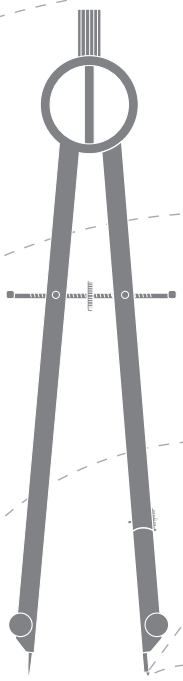


# TIMSS 1999 Benchmarking Technical Report


**edited by:**

**Michael O. Martin**
**Kelvin D. Gregory**
**Kathleen M. O'Connor**
**Steven E. Stemler**

**with contributors:**

Jean Fowler
Pierre Foy
Robert Garden
Eugenio J. Gonzalez
Dirk Hastedt
Marc Joncas
Edward Kulik
Barbara Malak
Dward Moore
Ina V.S. Mullis
Oliver Neuschmidt
Lou Rizzo
Keith Rust
Teresa A. Smith
Kentaro Yamamoto

**International Study Center • Boston College • Chestnut Hill, Massachusetts**

TIMSS 1999 Benchmarking Technical Report / edited by Michael O. Martin, Kelvin D. Gregory, Kathleen M. O'Connor and Steven E. Stemler

Publisher: International Study Center
Lynch School of Education
Boston College

Library of Congress Catalog Card Number: 2001096081

ISBN 1-889938-24-6

For more information about TIMSS contact:

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

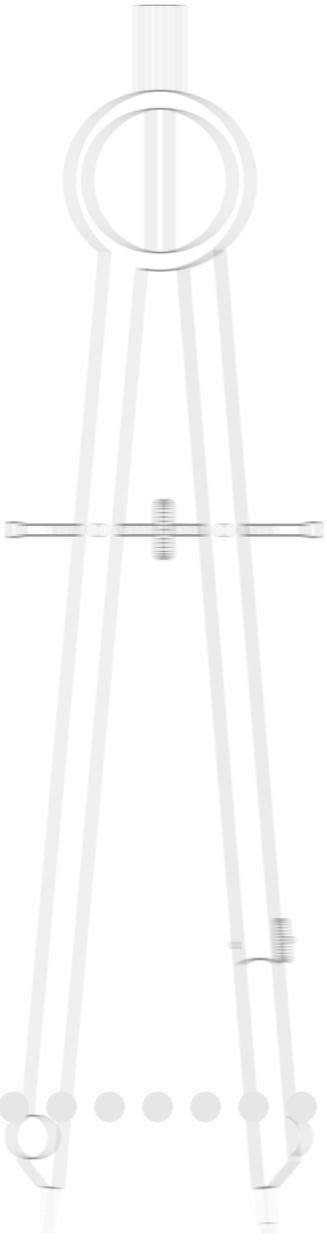
For information on ordering this report, write to the above address or call +1-617-552-1600

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

Funding for the TIMSS 1999 Benchmarking Study was provided by the National Center for Education Statistics and the Office of Educational Research and Improvement of the U.S. Department of Education, the U.S. National Science Foundation, and participating jurisdictions.

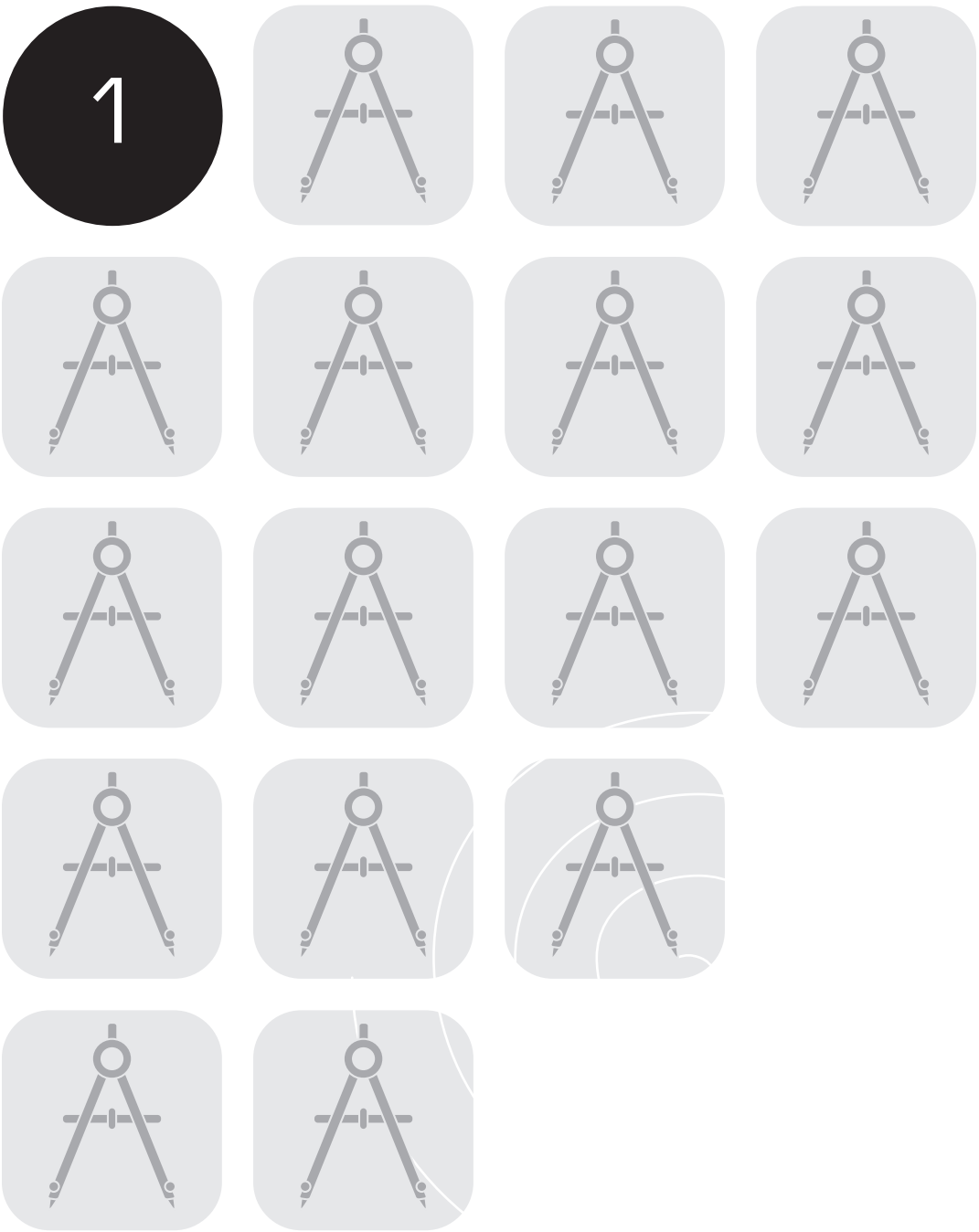


Printed and bound in the United States



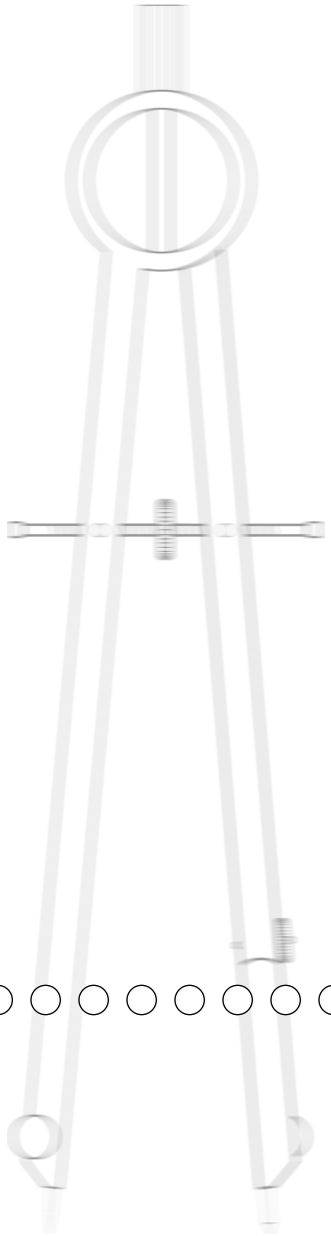
# Contents

# Contents




# 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 90$^{th}$, 75$^{th}$, 50$^{th}$, and 25$^{th}$ 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.

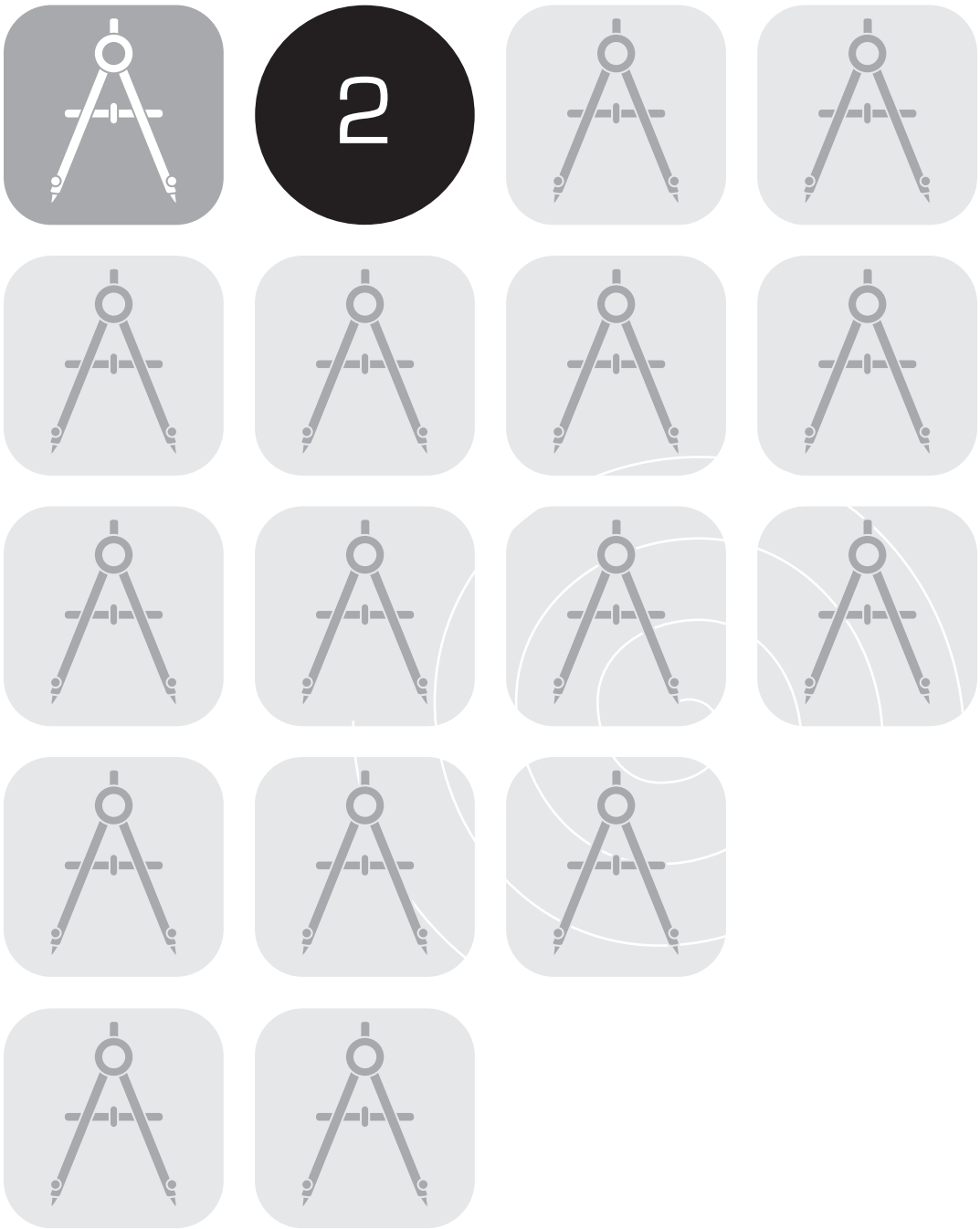
Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., Gregory, K.D., Garden, R.A., O'Connor, K.M., Chrostowski, S.J., & Smith, T.A. (2000). *TIMSS 1999 international mathematics report.* Chestnut Hill, MA: Boston College.

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.

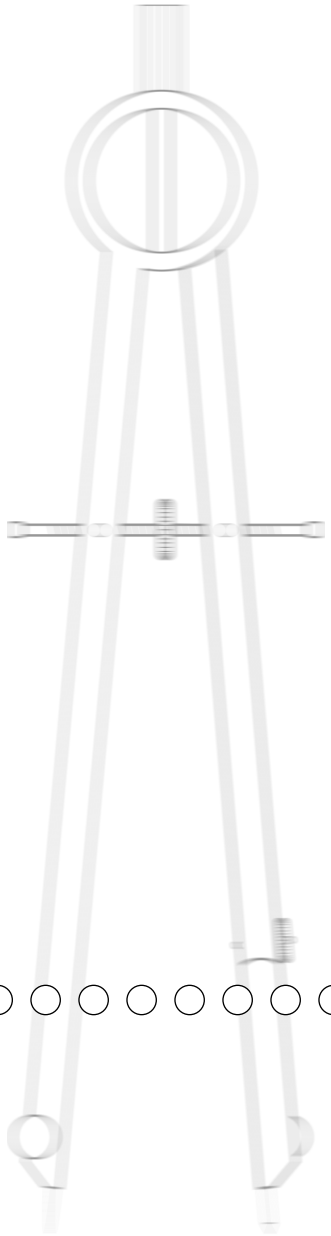
Robitaille, D.F., Schmidt, W.H., Raizen, S.A., McKnight, C.C., Britton, E., & Nicol, C. (1993). *TIMSS monograph No. 1: curriculum frameworks for mathematics and science.* Vancouver, Canada: Pacific Educational Press.

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




# TIMSS Test Development

Robert A. Garden
Teresa A. Smith



# 2 TIMSS Test Development[1]

Robert A. Garden
Teresa A. Smith

## 2.1 Overview

The mathematics and science achievement tests used to measure achievement in the TIMSS benchmarking project were identical to those used by the United States as part of the international data collection for TIMSS 1999. This chapter describes how the 1999 tests followed the same design and used many of the same items as the TIMSS 1995 assessment, how the items released into the public domain following the 1995 assessment were replaced for 1999, and the procedure for field testing the replacement items and reviewing the results.

To provide as much information as possible about the nature and scope of the 1995 TIMSS achievement tests, almost two-thirds of the test items 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 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 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.

○○○

1. This chapter was mainly reproduced from Garden & Smith (2000) from the international technical report for TIMSS 1999 (Martin, Gregory, & Stemler, 2000).

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 new 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 TIMSS framework[2] for the 1995 assessment.

## 2.2 Development of Replacement Items

The major goal of test development was to produce a test that would parallel TIMSS 1995 in overall structure and content. The strategy used was to treat the 1995 items as a representative sample from the "pool" of all possible items within the defined test domain and to select 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 (i.e., 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 TIMSS 1995. The approach is described in further detail in the following sections.

### 2.2.1 Replacement of Item Clusters

In the 1995 TIMSS assessment, mathematics and science items were organized into 26 clusters, labeled A to 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 2.1 (Adams & Gonzalez, 1996). The same booklet design was used in TIMSS 1999. Clusters A through H, of multiple-choice items only, each 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 each took 10 minutes of testing time.

○○○

2. 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.

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 retained the same format, assessed the same basic content area and performance expectation and, as nearly as possible matched the difficulty level of the 1995 items.

### 2.2.2 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 replacement item was to present students with a task similar to that in the corresponding 1995 item, care was taken not to make it so similar as to favor 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 were written in the same format as the original.[3] For multiple-choice items, when feasible, each distracter was designed to depend on the same faulty reasoning, miscalculation, or misconception as in the original item.

○○○

3. 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.

**Exhibit 2.1 Assignment of Item Clusters to Student Test Booklets* — 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 | | |

* 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.

Item-by-item matching of 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.

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*. This was done to expand the item pool and permit the results in these two content areas to be reported separately for TIMSS 1999 (see section 2.5 for a discussion of the final TIMSS 1999 science test).

### 2.2.3 Scoring Guides for Free-Response Items

The TIMSS 1999 item replacement efforts focused heavily on developing free-response items, questions to which 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 2.2 presents the number of free-response and multiple-choice questions included in the field test.

**Exhibit 2.2 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 approaches or 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 2.3.

**Exhibit 2.3 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 seven and eight 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 nine) at the appropriate score level.

### 2.2.4 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)[4], 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 each 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.

### 2.2.5 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 SMIRC consisted of prominent mathematics and science educators nominated by participating countries, and thus represented a variety of nations and cultures. 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

○○○

4. See Appendix A for a list of the members.

unambiguous in meaning, with appropriate reading demands, clear graphics, and a defensible key or scoring guide. For free-response 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 in both the items and scoring guides.

Selecting items for the field test also demanded the committee's expertise. 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.

## 2.3 Field Test

A total of 277 potential replacement items were 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. 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.

### 2.3.1 Field-Test Item Analyses

International item analysis of the field test results was used to help review and select the 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 for or against any particular country, or in favor of boys or girls. These statistics also included the response distributions across multiple-choice options or across diagnostic response codes for the free-response items. The results of these analyses were summarized in data almanacs that were used to review the field test results.

### International Study Center Review

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

### 2.3.2 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 chosen were used unchanged in the main survey, and only minor revisions were made to the others. Similarly, 75% of the science items chosen were left essentially unchanged. Revisions made included improving the clarity and print quality of graphics and drawings, clarifying item stems, and modifying distracters that were selected by very few students.

### 2.3.3 Revising the Scoring Guides

The TIMSS International Study Center used information collected in the field test to revise 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 few responses
- Adding categories with frequent responses as reported by the NRCs
- Clarifying or sometimes combining less reliable categories
- 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 improving scoring reliability. Consistent with the approach used in TIMSS 1995, some free-response items were awarded one point, others two points, and some had more than one part, each worth one or two points. In general, one 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, they usually required a one- or two-sentence explanation or factual description. In both subjects, two-point items were those demanding more than a numerical or short written response. In

○○○

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

mathematics, students were asked to show their work or explain their methods, and these explications were taken into account in scoring. In science, the two-point items required a fuller explanation demonstrating knowledge of science concepts. The distinction between the one- and two-point items was sometimes hazy in science, and for some two-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 two points vs. only one point. The generalized scoring guides for mathematics are presented in Exhibit 2.4 and those for science in Exhibit 2.5.

**Exhibit 2.4 TIMSS 1999 Mathematics Generalized Scoring Guide**

| Score Points for Extended-Response Items |
|---|
| **2 Points:**<br>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.<br>• Indicates that the student has completed the task, showing mathematically sound procedures<br>• Contains clear, complete explanations and/or adequate work when required |
| **1 Point:**<br>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.<br>• Addresses some elements of the task correctly but may be incomplete or contain some procedural or conceptual flaws<br>• May contain a correct solution with incorrect, unrelated, or no work and/or explanation when required<br>• May contain an incorrect solution but applies a mathematically appropriate process |
| **0 Points:**<br>A zero-point response is completely incorrect, irrelevant, or incoherent. |
| **Score Points for Short-Answer Items** |
| **1 Point:**<br>A one-point response is correct. The response indicates that the student has completed the task correctly. |
| **0 Points:**<br>A zero-point response is completely incorrect, irrelevant, or incoherent. |

**Exhibit 2.5 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 reflected their comments. A few last suggestions were made before the scoring guides were issued for use in training in the Southern Hemisphere countries in Wellington, New Zealand, in October 1998. During this first training session, a few additional revisions were made. These were incorporated in the final TIMSS 1999 scoring guides used during scoring training for the Northern Hemisphere countries in February 1999.

## 2.4 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 for the within-country reliability studies.

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

- Participants read the item and its scoring guide
- Trainers discussed the rationale and method 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 to provide 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 scoring guides, manuals, and packets of example and practice papers for each of item covered in the training. The training teams emphasized the need for the NRCs to prepare comparable materials for training in their own country, including all of the free-response items rather than only the sample 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. 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 done 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 content and clarity of the scoring guides. The revisions were made before the 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 three days, allotting one day for mathematics and two for science. This format permitted training on 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 more experience 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 NRCs.

## 2.5 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 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 2.6 and 2.7 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 with a maximum of two points, the number of score points may exceed the number of items.

**Exhibit 2.6 Number of TIMSS 1999 Test Items and Score Points by Type and Reporting Category —Mathematics**

| 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 2.7 Number of TIMSS 1999 Test Items and Score Points by Type and Reporting Category—Science**

| 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 of the original TIMSS study. Assignment to item clusters generally followed the original design, with most of the 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 in 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 2.8 and 2.9 present the distribution of items in each content area across the eight test booklets for mathematics and science, respectively.

**Exhibit 2.8 Number of TIMSS 1999 Items in Each Booklet by Subject Matter Content Category—Mathematics**

| 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 2.9 Number of TIMSS 1999 Items in Each Booklet by Subject Matter Content Category—Science**

| 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 2.10 and 2.11.

**Exhibit 2.10 Maximum Number of TIMSS 1999 Score Points in Each Booklet by Subject Matter Content Category—Mathematics**

| 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 2.11 Maximum Number of TIMSS 1999 Score Points in Each Booklet by Subject Matter Content Category—Science**

| 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* (pp. 3.1 - 3.36). Chestnut Hill, MA: Boston College.

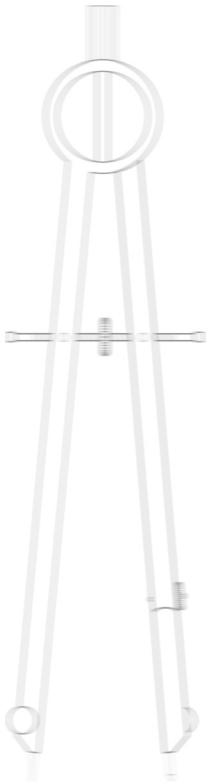
Garden, R.A., & Smith, T.A. (2000). TIMSS test development. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS 1999 technical report* (pp. 47 - 68). Chestnut Hill, MA: Boston College.

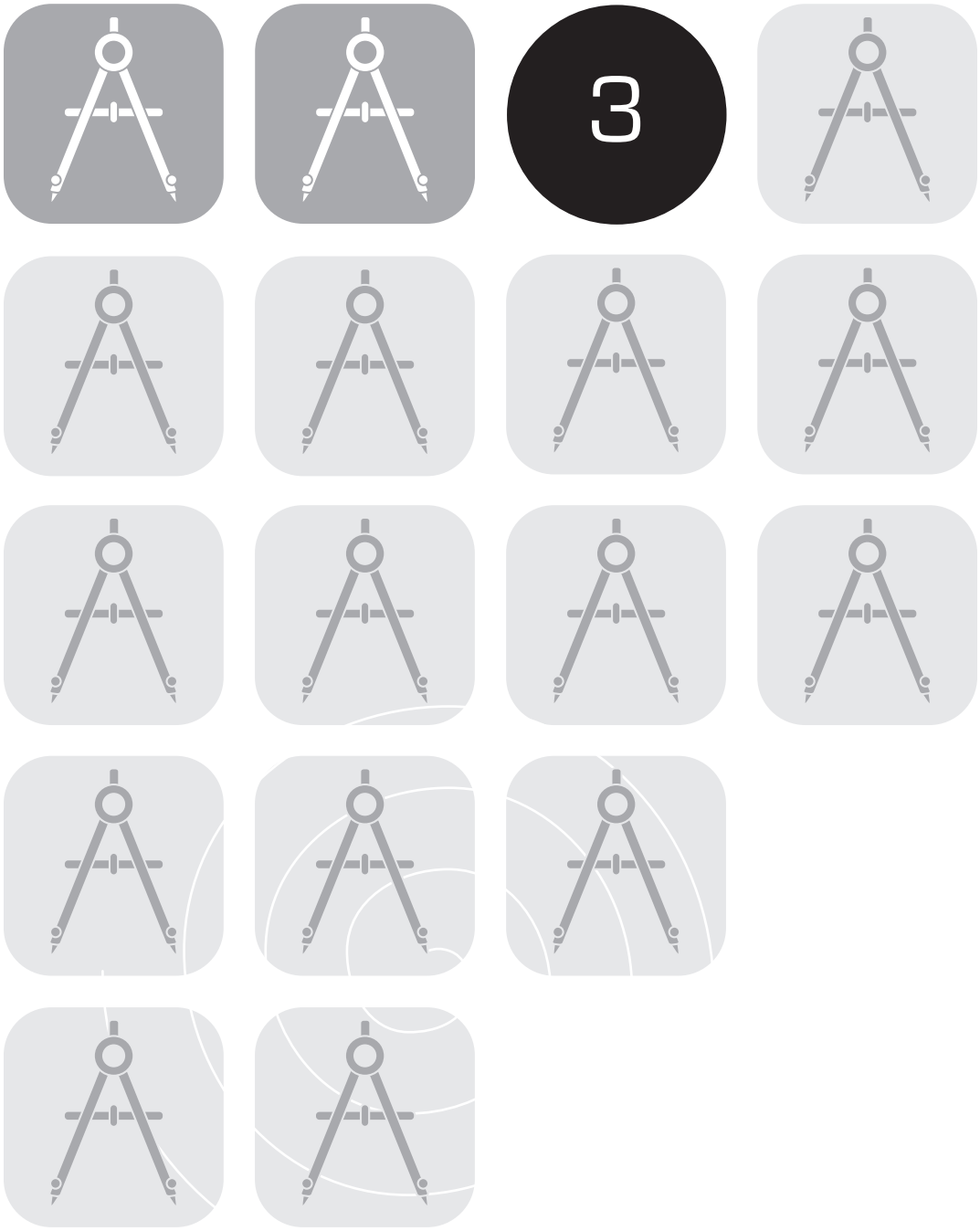
Lie, S., Taylor, A., & Harmon, M. (1996). Scoring techniques and criteria. In M.O., Martin & D.L. Kelly (Eds.), *Third international mathematics and science study technical report volume I: Design and development* (pp. 7.1 - 7.16). Chestnut Hill, MA: Boston College.

