-response items, and scanning of test booklets and questionnaires.

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 9, Kathleen O'Connor and Steven Stemler describe the program of quality control site visits to each of the Benchmarking states and districts. As part of this program,

TIMSS recruited and trained a team of quality control monitors to conduct the site visits. These monitors visited a sample of schools taking part in the study to interview the School Coordinator and Test Administrator and to observe the test administration.

To ensure the availability of comparable, high-quality data for analysis, TIMSS took rigorous quality control steps to create the international database and its Benchmarking counterpart. Upon arrival at the IEA Data Processing Center, 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. The Benchmarking data was subject to the same set of quality control checks. Throughout, the International Study Center monitored the process and managed the flow of data. In chapter 10, Dirk Hastedt, Oliver Neuschmidt, and Eugenio Gonzalez describe the procedures for cleaning and verifying the TIMSS international and Benchmarking data and for constructing the databases used for analysis and reporting.

The statistics presented in the TIMSS 1999 Benchmarking reports are estimates of student performance based on probability samples of eighth-grade students, with each student responding to just a segment of the whole mathematics and science assessment. In chapter 11, Eugenio Gonzalez and Pierre Foy describe the jackknife procedure used in TIMSS to estimate the standard errors associated with each statistic presented in the Benchmarking reports.

Before scaling the TIMSS data to produce achievement scores, summaries of students' responses to each individual item 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. In chapter 12, Ina Mullis and Michael Martin describe the procedures used at the International Study Center to review item statistics for every mathematics and science item in each country to identify potentially problematic items. Item statistics also were calculated for every item for each of the Benchmarking participants, and were subjected to the same review process.

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 methods in the analysis of the achievement results. In chapter 13, 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. The international item parameters and the same methodological approach were applied also to the data from the Benchmarking jurisdictions.

To enrich the description of student mathematics and science achievement, TIMSS identified the 90th, 75th, 50th, and 25th international percentiles as benchmarks with which student performance could be compared. In chapter 14, 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. The international percentiles were also used in reporting the Benchmarking data.

The data in the TIMSS 1999 international and Benchmarking reports are presented mainly using basic descriptive statistics such as averages and percentages. However, because of the complexity of the data, especially the use of plausible values as measures of student achievement, the calculation of even simple statistics is not straightforward. In chapter 15, Eugenio Gonzalez and Kelvin Gregory describe how these analyses were conducted, paying particular attention to multiple comparisons between average scores, standard errors for differences, and the relative performance of countries and jurisdictions across mathematics and science content areas. They also describe the calculation of the international percentiles that were used as international benchmarks, and how the percentages of students reaching each benchmark were computed.

TIMSS 1999 collected an enormous amount of data on educational contexts from students, teachers, and school principals, as well as information about the intended curriculum. In chapter 16, Teresa Smith describes the analysis and reporting of these background data in the Benchmarking reports - the development of the plans for the reports, the construction of composite indices, the review procedures, and special issues in reporting, such as response rates and reporting teacher data.

1.17 Summary

This technical report provides an overview of the main features of the TIMSS 1999 Benchmarking 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 Benchmarking 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., Gregory, K.D., & Stemler, S.E. (2000). *TIMSS 1999 technical report.* 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.* Chestnut Hill, MA: Boston College.

Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., O'Connor, K.M., Chrostowski, S.J., Gregory, K.D., Smith, T.A., & Garden, R.A. (2001). *Science benchmarking report: TIMSS 1999—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.* Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., O'Connor, K.M., Chrostowski, S.J., Gregory, K.D., Garden, R.A & Smith, T.A. (2001). *Mathematics benchmarking report: TIMSS 1999—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.

Travers, K.J., & Westbury, I. (1989). *The IEA study of mathematics I: Analysis of mathematics curricula.* Oxford: Pergamon Press.