Martin, M.O., Gregory, K.D., & Stemler, S.E. (Eds.). (2000). *TIMSS 1999 technical report.* Chestnut Hill, MA: Boston College.

Mullis, I.V.S., & Smith, T.A. (1996). Quality control steps for free-response scoring. In M.O. Martin & I.V.S. Mullis (Eds.), *Third international mathematics and science study: Quality assurance in data collection* (pp. 5.1 - 5.32). Chestnut Hill, MA: Boston College.

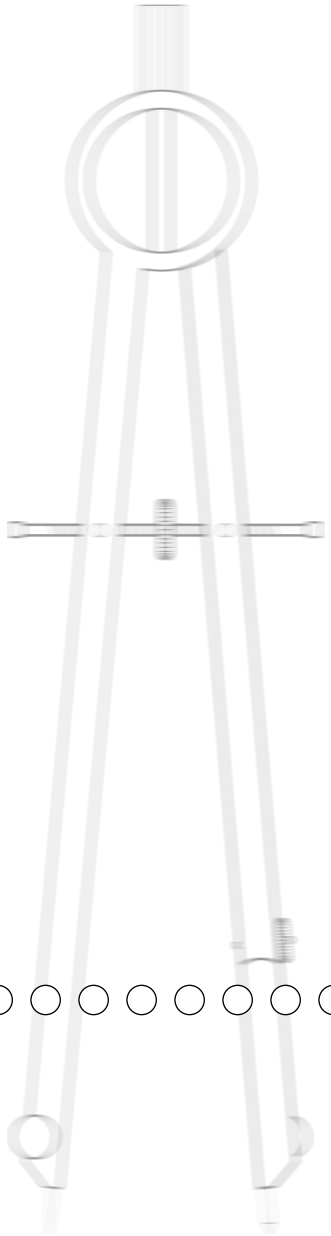
Robitaille, D.F., Schmidt, W.H., Raizen, S., McKnight, C., Britton, E., & Nicol, C. (1993). *Curriculum frameworks for mathematics and science.* Vancouver: Pacific Educational Press.






# TIMSS Questionnaire Development

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



# 3 TIMSS Questionnaire Development[1]

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

## 3.1 Overview

Just as the TIMSS Benchmarking study used the U.S. versions of the TIMSS 1999 mathematics and science tests to measure achievement, it also used the U.S. versions of the TIMSS questionnaires to gather information about the educational context in each participating jurisdiction. This chapter describes the design and development of the TIMSS questionnaires, and summarizes the content of each one.

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 to have not only 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 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 developing the international versions of the questionnaires 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

○○○

1. This chapter was based on Mullis, Martin, & Stemler (2000) from the international technical report for TIMSS 1999.

the field test, and was reviewed again in light of the field-test data. Items retained for the final versions of the questionnaires were those judged to yield the most information with the least burden to respondents.

Each country was permitted to include additional questions to the international version of the questionnaires. These additions were categorized as: international or national options. International options are questions provided by the International Study Center as having been found to be of interest in many countries but not required. National options, developed and included in the questionnaire at the individual country level, allowed participants the flexibility of collecting other useful data that pertained to their nation or education system.

Benchmarking participants used the same Student, Teacher, and School Questionnaires as those administered to the U.S. national sample. The Curriculum Questionnaire, however, was adapted to gather more specific information relevant to Benchmarking jurisdictions.

This chapter begins with an overview of the conceptual framework and research questions that guided the development of the questionnaires; it goes on to present the main issues addressed by each questionnaire as well as questions included as international options, U.S. national options, and international questions adapted for Benchmarking participants.

## 3.2 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 education.

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).

## 3.3 Research Questions

Consistent with TIMSS 1995, TIMSS 1999 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?

What students know and are expected to learn was addressed by questionnaires distributed to mathematics and science curriculum experts in participating countries. The characteristics and preparation of mathematics and science teachers were addressed by questionnaires distributed to school principals and teachers. The third question, on instructional approaches to the teaching of mathematics and science, was also addressed by questionnaires to principals and teachers, as well as to students. The fourth question, what students had learned, was examined by measuring performance on the TIMSS 1999 achievement tests.

The research questions cast a broad net for exploring factors potentially associated with achievement in mathematics and science. For example, in attempting to answer the question "Who provides the instruction?" the questionnaires tapped characteristics of the instructor, such as gender, age, years of experience, attitude toward the subject, and time spent preparing lessons. The national options also allowed the U.S. and Benchmarking jurisdictions to gain additional information on teachers' professional development activities. The background questionnaires enable researchers to investigate the most influential characteristics of the people, practices, and policies affecting student achievement.

## 3.4 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 Benchmarking jurisdiction, and about the topics intended to be covered up to the eighth grade. Coordinators in each country and jurisdiction were

asked to complete separate questionnaires about the mathematics and the science curriculum, drawing on the expertise of mathematics and science specialists in their country or jurisdiction as necessary.

The international 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.

A shortened version of these questionnaires was completed by Benchmarking coordinators in each jurisdiction, supplemented with follow-up surveys and interviews designed to put the international questions in the context of the states and districts. For example, while countries were asked to provide information on their national curriculum, states were asked to report on their content standards and curriculum frameworks, and districts and consortia were asked to report on the level at which curriculum is developed. These follow-up surveys and interviews focused on the issues that coordinators from the Benchmarking jurisdictions found to be interesting for international comparisons.

The interviews with national research coordinators (NRCs) and Benchmarking coordinators resolved ambiguities and assisted in developing a clear understanding of each entity's curriculum. Important research questions addressed by the questionnaires were:

- Is there a country-, state-, district-level curriculum or curriculum framework? If so, how is implementation monitored?
- What is the nature of system-wide assessments, if there are any?
- What content is emphasized in the curriculum or curriculum framework?

The contents of the national Mathematics and Science Curriculum Questionnaires are described further in Exhibits 3.1 and 3.2. Additional questions addressed in the Benchmarking curriculum surveys and interviews are described in Exhibit 3.3.

## 3.5 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?

In addition to questions asked internationally, the U.S. and Benchmarking School Questionnaire gathered information on the percentage of students at each school eligible to receive free or reduced-price lunches, to be used as an indicator of socio-economic status.

### 3.5.1 Changes from the 1995 Questionnaire

For the most part, the TIMSS 1999 international 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 3.4.

## 3.6 Teacher Questionnaires

Typically, a single mathematics class in each participating school 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 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 their views of subject matter were geared toward 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?

The U.S. version of the questionnaire included a national-option section that elicited information on teachers' participation in specific professional development activities.

### Changes from the 1995 Questionnaire

Several changes were made in the Mathematics and Science Teacher Questionnaires for the 1999 assessment. The originals were judged by most NRCs to be too lengthy, 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, which were resolved before the questionnaires for the main survey were administered.

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 3.5.

## 3.7 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 TIMSS 1995, two versions of the questionnaire were used internationally:

- The **General Science Version** was intended for systems where science is taught as a single integrated subject
- The **Separate Science Subject Version** was intended for systems where science is taught as separate subjects (i.e., biology, chemistry, earth science, and physics)

Countries administered the version that was consistent with the way in which science instruction was organized at the target grade. U.S. and Benchmarking entities administered the general science version. Although the two versions differed with respect to the science questions, the general background and mathematics-related questions were identical across the two. 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 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 toward mathematics and science?

#### Changes from the 1995 Questionnaire

Five questions from the TIMSS 1995 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 Student 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 TIMSS 1995 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 3.6.

## 3.8 Summary

The U.S. versions of the background questionnaires were very similar to the international versions; however, the U.S. chose to develop and include an additional section in the Teacher Questionnaire related to professional development activities. In addition, the mathematics and science international Curriculum Questionnaires were adapted to apply in the context of states and districts.

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 the 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 TIMSS 1995 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 permit 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 3.1 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 under way. |
| 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 3.2 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 under way. |
| 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 3.3 Contents of the Benchmarking Curriculum Survey and Interview**

| State-level Question | District-level Question | Item Content | Description |
|---|---|---|---|
| NA | 1 | Level of Curriculum Development | Indicates the administrative level at which the curriculum is developed - state, district, or school and whether it is based on state standards. |
| 1 | NA | Curriculum Frameworks/ Content Standards | Indicates the title, date and organization of the state curium framework or content standards |
| 2 | 2 | Status of Assessments | Indicates current status of development of new assessments. |
| 3 | 2 | Assessments | Indicates assessments administered including criterion-referenced assessments and norm-referenced assessments. |
| 3 | NA | Consequences of Assessments | Indicates whether the state requires students to pass an exam for graduation, as well as other consequences for the student, school, or district based on results (includes sanctions and rewards). |
| 4 | 3 | Textbook Selection | Indicates the policy for textbook selection. |
| 5 | 4 | Pedagogical Guide | Provides information on state (or local) pedagogical guides. |
| 6 | 5 | Accreditation | Indicates use of accreditation to support curriculum implementation. |
| 7 | 6 | 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). |
| 8 | 7 | Science Subjects Offered | Provides information on the science courses offered up to an including grade 8 (e.g., biology, chemistry, physics). |
| 9 | 8 | Policy on Calculator Use | Identifies the policy on calculator usage as well and any policy changes that on calculator use that occur as the students progress through school |
| CQ | CQ | Curricular Emphasis | Identifies the extent to which the curriculum emphasizes each of several approaches / processes (e.g., knowing basic science facts, performing science experiments). |

**Exhibit 3.4 Contents of the School Questionnaire**

| International Question Number | U.S. Question Number | Item Content | Description |
|---|---|---|---|
| 1 | 1 | Community | Situates the school within a community of a specific type. |
| 2-4 | 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 | 5 | Years Students Stay with Teacher | Indicates the number of years students typically stay with the same teacher. |
| 6 | 6 | Collaboration Policy | Identifies the existence of a school policy promoting teacher cooperation and collaboration. |
| 7 | 7 | Principal's Time | Indicates the amount of time the school's lead administrator typically spends on particular roles and functions. |
| 8 | 8 | School Decisions | Identifies who has the responsibility for various decisions for the school. |
| 9 | 9 | Curriculum Decisions | Identifies the amount of influence various individuals and educational and community groups have on curriculum decisions. |
| 10 | 10 | Formal Goals Statement | Indicates the existence of school-level curriculum goals for mathematics and science. |
| 11-12 | 11-12 | Instructional Resources | Describes the material factors limiting the school's instructional activities. |
| 13 | 13 | Students in the school | Provides total school enrollment and attendance data. |
| | 13 (e.) | Students in the school | Provides percentage of students receiving free or reduced-price lunches. |
| 14 | 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 | 15 | Number of Computers | Provides the number of computers for use by students in the target grade, by teachers, and in total. |
| 16 | 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 | 17 | Student Behaviors | Describes the frequency with which schools encounter various unacceptable student behaviors. |
| 18 | 18 | Instructional Time | Indicates the amount of instructional time scheduled for the target grade, according to the school's academic calendar. |
| 19 | 19 | Instructional Periods | Indicates the existence and length of weekly instructional periods for the target grade. |
| 20 | 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 | 21 | Program Decision Factors in Mathematics | Indicates how important various factors are in assigning students to different educational programs or tracks in mathematics. |
| 22 | 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 | 23 | Program Decision Factors in Science | Indicates how important various factors are in assigning students to different educational programs or tracks in science. |
| 24 | 24 | Admissions | Describes the basis on which students are admitted to the school. |
| 25 | 25 | Parental Involvement | Describes the kinds of activities in which parents are expected to participate (e.g., serve as teacher's aides, fundraising). |

**Exhibit 3.5 Contents of the Teacher Questionnaires**

| Question Number | U.S. Number | Item Content | Description |
|---|---|---|---|
| **Section A** | | | |
| 1-2 | 1-2 | Age and Sex | Identifies teacher's sex and age range. |
| | 2b | Race / Ethnicity | Identifies teacher's race/ethnicity |
| 3 | 3 | Teaching Experience | Describes the teacher's number of years of teaching experience. |
| 4-5 | 4-5 | Instructional Time | Identifies the number of hours per week the teacher devotes to teaching mathematics, science, and other subjects. |
| 6 | 6 | Administrative Tasks | Identifies the number of hours per week spent on administrative tasks such as student supervision and counseling. |
| 7 | 7 | Other Teaching-Related Activities | Describes the amount of time teachers are involved in various professional responsibilities *outside* the formally-scheduled school day. |
| 8 | 8 | Teaching Activities | Describes the total number of hours per week spent on teaching activities. |
| 9 | 9 | Meet with Other Teachers | Describes the frequency with which teachers collaborate and consult with their colleagues. |
| 10 | 10 | Teacher's Influence | Describes the amount of influence that teachers perceive they have on various instructional decisions. |
| 11 | 11 | Being Good at Mathematics / Science | Describes teacher's beliefs about what skills are necessary for students to be good at mathematics / science. |
| 12 | 12 | Ideas about Mathematics / Science | Describes teacher's beliefs about the nature of mathematics / science and how the subject should be taught. |
| 13 | 13 | Document Familiarity | Describes teacher's knowledge of curriculum guides, teaching guides, and examination prescriptions (country-specific options). |
| 14 | 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 | 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 | NA | Career Choices | Identifies whether teaching was a first choice and if the teacher would change careers if given the opportunity. |
| 21 | NA | Social Appreciation | Describes whether teachers believe society appreciates their work. |
| 22 | NA | Student Appreciation | Describes whether teachers believe students appreciate their work. |
| 23 | NA | Books in Home | Provides an indicator of teacher's cultural capital. |

**Exhibit 3.5 (continued) Contents of the Teacher Questionnaires**

| Question Number | U.S. Number | Item Content | Description |
|---|---|---|---|
| **Section B** | | | |
| 1 | 1 | Target Class | Identifies the number of students in the TIMSS 1999 tested class, by gender. |
| 2 | 2 | Instructional Emphasis | Identifies the subject matter emphasized most in the target mathematics / science class. |
| 3 | 3 | Instructional Time | Identifies the number of minutes per week the class is taught. |
| 4 | 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 | 5-7 | Calculators | Describes the availability of calculators and how they are used in the target class. |
| 8 | 8 | Computers | Describes the availability of computers and whether they are used to access the internet. |
| 9 | 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 | 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 | 11 | Student's Work Arrangements | Describes how often students work in various group arrangements. |
| 12 | 12 | Time Allocation | Describes the percentage of time spent on each of several activities associated with teaching (e.g., homework review, tests). |
| 13 | 13 | Mathematics / Science Topic Coverage | Indicates the extent of teacher's coverage in target class of mathematics / science topics included in the assessment. |
| 14 | 14 | Classroom Factors | Identifies the extent to which teachers perceive that various factors limit classroom instructional activities. |
| 15-16 | 15-16 | Amount of Homework Assigned | Describes the frequency and amount of homework assigned to the target class. |
| 17-18 | 17-18 | Type and Use of Homework | Describes the homework assignments and how the homework is used by the teacher. |
| 19-20 | 19-20 | Assessment | Describes the kind and use of various forms of student assessment in the target class. |
| **Question Number** | **U.S. Number** | **Item Content** | **Description** |
| **U.S. National Option: Professional Development Activities** | | | |
| NA | 1-2 | Classroom Observations | Indicates the number of class periods spent observing other teachers and being observed by other teachers. |
| NA | 3 | Participation in Professional Development Activities | Indicates hours spent on various types of professional development activities. |
| NA | 4 | Participation in Individual Activities | Indicates hours spent on various types of individual professional development activities. |
| NA | 5 | Focus of Activities | Indicates the extent to which professional development activities focused on certain topics (i.e. pedagogy, curriculum, assessment, leadership, etc.) |
| NA | 6 | Focus on Content Areas | Indicates whether the teacher participated in professional development activities related to specific mathematics topics or science content areas covered in the assessment. |
| NA | 7-8 math only | Effect on Student Learning | Identifies the extent to which the teacher believes that student learning was improved as a result of his or her professional development. |

**Exhibit 3.6 Contents of the Student Questionnaires**

| Question Number | | Item Content | Description |
|---|---|---|---|
| General Version (U.S.) | 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. (U.S. version includes a question on race/ethnicity.) |
| 5 | 5 | Academic Activities Outside of School | Provides information on students' 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, dictionary, study desk, and 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 | Describes typical student behavior during mathematics lessons. |
| 14 | 14 | Peers' Values | Provides information about the student's perception of the degree of importance 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 | Describes 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. |

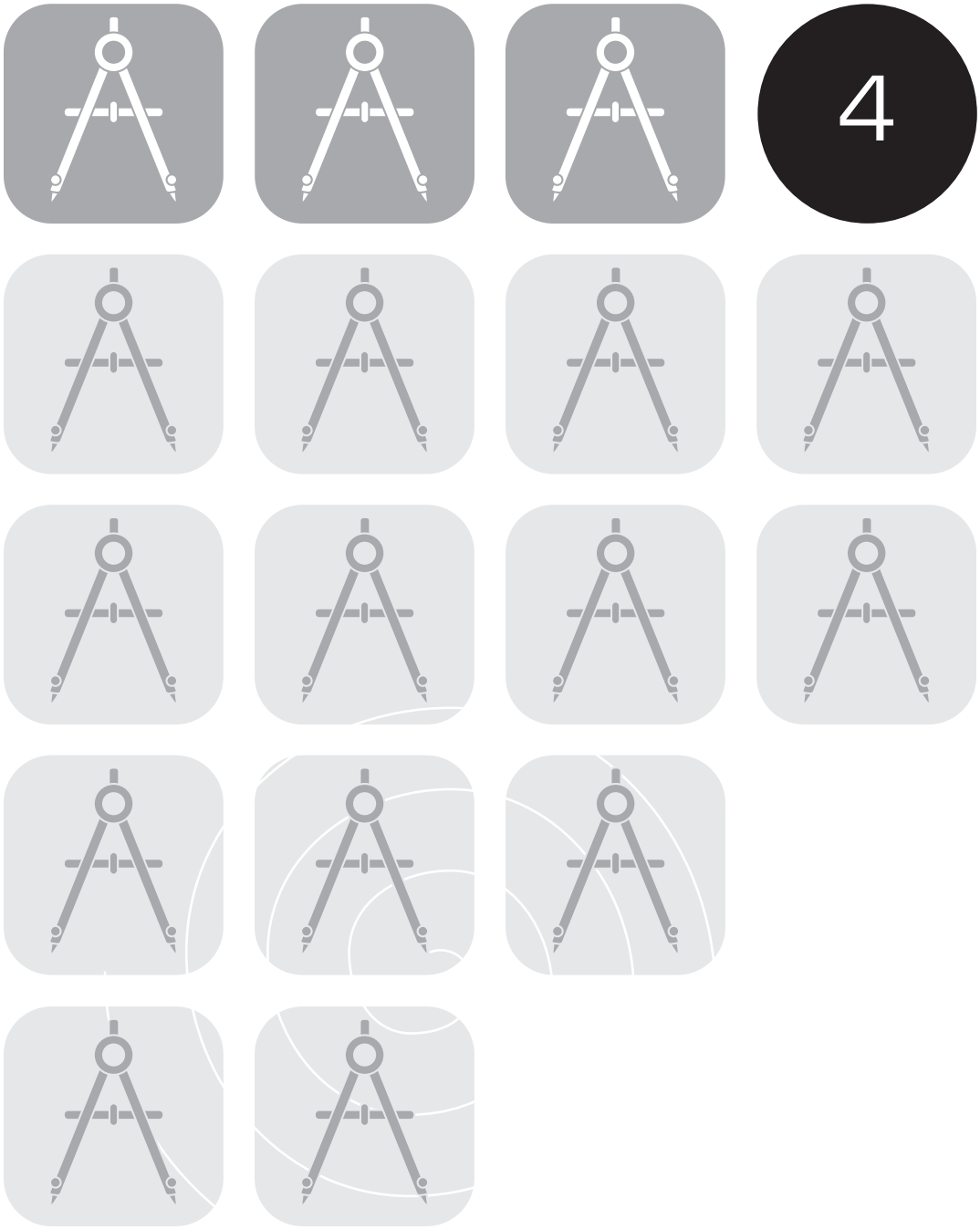
**Exhibit 3.6 (continued) Contents of the Student Questionnaire**

| 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 | Describes 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 | Describes 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 | Describes 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 | Describes 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 | Describes student's involvement in cultural events or programming such as plays or concerts. |
| 37 | 52 | Report on Student Behaviors | Indicates the student's perspective of the existence of specific problematic student behaviors at school. |
| 38 | 53 | Environmental Issues | Indicates the 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

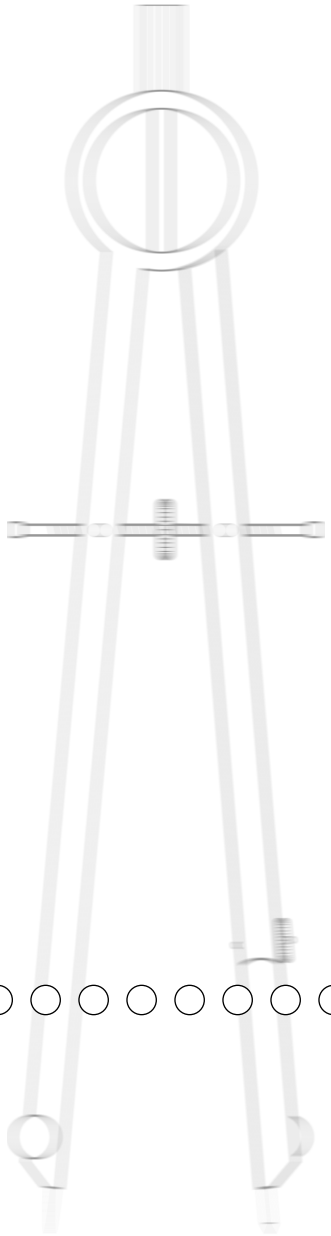
Mullis, I.V.S., Martin, M.O., & Stemler, S.E. (2000). TIMSS questionnaire development. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS 1999 technical report* (pp. 69 - 86). Chestnut Hill, MA: Boston College.

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 I: Design and development* (pp. 5.1 - 5.22). Chestnut Hill, MA: Boston College.




# Translation and Cultural Adaptation of the TIMSS Instruments

Kathleen M. O'Connor
Barbara Malak



# 4 Translation and Cultural Adaptation of the TIMSS Instruments[1]

Kathleen M. O'Connor
Barbara Malak

## 4.1 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). Even though the United States and the Benchmarking participants collected data in English, it was nonetheless necessary to make minor cultural adaptations to reflect U.S. language usage. This chapter describes the extensive procedures for translating and adapting the TIMSS instruments.

Countries that administered the instruments in English followed the same procedures for cultural adaptation and verification as countries that translated them into a different target language. For the TIMSS 1999 main survey, each country had to translate and adapt 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 and adaptation 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). Upon completion, the translated instruments were checked by an international translation company against the TIMSS 1999 international version to assess the faithfulness of

○○○

1. This chapter was based on O'Connor & Malak (2000) from the international technical report for TIMSS 1999.

translation. The NRCs then received suggestions from the translation company and the International Study Center for additional revisions. After these had been made, the final version was checked by the International Study Center at Boston College.

## 4.2 Translation and Adaptation 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 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 (see O'Connor & Malak, 2000 for further details of the international translation process).

Each NRC identified the language or languages to be used in testing and the geographical or political areas involved. Most countries tested in just one language, but nine tested in two languages. The testing languages used in the participating countries are presented in Exhibit 4.1.

The United States administered the assessment in English. As mentioned previously, the U.S.(English) versions of the instruments were used in the Benchmarking study.

**Exhibit 4.1 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 the 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.

### 4.2.1 Guidelines for Translation and Cultural Adaptation of Instruments

Translators were given guidelines to follow in translating the 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 versions in style and meaning, while allowing for cultural adaptations where necessary. Even countries that tested in English, such as Australia, Canada, England, New Zealand, the Philippines, Singapore, South Africa, and the United States, found it necessary to adapt to national usage with regard to spelling and usage. 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
- Ensuring that there was no other possible correct answer to an adapted item
- 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. Thus they could change the names of seasons, people, places, animals, plants, currencies, and the like. Exhibit 4.2 shows a list provided to translators detailing the types of adaptations that were acceptable.

**Exhibit 4.2 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 |

### 4.2.2 Adaptation of the U.S Instruments

The review of the US instruments for cultural adaptation was led by Westat with the work conducted by Educational Testing Services (ETS) under subcontract. No translation was necessary as the international versions of the instruments were in American English, thus the purpose of the review was to identify changes necessary due to cultural context. Westat suggested a number of changes that were to be made in the U.S. versions. These included adding commas in numbers to denote thousands, millions, etc.; spelling out units of measurement; and changing unit terms from the International System of Units (metric units) to U.S. inch-pound units when the measure was not integral to the task.

International procedures required that the International Study Center be notified, and a corresponding statement included in the NRC Survey Activities Report, of any items that proved to be problematic for translators. To identify problematic items, Westat contracted with Educational Testing Service to conduct a sensitivity and fairness review. Reviewers indicated that no items were found to be problematic and that the items were of excellent quality.

### 4.2.3 Recording Deviations from the International Version

After a single translation had been agreed upon, the Translation Deviation Form was used to record all changes in test and questionnaire items. Translators were asked to document 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 in or problems with the translation. This record was used in translation verification and during the item analysis and review.

## 4.3 Verification of Instruments

Each country's translated documents went through a rigorous process that included 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.

### 4.3.1 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, NRCs were asked to check for items that might have been mistranslated. NRCs were also notified of any potentially problematic items and asked to verify that the translation was sound.

### 4.3.2 External Verification

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

- A copy of the test item clusters (A through Z) and the accompanying instructions for students
- A set of test booklets (1 through 8)
- A copy of the School Questionnaire, Student Questionnaire, and Teacher Questionnaires

All 38 countries that participated in the TIMSS 1999 main survey submitted their national versions of instruments for external verification.

### 4.3.3 International Verification

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 to make suggestions for improvements so that NRCs could review and revise their instruments.

Verifiers received general information about the study and instrument design. They also received materials describing the translation procedures used by the national centers along with detailed instructions for reviewing the instruments (TIMSS, 1998b). 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 instrument
- Translation Verification Report Forms

The main task of the translation verifier 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 the United States and other 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

### 4.3.4 Translation Verification Reports

The translation verifier prepared two types of reports. The first was 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 number)
- The severity of the deviation (using the severity code below)
- A description of the change
- A suggested alternative translation

These records were 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)[2] as described below:

○○○

2. 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; 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.

**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 4.3 details the layout issues to be considered and checked for each survey instrument.

**Exhibit 4.3 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 |

For any deviation from the original international version, an entry was made in the Translation Verification Report Form indicating location and severity 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. In the United States, Westat was responsible for reviewing the report forms and reevaluating the instruments based on the translation verifiers' suggestions. Necessary changes were sent by Westat to a subcontractor, National Computer Systems, who produced the assessment materials.

### 4.3.5 International Study Center Item Review

As a final review, when the suggestions of the verifiers had been acted upon, a print-ready copy of the achievement test booklets and questionnaires was submitted to the International Study Center at Boston College. This was 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 like the U.S., 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 and Verification Forms. 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.

#### 4.3.6 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 9 for a description of their work). An important task for the QCMs was to review the translation of the test items. QCMs examined 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.

## 4.4 Summary

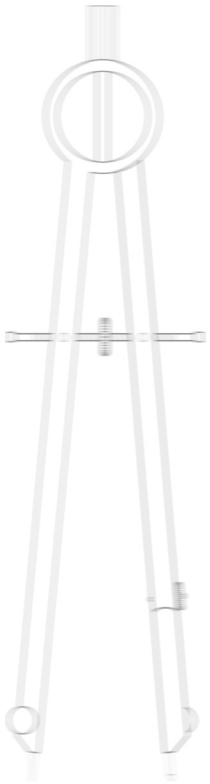
The rigorous procedures for translation, cultural adaptations, translation verification, and review of the instruments put in place for TIMSS 1999 resulted in comparable translations across participating countries. Verification by internal statistical review, external translation verification by bilingual judges, and review by the International Study Center and QCMs proved to be a comprehensive way to check and document deviations and review anomalies, ensuring accuracy in the analysis and reporting of the main survey data.

## References

O'Connor, K.M., & Malak, B. (2000). Translation and cultural adaptation of the TIMSS instruments. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.). *TIMSS 1999 technical report* (pp. 89-100). Chestnut Hill, MA: Boston College.

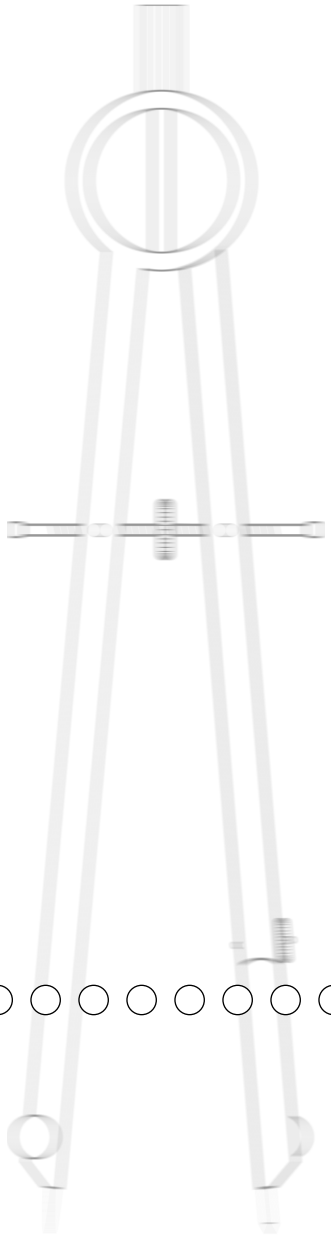
TIMSS (1998a). *Survey operations manual* (TIMSS 1999 Doc. Ref. No. 98-0026). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

TIMSS (1998b). *Guidelines for the translation verification of the TIMSS-R main survey instruments* (TIMSS 1999 Doc. Ref. No. 98-0042). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.




# Sampling Design and Implementation for TIMSS 1999 Countries

Pierre Foy
Marc Joncas



# 5 Sampling Design and Implementation for TIMSS 1999 Countries[1]

Pierre Foy
Marc Joncas

## 5.1 Overview

This chapter describes the TIMSS 1999 sampling procedures as they were implemented in the TIMSS countries. The next chapter describes sampling activities in the Benchmarking jurisdictions. 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 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

○○○

1. This chapter describes the design and implementation of the TIMSS sampling plan for participating countries, and is based mainly on Foy & Joncas (2000a,2000b) and Foy (2000). The following chapter (Chapter 6) provides details of the sampling activities for the benchmarking jurisdictions.

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, Keith Rust, Westat, 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.

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, NRCs provided documentation for all phases of sampling. This documentation was used by the International Study Center 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 of the TIMSS 1999 Technical Report (Martin, Gregory, & Stemler, 2000).

## 5.2 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 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[2] 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.

### 5.2.1 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 TIMSS 1999 International Reports (Martin et al., 2000; Mullis et al., 2000) document any deviations from the international definition of the TIMSS 1999 target population.

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*, was 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.

○○○

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

- 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.

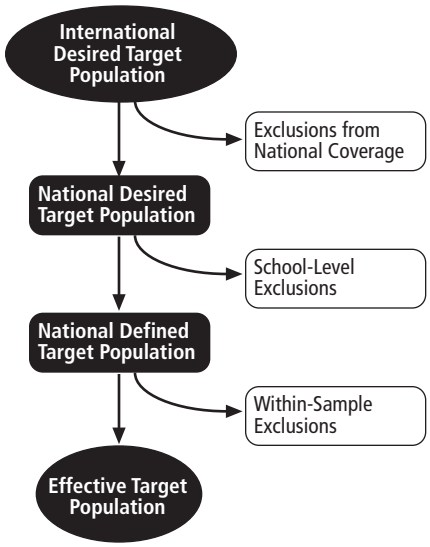
The general TIMSS 1999 rules for defining within-school exclusions included:

- **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 the tasks required for 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 5.1 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 5.1 Relationship Between the Desired Populations and Exclusions**




## 5.3 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[3], 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.

### 5.3.1 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.

○○○

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

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.

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.

### 5.3.2 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: $m \pm 0.1s$ (where m is the mean estimate and s is the estimated standard deviation for students)
- Percentages: $p \pm 5.0\%$ (where p is a percentage estimate)
- Correlations: $r \pm 0.1$ (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 5.2 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 5.2 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 the following section on Small Schools).

**Exhibit 5.2: 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).

### 5.3.3 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.

### 5.3.4 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.

### 5.3.5 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 each 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 5.3). 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 one 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 5.3 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 5.3: Application of the PPS Systematic Sampling Method**

| | | | |
|---|---|---|---|
| Total MOS: | 392154 | Sampling Interval: | 2614 |
| School Sample: | 150 | Random Start: | 1135 |

| 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

### 5.3.6 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. 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.

In TIMSS 1999, small schools were handled differently than in TIMSS 1995. In TIMSS 1999, two options were available for dealing with small schools:

- **Exclusion**. 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 10% criterion.
- **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.

### 5.3.7 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.

### 5.3.8 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 randomly selected classrooms using PPS.

### 5.3.9 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.

### 5.3.10 Required Participation Rates

#### 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 as described in section 5.6.

### 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.

### Overall Participation Rates

The minimum acceptable overall participation 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.

## 5.4 Implementation of the Sample Design

### 5.4.1 Target Population Grades

Exhibit 5.4 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.[4] 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 Cyprus and Finland to 15.5 in South Africa.

○○○

4. 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 their 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.

**Exhibit 5.4 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 |

### Coverage And Exclusions

Exhibit 5.5 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.[5] 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 of the TIMSS 1999 Technical Report (Martin, Gregory, & Stemler, 2000). 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 TIMSS 1999 International Reports (Martin et al., 2000; Mullis et al., 2000). A few countries had no within-school exclusions.

○○○

5. 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 5.5 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% |

### 5.4.2 Sampling of Schools and Students

#### Target Population Sizes

Exhibit 5.6 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 section 5.5), TIMSS derived an estimate of the student population size, which matched closely the student population size from the sampling frame (see Exhibit 5.6).

**Exhibit 5.6 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 |

## 5.5 Sampling Weights

The multi-stage nature of the TIMSS sampling design meant that students were sampled with varying probabilities. 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. The procedures for calculating sampling weights are described fully in Foy (2000).

### 5.5.1 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). 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.

### 5.5.2 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[6] were removed from the calculation of this adjustment. The school-level participation adjustment was calculated separately for each explicit stratum.

○○○

6. 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.

The adjustment was calculated as follows:

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

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 $i^{th}$ school, corrected for non-participating schools, thus became:

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

### 5.5.3 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 i$^{th}$ 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 i$^{th}$ 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 $i^{th}$ school, let $k^{i,j}$ be the size of the $j^{th}$ classroom. Using PPS sampling, the final second stage weight assigned to the $j^{th}$ sampled classroom in the $i^{th}$ 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 $i^{th}$ school, as defined earlier, and

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

Again, usually $c^i$ took the values one or two 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 two.

### 5.5.4 The Third Stage (Student) Weight

The third stage weight represented the inverse of the third stage selection probability attached to a sampled student.

**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 eight achievement test booklets[7], 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

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

○○○

7. See chapter 2 for a description of the TIMSS test design.

where

$k^{i,j}_{TIMSS\ 1999}$ = number of students assigned a TIMSS 1999 booklet in the $j^{th}$ classroom of the $i^{th}$ school,

$k^{i,j}_{natl}$ = number of students assigned a national booklet in the $j^{th}$ classroom of the $i^{th}$ school, and

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

where $k^{i,j}_{ex}$ was the number of excluded students[8] that were not assigned any booklet. Note that this number could be zero if there were no excluded students in the classroom.

### 5.5.5 Adjustment for Student Non-Participation

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

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

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

$$A^{i,j}_{st} = \frac{1}{R^{i,j}_{st}}$$

The third and final stage weight for the $j^{th}$ classroom in the $i^{th}$ school thus became

$$FW^{i,j}_{st1} = A^{i,j}_{st} \cdot BW^{i,j}_{st1}$$

when intact classrooms were sampled.

○○○

8. Criteria for excluding students from the data collection are presented in chapter 2.

### 5.5.6 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

$$W^{i,j} = A^{i,j}_{sc} \cdot BW^{i}_{sc} \cdot FW^{i,j}_{cl1} \cdot A^{i,j}_{st} \cdot BW^{i,j}_{st1}$$

or

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

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

$$W^{i,j} = A^{i,j}_{sc} \cdot BW^{i}_{sc} \cdot FW^{i,j}_{cl2} \cdot A^{i,j}_{st} \cdot BW^{i,j}_{st2}$$

or

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

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.

## 5.6 Calculating 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: (1) using originally sampled schools only; (2) using sampled and first replacement schools; and (3) using sampled and both first and second replacement schools. Student participation rates were also computed, as were overall participation rates.

### 5.6.1 Unweighted School Participation Rates

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

$R^{sc-s}_{unw}$ = 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:

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

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

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

### 5.6.2 Unweighted Student Participation Rate

The unweighted student participation rate was computed as follows:

$$R_{unw}^{st} = \frac{\sum_{i,j} s_{rs}^{i,j}}{\sum_{i,j} s_{rs}^{i,j} + \sum_{i,j} s_{nr}^{i,j}}$$

### 5.6.3 Unweighted Overall Participation Rates

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

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

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

$R_{unw}^{ov-r2}$ = 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:

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

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

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

### 5.6.4 Weighted School Participation Rates

In TIMSS 1995, the weighted school-level participation rates were computed using school sampling frame information. 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 TIMSS 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:

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

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

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

The weighted school participation rates were calculated as follows:

$$R^{sc-s}_{wtd} = \frac{\sum_{i,j}^{s} BW^{i}_{sc} \cdot FW^{i,j}_{clx} \cdot FW^{i,j}_{stx}}{\sum_{i,j}^{s+r1+r2} FW^{i}_{sc} \cdot FW^{i,j}_{clx} \cdot FW^{i,j}_{stx}}$$

$$R^{sc-r1}_{wtd} = \frac{\sum_{i,j}^{s+r1} BW^{i}_{sc} \cdot FW^{i,j}_{clx} \cdot FW^{i,j}_{stx}}{\sum_{i,j}^{s+r1+r2} FW^{i}_{sc} \cdot FW^{i,j}_{clx} \cdot FW^{i,j}_{stx}}$$

$$R^{sc-r2}_{wtd} = \frac{\sum_{i,j}^{s+r1+r2} BW^{i}_{sc} \cdot FW^{i,j}_{clx} \cdot FW^{i,j}_{stx}}{\sum_{i,j}^{s+r1+r2} FW^{i}_{sc} \cdot FW^{i,j}_{clx} \cdot FW^{i,j}_{stx}}$$

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.

### 5.6.5 Weighted Student Participation Rates

The weighted student response rate was computed as follows:

$$R^{st}_{wtd} = \frac{\sum_{i,j}^{s+r1+r2} BW^{i}_{sc} \cdot FW^{i,j}_{clx} \cdot BW^{i,j}_{stx}}{\sum_{i,j}^{s+r1+r2} BW^{i}_{sc} \cdot FW^{i,j}_{clx} \cdot FW^{i,j}_{stx}}$$

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.

### 5.6.6 Weighted Overall Participation Rates

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

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

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

$R^{ov-r2}_{wtd}$ = 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:

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

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

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

## 5.7 Final Participation Rates

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 5.7). Countries in category 1 were considered to have met the TIMSS sampling requirements and to have an acceptable partici-

pation 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 5.7 Categories of Sampling Participation**

| | |
|---|---|
| Category 1 | Acceptable sampling participation rate **without** the use of replacement schools.<br>In order to be placed in this category, a country had to have:<br>• 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%<br>OR<br>• 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%<br>OR<br>• 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).<br>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 after replacement schools were included**. A country was placed in category 2 if:<br>• 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)<br>AND HAD EITHER<br>• 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%<br>OR<br>• 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%<br>OR<br>• 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).<br>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.<br>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 5.8 through 5.11 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 in the achievement exhibits in the TIMSS 1999 International Reports. In TIMSS 1999, no country was assigned to category 3.

**Exhibit 5.8 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 5.9 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 5.10 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 5.11 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% |

## 5.8 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 of the TIMSS 1999 Technical Report (Martin, Gregory, & Stemler, 2000).

## References

Foy, P. (2000). Sampling weights. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS 1999 technical report* (pp. 189-202). Chestnut Hill, MA: Boston College.

Foy, P., & Joncas, M. (2000a). TIMSS sample design. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS 1999 technical report* (pp. 29-48). Chestnut Hill, MA: Boston College.

Foy, P., & Joncas, M. (2000b). Implementation of the sample design. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS 1999 technical report* (pp. 157-170). Chestnut Hill, MA: Boston College.

Foy, P., Rust, K., & Schleicher, A. (1996). Sample design. In M.O. Martin & D.L. Kelly (Eds.), *Third international mathematics and science study technical report volume I: Design and development* (pp. 4.1 - 4.17). Chestnut Hill, MA: Boston College.

Martin, M. O., Gregory, K. D. & Stemler, S. E. (Eds.). (2000). *TIMSS 1999 technical report.* Chestnut Hill, MA: Boston College.

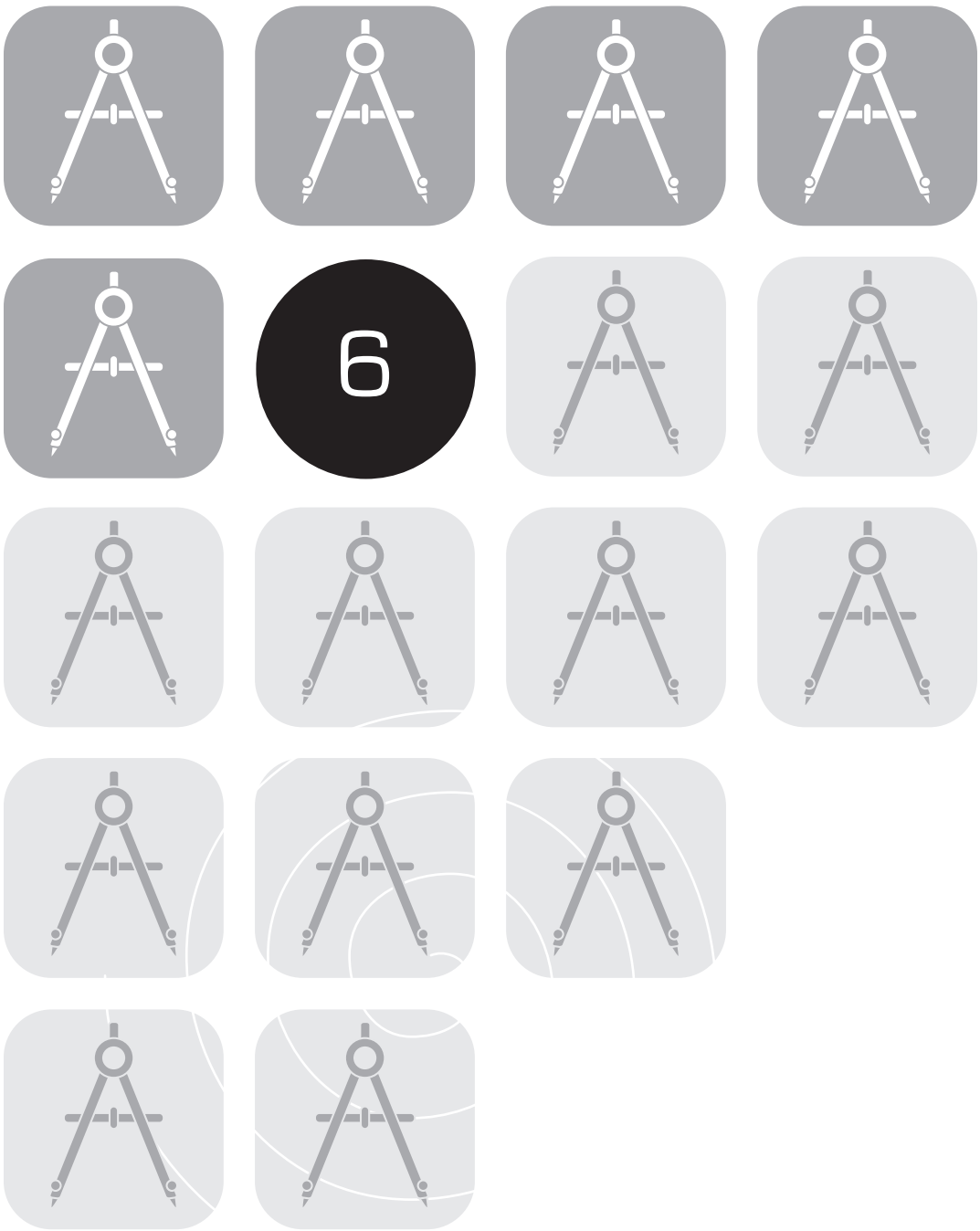
Martin, M.O., Mullis, I.V.S., Gonzalez, E.G., Gregory, K.D., Smith, T.A, Chrostowski, S.J., Garden, R.A., & O'Connor, K.M. (2000). *TIMSS 1999 international science report.* Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.G., 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.

TIMSS (1997). *TIMSS 1999 school sampling manual–version 2* (TIMSS 1999 Doc. Ref. No. 97-0012). Prepared by Pierre Foy, Statistics Canada. Chestnut Hill, MA: Boston College.

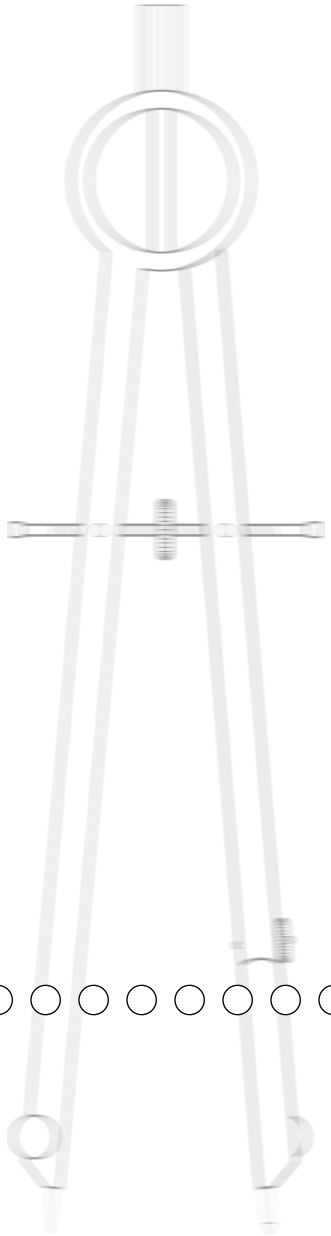
TIMSS (1998a). *Survey operations manual* (TIMSS 1999 Doc. Ref. No. 98-0026). Prepared by the International Study Center. Chestnut Hill, MA: Boston College.

TIMSS (1998b). *School coordinator manual* (TIMSS 1999 Doc. Ref. No. 98-0024). Prepared by the International Study Center. Chestnut Hill, MA: Boston College.




# Sampling Design and Implementation for TIMSS 1999 Benchmarking

Jean Fowler
Lou Rizzo
Keith Rust



# 6 Sampling Design and Implementation for TIMSS 1999 Benchmarking

Jean Fowler
Lou Rizzo
Keith Rust

## 6.1 Overview

The previous chapter described the design and implementation of the TIMSS samples for the participating countries, including the United States. This chapter describes the sampling procedures for the 27 Benchmarking participants.

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.

## 6.2 School Sample

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 nonreligious private schools were "private". Strata were also defined to take into account selection of the TIMSS 1999 national sample primary sampling units (PSUs). A PSU is a consolidated metropolitan statistical area, a metropolitan statistical area, a county, or a group of contiguous counties. Benchmarking PSUs were grouped according to whether or not they had been selected for the TIMSS 1999 national sample, thus defining "overlap" and "nonoverlap" strata.

## 6.3 Target School Sample Size

The initial public school target sample size was 50 for states, 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. According to this scheme the sampling strata fell into three classes:

- Overlap strata where the TIMSS 1999 international sample met or exceeded the Benchmarking target stratum sample size. No additional schools were selected from these strata for the Benchmarking sample.
- Overlap strata where the TIMSS 1999 international sample was smaller than the Benchmarking target stratum sample size. A supplementary sample was drawn so that the final stratum sample size would meet the Benchmarking target.
- Nonoverlap strata. A sample was drawn, with target sample size equal to the Benchmarking target.

## 6.4 Selecting Schools

Within each stratum, the school frame was ordered according to eighth grade enrollment. Using a random start and an interval determined by total enrollment and desired sample size, schools were systematically selected. Thus a school's probability of selection was proportional to its share of the target population, that is, the eighth grade enrollment. All schools were selected with certainty in districts and consortia having 25 or fewer members. Final sample sizes ranged from 4 to 71 schools.

Since TIMSS 1999 national sample schools were not removed from the frame, the possibility existed in the overlap strata that some of these schools would be selected into the supplementary sample. Expected overlap was calculated for each sampling frame. For all jurisdictions but Miami Dade County this was less than two schools. Based on an expected overlap of about four schools, the Miami Dade County supplementary sample target size was set to 19. Four of the ten Miami Dade County TIMSS 1999 national sample schools were in fact selected, resulting in a final Benchmarking sample size equal to the target of 25 schools. Two TIMSS 1999 national sample schools were selected into the Massachusetts supplementary sample, reducing the final Benchmarking supplementary sample size from the target of 61 schools to 59. Otherwise, the TIMSS 1999 national and supplementary samples did not overlap.

States were offered the option of sampling private schools, with target sample sizes proportional to the private share of total eighth grade enrollment. Idaho, Indiana, Michigan, and Pennsylvania chose to sample private schools. Consortia might include private schools, but there was no provision to sample these schools independently. The exception to this scheme was the SW Pennsylvania Regional Math & Science Collaborative,

with a sample size of 50, split in proportion to enrollment and sampled independently: 44 public schools and 6 private. Private schools sampled in TIMSS 1999 Benchmarking were included in the final samples for these jurisdictions in the same manner as TIMSS 1999 public schools, described above.

## 6.5 Substitute Schools

When possible, two substitutes were identified for each Benchmarking sample school. The general rule was to assign as substitutes the two schools neighboring the sampled school on the frame, with the preceding school in the frame order as the first substitute, and the succeeding school as the second. The other conditions were that a TIMSS 1999 national sample school could not serve as a Benchmarking substitute, and that a substitute had to be in the same sampling stratum as the school to which it was assigned.

Exhibit 6.1 summarizes the Benchmarking school samples. Final sample sizes are shown for each jurisdiction, including the numbers of TIMSS 1999 original selections and substitutes. Counts are also broken down by sampling stratum, which are identified according to overlap status. This table reflects the sampling procedure described above by which states and the districts and consortia within them were sampled independently. Final state samples incorporated the district and consortium samples. The Illinois sample included Chicago Public Schools, First in the World Consortium, and Naperville Community Unit School District #203; the Maryland sample included Montgomery County Public Schools; the North Carolina sample included Guilford County Public Schools; the Pennsylvania sample included Southwest Pennsylvania Regional Math & Science Collaborative.

## Exhibit 6.1 TIMSS 1999 Benchmarking School Sample Summary

| State | Sample or Census | Jurisdiction | Number of Schools in TIMSS 1999 Benchmarking Sample | Stratum | N | Schools in National Sample | | Type / Entity | Sampling Stratum |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Orig | Sub | | |
| CO | Census | Academy | 4 | | 4 | | | District | Rem[1] |
| CT | Sample | | 54 | PU3 | 12 | 4 | 1 | State | Ovp |
| | | | | PU4 | 42 | | | State | Rem |
| DE | Census | DE Sci Coal | 25 | | 25 | | | Consortium | Ovp/Rem |
| FL | Sample | Dade Co | 25 | | 25 | 4 | 3 | District | Ovp |
| ID | Sample | | 54 | PR1 | 0 | | | State | Ovp |
| | | | | PR2 | 2 | | | State | Rem |
| | | | | PU3 | 2 | 2 | | State | Ovp |
| | | | | PU4 | 50 | | | State | Rem |
| IL | Sample | | 41 | PU1 | 21 | 4 | 1 | State | Ovp |
| | | | | PU2 | 3 | | | State | Ovp |
| | | | | PU3 | 17 | | | State | Rem |
| IL | Sample | Chicago PS | 27 | | 27 | 2 | 1 | District | Ovp |
| IL | Census | 1st in World | 17 | | 17 | | 1 | Consortium | Ovp |
| IL | Census | Naperville | 5 | | 5 | | | District | Ovp |
| IN | Sample | | 61 | PR1 | 2 | 1 | | State | Ovp |
| | | | | PR2 | 5 | | | State | Rem |
| | | | | PU3 | 6 | | | State | Ovp |
| | | | | PU4 | 0 | | | State | Ovp |
| | | | | PU5 | 13 | 4 | 1 | State | Ovp |
| | | | | PU6 | 35 | | | State | Rem |
| MD | Sample | Mont Co | 25 | PU3 | 25 | | 1 | District | Ovp |
| MD | Sample | | 54 | PU4 | 17 | 1 | 1 | State | Ovp |
| | | | | PU5 | 30 | 3 | 2 | State | Ovp |
| | | | | PU6 | 7 | | | State | Rem |
| MA | Sample | | 59 | PU3 | 2 | | | State | Rem |
| | | | | PU4 | 35 | 2 | 1 | State | Ovp |
| | | | | PU5 | 8 | 3 | 1 | State | Ovp |
| | | | | PU6 | 5 | 4 | | State | Ovp |
| | | | | PU7 | 9 | | | State | Rem |

1 "Ovp" means that some of the benchmark sample schools from this stratum were also in the national sample. "Rem" means that none of the benchmark sample schools from this stratum were part of the national sample..

**Exhibit 6.1 (continued) TIMSS 1999 Benchmarking School Sample Summary**

| State | Sample or Census | Jurisdiction | Number of schools from TIMSS 1999 National Sample | Stratum | N | Schools in National Sample | | Type Entity | Type Sampling Stratum |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Orig | Sub | | |
| MI | Census | Invit Group | 21 | | 21 | | | Consortium | n/a |
| MI | Sample | | 66 | PR1 | 6 | 3 | | State | Ovp |
| | | | | PR2 | 3 | | | State | Rem |
| | | | | PU3 | 26 | 3 | | State | Ovp |
| | | | | PU4 | 4 | 4 | | State | Ovp |
| | | | | PU5 | 27 | | | State | Rem |
| MO | Sample | | 57 | PU1 | 3 | 3 | | State | Ovp |
| | | | | PU2 | 18 | 4 | 2 | State | Ovp |
| | | | | PU3 | 36 | | | State | Rem |
| NC | Census | Guilford Co | 17 | PU3 | 17 | | | District | Rem |
| NC | Sample | | 54 | PU4 | 4 | 4 | | State | Ovp |
| | | | | PU5 | 50 | | | State | Rem |
| NE | Census | Lincoln/ Fremont/ WestSide PS | 12 | | 12 | | | Consortium | Rem |
| NJ | Census | Jrsy City PS | 25 | | 25 | 1 | | District | Ovp |
| NY | Census | Rochester PS | 7 | | 7 | | | District | Rem |
| OH | Census | Prj SMART | 24 | | 24 | 1 | 1 | Consortium | Ovp/Rem |
| OR | Sample | | 51 | PU3 | 1 | 1 | | State | Ovp |
| | | | | PU4 | 50 | | | State | Rem |
| PA | Sample | | 66 | PR2 | 6 | 2 | | State | Ovp |
| | | | | PR3 | 7 | | | State | Rem |
| | | | | PU5 | 19 | 3 | 1 | State | Ovp |
| | | | | PU6 | 34 | | | State | Rem |
| PA | Sample | SW PA Sci& Math Coll | 50 | PR1 | 6 | | | Consortium | Rem |
| | | | | PU4 | 44 | | | Consortium | Rem |
| SC | Sample | | 53 | PU3 | 3 | 3 | | State | Ovp |
| | | | | PU4 | 50 | | | | Rem |
| TX | Sample | | 71 | PU3 | 28 | 9 | 2 | State | Ovp |
| | | | | PU4 | 7 | 7 | | State | Ovp |
| | | | | PU5 | 5 | 5 | | State | Ovp |
| | | | | PU6 | 31 | | | State | Rem |

## 6.6 School Participation Rates

School participation rates are shown for all schools and by school type in Exhibits 6.2 and 6.3. Four states used replacement schools; this choice considerably improved school participation rates in two of them: Indiana and Missouri. Five jurisdictions sampled private schools, with unweighted participation rates ranging from 50 to 100 percent. Only in Indiana were public and private school participation rates about the same.

The three unweighted school participation rates were computed as in section 5.6.1. The weighted school participation rates shown in Exhibit 6.2 and 6.3 were calculated as follows:

$$R^{sc-s}_{wtd} = \frac{\sum_{i,j}^{s} BW^{i}_{sc} \cdot MOS_i}{\sum_{i,j}^{s+r1+r2} BW^{i}_{sc} \cdot MOS_i}$$

$$R^{sc-r1}_{wtd} = \frac{\sum_{i,j}^{s+r1} BW^{i}_{sc} \cdot MOS_i}{\sum_{i,j}^{s+r1+r2} BW^{i}_{sc} \cdot MOS_i}$$

$$R^{sc-r2}_{wtd} = \frac{\sum_{i,j}^{s+r1+r2} BW^{i}_{sc} \cdot MOS_i}{\sum_{i,j}^{s+r1+r2} BW^{i}_{sc} \cdot MOS_i}$$

where $BW^{i}_{sc}$ is the basic school weight defined in Section 5.5.1 and represents the inverse of the first stage selection probability assigned to a sample school. $MOS_i$ is the estimated eighth enrollment of the sampled school.

### 6.6.1 Alternate Method for Weighted School Participation Rates

Three weighted school-level participation rates were computed using the alternate method with similar results. This method is described in section 5.6.4 and is identical to the method used in the TIMSS 1999 International Reports. These rates are shown in Exhibits 6.4 and 6.5.

## Exhibit 6.2 TIMSS 1999 Benchmarking School Participation Rates

| Jurisdiction | Number of Schools | | | | | | Unweighted Participation Rate | | Weighted Participation Rate | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Selected | Ineligible | Refusing | Participating | | | Substitutes Not Included | Substitutes Included | Substitutes Not Included | Substitutes Included |
| | | | | Originals | Substitutes | Total | | | | |
| Connecticut | 54 | 0 | 2 | 52 | 0 | 52 | 96.30 | 96.30 | 95.99 | 95.99 |
| Idaho | 54 | 0 | 7 | 47 | 0 | 47 | 87.04 | 87.04 | 87.16 | 87.16 |
| Illinois | 90 | 0 | 5 | 85 | 0 | 85 | 94.44 | 94.44 | 95.48 | 95.48 |
| Indiana | 61 | 0 | 22 | 39 | 13 | 52 | 63.93 | 85.25 | 62.42 | 83.01 |
| Maryland | 79 | 2 | 4 | 73 | 0 | 73 | 94.81 | 94.81 | 93.54 | 93.54 |
| Massachusetts | 59 | 1 | 1 | 57 | 0 | 57 | 98.28 | 98.28 | 98.22 | 98.22 |
| Michigan | 66 | 4 | 7 | 55 | 2 | 57 | 88.71 | 91.94 | 88.67 | 91.93 |
| Missouri | 57 | 2 | 12 | 43 | 8 | 51 | 78.18 | 92.73 | 78.73 | 93.39 |
| North Carolina | 71 | 3 | 1 | 67 | 0 | 67 | 98.53 | 98.53 | 98.01 | 98.01 |
| Oregon | 51 | 0 | 6 | 45 | 0 | 45 | 88.24 | 88.24 | 88.93 | 88.93 |
| Pennsylvania | 116 | 3 | 33 | 80 | 0 | 80 | 70.80 | 70.80 | 66.12 | 66.12 |
| South Carolina | 53 | 0 | 4 | 49 | 0 | 49 | 92.45 | 92.45 | 92.25 | 92.25 |
| Texas | 71 | 1 | 19 | 51 | 1 | 52 | 72.86 | 74.29 | 72.39 | 73.94 |
| Academy #20, CO | 4 | 0 | 0 | 4 | 0 | 4 | 100.00 | 100.00 | 100.00 | 100.00 |
| Delaware Math & Sci., DE | 25 | 0 | 0 | 25 | 0 | 25 | 100.00 | 100.00 | 100.00 | 100.00 |
| Dade County, FL | 25 | 0 | 0 | 25 | 0 | 25 | 100.00 | 100.00 | 100.00 | 100.00 |
| Chicago Public Schools, IL | 27 | 0 | 1 | 26 | 0 | 26 | 96.30 | 96.30 | 96.30 | 96.30 |
| First in the World, IL | 17 | 0 | 2 | 15 | 0 | 15 | 88.24 | 88.24 | 93.64 | 93.64 |
| Naperville #203, IL | 5 | 0 | 0 | 5 | 0 | 5 | 100.00 | 100.00 | 100.00 | 100.00 |
| Montgomery County, MD | 25 | 0 | 0 | 25 | 0 | 25 | 100.00 | 100.00 | 100.00 | 100.00 |
| Invitational Group, MI | 21 | 0 | 0 | 21 | 0 | 21 | 100.00 | 100.00 | 100.00 | 100.00 |
| Fremont/Lincoln/ WestSide P.S., NE | 12 | 0 | 0 | 12 | 0 | 12 | 100.00 | 100.00 | 100.00 | 100.00 |
| Jersey City Public Schools, NJ | 25 | 0 | 1 | 24 | 0 | 24 | 96.00 | 96.00 | 96.57 | 96.57 |
| Rochester City Sch. Dist., NY | 7 | 0 | 0 | 7 | 0 | 7 | 100.00 | 100.00 | 100.00 | 100.00 |
| Guilford County, NC | 17 | 0 | 0 | 17 | 0 | 17 | 100.00 | 100.00 | 100.00 | 100.00 |
| Project SMART, OH | 24 | 0 | 0 | 24 | 0 | 24 | 100.00 | 100.00 | 100.00 | 100.00 |
| SW PA Math & Sci. Collaborative, PA | 50 | 1 | 10 | 39 | 0 | 39 | 79.59 | 79.59 | 79.43 | 79.43 |
| TOTAL SCHOOLS | 1025 | 16 | 124 | 885 | 24 | 909 | | | | |

**Exhibit 6.3 TIMSS 1999 Benchmarking Participation Rates by School Type**

| Jurisdiction | School Type | Number of Schools | | | | | | Unweighted Participation Rates | | Weighted Participation Rates | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Selected | Ineligible | Refusing | Participating | | | Substitutes Not Included | Substitutes Included | Substitutes Not Included | Substitutes Included |
| | | | | | Originals | Substitutes | Total | | | | |
| Idaho | Private | 2 | 0 | 1 | 1 | 0 | 1 | 50.00 | 50.00 | 50.00 | 50.00 |
| | Public | 52 | 0 | 6 | 46 | 0 | 46 | 88.46 | 88.46 | 88.46 | 88.46 |
| Indiana | Private | 7 | 0 | 1 | 6 | 0 | 6 | 85.71 | 85.71 | 74.72 | 74.72 |
| | Public | 54 | 0 | 21 | 33 | 13 | 46 | 61.11 | 85.19 | 60.94 | 84.01 |
| Michigan | Private | 9 | 1 | 0 | 8 | 0 | 8 | 100.00 | 100.00 | 100.00 | 100.00 |
| | Public | 57 | 3 | 7 | 47 | 2 | 49 | 87.04 | 90.74 | 87.13 | 90.83 |
| Pennsylvania | Private | 19 | 1 | 9 | 9 | 0 | 9 | 50.00 | 50.00 | 35.02 | 35.02 |
| | Public | 97 | 2 | 24 | 71 | 0 | 71 | 74.74 | 74.74 | 73.25 | 73.25 |
| SW PA Math & Sci. Collaborative, PA | Private | 6 | 0 | 0 | 6 | 0 | 6 | 100.00 | 100.00 | 100.00 | 100.00 |
| | Public | 44 | 1 | 10 | 33 | 0 | 33 | 76.74 | 76.74 | 76.74 | 76.74 |
| TOTAL SCHOOLS | Private | 56 | 3 | 20 | 33 | 0 | 33 | | | | |
| | Public | 969 | 13 | 104 | 852 | 24 | 876 | | | | |

**Exhibit 6.4 TIMSS 1999 Benchmarking Weighted School Participation Rates: Alternate Method**

| Jurisdiction | Substitutes Not Included | Substitutes Included |
|---|---|---|
| Connecticut | 96% | 96% |
| Idaho | 88% | 88% |
| Illinois | 95% | 95% |
| Indiana | 61% | 83% |
| Maryland | 94% | 94% |
| Massachusetts | 98% | 98% |
| Michigan | 89% | 92% |
| Missouri | 79% | 94% |
| NC, combined | 98% | 98% |
| Oregon | 89% | 89% |
| PA, combined | 66% | 66% |
| South Carolina | 92% | 92% |
| Texas | 73% | 74% |
| Academy #20, CO | 100% | 100% |
| Delaware Math & Sci., DE | 100% | 100% |
| Dade County, FL | 100% | 100% |
| Chicago Public Schools, IL | 95% | 95% |
| First in the World, IL | 93% | 93% |
| Naperville #203, IL | 100% | 100% |
| Montgomery County, MD | 100% | 100% |
| Invitational Group, MI | 100% | 100% |
| Fremont/Lincoln/ WestSide P.S., NE | 100% | 100% |
| Jersey City Public Schools, NJ | 97% | 97% |
| Rochester City Sch. Dist., NY | 100% | 100% |
| Guilford County, NC | 100% | 100% |
| Project SMART, OH | 100% | 100% |
| SW PA Math & Sci. Collaborative, PA | 78% | 78% |

**Exhibit 6.5 TIMSS 1999 Benchmarking Weighted School Participation Rates by School Type: Alternate Method**

| Jurisdiction | School Type | Substitutes Not Included | Substitutes Included |
|---|---|---|---|
| Idaho | Private | 50% | 50% |
| | Public | 89% | 89% |
| Indiana | Private | 75% | 75% |
| | Public | 59% | 84% |
| Michigan | Private | 100% | 100% |
| | Public | 87% | 91% |
| PA, combined | Private | 36% | 36% |
| | Public | 72% | 72% |
| SW PA Math & Sci. Collaborative, PA | Private | 100% | 100% |
| | Public | 76% | 76% |

## 6.7 Selecting Classes

Classes were randomly selected within schools. All eighth grade mathematics classes were listed in order of increasing difficulty, with a provision for grouping classes having nine or fewer students into "pseudo classes" of up to 20 students. Using a random start and an interval determined by the desired class sample size and the total number of classes on the list, classes were systematically selected for assessment. When the school sample size was 25 or greater, the number of classes sampled was two. For smaller school samples, the classroom sample was allocated among the schools in proportion to enrollment, so that the number of students assessed would be approximately 1000. In Academy School District 20, Colorado, with an estimated eighth grade enrollment of 1318, all classes were selected with certainty for assessment.

## 6.8 Student Sample

The student sample consisted of all eligible students within the selected classes. The exception to this plan was Montgomery County, Maryland, where students were sampled, not classes. Using a random start, 60 students were systematically selected in each school from a list of eighth grade math students. The selected students were randomly assigned to two groups, which were treated as classes for weighting.

Exhibit 6.6 shows the number of students sampled by jurisdiction and school type.

**Exhibit 6.6 TIMSS 1999 Benchmarking Student Sample Size by Jurisdiction and School Type**

| Jurisdiction | School Type | Student Population | Estimated Student Population | Number of Sampled Schools | Number of Sampled Students |
|---|---|---|---|---|---|
| Connecticut | Public | 36775 | 38742 | 54 | 2190 |
| Idaho | Private | 747 | 729 | 2 | 26 |
| | Public | 19430 | 18185 | 52 | 1942 |
| | All | 20177 | 18914 | 54 | 1968 |
| Illinois | Public | 144323 | 147621 | 90 | 5144 |
| Indiana | Private | 8684 | 10934 | 7 | 135 |
| | Public | 76504 | 66650 | 54 | 2040 |
| | All | 85188 | 77584 | 61 | 2175 |
| Maryland | Public | 60756 | 59789 | 79 | 3877 |
| Massachusetts | Public | 65981 | 67531 | 59 | 2538 |
| Michigan | Private | 16375 | 15974 | 9 | 238 |
| | Public | 121972 | 124773 | 57 | 2573 |
| | All | 138347 | 140747 | 66 | 2811 |
| Missouri | Public | 67278 | 65074 | 57 | 2147 |
| North Carolina | Public | 92684 | 84685 | 71 | 3502 |
| Oregon | Public | 41762 | 40847 | 51 | 2044 |
| Pennsylvania | Private | 31014 | 23915 | 19 | 282 |
| | Public | 132795 | 130658 | 97 | 3181 |
| | All | 163809 | 154573 | 116 | 3463 |
| South Carolina | Public | 51632 | 50165 | 53 | 2177 |
| Texas | Public | 284146 | 283538 | 71 | 2189 |
| Academy #20, CO | Public | 1588 | 1318 | 4 | 1329 |
| Delaware Math & Sci., DE | Public | 6753 | 7861 | 25 | 1389 |
| Dade County, FL | Public | 24485 | 22040 | 25 | 1356 |
| Chicago Public Schools, IL | Public | 33355 | 26118 | 27 | 1227 |
| First in the World, IL | Public | 2533 | 2611 | 17 | 782 |
| Naperville #203, IL | Public | 1430 | 1472 | 5 | 1343 |
| Montgomery County, MD | Public | 8787 | 9432 | 25 | 1481 |
| Invitational Group, MI | Public | 3156 | 3039 | 21 | 994 |
| Fremont/Lincoln/ West Side P.S., NE | Public | 3105 | 3044 | 12 | 1178 |
| Jersey City Public Schools, NJ | Public | 2365 | 1749 | 25 | 1116 |
| Rochester City Sch. Dist., NY | Public | 2669 | 2001 | 7 | 1165 |

**Exhibit 6.6 (continued) TIMSS 1999 Benchmarking Student Sample Size by Jurisdiction and School Type**

| Jurisdiction | School Type | Student Population | Estimated Student Population | Number of Sampled Schools | Number of Sampled Students |
|---|---|---|---|---|---|
| Guilford County, NC | Public | 4396 | 5155 | 17 | 1215 |
| Project SMART, OH | Public | 5940 | 5956 | 24 | 1188 |
| SW PA Math & Sci Collaborative, PA | Private | 3661 | 3181 | 6 | 166 |
| | Public | 28648 | 26895 | 44 | 1472 |
| | All | 32309 | 30076 | 50 | 1638 |
| TOTAL | All | 1764489 | 1723486 | 1025 | 45940 |

## 6.9 Student Participation Rates

Student participation rates were calculated as shown in sections 5.6. Exhibits 6.7 and 6.8 show the weighted and unweighted student participation rates overall and by school type.

## Exhibit 6.7 TIMSS 1999 Benchmarking Student Participation Rates

| Jurisdiction | Number of Students | | | | | | Participation Rates | |
|---|---|---|---|---|---|---|---|---|
| | Population | Est. Population | Sampled | Excluded | Absent | Participating | Unweighted | Weighted |
| Connecticut | 36775 | 38742 | 2190 | 43 | 124 | 2023 | 94% | 94% |
| Idaho | 20177 | 18914 | 1968 | 27 | 94 | 1847 | 95% | 95% |
| Illinois | 144323 | 147621 | 5144 | 136 | 227 | 4781 | 95% | 96% |
| Indiana | 85188 | 77584 | 2175 | 27 | 102 | 2046 | 95% | 95% |
| Maryland | 60756 | 59789 | 3877 | 339 | 221 | 3317 | 94% | 94% |
| Massachusetts | 65981 | 67531 | 2538 | 54 | 131 | 2353 | 95% | 95% |
| Michigan | 138347 | 140747 | 2811 | 45 | 143 | 2623 | 95% | 96% |
| Missouri | 67278 | 65074 | 2147 | 40 | 128 | 1979 | 94% | 94% |
| North Carolina | 92684 | 84685 | 3502 | 191 | 214 | 3097 | 94% | 94% |
| Oregon | 41762 | 40847 | 2044 | 29 | 126 | 1889 | 94% | 93% |
| Pennsylvania | 163809 | 154573 | 3463 | 60 | 167 | 3236 | 95% | 95% |
| South Carolina | 51632 | 50165 | 2177 | 36 | 130 | 2011 | 94% | 94% |
| Texas | 284146 | 283538 | 2189 | 44 | 149 | 1996 | 93% | 93% |
| Academy #20, CO | 1588 | 1318 | 1329 | 15 | 81 | 1233 | 94% | 94% |
| Delaware Math & Sci., DE | 6753 | 7861 | 1389 | 18 | 103 | 1268 | 92% | 92% |
| Dade County, FL | 24485 | 22040 | 1356 | 10 | 117 | 1229 | 91% | 91% |
| Chicago Public Schools, IL | 33355 | 26118 | 1227 | 21 | 74 | 1132 | 94% | 94% |
| First in the World, IL | 2533 | 2611 | 782 | 2 | 30 | 750 | 96% | 96% |
| Naperville #203, IL | 1430 | 1472 | 1343 | 84 | 47 | 1212 | 96% | 96% |
| Montgomery County, MD | 8785 | 9432 | 1481 | 254 | 72 | 1155 | 94% | 94% |
| Invitational Group, MI | 3156 | 3039 | 994 | 11 | 80 | 903 | 92% | 91% |
| Fremont/Lincoln/ WestSide P.S., NE | 3105 | 3044 | 1178 | 25 | 60 | 1093 | 95% | 95% |
| Jersey City Public Schools, NJ | 2365 | 1749 | 1116 | 47 | 65 | 1004 | 94% | 94% |
| Rochester City Sch. Dist., NY | 2669 | 2001 | 1165 | 9 | 190 | 966 | 84% | 84% |
| Guilford County, NC | 4396 | 5155 | 1215 | 121 | 76 | 1018 | 93% | 92% |
| Project SMART, OH | 5940 | 5956 | 1188 | 18 | 74 | 1096 | 94% | 94% |
| SW PA Math & Sci. Collaborative, PA | 32309 | 30076 | 1638 | 21 | 79 | 1538 | 95% | 95% |
| TOTAL STUDENTS | 1764489 | 1723486 | 45940 | 1224 | 2726 | 41990 | | |

**Exhibit 6.8 TIMSS 1999 Benchmarking Student Participation Rates by School Type**

| Jurisdiction | School Type | Number of Students | | | | | | Participation Rates | |
|---|---|---|---|---|---|---|---|---|---|
| | | Population | Est. Population | Sampled | Excluded | Absent | Participating | Unweighted | Weighted |
| Idaho | Private | 747 | 729 | 26 | 0 | 1 | 25 | 96% | 96% |
| | Public | 19430 | 18185 | 1942 | 27 | 93 | 1822 | 95% | 95% |
| Indiana | Private | 8684 | 10934 | 135 | 0 | 9 | 126 | 93% | 95% |
| | Public | 76504 | 66650 | 2040 | 27 | 93 | 1920 | 95% | 95% |
| Michigan | Private | 16375 | 15974 | 238 | 0 | 9 | 229 | 96% | 97% |
| | Public | 121972 | 124773 | 2573 | 45 | 134 | 2394 | 95% | 95% |
| Pennsylvania | Private | 31014 | 23915 | 282 | 1 | 10 | 271 | 96% | 96% |
| | Public | 132795 | 130658 | 3181 | 59 | 157 | 2965 | 95% | 95% |
| SW PA Math & Sci. Collaborative, PA | Private | 3661 | 3181 | 166 | 1 | 3 | 162 | 98% | 98% |
| | Public | 28648 | 26895 | 1472 | 20 | 76 | 1376 | 95% | 95% |
| TOTAL STUDENTS | Private | 87834 | 75466 | 681 | 1 | 29 | 651 | | |
| | Public | 1676655 | 1648020 | 45259 | 1223 | 2697 | 41339 | | |

## 6.10 Combined Participation Rates

The combined school and student Benchmarking participation rates are shown in Exhibits 6.9 through 6.11. The combined rates are the product of the school and student participation rates.

**Exhibit 6.9 TIMSS 1999 Benchmarking Combined Participation Rates**

| Jurisdiction | Unweighted Rate | | Weighted Rate | |
|---|---|---|---|---|
| | Including Substitutes | Not Including Substitutes | Including Substitutes | Not Including Substitutes |
| Connecticut | 91% | 91% | 90% | 90% |
| Idaho | 83% | 83% | 83% | 83% |
| Illinois | 90% | 90% | 91% | 91% |
| Indiana | 61% | 81% | 59% | 79% |
| Maryland | 89% | 89% | 88% | 88% |
| Massachusetts | 93% | 93% | 93% | 93% |
| Michigan | 84% | 87% | 85% | 88% |
| Missouri | 74% | 87% | 74% | 88% |
| North Carolina | 92% | 92% | 92% | 92% |
| Oregon | 83% | 83% | 83% | 83% |
| Pennsylvania | 67% | 67% | 63% | 63% |
| South Carolina | 87% | 87% | 87% | 87% |
| Texas | 68% | 69% | 67% | 69% |
| Academy #20, CO | 94% | 94% | 94% | 94% |
| Delaware Math & Sci., DE | 92% | 92% | 92% | 92% |
| Dade County, FL | 91% | 91% | 91% | 91% |
| Chicago Public Schools, IL | 90% | 90% | 91% | 91% |
| First in the World, IL | 85% | 85% | 90% | 90% |
| Naperville #203, IL | 96% | 96% | 96% | 96% |
| Montgomery County, MD | 94% | 94% | 94% | 94% |
| Invitational Group, MI | 92% | 92% | 91% | 91% |
| Fremont/Lincoln/ WestSide P.S., NE | 95% | 95% | 95% | 95% |
| Jersey City Public Schools, NJ | 90% | 90% | 91% | 91% |
| Rochester City Sch. Dist., NY | 84% | 84% | 84% | 84% |
| Guilford County, NC | 93% | 93% | 92% | 92% |
| Project SMART, OH | 94% | 94% | 94% | 94% |
| SW PA Math & Sci. Collaborative, PA | 76% | 76% | 76% | 76% |

**Exhibit 6.10 TIMSS 1999 Benchmarking Combined Participation Rates by School Type**

| Jurisdiction | School Type | Unweighted Rate | | Weighted Rate | |
|---|---|---|---|---|---|
| | | Not Including Substitutes | Including Substitutes | Not Including Substitutes | Including Substitutes |
| Idaho | Private | 48% | 48% | 48% | 48% |
| | Public | 84% | 84% | 84% | 84% |
| Indiana | Private | 80% | 80% | 71% | 71% |
| | Public | 58% | 81% | 58% | 80% |
| Michigan | Private | 96% | 96% | 97% | 97% |
| | Public | 82% | 86% | 83% | 87% |
| Pennsylvania | Private | 48% | 48% | 34% | 34% |
| | Public | 71% | 71% | 70% | 70% |
| SW PA Math & Sci. Collaborative, PA | Private | 98% | 98% | 98% | 98% |
| | Public | 73% | 73% | 73% | 73% |

**Exhibit 6.11 TIMSS 1999 Benchmarking Weighted Combined Participation Rates Alternate Method**

| Jurisdiction | Substitutes Not Included | Substitutes Included |
|---|---|---|
| Connecticut | 90% | 90% |
| Idaho | 83% | 83% |
| Il, combined | 91% | 91% |
| Indiana | 58% | 79% |
| MD, combined | 88% | 88% |
| Massachusetts | 93% | 93% |
| Michigan | 85% | 88% |
| Missouri | 75% | 88% |
| NC, combined | 92% | 92% |
| Oregon | 83% | 83% |
| PA, combined | 63% | 63% |
| South Carolina | 86% | 86% |
| Texas | 67% | 67% |
| Academy #20, CO | 94% | 94% |
| Delaware Math & Sci., DE | 92% | 92% |
| Dade County, FL | 91% | 91% |
| Chicago Public Schools, IL | 90% | 90% |

**Exhibit 6.11 (continued) TIMSS 1999 Benchmarking Weighted Combined Participation Rates Alternate Method**

| Jurisdiction | Substitutes Not Included | Substitutes Included |
|---|---|---|
| First in the World, IL | 90% | 90% |
| Naperville #203, IL | 96% | 96% |
| Montgomery County, MD | 94% | 94% |
| Invitational Group, MI | 91% | 91% |
| Fremont/Lincoln/ WestSide P.S., NE | 95% | 95% |
| Jersey City Public Schools, NJ | 91% | 91% |
| Rochester City Sch. Dist., NY | 84% | 84% |
| Guilford County, NC | 92% | 92% |
| Project SMART, OH | 94% | 94% |
| SW PA Math & Sci. Collaborative, PA | 75% | 75% |

**Exhibit 6.12 TIMSS 1999 Benchmarking Weighted Combined Participation Rates: Alternate Method**

| Jurisdiction | School Type | Substitutes Not Included | Substitutes Included |
|---|---|---|---|
| Idaho | Private | 48% | 48% |
| | Public | 85% | 85% |
| Indiana | Private | 71% | 71% |
| | Public | 56% | 80% |
| Michigan | Private | 97% | 97% |
| | Public | 83% | 87% |
| PA, combined | Private | 34% | 34% |
| | Public | 69% | 69% |
| SW PA Math & Sci. Collaborative, PA | Private | 98% | 98% |
| | Public | 72% | 72% |

## 6.11 TIMSS 1999 Benchmarking Sample Weights

Benchmarking sample weights have four components:

1. **The school base weight** is the reciprocal of the school's selection probability;
2. **A school nonresponse adjustment** is an adjustment to the school base weight for schools that did not participate;
3. **The student base weight** is the product of the adjusted school weight and the reciprocal of the student's selection probability;
4. **A student nonresponse adjustment** is an adjustment to the student base weight for eligible students that did not participate.

Sample weights were computed by the same general methodology for all Benchmarking jurisdictions. The following sections discuss: computation of school base weights for the Benchmarking samples, school-level non-response adjustment, non-response adjustment at the student level, computation of final student weights, and the creation of variance estimation strata and replicates for jackknife variance estimators.

### 6.11.1 School Base Weights

The school base weight is the inverse of the sampled school's probability of selection into the TIMSS 1999 Benchmarking sample. (see Section 5.5.1):

$$BW_{sc}^{i} = \frac{M}{n \bullet m_i} = (p_i^{(B)})^{-1} .$$

#### TIMSS 1999 overlap strata where no supplementary Benchmarking sample was selected.

The only sample schools in these strata were TIMSS 1999 national sample schools. The probability of selection into the Benchmarking sample was the conditional probability of selection into the TIMSS 1999 national sample, given that the PSU had been selected:

$$p_i^{(B)} = p_i^{(N)}$$

#### TIMSS 1999 overlap strata where a supplementary Benchmarking sample was selected

Any school in these strata had a chance of selection into both samples: the TIMSS 1999 national sample $(p_i^{(N)})$ and the Benchmarking supplementary sample $(p_i^{(S)})$. Since the final Benchmarking sample was composed of schools in either sample, the probability of selection for these schools was:

$p_i^{(B)} = p_i^{(N)} + p_i^{(S)} - p_i^{(N)} p_i^{(S)}$.

### Nonoverlap strata

These strata were composed of PSUs that had not been selected for the TIMSS 1999 national sample. Thus the final sample was composed entirely of schools selected into the Benchmarking sample with probability $p_i^{(B)}$.

Each participating substitute school was assigned the weight $w_i$ of the sample school it replaced.

### Adjustment for school nonresponse

The school base weights were adjusted for nonresponse by a factor equal to the reciprocal of the weighted school response rates:

$$SCNRA_a = \frac{\sum_{sampled\ schools} w_i \cdot G_i}{\sum_{i participating\ schools} w_i \cdot G_i}$$

where $w_i$ is the school base weight defined in Section 6.11.1, $G_i$ is the estimated eighth grade enrollment, and $a$ is the school nonresponse cell. Sampled schools included eligible participating and refusing originally selected schools; participating schools included originally selected schools and substitutes. Nonresponse cells were defined within private and public sampling strata by zip code.

## 6.11.2 Student Base Weights

Within each sampled school, eighth grade math classes were selected with equal probability and all students in the selected classes were sampled. The calculation of the student base weights is shown in section 5.5.4.

### Student Nonresponse Adjustments

Student nonresponse cells were defined by classes within schools. This is described in section 5.5.5.

### Final Student Weights

The final weight assigned to each student is the nonresponse-adjusted student weight shown in section 5.6.5. Exhibit 6.12 shows the distribution of the final student sampling weights for each Benchmarking jurisdiction.

**Exhibit 6.12 Distribution of TIMSS 1999 Benchmarking Final Student Weights**

| Jurisdiction | Minimum | 25th percentile | Median | 75th percentile | Maximum |
|---|---|---|---|---|---|
| Connecticut | 4.7803 | 15.3726 | 17.8114 | 20.3611 | 39.1346 |
| Idaho | 6.5487 | 7.3725 | 8.5156 | 10.7137 | 30.5891 |
| Chicago Public Schools, IL | 3.2342 | 17.3196 | 22.1894 | 27.5666 | 42.6459 |
| First in the World, IL | 1.0000 | 2.9268 | 3.3951 | 3.7372 | 6.6755 |
| Naperville #203, IL | 1.0000 | 1.0256 | 1.1818 | 1.2273 | 1.3016 |
| Illinois | 1.0000 | 1.3016 | 18.2931 | 56.3814 | 154.3068 |
| Indiana | 15.9424 | 30.3584 | 33.2721 | 38.8407 | 261.3641 |
| Montgomery County, MD | 2.5783 | 5.4959 | 6.7896 | 7.6230 | 11.4781 |
| Maryland | 2.5783 | 7.4833 | 19.3411 | 22.6094 | 37.7517 |
| Massachusetts | 10.7310 | 21.3892 | 26.4631 | 32.2549 | 57.6235 |
| Michigan | 12.9524 | 43.7418 | 49.8401 | 57.5453 | 302.1111 |
| Missouri | 13.7907 | 26.3760 | 29.4220 | 34.8685 | 94.7381 |
| North Carolina | 6.0000 | 33.3203 | 37.1670 | 44.3448 | 87.3830 |
| Guilford County, NC | 2.6744 | 3.4690 | 4.4103 | 5.3191 | 10.0000 |
| NC, combined | 2.6744 | 5.3191 | 33.3745 | 41.1138 | 87.3830 |
| Oregon | 13.5971 | 15.1030 | 18.1235 | 23.3453 | 68.5553 |
| Pennsylvania | 8.2000 | 48.4389 | 59.4357 | 82.3808 | 298.4658 |
| SW PA Math & Sci Collaborative, PA | 8.9883 | 14.2627 | 18.5946 | 25.7996 | 36.2519 |
| PA, combined | 8.2000 | 16.4507 | 32.6016 | 66.0394 | 298.4658 |
| South Carolina | 4.0663 | 20.2412 | 24.2094 | 28.0881 | 58.3424 |
| Texas | 27.5546 | 112.7242 | 133.6627 | 171.0004 | 386.1602 |
| Academy #20, CO | 1.0000 | 1.0333 | 1.0435 | 1.0833 | 1.2667 |
| Delaware Math & Sci, DE | 2.6563 | 4.5776 | 6.0000 | 7.5122 | 9.7347 |
| Dade County, FL | 7.5118 | 13.4984 | 17.5315 | 20.9744 | 30.4205 |
| Invitational Group, MI | 1.0000 | 2.2623 | 3.0000 | 3.4167 | 6.7273 |
| Lincoln/Fremont/West Side P.S., NE | 1.0000 | 1.0455 | 1.0952 | 4.2857 | 10.0000 |
| Jersey City Public Schools, NJ | 1.0357 | 1.1081 | 1.6216 | 2.1053 | 2.6500 |
| Rochester City Sch. Dist., NY | 1.5039 | 1.8107 | 1.9402 | 2.2279 | 3.2464 |
| Project SMART, OH | 1.5882 | 4.2927 | 5.6667 | 6.3750 | 8.8000 |

## 6.12 Defining Variance Estimation Strata and Creating Replicates

The sampling variability of statistics based on TIMSS 1999 Benchmarking data was estimated by the jackknife repeated replication method, as described by Gonzalez & Foy in chapter 11 of this volume. This method requires repeatedly dividing the full sample into subsamples, or replicates, and calculating the statistic of interest for each replicate. The jackknife variance estimator is then:

$$v(p) = \sum_{k=1}^{K} (p_k - p)^2,$$

where

$p$ = the full-sample statistic of interest

$p_k$ = the statistic of interest for the $k^{th}$ replicate

$K$ = the number of replicates

Replicates are created by randomly deleting first-stage sampling units from the full sample, which for the TIMSS 1999 Benchmarking samples were schools, classes (or pseudo classes), or sets of students.

Replicates for the TIMSS 1999 Benchmarking samples corresponded to variance strata that in most cases were defined by pairs (or triples) of schools or classes. Within these variance strata the variance unit was a school or a class, respectively. In some cases, variance strata were defined by single classes. This occurred when a school had been selected with certainty and all classes within that school were selected for assessment. In such cases students were systematically assigned to two groups within each class, and variance strata were defined by these "half-class" pairs; the variance unit was a half-class. Variance strata were assigned within sampling strata after sorting each sample in selection order. They were numbered sequentially within each sample across the sampling strata. The Benchmarking samples were classified into three groups for replication. Exhibit 6.13 shows this classification and identifies the variance strata and variance units for each sample.

### 6.12.1 Group A: districts and consortia having fewer than 25 schools

All schools were selected with certainty in these small self-defined jurisdictions. Variance strata were defined by half-class pairs when classes had been selected with certainty, or by class pairs (or triples) otherwise. Variance units were half-classes for certainty selections and classes for noncertainties.

Pseudo classes that had been created for sampling were defined as classes, and each sample was sorted by certainty status, school ID, (pseudo) class ID, and student ID. Variance strata and variance units were then assigned in order at the appropriate level. Five of these jurisdictions had at least one school where some classes were selected with certainty; all students were selected with certainty in Academy School District # 20, Colorado (see Exhibit 6.13).

### 6.12.2 Group B: districts and consortia having at least 25 schools

Three of the jurisdictions in this group were public school districts: Miami Dade County, FL; Chicago, IL; and Montgomery County, MD. The fourth was a consortium of public and private schools: Southwest Pennsylvania Regional Mathematics and Science Collaborative. The Miami Dade County, Chicago, and Southwest Pennsylvania samples were composite samples, that is, they were composed of schools that had been selected for the TIMSS 1999 national assessment, in addition to those selected for their respective Benchmarking assessments. There were no explicit sampling strata in Miami Dade County, Chicago, or Montgomery County. Southwest Pennsylvania, however, had public and private, overlap and nonoverlap sampling strata. "Overlap" refers to PSUs within a Benchmarking jurisdiction that were also TIMSS 1999 national PSUs. TIMSS 1999 national sample schools in Pennsylvania were assigned to appropriate Southwest Pennsylvania Benchmarking sampling strata for the purpose of defining variance strata.

Eight schools were selected with certainty in Montgomery County; these schools defined variance strata. Since students, not classes, had been sampled in Montgomery County schools, the sampled students within each school were systematically assigned to two groups, treated as classes. These classes defined variance units in the Montgomery County certainty schools. In all four of these samples, school pairs were variance strata and schools were variance units for noncertainty selections.

Each sample was sorted within sampling strata by certainty status, enrollment, and class ID. Variance strata and variance units were then assigned in order at the appropriate level; they are shown in Exhibit 6.13.

### 6.12.3 Group C: States

All TIMSS 1999 Benchmarking state samples were composite samples consisting of schools that had been selected for the TIMSS 1999 national assessment, in addition to those selected for the state Benchmarking assessments. Idaho, Indiana, Michigan, and Pennsylvania sampled both private and public schools; all others sampled only public schools. Thus, there were private and public, overlap and nonoverlap state Benchmarking sampling strata. Overlap sampling strata were defined by TIMSS 1999 national PSUs.

Five schools were selected with certainty in Idaho, two in North Carolina; these schools defined variance strata, and classes within them were variance units. All other state Benchmarking sample schools were noncertainty selections. Variance strata were defined in these samples by school pairs (or triples); the schools were variance units. Each sample was sorted within sampling strata by certainty status, enrollment, and class ID. Variance strata and variance units were then assigned in order at the appropriate level.

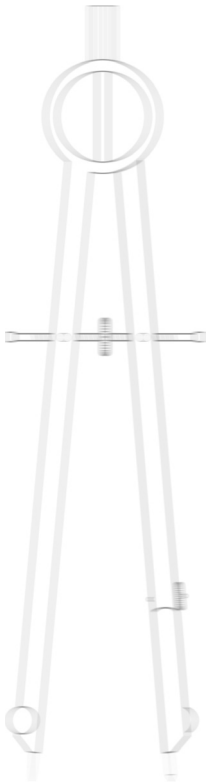
School districts and consortia undertook independent Benchmarking assessments in four states: Illinois, Maryland, North Carolina, and Pennsylvania. The records for these district and consortium samples (Groups A and B) were appended to the appropriate state samples (Group C), and their variance strata were renumbered. These renumbered variance strata are shown in Exhibit 6.13.

**Exhibit 6.13 TIMSS 1999 Benchmarking Variance Strata**

| Group | IDCNTRY | Entity | Variance Stratum | | Variance Unit |
|---|---|---|---|---|---|
| A | 10801 | Academy CO | 1-49 | Class (certainty) | Half-class |
| A | 11001 | DE Sci Coal | 1-25 | Class pair | Class |
| A | 11701 | Naperville IL | 1-21<br>22-34 | Class (certainty)<br>Class pair | Half-class<br>Class |
| A | 11702 | 1st in World IL | 1<br>2-15 | Class (certainty)<br>Class pair | Half-class<br>Class |
| A | 12601 | MI Invitational Group | 1-7<br>8-24 | Class (certainty)<br>Class pair | Half-class<br>Class |
| A | 13101 | Lincoln/Fremont/<br>West Side PS NE | 1-33<br>34-43 | Class (certainty)<br>Class pair | Half-class<br>Class |
| A | 13401 | Jersey City PS NJ | 1-22<br>23-35 | Class (certainty)<br>Class pair | Half-class<br>Class |
| A | 13601 | Rochester PS NY | 1-24 | Class pair | Class |
| A | 13701 | Guilford Co NC | 1-21 | Class pair | Class |
| A | 13901 | Project SMART OH | 1-24 | Class pair | Class |
| B | 11201 | Dade Co FL | 1-12 | School pair | School |
| B | 11703 | Chicago PS IL | 1-13 | School pair | School |
| B | 12401 | Montgomery Co MD | 1-8<br>9-16 | School (certainty)<br>School pair | Class<br>School |
| B | 14201 | SW PA Science & Math<br>Collaborative | 1-3<br>4-19 | School pair (private)<br>School pair (public) | School<br>School |
| C | 10900 | CT | 1-26 | School pair | School |
| C | 11600 | ID | 1<br>2-5<br>6-25 | School pair (private)<br>School (certainty; public)<br>School pair (public) | School<br>Class<br>School |
| C | 11700 | IL | 1-6<br>1-6<br>7-32<br>33<br>34-47<br>48-68<br>69-75 | School pair (IDSTRATE=1)<br>Class pair (IDSTRATE=5)<br>School pair<br>Class (certainty)<br>Class pair<br>Class (certainty)<br>Class pair | School<br>Class<br>School<br>Half-class<br>Class<br>Half-class<br>Class |
| C | 11800 | IN | 1-3<br>4-26 | School pair (private)<br>School pair (public) | School<br>School |
| C | 12400 | MD | 1-24<br>25-32<br>33-40 | School pair<br>School (certainty)<br>School pair | School<br>Class<br>School |
| C | 12500 | MA | 28 | School pair | School |
| C | 12600 | MI | 1-4<br>5-28 | School pair (private)<br>School pair (public) | School<br>School |
| C | 12900 | MO | 1-25 | School pair | School |
| C | 13700 | NC | 1-2<br>3-25<br>26-47 | School (certainty)<br>School pair<br>Class pair | Class<br>School<br>class |

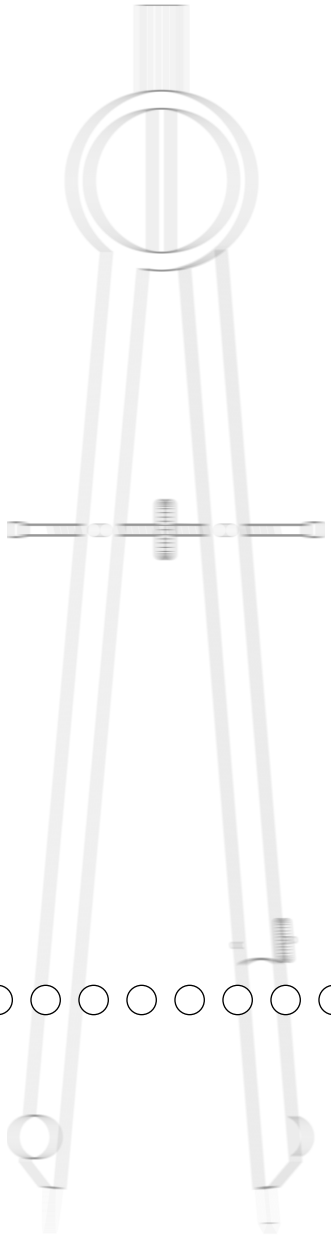
**Exhibit 6.13 (continued) TIMSS 1999 Benchmarking Variance Strata**

| Group | IDCNTRY | Entity | Variance Stratum | | Variance Unit |
|---|---|---|---|---|---|
| C | 14100 | OR | 1-22 | School pair | School |
| C | 14200 | PA | 1<br>2-20<br>21-23<br>24-39 | School pair (private)<br>School pair (public)<br>School pair (private)<br>School pair (public) | School<br>School<br>School<br>School |
| C | 14500 | SC | 1-24 | School pair | School |
| C | 15800 | TX | 1-26 | School pair | School |




# Data Collection and Data Preparation for TIMSS 1999 Countries

Eugenio J. Gonzalez
Dirk Hastedt



# 7 Data Collection and Data Preparation for TIMSS 1999 Countries

Eugenio J. Gonzalez
Dirk Hastedt

## 7.1 Overview

This chapter, which is based on Gonzalez & Hastedt (2000), describes the procedures for administering the TIMSS 1999 tests and questionnaires in the participating countries, and for scoring the free-response achievement items and preparing the computer files. Chapter 8 describes the data collection and data preparation for the Benchmarking participants.

The TIMSS 1999 data collection in each country was a very demanding exercise, requiring close cooperation between the national research coordinator (NRC), school personnel, and students. The first part of this chapter describes the international 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 in the participating countries. The second part describes the activities involved in preparing the data files at the national centers throughout the world, 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. Chapter 8 describes the field operations within the United States.

## 7.2 TIMSS 1999 Field Operations

The TIMSS 1999 international 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. The TIMSS 1999 field testing took place in 31 countries around the world (for more details see O'Connor, 2000).

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

### 7.2.2 Documentation and Software

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

- The *Survey Operations Manual* (TIMSS, 1997b) was the essential handbook of the NRC, 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, 1997d) defined the TIMSS 1999 target population and sampling goals and described the procedures for the sampling of schools.
- The *School Coordinator Manual* (TIMSS, 1997a) 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, 1997c) 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.

○○○

1. See chapter 4 for details of the translation and cultural adaptation task and chapter 5 for information about sampling of schools by participating countries.

- The *Scoring Guides for Mathematics and Science Free-Response Items* (TIMSS, 1998c) contained instructions for scoring the short-answer and extended-response test items.
- The *Manual for Entering the TIMSS-R Data* (TIMSS, 1998a) provided the NRC with instructions for coding, entering, and verifying the data. This manual included the codebook, which defined the variables and file formats in the data files.
- The *Manual for National Quality Control Observers* (TIMSS, 1998b) provided instructions for conducting classroom observations in a sample of participating schools.

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 sampling unit, the 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 of those students be identified and asked to complete a Teacher Questionnaire.

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 together 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.

○○○

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.

- During testing, the test administrator and school coordinator used the tracking forms to record student and teacher participation, and returned them to the NRC after the testing 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 5). 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 linking schools, classes, teachers, and students; 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, an 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 number for the class.

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 on 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 an ID that consisted of the three-digit school ID plus a two-digit number for the teacher. Since a teacher could be teaching both mathematics and science to some or all of the students in a class, a unique identification number was needed 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. These procedures had to be carefully followed 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, and 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. 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 tasks. For example, 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 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, which involved arranging rooms, times, classes and materials.

Further details about the school coordinator's responsibilities are detailed in the *School Coordinator Manual* (TIMSS, 1997c).

The test administrator's responsibilities in administering the TIMSS tests and Student Questionnaires were described in the *Test Administrator Manual* (TIMSS, 1997d) and included:

- Ensuring that each student received the correct testing materials that 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. The national program required quality control monitors to document data collection activities in their country. They visited a 10% sample of TIMSS 1999 schools, observed testing sessions, and recorded compliance with prescribed procedures.

The International Study Center prepared the *Manual for National Quality Control Monitors* (TIMSS, 1998b), which contained information about TIMSS 1999 and detailed the role and responsibilities of the national quality control monitors.

## 7.4 Preparation of Materials for Scoring and Data Entry

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.

## 7.5 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, 1997b) 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 score 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 eight 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 eight 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 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 who had educational backgrounds in the mathematics and science curriculum areas or 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.

### 7.5.1 Preparing Materials to Train the Scorers

The success of assessments containing free-response questions depends upon reliability in scoring 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 Scoring Guides for Mathematics and Science Free-Response Items (TIMSS, 1998c). 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.

○○○

3. The TIMSS scoring scheme for free-response items is described in chapter 2.

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.

### 7.5.2 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. Neither scorer knew the scores assigned by the other.

Since the purpose of the double scoring was to document scoring consistency, the procedure used 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.

### 7.5.3 Implementing the Scoring Procedures

TIMSS 1999 recommended that scorers be organized into teams of about six, headed by a team leader. The leader's primary responsibility was to continually check and recheck 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), it had to be done before the main data were scored, and the reliability scores had to be recorded on a separate scoring sheet rather than in the booklets.

To implement the scoring plan effectively the scorers were divided into two equivalent teams (Team A and Team B) and booklets 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, Team B scored the reliability sample (every fourth booklet) in the Set A batches, while Team A 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.

## 7.6 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 (1) identify problems in the identification variables, and invalid codes; and (2) identify problems in the structure of the data files, which could then be corrected 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 had 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.

## 7.7 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 document described 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, scoring the free-response items, or creating and checking the data files.

## 7.8 Summary

This chapter summarizes the design and implementation of the TIMSS 1999 international 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. Chapter 8 describes the implementation of the field operation procedures within the United States in the TIMSS Benchmarking jurisdictions.

## References

Gonzalez, E.G., & Hastedt, D. (2000). TIMSS field operations and data preparation (pp. 119-134). In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS 1999 technical report*. Chestnut Hill, MA: Boston College.

O'Connor, K.M. (2000). TIMSS field test (pp. 103-118). In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS technical report*. Chestnut Hill, MA: Boston College.

TIMSS (1997a). *School coordinators manual* (TIMSS 1999 Doc. Ref. No. 98-0024). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

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

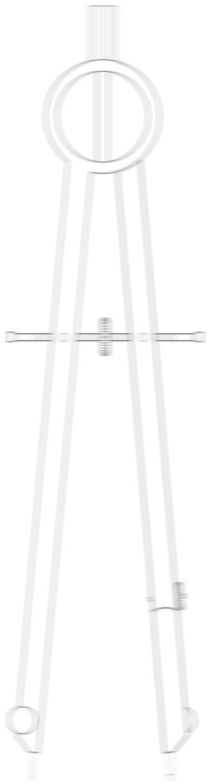
TIMSS (1997c). *Test administrator manual* (TIMSS 1999 Doc. Ref. No. 98-0025). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.

TIMSS (1997d). *TIMSS-R school sampling manual* (TIMSS 1999 Doc. Ref. No. 97-0012). Prepared by Pierre Foy, Statistics Canada. Chestnut Hill, MA: Boston College.

TIMSS (1998a). *Manual for entering the TIMSS-R data* (TIMSS 1999 Doc. Ref. No. 98-0028). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

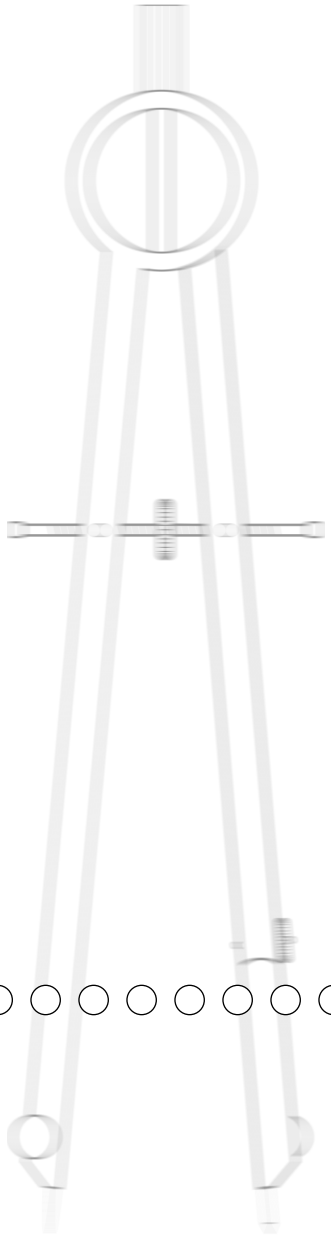
TIMSS (1998b). *Manual for national quality control monitors* (TIMSS 1999 Doc. Ref. No. 98-0044). Prepared jointly by the International Study Center at Boston College and the IEA's Data Processing Center. Chestnut Hill, MA: Boston College.

TIMSS (1998c). *Scoring guides for mathematics and science free-response items* (TIMSS 1999 Doc. Ref. No. 98-0049). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.




# Data Collection and Data Preparation for TIMSS 1999 Benchmarking

Dward Moore



# 8 Data Collection and Data Preparation for TIMSS 1999 Benchmarking

Dward Moore

## 8.1 Overview

This chapter discusses the data collection and processing activities conducted for the TIMSS 1999 Benchmarking study. Under contract with the International Study Center at Boston College, Westat was responsible for data collection and preparation. In particular, Westat coordinated the within-school sampling and data collection activities. Westat subcontracted National Computer Systems (NCS) to process the data, produce the data collection forms, receive and score the completed assessment materials, and prepare the database for final analysis.

## 8.2 Field Operations and Data Collection

Data collection for the TIMSS 1999 Benchmarking study occurred March 15 through June, 1999, concurrently with data collection for the U.S. national sample for TIMSS 1999 and for most of the TIMSS 1999 Northern Hemisphere countries. A number of people were involved in the data collection effort in the U.S., including the Westat field manager, jurisdiction coordinators, field supervisors, school coordinators, and test administrators. These individuals assisted in gaining access, gathering necessary sampling information, and scheduling and administering the assessment in the sampled schools.

### 8.2.1 Field Manager

The TIMSS 1999 Benchmarking data collection activities were directed by a Westat field manager who oversaw the work of 20 field supervisors. The field manager also visited sites to assess whether test administration procedures were implemented correctly and uniformly.

### 8.2.2 Jurisdiction Coordinators

Each participating jurisdiction selected a contact person to work with the Westat field staff in gaining participation of the sampled schools and to act as the conduit in collecting the necessary information from the schools.

### 8.2.3 Field Supervisors

The field supervisors were mainly responsible for coordinating data collection activities in the schools assigned to them within a geographic area. They attended a training session and received a manual developed by Westat. Field supervisors contacted school coordinators to confirm schedules, ascertain that the testing materials had arrived at the schools, and finalize assessment arrangements. They also trained the test administrators. Westat provided a script for the field supervisors to use in the training sessions.

### 8.2.4 School Coordinators

Each participating school appointed a school-level coordinator, usually a teacher or principal, to be responsible for obtaining information on sampling classes within a school and making preparations for the assessment: scheduling test administration, receiving testing materials, distributing the Teacher Questionnaires, and completing (or selecting a designee who would complete) the School Questionnaire.

### 8.2.5 Test Administrators

Test administrators were responsible for preparing for test administration and for going to the schools on the agreed-upon assessment date to collect the Teacher and School Questionnaires and to administer the assessments. Test administrators were expected to spend up to four hours studying the *TIMSS 1999 Test Administrator Manual* (TIMSS, 1998) and attending the training session run by the field supervisor before the assessment day. Adherence to the standard procedures set forth in the *TIMSS 1999 Test Administrator Manual* was emphasized.

## 8.3 Benchmarking Manuals

Westat followed the data collection procedures detailed in the international manuals (see chapter 7 for a list of manuals provided to NRCs). Based on these materials, Westat developed a guide for field supervisors during data collection. In addition, Westat adapted the *TIMSS 1999 Test Administrator Manual* (TIMSS, 1998) to apply to the Benchmarking context.

### 8.3.1 Supervisors Manual

Westat developed the *TIMSS 1999 Benchmarking Supervisors Manual* (Westat, 1998a) to guide supervisors in conducting within-school sampling; recruiting, hiring and training test administrators; contacting schools to secure participation and scheduling; and observing at least one testing session conducted by each test administrator.

### 8.3.2 Benchmarking Test Administrator Manual

The *TIMSS 1999 Benchmarking Test Administrators Manual* (Westat, 1998b) detailed the standard international procedures and included additional information to help test administrators conduct a high-quality session. The manual included an overview of the Benchmarking study, a section on classroom management, question-by-question specifications for the Student Questionnaire, and administrative procedures such as completing the expense reports.

## 8.4 Within-School Sampling

Once jurisdiction coordinators secured cooperation of the sampled schools, each school coordinator obtained the information necessary for sampling classes within the school. Coordinators followed the standard international procedure for sampling classes and teachers. The process and forms used were as follows:

- The school coordinator sent a list of the eighth-grade mathematics classes and their teachers' names to the jurisdiction coordinator using the Class Listing Form. This form, developed by Westat, included instructions and definitions regarding the classes to be listed.
- The jurisdiction coordinator forwarded the class listing to Westat staff who selected the sample of two intact mathematics classrooms using the Class Sampling Form. The Class Sampling Form used for Benchmarking was identical to the main survey form used in the United States. It differed from the international version in that it specified minimum class size rather than minimum cluster size. (Minimum class size, 10, is given as half the minimum cluster size). In the Class Sampling Forms, pseudoclasses or classes of students "not taking math" were formed as necessary following the international procedures required by the International Study Center.

- The selected mathematics classes were each specified on a Student-Teacher Worksheet. The jurisdiction coordinator mailed the Student-Teacher Worksheet to the school coordinator along with a cover letter asking for lists of the students from the selected mathematics classes (their names, gender, and birth dates) and each student's science class and science teacher's name.
- Westat staff prepared the Teacher-Student Linkage Forms, Teacher Tracking Forms, and Student Tracking Forms based on the information submitted on the Student-Teacher Worksheet. They passed the completed forms to jurisdiction coordinators who mailed them to the school for use by the Westat test administrators on assessment day. In addition, Westat sent test administrators the prepared School Questionnaire and Teacher Questionnaires with a letter asking them to distribute these so that the completed responses could be picked up by the Westat assessment staff on testing day.
- The "exclusion" code, if any, was recorded only on the Student Tracking Forms, rather than entering it on both the Linkage and Student Tracking Forms as indicated in the international procedural manual. Schools were provided with a printed copy of the international "exclusion" definitions, and decisions on excluding students were made by school staff. The Westat assessment staff recorded exclusions on the Student Tracking Forms on the assessment day before testing began.
- Since the sampling activities for the TIMSS 1999 Benchmarking study in the U.S. were not centrally organized, it was decided not to use the TIMSS sampling software.

## 8.5 Test Administration

In most schools, students were assessed in two groups (the mathematics classes that were sampled). Due to class scheduling, students in some schools were assessed as one group in a large facility such as the cafeteria or library.

Test administrators followed the standard script for administering the assessment based on internationally prescribed procedures. The responsibilities of the test administrator are detailed in the *TIMSS 1999 Benchmarking Test Administrators Manual* (Westat 1998b) and included:

- Ensuring that each student received the correct testing materials that were specifically prepared for him or her

- Administering the test in accordance with the instructions in the *TIMSS 1999 Benchmarking Test Administrator Manual* and the session script
- Ensuring the correct timing of the testing session
- Recording student participation on the Student Tracking Form
- Copying student demographic information onto booklet covers
- Recording session details on the Test Administration Form and Student Response Rate Form.

Make-up sessions were requested whenever attendance in the school's sessions was below the target 90% participation. In a few schools, it was not possible to schedule make-up sessions. As each testing session was completed, test administrators sent all assessment materials to NCS for processing.

## 8.6 Quality Control Monitoring

Site visits were made by field managers and by Westat home office staff to make sure that all procedures were being implemented correctly and uniformly. Test administrators were observed by their field supervisors in at least one assessment for quality control. In addition, about 10% of schools in each jurisdiction were randomly selected for a visit by a Quality Control Monitor from the International Study Center (see chapter 9 for details).

## 8.7 Data Processing

NCS was subcontracted by Westat to process and score the TIMSS 1999 Benchmarking data for the states and districts. NCS used the same receipt control procedures that were used to process the TIMSS 1999 national assessment materials and carried out all scoring procedures for TIMSS 1999 Benchmarking simultaneously with the TIMSS 1999 national assessment. Student papers from the TIMSS 1999 national and state/district Benchmarking samples were intermingled and scored simultaneously.

When TIMSS 1999 national and Benchmarking materials came into NCS from the field, they were checked using the same specifications and were processed and scanned on parallel tracks. Each group of materials was given a letter code (TR for national and DS for Benchmarking districts and states) and a sequential number to identify the batch and document. After conducting receipt control procedures, NCS scanned the test booklets and sorted the open-ended items by item in preparation for scoring.

## 8.8 Data Entry

The data on the scannable documents were collected using NCS optical-scanning equipment that also captured images of the constructed-response items and intelligent character recognition (ICR) fields. The School and Teacher Questionnaires were entered into a data file using key-entry methods. A second person keyed the same data into a verification file. Both sets of data were then programmatically compared and discrepancies were corrected. The data were run through the NCS Pre-Edit Program and edited according to specifications that the development staff created. This ensured that the data would conform to the codebook specifications. If there were problems, the editing staff corrected them and then ran the batch through the NCS Post-Edit Program to insure that the changes were entered correctly and the data were clean.

## 8.9 Image Scoring

Because of the economy of scale due to the increased sample size and the good rate of participation by states and districts, NCS used image processing and on-line scoring for the TIMSS 1999 national and Benchmarking studies. Two of the significant advantages of this on-line system were the ease of regulating the flow of work to scorers and the ease of monitoring scoring. The system allowed for item-by-item, rather than book-by-book, scoring. Item-by-item scoring increased efficiency because scorers and trainers could focus on one item at a time, thus improving scoring reliability and validity.

On-line scoring was a valuable innovation in the TIMSS 1999 project. Training for a particular item occurs before the scoring of all responses for the item. Student papers from both the TIMSS 1999 national sample and district/state Benchmarking samples were intermingled and scored simultaneously by scorers trained in the scoring procedures for the TIMSS 1999 open-ended items. Coupled with the fact that responses were scored item by item, this meant that all occurrences of the items, whether in the national sample, in the district/state sample, or in different booklet types, were scored together following the training for that item.

## 8.10 Scoring Training

NCS prepared training materials for scorers using the international training materials distributed at the TIMSS 1999 international scoring training meeting in February 1999 as well as sample papers from the responses received from the U.S. assessment in Spring 1999. NCS scoring center staff selected a random sample of about

50 responses for each item to be scored. NCS trainers who attended the international training reviewed the responses to ensure that a wide range of responses were represented for each score point. A score was assigned to each response, and sets of papers that exemplified score points were created. Training involved:

- Presenting and discussing the item to be scored along with the item rationale
- Explaining the scoring guide to the team and discussing the anchor papers, which contain the scoring guide, the item, its scoring rationale, and the student responses that represent the various score points in the guide
- Discussing the rationale behind the scoring guide, focusing on the criteria that differentiate the levels in the guide
- Practicing scoring on a shared set of student responses
- Discussing the responses in the shared packet
- Scoring and discussing 10 to 20 practice papers, which represented the entire range of score points for that item.

After the trainer and the scoring supervisor determined that the team had reached consensus, the scoring supervisor released the work electronically to scorers through the on-line scoring system. When scorers first received live responses, they either took turns scoring these responses or worked in pairs as a final quality check before they began working on their own.

Training sets were created by May 24, 1999. The trainers received their materials June 1. Scoring began on June 14 and was completed July 9.

## 8.11 Monitoring Scoring

Trainers and scoring supervisors monitored the scoring process in three ways: (1) using a software feature that allowed for ongoing checking of scorer agreement rates; (2) using a feature that allowed for backreading the papers read by scorers; and (3) using a feature that monitored scoring rates. These software features are discussed below.

Ten percent of all of the constructed responses were scored by a second reader to obtain statistics on inter-rater reliability. The responses chosen for "second scoring" were determined systematically. Second scorers did not know that they were second scoring

or what score the first rater had assigned. NCS scoring supervisors used this information to monitor the capabilities of all scorers, to maintain uniformity of scoring, and to ensure that scorer agreement rates met minimum standards.

The inter-rater reliability feature produced on-demand reports in either of two modes—aggregate information of all first scorings versus all second scorings, or overall agreement percentage for each individual scorer. The information was displayed in a matrix format showing the instances of exact agreement, which fell along the diagonal of the matrix. Data in each cell of the matrix provided the number and percentage of cases of agreement (or disagreement). The display also contained information on the total number of second scorings and the overall percentage of cases in which two scorers agreed on a rating.

The scoring supervisor monitored each scorer's progress using the system's backreading feature. Typically, a scoring supervisor looked at about 10% of all responses graded by each scorer. With this feature, scoring supervisors could see all the papers a given rater graded, and the scores that were assigned. Where scoring supervisors disagreed with the score assigned by the first scorer, they assigned a new score, which then became the reported score.

In addition to checking rater agreement and backreading individual responses, scoring supervisors monitored workflow using a status report that displayed the number of responses scored, the number of responses scored a second time, and the number of responses remaining to be scored. This facility allowed the scoring supervisor and trainer to monitor the scoring rate accurately and to estimate the time needed to complete the scoring.

## 8.12 Completing Data Files

After open-ended scoring was completed, scores were merged with the demographic, gridded, and key-entered data. At this time, final output files were produced for each file type. The final files were checked by the Software Quality Specialists to ensure that the data adhered to the international format. In earlier editing functions, data were checked for completeness and compliance with international codebook specifications. In addition, a check was performed to verify correct linking and matching of student, teacher, and school data files. To maintain confidentiality, all personal identification in the data set was removed before the data were sent to the IEA DPC and the International Study Center at Boston College. Further checking and cleaning took place at the DPC using international procedures for data cleaning.

## References

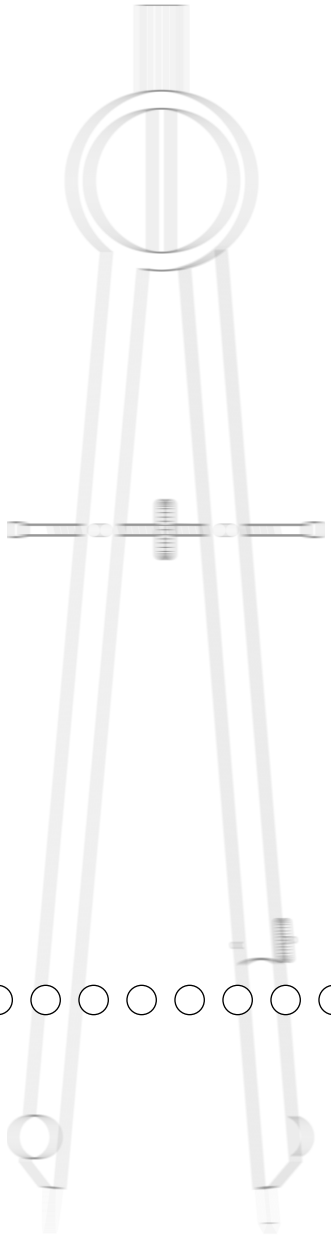
TIMSS (1998). *TIMSS 1999 test administrator manual* (TIMSS 1999 Doc Ref. No. 98-0025). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

Westat (1998a). *TIMSS 1999 benchmarking supervisors manual*. Prepared by Westat, Inc. Rockville, MD: Westat, Inc.

Westat (1998b). *TIMSS 1999 benchmarking test administrators manual*. Prepared by Westat, Inc. Rockville, MD: Westat, Inc.


# Quality Control in Data Collection for TIMSS 1999 Benchmarking

Kathleen M. O'Connor
Steven E. Stemler



# 9 Quality Control in Data Collection for TIMSS 1999 Benchmarking

Kathleen M. O'Connor
Steven E. Stemler

## 9.1 Overview

To verify that the TIMSS international data-collection procedures were applied uniformly in each of the benchmarking jurisdictions, the International Study Center instituted a program for quality assurance in data collection[1]. Quality Control Monitors (QCMs) were recruited by the International Study Center to document procedures in a sample of schools in each participating Benchmarking entity. The International Study Center selected approximately five schools in each state and two or three schools in each district/consortium to take part in the quality control program.

The major responsibility of Benchmarking QCM was to observe the TIMSS test administration in selected schools. QCMs were assigned to schools and completed a Classroom Observation Record documenting test administration procedures for each session they observed.

In preparation for their task, QCMs were given an overview of the TIMSS 1999 survey operations procedures and were trained by the staff of the International Study Center in how to conduct and document their quality control task. In order to facilitate the training, the TIMSS International Study Center developed a manual to inform QCMs about the TIMSS Benchmarking project and to describe in detail their roles and responsibilities.

QCMs were provided with the following materials to conduct their task: the *Manual for Quality Control Monitors* (TIMSS, 1998), the *TIMSS 1999 Benchmarking Test Administrator's Manual* (Westat, 1999), the Classroom Observation Forms, the list of schools selected for site visits, and contact information for the region's Westat supervisor.

○○○

1. See O'Connor, K.M., and Stemler, S.E. (2000) for information about the international quality control effort, which documented the data collection in the TIMSS 1999 countries.

Before beginning their task, QCMs assembled in Boston for a training meeting. At the meeting, QCMs were trained to do the following:

- Contact the supervisors to gather information about sites being observed, including the contact information for the school coordinator
- Contact the school coordinators to explain the site visits
- Review the condition of the achievement test booklets prior to testing
- Collect and review the tracking forms being used to sample students and to document participation status
- Observe the testing sessions
- Interview the school coordinator
- Document findings.

In total, 18 Benchmarking quality control monitors were recruited and trained. They observed a total of 98 testing sessions.

## 9.2 Observing the TIMSS Test Administration

The Classroom Observation Record was designed to allow the QCM 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.

### 9.2.1 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 9.1 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, appropriate corrections were made. In the few instances where QCMs reported a discrepancy between information on the Student

Tracking Form and information listed on the Student Identification Form, the errors were usually limited to one student in the group and consisted of a mismarking of the student's gender or a mismarked digit on the student ID number.

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, the desks were too narrow, the students sat at round tables).

The absence of a visible wall clock was also considered an environmental restriction more than a limitation of the implementation of the testing procedures. In many of the cases the room had a clock, but not all students were able to see 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 9.1 Preliminary Activities of the Test Administrator**

| Question | Yes | No | N/A |
|---|---|---|---|
| Had the test administrator verified adequate supplies of the test booklets? | 97* | 1** | - |
| Had the test administrator familiarized himself or herself with the script prior to testing? | 94* | 4** | - |
| Were all the seals intact on the test booklets prior to distribution? | 41 | 1 | 55+ |
| Did the Student Identification information on test booklet correspond with the Student Tracking Form? | 85 | 10 | 3 |
| Was there adequate seating space for the students to work without distractions? | 85 | 13 | - |
| Was there adequate room for the test administrator to move about the room during testing? | 93 | 5 | - |
| Did the test administrator have a stopwatch or timer for accurately timing testing sessions? | 96 | 2 | - |
| Did the test administrator have an adequate supply of pencils and other materials? | 97 | - | 1 |
| Was there a wall clock visible for the students to check their timing during the testing? | 83 | 14 | 1 |

+ Seals were not used on the booklets in these states, districts, or consortia

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

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

### 9.2.2 Test Session Activities

Section B of the classroom observation record dealt with the test session activities themselves. These 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 9.2 documents the activities associated with the first testing session and shows that at least 70% of the test administrators followed their script exactly when preparing the students and delivering instructions for Session 1. Where changes were made, they tended to be additions to the script.

Further examination of Exhibit 9.2 shows that in more than 75% of the sessions, the test administrator collected booklets one at a time from students. In the remaining sessions, students laid their booklets down on their desks during a brief 1 to 2 minute break.

Note that in 35 of 98 testing sessions (36%), the length of the testing session did not equal the time allowed. In each instance, all students had finished early.

Finally, booklets were rarely collected at the end of Session 1; rather, students were given a very short 1 to 2 minute break while the books remained on their desks.

**Exhibit 9.2 Test Administrator's Activities—Testing Session 1**

| Question | Yes | No | N/A |
|---|---|---|---|
| Did the test administrator follow the test administrator's script exactly in… | | | |
| …preparing the students? | 72 | 23 (minor changes)<br>2 (major changes) | 1 |
| …distributing the materials? | 58 | 30 (minor changes)<br>10 (major changes) | - |
| …giving General Directions? | 62 | 30 (minor changes)<br>6 (major changes) | - |
| …giving instructions for Part I? | 79 | 11 (minor changes)<br>8 (major changes) | - |
| If the test administrator made changes to the script, would you describe them as… | | | |
| …additions? | 39 | 20 | 39 |
| …revisions? | 24 | 28 | 46 |
| …deletions? | 21 | 23 | 54 |
| Did the test administrator distribute test booklets one-at-a-time to students? | 76 | 22 | - |
| Did the test administrator distribute the test booklets according to the booklet assignment on the Student Tracking Form? | 94 | 4 | - |
| Did the test administrator record attendance correctly on the Student Tracking Form? | 91 | 1 | 6 |
| Did the total testing time for Session 1 equal the time allowed? | 63 | 35 | - |
| Did the test administrator announce "you have 10 minutes left" prior to the end of Session 1? | 94 | 4 | - |
| Were any other "time remaining" announcements made during Session 1? | 10 | 88 | - |
| At the end of Session 1, did the test administrator collect the test booklets one at a time from students? | 2 | 96 | - |

Exhibit 9.3 summarizes QCMs' observations from the second testing session. The amount of time it took to restart the testing sessions ranged from 0 to 23 minutes; however, the vast majority of sessions were restarted in five minutes or less. In fact, because booklets were rarely collected during the break, testing typically resumed in 1-2 minutes.

Exhibit 9.3 shows that in about 30% of the testing sessions, the time used for Session 2 was less than the full time allowed. The test administrators reported that all of the students in those sessions had finished the exam early, had finished reviewing their work, and in many cases "were becoming unruly and impatient."

In 47 of 98 sessions (48%), booklets were collected one at a time from students. When they were not, students were simply asked to pass them to the front of their rows.

Exhibit 9.3 also reveals that in about two-thirds of the sessions observed, no break was given between testing and the administration of Student Questionnaires. Administrators often reported that students already had time to rest due to the fact that students often finished testing early. As a result, no official break preceded administration of the Student Questionnaires at many of the observed testing sessions.

A final statistic from Exhibit 9.3 worth noting is that in about two-thirds of the testing sessions, students requested additional time to complete the Student Questionnaire. It was almost always the case that these students were given an extra 5-10 minutes to complete the questionnaires.

**Exhibit 9.3 Test Administrator's Activities—Testing Session 2**

| Question | Yes | No | N/A |
|---|---|---|---|
| Was the time spent to restart the testing in Session 2 equal to 5 minutes? | 1 | 97 | - |
| Did the total testing time for Session 2 equal the time allowed? | 65 | 33 | - |
| Did the test administrator announce "you have 10 minutes left" prior to the end of Session 2? | 94 | 4 | - |
| Were any other "time remaining" announcements made during Session 2? | 7 | 91 | - |
| At the end of Session 2, did the test administrator collect the test booklets one at a time from the students? | 47 | 51 | - |
| 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*? | 32 | 61 | 5 |
| How accurately did the test administrator follow the script to end the testing and signal a break? | 40 (no changes) | 33 (minor changes) 21 (major changes) | 4 |
| If there were any changes, would you describe them as… | | | |
| …additions? | 20 | 24 | 54 |
| …some minor changes? | 32 | 20 | 46 |
| …omissions? | 21 | 24 | 53 |
| At the end of the break, did the test administrator distribute the Student Questionnaires and give directions as specified in the script? | 56 | 29 | 13 |
| Did the students ask for additional time to complete the questionnaire? | 66 | 26 | 6 |
| At the end of the session, prior to dismissing the students, did the test administrator thank the students for participating in the study? | 86 | 8 | 4 |

Exhibit 9.4 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. Additionally, in nearly 70% of the sessions students were given extra time to complete the Student Questionnaire.

**Exhibit 9.4 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"? | 93 | 2 | 2 | - | 1 |
| When the test administrator ended Session 2, how well did the students comply with the instructions to "stop work"? | 93 | 4 | - | - | 1 |
| | **Exactly** | **Longer** | **Shorter** | **N/A** | |
| How does the total time allocated for the administration of the *Student Questionnaire* compare with the time specified in the script? | 13 | 70 | 6 | 9 | |
| | **Very orderly** | **Somewhat orderly** | **Not orderly at all** | **N/A** | |
| How orderly was the dismissal of students? | 60 | 29 | 5 | 4 | |

### 9.2.3 General Impressions

Section C dealt with the quality control monitors 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 9.5 shows that in almost all sessions, the testing took place without any problems. In roughly 7% of sessions, QCMs reported seeing evidence of students attempting to cheat on the test. When asked to expand on this, QCMs generally indicated that students were either whispering to each other after they were done or were looking around at their neighbors to see whether their test booklets were indeed different. Because the TIMSS test design involves eight 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 deterred by the test design.

Finally, a large proportion of testing sessions had one or more students leave the room for an "emergency" during testing. Typically these emergencies were bathroom breaks. In many of these instances, booklets were not collected from the student; instead, the students left the booklets on their desk.

**Exhibit 9.5 Summary Observations of the QCMs**

| 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? | 93 | 4 | 1 |
| In your opinion, did the test administrator address students' questions appropriately? | 94 | 3 | 1 |
| Did you see any evidence of students attempting to cheat on the tests (e.g., by copying from a neighbor)? | 6 | 90 | 2 |
| Were any defective booklets detected and replaced before the testing began? | - | 95 | 3 |
| Were any defective booklets detected and replaced after the testing began? | - | 97 | 1 |
| If any defective test booklets were replaced, did the test administrator replace them appropriately? | - | 4 | 94 |
| Did any students refuse to take the test either prior to the testing or during the testing? | 9 | 86 | 3 |
| 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)? | 2 | 2 | 94 |
| Did any students leave the room for an "emergency" during the testing? | 43 | 52 | 3 |
| 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)? | 13 | 25 | 60 |

Finally, Exhibit 9.6 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 9.6 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? | 65 | 26 | 6 | - | 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? | 74 | 2 | 8 | 13 | 1 |
| | **Excellent** | **Very good** | **Good** | **Fair** | **Poor** |
| In general, how would you describe the overall quality of the testing session? | 52 | 24 | 12 | 5 | 4 |
| | **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? | 23 | 9 | 5 | 61 | |

### 9.2.4 Interview with the School Coordinator

In Section D of the Classroom Observation Record, the QCMs 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 Westat 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 9.7 show that TIMSS 1999 was an administrative success in the eyes of the school coordinators. In 80% or more of the cases, school coordinators reported that Westat was 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. Of the school coordinators who had administered the Teacher Questionnaire at the time of the interview, about 61% indicated that the estimate was about right, while about 11% reported that the questionnaires took longer and about 28% 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. Students were given special instructions more often than motivational talks or special incentives, and most frequently these were contained in a letter sent home to the students' parents.

**Exhibit 9.7 Interview with the School Coordinator**

| Question | Yes | No | N/A |
|---|---|---|---|
| Was Westat responsive to your questions or concerns? | 86 | 4 | 8 |
| Were you able to collect completed *Teacher Questionnaires* prior to the test administration? | 52 | 39 | 7 |
| It was expected that the *Teacher Questionnaire* would require about 60 minutes to complete. In your opinion, was that estimate correct? | 43 | 8 (longer) 20 (less time) | 27 |
| Were you satisfied with the accommodations (testing room) you were able to arrange for the testing? | 86 | 8 | 4 |
| Did the students receive any special instructions, motivational talk, or incentives to prepare them for the assessment? | 55 | 40 | 3 |
| Were students given any opportunity to practice on questions like those in the tests before the testing session? | 4 | 92 | 2 |
| Is this a complete list of the mathematics classes in this grade in this school? | 85 | 4 | 9 |
| To the best of your knowledge, are there any students in this grade level who are *not* in any of these mathematics classes? | 18 | 76 | 4 |
| To the best of your knowledge, are there any students in this grade level in more than one of these mathematics classes? | 6 | 89 | 3 |
| If there were another TIMSS Benchmarking assessment, would you be willing to serve as a school coordinator? | 92 | 4 | 2 |

Perhaps the biggest tribute to the successful planning and implementation of TIMSS 1999 was the fact that nearly 94% of respondents said that if there were to be another TIMSS Benchmarking assessment, they would be willing to serve as the school coordinator. Furthermore, the results shown in Exhibit 9.8 suggest that practically all of the school coordinators thought the testing sessions went well, and that most thought that staff members in their school felt positive about the TIMSS 1999 testing.

**Exhibit 9.8 Interview with the School Coordinator (continued)**

| Question | Very well | Satisfactory | Unsatisfactory | N/A |
|---|---|---|---|---|
| Overall, how would you say the session went? | 61 | 30 | 1 | 6 |
| | **Positive** | **Neutral** | **Negative** | **N/A** |
| Overall, how would you rate the attitude of the other school staff members towards the TIMSS testing? | 57 | 34 | 4 | 3 |

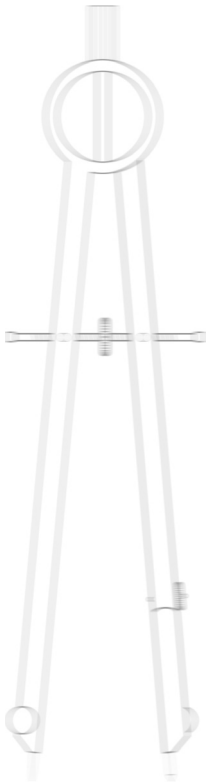
## 9.3 Summary

In summary, the observations by the quality control monitors indicate that the data collected in the TIMSS 1999 Benchmarking study met strict standards for quality, and that as a result there can be a high level of confidence in the findings.

## References

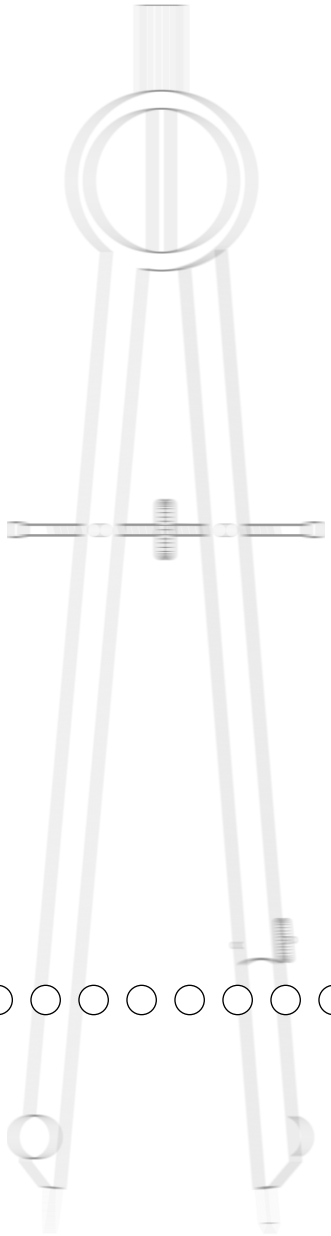
TIMSS (1998). *Manual for quality control monitors* (TIMSS 1999 Doc. Ref. No. 98-0023). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

Westat (1999). *TIMSS 1999 Benchmarking test administrator manual.* Prepared by Westat, Inc. Rockville, MD: Westat, Inc.




# Data Management and Database Construction for TIMSS 1999 Benchmarking

Dirk Hastedt
Oliver Neuschmidt
Eugenio J. Gonzalez



# 10 Data Management and Database Construction for TIMSS 1999 Benchmarking

Dirk Hastedt
Oliver Neuschmidt
Eugenio J. Gonzalez

## 10.1 Overview

The achievement test and questionnaire data from the countries (including the United States) that participated in TIMSS 1999 were processed through a closely cooperative process involving the TIMSS International Study Center at Boston College, the IEA Data Processing Center (DPC), the Educational Testing Service (ETS), 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. This process is described in Hastedt & Gonzalez (2000). The present chapter, which is based on Hastedt & Gonzalez, describes those aspects of the database construction process that were relevant to the Benchmarking database.

In general, data processing for TIMSS consisted of six basic tasks: (1) data entry, (2) creation of the international database, (3) calculation of sampling weights, (4) scaling of achievement data, (5) analysis of the background data, and (6) creation of the exhibits for the TIMSS 1999 reports. This chapter describes the data checking and database creation that was implemented by the DPC, and the steps taken to ensure the quality and accuracy of the Benchmarking database.[1] It discusses the responsibilities of each participant in creating the database; the flow of the data files among the centers involved in the data processing; the structure of the data files submitted by Westat for processing, and the resulting files that are part of the database; the rules, methods, and procedures used for data verification and manipulation; the data products created during data cleaning; and the computer software used in that process.

○○○

1. Data entry for the Benchmarking data is described in Chapter 8. The weighting, scaling, and analysis procedures are described in chapters 5, 13, and 15, respectively.

## 10.2 Data Flow

The data collected in the TIMSS 1999 Benchmarking survey were transferred to data files in the international format by NCS and Westat, as described in chapter 8. The files were then submitted to the DPC for cleaning and verification. The major responsibilities of the DPC were to check that the data files matched the international standard and to make modifications as necessary; to apply standard cleaning rules to the data to verify their consistency and accuracy; to interact with Westat to resolve any problems that arose; to produce summary statistics of the background and achievement data for review by the TIMSS International Study Center; and finally, upon feedback from Westat and the TIMSS International Study Center, to construct the Benchmarking database. The DPC was also responsible for returning validated data files to Westat.

Once the achievement data had been checked for format and internal consistency, they were sent to the International Study Center where basic item statistics were produced and reviewed.[2] The sampling weights, which were produced by Westat, were reviewed by Statistics Canada and the TIMSS International Study Center before being forwarded to ETS for use in scaling the student achievement data. Once the sampling weights and the scaled scores for mathematics and science achievement were verified at the International Study Center, they were sent to the DPC for inclusion in the Benchmarking database. The International Study Center prepared the exhibits for the TIMSS 1999 Benchmarking reports and published the results of the study.

## 10.3 Data Cleaning at the IEA Data Processing Center

Once the Benchmarking data were received from Westat, they were submitted to the DPC for checking before being incorporated into the Benchmarking database. This process is generally referred to by the DPC as data cleaning. The goals of the TIMSS international 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

○○○

2. The item review process is described in chapter 12.

within each participating country. All of the international data-cleaning steps also were applied to the Benchmarking data to ensure that these data achieved the same quality standards as the international data.

Data cleaning involved three main steps. First, all incoming data files were checked, and reformatted as necessary, so that their 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. Thirdly, the distribution of 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. In this third stage, data summary reports were generated for each country. They listed the 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 the case of the Benchmarking participants these reports were sent to Westat, the contractor for field operations, for review.

In data cleaning, two main procedures were used to make any changes to the data that were necessary. Inconsistencies that could be resolved unambiguously were corrected automatically by a program applying standard cleaning routines. Other errors were corrected case by case by the DPC staff. All changes made to the data were recorded in an editing database, so that it was possible to reconstruct the original database received from any country. The three main steps in the data cleaning process are described in more detail below.

### 10.3.1 Standardization of the National File Structure

The first step in data checking at the international level was to verify the compatibility of the datasets received from countries with the international file structure as defined in the TIMSS 1999 international codebook.

Although the TIMSS 1999 codebooks distributed with the data entry software gave clear and detailed instructions about the structure and format of the files to be submitted to the DPC, some countries opted to use other formats that differed from the international standard. For the most part, these differences were due to specific national circumstances. For example, the U.S. added questions to the School, Teacher, and Student Questionnaires. These items had to be added to the codebooks to ensure that they corresponded to the data files.

After the national files were converted into the extended dBase format required by the DPC, the structure of the files was inspected and any deviations from the international file structure identified. A custom-designed program was used to scan the structure of the files for each country to identify 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

At the same time, the data management and tracking forms submitted by each NRC to document such deviations were reviewed. Following these checks, the DPC made any changes necessary to make the files compatible with the international format. In most cases specific programs had to be customized 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 the files no longer needed to correspond directly to the data-collection instruments. At this time also, the Student Background Files and Student Achievement Files were merged into a single file. Variables created during data entry solely for the purpose of cross-checking the data were omitted from all files at this time, and new variables were added (i.e., reporting variables, derived variables, sampling weights, and achievement scores).

### 10.3.2 Cleaning Rules and Procedures

After the data files received from the countries were transformed into the international format, a set of standard checks and cleaning rules were applied.

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.

All problems were recorded in an error database containing one error file for each file that was checked. The cleaning program labelled each problem with an identification number, and provided a description of the problem, and the action taken. 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 DPC 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 the NRC could not solve problems by inspecting the instruments or forms, a general cleaning program was applied.

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

### 10.3.3 National Cleaning Documentation

NRCs 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 documenting data errors and updates, the DPC provided each NRC with new data files that incorporated all agreed updates. In the case of the Benchmarking data, these reports were sent to Westat. 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.

## 10.4 Data Products

Data products sent by the DPC to NRCs included both data almanacs and data files.

### 10.4.1 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 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. These almanacs were also produced for participating states and districts. Example of such data displays are presented in Exhibits 10.1 and 10.2. 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 10.1 Example Data Almanac Display for a Categorical Student Background Variable**

Trends in International Mathematics and Science Study - 1999 Benchmarking Assessment 19:10 Thursday, May 17, 2001 1
Student Background Data Almanac by Mathematics Achievement - 8th Grade

Question : Are you a girl or a boy?
Location : SQ2-02 / SQ2S-02 (BSBGSEX)

| Entity | Sample | Valid N | 1.GIRL % | 2.BOY % | NOT ADMINISTERED % | OMIT % | 1.GIRL Mean | 2.BOY Mean | NOT ADMINISTERED Mean | OMIT Mean |
|---|---|---|---|---|---|---|---|---|---|---|
| Connecticut | 2023 | 2012 | 52.7 | 47.3 | 0.2 | 0.3 | 505.6 | 520.6 | 464.2 | 441.7 |
| Idaho | 1847 | 1832 | 49.6 | 50.4 | 0.5 | 0.3 | 494.2 | 495.7 | 523.7 | 415.0 |
| Illinois | 4781 | 4762 | 50.7 | 49.3 | 0.3 | 0.2 | 504.6 | 515.1 | 474.8 | 412.8 |
| Indiana | 2035 | 2021 | 50.4 | 49.6 | 0.6 | 0.1 | 509.9 | 519.9 | 483.7 | 419.0 |
| Maryland | 3317 | 3297 | 53.2 | 46.8 | 0.3 | 0.2 | 490.9 | 499.3 | 428.7 | 491.7 |
| Massachusetts | 2353 | 2335 | 50.2 | 49.8 | 0.4 | 0.2 | 509.8 | 517.0 | 534.1 | 513.7 |
| Michigan | 2623 | 2596 | 51.2 | 48.8 | 0.5 | 0.3 | 512.3 | 522.4 | 406.6 | 489.6 |
| Missouri | 1979 | 1963 | 51.2 | 48.8 | 0.5 | 0.3 | 488.2 | 492.3 | 437.8 | 434.2 |
| North Carolina | 3089 | 3079 | 53.5 | 46.5 | 0.1 | 0.2 | 493.6 | 497.5 | 428.6 | 438.5 |
| Oregon | 1889 | 1864 | 50.6 | 49.4 | 0.8 | 0.4 | 514.1 | 515.2 | 463.4 | 471.7 |
| Pennsylvania | 3236 | 3207 | 50.1 | 49.9 | 0.6 | 0.2 | 503.1 | 512.0 | 500.0 | 489.1 |
| South Carolina | 2011 | 1996 | 52.2 | 47.8 | 0.5 | 0.1 | 500.8 | 502.9 | 474.9 | 457.6 |
| Texas | 1996 | 1844 | 50.6 | 49.4 | 3.5 | 0.3 | 516.2 | 522.3 | 441.0 | 529.0 |
| Academy School Dist. #20, | 1233 | 1207 | 48.6 | 51.4 | 1.5 | 0.7 | 526.0 | 532.5 | 510.7 | 434.9 |
| Chicago Public Schools, I | 1132 | 1121 | 52.9 | 47.1 | 0.2 | 0.8 | 460.7 | 465.8 | 263.5 | 424.3 |
| Delaware Science Coalitio | 1268 | 1247 | 52.7 | 47.3 | 1.0 | 0.8 | 474.6 | 486.5 | 438.1 | 443.5 |
| First in the World Consor | 750 | 749 | 50.1 | 49.9 | 0.1 | 0.0 | 556.1 | 563.1 | 600.2 | . |
| Fremont/Lincoln/WestSide | 1093 | 1079 | 49.1 | 50.9 | 1.4 | 0.3 | 484.0 | 492.0 | 514.0 | 370.7 |
| Guilford County, NC | 1018 | 1014 | 53.5 | 46.5 | 0.1 | 0.2 | 506.6 | 522.0 | 493.1 | 424.6 |
| Jersey City Public School | 1004 | 994 | 52.9 | 47.1 | 0.9 | 0.2 | 472.6 | 479.8 | 362.3 | 395.4 |
| Miami-Dade County PS, FL | 1229 | 1207 | 49.2 | 50.8 | 0.9 | 0.8 | 420.5 | 424.5 | 391.8 | 300.9 |
| Michigan Invitational Gro | 903 | 895 | 50.3 | 49.7 | 0.5 | 0.5 | 535.0 | 529.5 | 483.0 | 472.4 |
| Montgomery County, MD | 1155 | 1145 | 51.8 | 48.2 | 0.2 | 0.7 | 533.8 | 542.2 | 473.0 | 483.1 |
| Naperville Sch. Dist. #20 | 1212 | 1210 | 51.0 | 49.0 | 0.1 | 0.1 | 566.2 | 572.9 | 318.3 | 495.9 |
| Project SMART Consortium, | 1096 | 1096 | 50.7 | 49.3 | 0.0 | 0.0 | 518.6 | 522.7 | . | . |
| Rochester City Sch. Dist. | 966 | 938 | 52.6 | 47.4 | 2.4 | 0.9 | 439.8 | 452.7 | 414.0 | 365.7 |
| SW Math/Sci. Collaborativ | 1538 | 1518 | 51.2 | 48.8 | 1.3 | 0.0 | 509.2 | 524.3 | 527.5 | . |
| Belgium (Flemish) | 5259 | 5218 | 49.7 | 50.3 | 0.9 | 0.0 | 560.8 | 556.1 | 504.6 | 594.3 |
| Canada | 8770 | 7558 | 50.4 | 49.6 | 0.9 | 12.6 | 531.4 | 535.2 | 513.6 | 514.8 |
| Chinese Taipei | 5772 | 5765 | 50.2 | 49.8 | 0.0 | 0.1 | 583.6 | 587.1 | 567.9 | 401.9 |
| Czech Republic | 3453 | 3448 | 51.5 | 48.5 | 0.2 | 0.0 | 511.8 | 528.5 | 499.2 | . |
| England | 2960 | 2841 | 49.0 | 51.0 | 4.1 | 0.1 | 487.2 | 507.1 | 474.9 | 358.4 |
| Hong Kong, SAR | 5179 | 5098 | 49.4 | 50.6 | 0.4 | 1.1 | 583.7 | 582.3 | 492.1 | 536.3 |
| Italy | 3328 | 3328 | 51.1 | 48.9 | 0.0 | 0.0 | 474.9 | 484.2 | . | . |
| Japan | 4745 | 4686 | 49.5 | 50.5 | 1.3 | 0.1 | 574.8 | 582.5 | 574.0 | 523.3 |
| Korea, Rep. of | 6114 | 6113 | 49.2 | 50.8 | 0.0 | 0.0 | 584.4 | 589.9 | 604.0 | . |
| Netherlands | 2962 | 2883 | 52.3 | 47.7 | 2.5 | 0.2 | 538.3 | 544.2 | 506.7 | 377.5 |
| Russian Federation | 4332 | 4329 | 52.1 | 47.9 | 0.1 | 0.0 | 525.7 | 526.4 | 509.9 | . |
| Singapore | 4966 | 4964 | 48.5 | 51.5 | 0.0 | 0.0 | 603.3 | 605.5 | 551.9 | . |
| United States | 9072 | 8797 | 50.2 | 49.8 | 2.2 | 0.3 | 499.2 | 507.1 | 436.9 | 489.7 |
| International Avg. | 4755 | 4678 | 50.0 | 50.0 | 0.8 | 0.6 | 485.2 | 490.0 | 451.8 | 421.1 |

Trends in International Mathematics and Science Study - 1999 Benchmarking Assessment 19:10 Thursday, May 17, 2001 253
Student Background Data Almanac by Mathematics Achievement - 8th Grade

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

| Entity | Sample | Valid N | N.A. % | Omit % | Mean | Mode | Min | P5 | P10 | Q1 | Median | Q3 | P90 | P95 | Max |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Connecticut | 2023 | 2018 | 0.2 | 0.0 | 14.0 | 14.0 | 9.5 | 13.3 | 13.5 | 13.7 | 14.0 | 14.3 | 14.6 | 14.8 | 17.9 |
| Idaho | 1847 | 1838 | 0.5 | 0.0 | 14.2 | 13.9 | 9.3 | 13.7 | 13.8 | 13.9 | 14.2 | 14.4 | 14.7 | 14.9 | 17.3 |
| Illinois | 4781 | 4773 | 0.3 | 0.0 | 14.2 | 13.9 | 9.3 | 13.7 | 13.8 | 13.9 | 14.2 | 14.5 | 14.8 | 15.0 | 18.3 |
| Indiana | 2035 | 2023 | 0.6 | 0.0 | 14.4 | 14.6 | 9.3 | 13.8 | 13.8 | 14.1 | 14.3 | 14.7 | 15.0 | 15.3 | 17.3 |
| Maryland | 3317 | 3308 | 0.3 | 0.0 | 13.9 | 13.8 | 9.6 | 13.3 | 13.4 | 13.7 | 13.9 | 14.2 | 14.4 | 14.8 | 17.4 |
| Massachusetts | 2353 | 2342 | 0.4 | 0.0 | 14.1 | 14.0 | 9.7 | 13.4 | 13.6 | 13.8 | 14.1 | 14.4 | 14.7 | 14.9 | 18.2 |
| Michigan | 2623 | 2606 | 0.5 | 0.0 | 14.1 | 14.1 | 9.6 | 13.5 | 13.7 | 13.8 | 14.1 | 14.4 | 14.7 | 14.9 | 17.8 |
| Missouri | 1979 | 1969 | 0.5 | 0.0 | 14.3 | 13.9 | 9.3 | 13.7 | 13.8 | 14.0 | 14.3 | 14.6 | 14.9 | 15.2 | 17.9 |
| North Carolina | 3089 | 3084 | 0.1 | 0.0 | 14.2 | 13.9 | 9.4 | 13.5 | 13.6 | 13.8 | 14.1 | 14.4 | 14.8 | 15.2 | 18.3 |
| Oregon | 1889 | 1871 | 0.8 | 0.0 | 14.2 | 14.2 | 9.3 | 13.7 | 13.8 | 13.9 | 14.2 | 14.5 | 14.7 | 14.9 | 18.0 |
| Pennsylvania | 3236 | 3209 | 0.6 | 0.0 | 14.2 | 14.0 | 9.4 | 13.6 | 13.7 | 13.9 | 14.2 | 14.5 | 14.8 | 15.1 | 17.4 |
| South Carolina | 2011 | 2000 | 0.5 | 0.0 | 14.2 | 14.3 | 9.5 | 13.6 | 13.7 | 13.8 | 14.2 | 14.4 | 14.9 | 15.3 | 17.6 |
| Texas | 1996 | 1850 | 3.5 | 0.0 | 14.3 | 14.3 | 12.9 | 13.7 | 13.8 | 14.0 | 14.3 | 14.6 | 15.0 | 15.3 | 18.0 |
| Academy School Dist. #20, | 1233 | 1215 | 1.5 | 0.0 | 14.2 | 14.0 | 9.6 | 13.6 | 13.7 | 13.9 | 14.2 | 14.5 | 14.7 | 14.9 | 16.3 |
| Chicago Public Schools, I | 1132 | 1130 | 0.2 | 0.0 | 14.2 | 13.9 | 9.3 | 13.6 | 13.7 | 13.8 | 14.2 | 14.5 | 14.8 | 15.1 | 16.1 |
| Delaware Science Coalitio | 1268 | 1256 | 1.0 | 0.0 | 14.1 | 13.9 | 9.6 | 13.4 | 13.6 | 13.8 | 14.1 | 14.4 | 14.8 | 15.2 | 17.3 |
| First in the World Consor | 750 | 749 | 0.1 | 0.0 | 14.2 | 14.5 | 12.4 | 13.7 | 13.8 | 13.9 | 14.2 | 14.5 | 14.6 | 14.8 | 15.4 |
| Fremont/Lincoln/WestSide | 1093 | 1082 | 1.4 | 0.0 | 14.2 | 14.1 | 9.3 | 13.6 | 13.7 | 13.8 | 14.1 | 14.4 | 14.7 | 14.9 | 16.8 |
| Guilford County, NC | 1018 | 1016 | 0.1 | 0.0 | 14.2 | 13.8 | 13.1 | 13.6 | 13.7 | 13.8 | 14.2 | 14.5 | 14.8 | 15.3 | 16.4 |
| Jersey City Public School | 1004 | 995 | 1.1 | 0.0 | 14.3 | 13.8 | 9.6 | 13.4 | 13.4 | 13.7 | 14.2 | 14.8 | 15.3 | 15.8 | 18.0 |
| Miami-Dade County PS, FL | 1229 | 1216 | 0.9 | 0.0 | 14.3 | 14.3 | 9.4 | 13.5 | 13.7 | 13.8 | 14.3 | 14.6 | 15.0 | 15.3 | 18.3 |
| Michigan Invitational Gro | 903 | 899 | 0.5 | 0.0 | 14.1 | 14.2 | 9.3 | 13.6 | 13.7 | 13.8 | 14.1 | 14.4 | 14.7 | 14.9 | 17.0 |
| Montgomery County, MD | 1155 | 1153 | 0.2 | 0.0 | 14.0 | 13.8 | 9.8 | 13.4 | 13.5 | 13.7 | 13.9 | 14.3 | 14.5 | 14.8 | 17.4 |
| Naperville Sch. Dist. #20 | 1212 | 1211 | 0.1 | 0.0 | 14.1 | 14.1 | 9.8 | 13.7 | 13.7 | 13.9 | 14.2 | 14.4 | 14.6 | 14.8 | 15.8 |
| Project SMART Consortium, | 1096 | 1096 | 0.0 | 0.0 | 14.2 | 14.1 | 9.5 | 13.6 | 13.8 | 13.9 | 14.2 | 14.5 | 14.8 | 15.1 | 16.3 |
| Rochester City Sch. Dist. | 966 | 947 | 2.4 | 0.0 | 14.2 | 13.8 | 9.3 | 13.4 | 13.5 | 13.8 | 14.1 | 14.7 | 15.1 | 15.3 | 17.0 |
| SW Math/Sci. Collaborativ | 1538 | 1518 | 1.3 | 0.0 | 14.2 | 14.0 | 9.8 | 13.5 | 13.7 | 13.9 | 14.2 | 14.5 | 14.8 | 15.0 | 16.8 |
| Belgium (Flemish) | 5259 | 5221 | 0.9 | 0.0 | 14.1 | 14.0 | 11.6 | 13.4 | 13.5 | 13.8 | 14.0 | 14.3 | 14.9 | 15.3 | 17.1 |
| Canada | 8770 | 8711 | 0.8 | 0.0 | 14.0 | 13.8 | 11.2 | 13.4 | 13.5 | 13.7 | 14.0 | 14.3 | 14.5 | 14.9 | 17.3 |
| Chinese Taipei | 5772 | 5772 | 0.0 | 0.0 | 14.2 | 14.4 | 10.3 | 13.7 | 13.8 | 13.9 | 14.3 | 14.5 | 14.6 | 14.7 | 17.0 |
| Czech Republic | 3453 | 3448 | 0.2 | 0.0 | 14.4 | 14.6 | 13.6 | 13.8 | 13.8 | 14.1 | 14.3 | 14.7 | 15.0 | 15.3 | 16.3 |
| England | 2960 | 2842 | 4.2 | 0.0 | 14.2 | 14.0 | 9.5 | 13.8 | 13.8 | 13.9 | 14.2 | 14.5 | 14.6 | 14.7 | 15.8 |
| Hong Kong, SAR | 5179 | 5156 | 0.4 | 0.0 | 14.2 | 13.6 | 9.7 | 13.4 | 13.5 | 13.7 | 14.0 | 14.4 | 15.3 | 16.2 | 19.2 |
| Italy | 3328 | 3328 | 0.0 | 0.0 | 14.0 | 14.0 | 12.5 | 13.4 | 13.4 | 13.7 | 13.9 | 14.2 | 14.3 | 14.9 | 18.3 |
| Japan | 4745 | 4684 | 1.4 | 0.0 | 14.4 | 14.5 | 13.2 | 13.9 | 14.0 | 14.2 | 14.4 | 14.6 | 14.8 | 14.8 | 16.0 |
| Korea, Rep. of | 6114 | 6113 | 0.0 | 0.0 | 14.4 | 14.1 | 11.3 | 14.0 | 14.0 | 14.2 | 14.4 | 14.7 | 14.8 | 14.9 | 17.3 |
| Netherlands | 2962 | 2890 | 2.4 | 0.0 | 14.2 | 14.0 | 12.6 | 13.5 | 13.6 | 13.8 | 14.2 | 14.5 | 15.0 | 15.3 | 17.8 |
| Russian Federation | 4332 | 4328 | 0.1 | 0.0 | 14.1 | 13.9 | 12.3 | 13.4 | 13.6 | 13.8 | 14.0 | 14.3 | 14.7 | 15.1 | 18.0 |
| Singapore | 4966 | 4966 | 0.0 | 0.0 | 14.4 | 14.0 | 13.0 | 13.8 | 13.9 | 14.0 | 14.3 | 14.6 | 14.8 | 15.1 | 18.8 |
| United States | 9072 | 8776 | 3.1 | 0.0 | 14.2 | 14.0 | 9.3 | 13.5 | 13.7 | 13.8 | 14.2 | 14.4 | 14.8 | 15.1 | 18.3 |
| International Avg. | 4755 | 4692 | 1.3 | 0.0 | 14.4 | 14.2 | 11.6 | 13.6 | 13.7 | 13.9 | 14.3 | 14.7 | 15.1 | 15.5 | 18.6 |

### 10.4.2 Versions of the National Data Files

Building the international database was an iterative process. The DPC 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 DPC 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.

### 10.4.3 Reports

Several reports were produced during data processing at the DPC 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.

## 10.5 Computer Software

The standard database program for handling the incoming data was dBase. 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 DPC used the QUEST software (Adams and Khoo, 1993).

The main programs that were developed by the DPC 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.

### 10.5.1 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 DPC'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 developed using CLIPPER in order to manipulate DataEntryManager files.

### 10.5.2 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 DPC.

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 errors 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.

### 10.5.3 Programs Creating Meta Databases

Using SAS, several programs were developed by the DPC 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).

### 10.5.4 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.

## 10.6 Summary

The structures and procedures designed for processing the TIMSS 1999 data were applied to the Benchmarking data files to ensure that the Benchmarking data conformed to the same format and quality standards as the TIMSS international data.

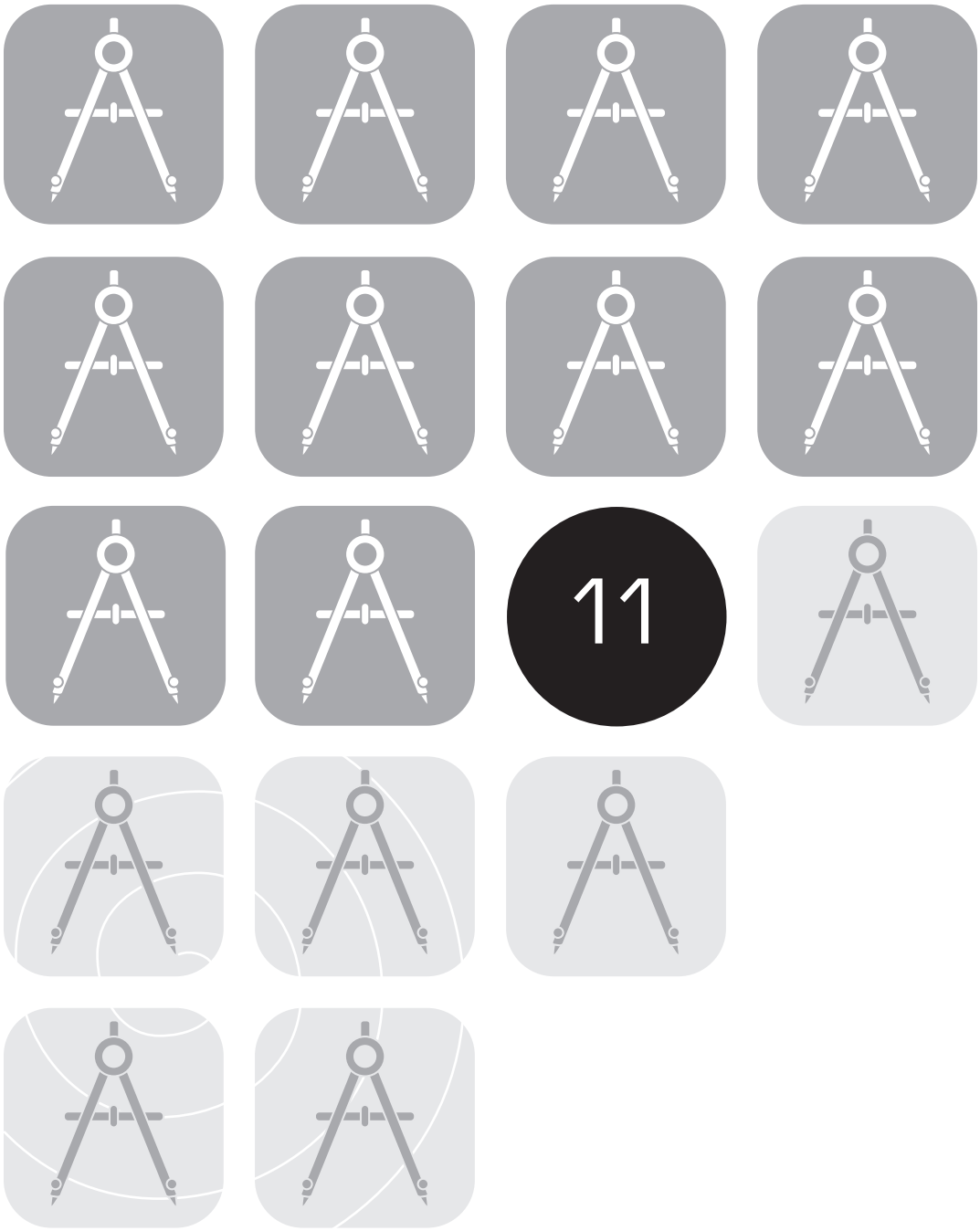
## References

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

Hastedt, D., & Gonzalez, E.J. (2000). Data management and database construction. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS 1999 technical report* (pp. 171-188). Chestnut Hill, MA: Boston College.

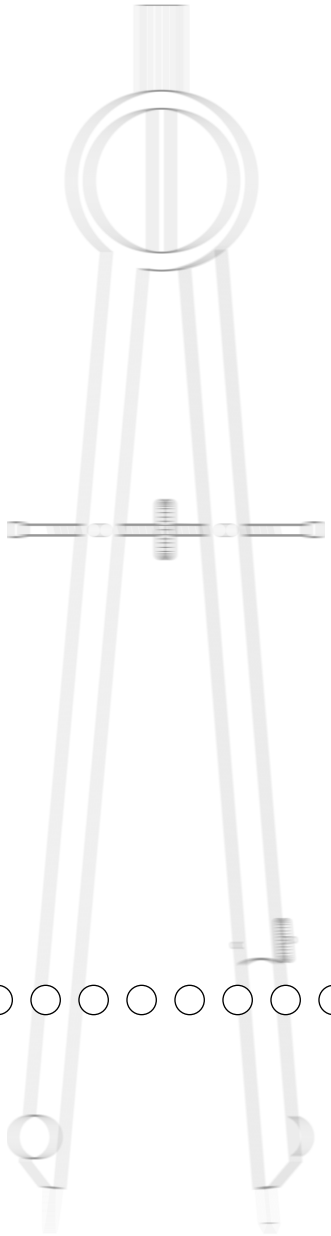
Gonzalez, E.J., & Smith, T.A., (Eds.). (1997). *User guide for the TIMSS international database: Primary and middle school years – 1995 assessment.* Chestnut Hill, MA: Boston College.

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.




# Estimation of Sampling and Imputation Variance for TIMSS 1999 Benchmarking

Eugenio J. Gonzalez
Pierre Foy



# 11 Estimation of Sampling and Imputation Variance for TIMSS 1999 Benchmarking[1]

Eugenio J. Gonzalez
Pierre Foy

## 11.1 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 and from the Benchmarking jurisdictions.[2] 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 and Benchmarking Reports (Martin et al., 2000; Martin et al., 2001; Mullis et al., 2000; Mullis et al., 2001) 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.[3] 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 and Benchmarking Reports the imputation error for each variable has been combined with the sampling error for that variable to provide a standard error incorporating both.

○○○

1. This chapter is based on Gonzalez & Foy (2000) from the international technical report for TIMSS 1999 (Martin, Gregory, and Stemler, 2000).
2. The TIMSS sample design is presented in chapters 5 & 6.
3. Details of the TIMSS test design can be found in chapter 2.

## 11.2 Estimating Sampling Variance

The TIMSS 1999 sampling design applied to the problem of selecting student samples a stratified multistage cluster-sampling technique that permitted efficient and economical data collection while working with schools and classes. This design capitalized effectively on the structure of the student population (i.e., students grouped in classes within schools) but complicated the task of computing standard errors to quantify sampling variability.

When sampling involves multistage cluster techniques, sampling errors can be estimated in several ways 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 and the Benchmarking 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 member of each pair of schools to have its contribution doubled and that of the other member zeroed, thus forming 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 replicate sample and the original sample is the jackknife estimate of the sampling error of the statistic.

### 11.2.1 Construction of Sampling Zones for Sampling Variance Estimation

To apply the JRR technique, the sampled schools had to be paired and assigned to a series of groups known as sampling zones. For the TIMSS 1999 countries, 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. For the Benchmarking data, the sampling zones were constructed by Westat, as part of

their data collection activities. 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. Among the Benchmarking jurisdictions, the number of zones was often less than 75.

In general, 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 11.1 shows the range of sampling zones used in each country and Benchmarking jurisdiction.

### 11.2.2 Computing Sampling Variance Using the JRR Method

The JRR algorithm used in TIMSS 1999 assumes that there are $H$ sampling zones within each country or jurisdiction, each containing two sampled schools selected independently. To compute a statistic $t$ from the sample for a country or jurisdiction, 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 or jurisdiction. 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.

**Exhibit 11.1 Range of Sampling Zones**

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

| States | Zones |
|---|---|
| Connecticut | 26 |
| Idaho | 25 |
| Illinois | 75 |
| Indiana | 26 |
| Maryland | 40 |
| Massachusetts | 28 |
| Michigan | 28 |
| Missouri | 25 |
| North Carolina | 47 |
| Oregon | 22 |
| Pennsylvania | 39 |
| South Carolina | 24 |
| Texas | 26 |

| Districts and Consortia | Zones |
|---|---|
| Academy School Dist. #20, CO | 49 |
| Chicago Public Schools, IL | 13 |
| Delaware Science Coalition, DE | 25 |
| First in the World Consort., IL | 15 |
| Fremont/Lincoln/WestSide PS, NE | 43 |
| Guilford County, NC | 21 |
| Jersey City Public Schools, NJ | 35 |
| Miami-Dade County PS, FL | 12 |
| Michigan Invitational Group, MI | 24 |
| Montgomery County, MD | 16 |
| Naperville Sch. Dist. #203, IL | 34 |
| Project SMART Consortium, OH | 24 |
| Rochester City Sch. Dist., NY | 24 |
| SW Math/Sci. Collaborative, PA | 19 |

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 or Benchmarking jurisdiction: 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 done by creating replicate weights that were then used in the calculations. This approach requires the user to 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, 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 weight, which is simply the product of the final school weight, the appropriate final classroom weight, and the appropriate final student weight, as described in chapters 5 and 6.

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 and jurisdiction regardless of the number of actual zones within that country or jurisdiction. 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.

Standard errors presented in the TIMSS 1999 International and Benchmarking Reports were computed using SAS programs developed at the International Study Center. They were then verified against results produced by the WesVarPC software (Westat, 1997) as an additional quality control check.

## 11.3 Estimating Imputation Variance

The general procedure for estimating the imputation variance using plausible values is discussed by Mislevy, Beaton, Kaplan, & Sheehan (1992) and is summarized here. 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 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.

## 11.4 Combining Sampling and Imputation Variance

When reporting standard errors for proficiency estimates using plausible values, it is 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 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. The error variance component for a statistic was computed using the following formula:

$$Var(t_{pv}) = Var_{jrr}(t_1) + Var_{imp}$$

where $Var_{jrr}(t_1)$ is the sampling variance for the first plausible value. The User Guide for the TIMSS 1999 International Database (Gonzalez & Miles, 2001) contains programs in SAS and SPSS that compute each of these variance components for the TIMSS 1999 data.

Exhibits 11.2 through 11.14 show for the TIMSS countries and the Benchmarking jurisdictions 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 11.2 Summary Statistics and Standard Errors for Overall 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 11.2 (continued) Summary Statistics and Standard Errors for Overall Mathematics Proficiency**

| Participants | Sample Size | Mean Proficiency[a] | Standard Deviation[a] | Jackknife Sampling Error | Overall Standard Error[b] |
|---|---|---|---|---|---|
| **States** | | | | | |
| Connecticut | 2023 | 512 | 85 | 9.0 | 9.1 |
| Idaho | 1847 | 495 | 82 | 7.1 | 7.4 |
| Illinois | 4781 | 509 | 82 | 6.6 | 6.7 |
| Indiana | 2046 | 515 | 76 | 7.1 | 7.2 |
| Maryland | 3317 | 495 | 88 | 6.2 | 6.2 |
| Massachusetts | 2353 | 513 | 82 | 5.8 | 5.9 |
| Michigan | 2623 | 517 | 81 | 7.4 | 7.5 |
| Missouri | 1979 | 490 | 77 | 5.0 | 5.3 |
| North Carolina | 3089 | 495 | 84 | 6.7 | 7.0 |
| Oregon | 1889 | 514 | 83 | 5.8 | 6.0 |
| Pennsylvania | 3236 | 507 | 82 | 6.1 | 6.3 |
| South Carolina | 2011 | 502 | 90 | 7.3 | 7.4 |
| Texas | 1996 | 516 | 90 | 9.0 | 9.1 |
| **Districts and Consortia** | | | | | |
| Academy School Dist. #20, CO | 1233 | 528 | 74 | 1.3 | 1.8 |
| Chicago Public Schools, IL | 1132 | 462 | 76 | 6.0 | 6.1 |
| Delaware Science Coalition, DE | 1268 | 479 | 90 | 8.9 | 8.9 |
| First in the World Consort., IL | 750 | 560 | 77 | 5.5 | 5.8 |
| Fremont/Lincoln/WestSide PS, NE | 1093 | 488 | 89 | 8.0 | 8.2 |
| Guilford County, NC | 1018 | 514 | 85 | 7.7 | 7.7 |
| Jersey City Public Schools, NJ | 1004 | 475 | 87 | 8.6 | 8.6 |
| Miami-Dade County PS, FL | 1229 | 421 | 99 | 9.4 | 9.4 |
| Michigan Invitational Group, MI | 903 | 532 | 73 | 5.8 | 5.8 |
| Montgomery County, MD | 1155 | 537 | 86 | 3.2 | 3.5 |
| Naperville Sch. Dist. #203, IL | 1212 | 569 | 69 | 2.6 | 2.8 |
| Project SMART Consortium, OH | 1096 | 521 | 77 | 7.4 | 7.5 |
| Rochester City Sch. Dist., NY | 966 | 444 | 82 | 6.1 | 6.5 |
| SW Math/Sci. Collaborative, PA | 1538 | 517 | 82 | 7.5 | 7.5 |

a. Average across the five plausible values.
b. Includes error due to sampling and imputation.

**Exhibit 11.3 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 11.3 (continued) Summary Statistics and Standard Errors for Geometry Proficiency**

| Participants | Sample Size | Mean Proficiency[a] | Standard Deviation[a] | Jackknife Sampling Error | Overall Standard Error[b] |
|---|---|---|---|---|---|
| **States** | | | | | |
| Connecticut | 2023 | 470 | 97 | 6.2 | 7.7 |
| Idaho | 1847 | 465 | 91 | 6.0 | 6.5 |
| Illinois | 4781 | 483 | 89 | 4.5 | 6.8 |
| Indiana | 2046 | 476 | 92 | 5.9 | 7.6 |
| Maryland | 3317 | 466 | 98 | 4.3 | 6.0 |
| Massachusetts | 2353 | 477 | 90 | 4.7 | 6.1 |
| Michigan | 2623 | 486 | 93 | 6.2 | 8.0 |
| Missouri | 1979 | 466 | 86 | 3.8 | 5.6 |
| North Carolina | 3089 | 475 | 89 | 5.1 | 5.6 |
| Oregon | 1889 | 486 | 92 | 5.0 | 6.8 |
| Pennsylvania | 3236 | 473 | 91 | 3.6 | 4.7 |
| South Carolina | 2011 | 476 | 97 | 6.5 | 7.8 |
| Texas | 1996 | 486 | 89 | 6.7 | 7.9 |
| **Districts and Consortia** | | | | | |
| Academy School Dist. #20, CO | 1233 | 499 | 93 | 2.2 | 5.0 |
| Chicago Public Schools, IL | 1132 | 457 | 92 | 4.5 | 6.4 |
| Delaware Science Coalition, DE | 1268 | 457 | 96 | 6.0 | 6.2 |
| First in the World Consort., IL | 750 | 519 | 99 | 8.5 | 8.6 |
| Fremont/Lincoln/WestSide PS, NE | 1093 | 467 | 92 | 5.1 | 5.6 |
| Guilford County, NC | 1018 | 491 | 95 | 5.9 | 7.5 |
| Jersey City Public Schools, NJ | 1004 | 458 | 94 | 6.0 | 7.6 |
| Miami-Dade County PS, FL | 1229 | 423 | 98 | 6.5 | 7.8 |
| Michigan Invitational Group, MI | 903 | 495 | 94 | 5.8 | 8.3 |
| Montgomery County, MD | 1155 | 501 | 95 | 3.3 | 4.5 |
| Naperville Sch. Dist. #203, IL | 1212 | 528 | 87 | 3.6 | 4.2 |
| Project SMART Consortium, OH | 1096 | 477 | 96 | 7.3 | 8.1 |
| Rochester City Sch. Dist., NY | 966 | 433 | 100 | 6.1 | 6.3 |
| SW Math/Sci. Collaborative, PA | 1538 | 482 | 97 | 6.4 | 8.9 |

a. Average across the five plausible values.
b. Includes error due to sampling and imputation.

**Exhibit 11.4 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 11.4 (continued) Summary Statistics and Standard Errors for Data Representation, Analysis and Probability Proficiency**

| Participants | Sample Size | Mean Proficiency[a] | Standard Deviation[a] | Jackknife Sampling Error | Overall Standard Error[b] |
|---|---|---|---|---|---|
| **States** | | | | | |
| Connecticut | 2023 | 516 | 105 | 8.2 | 9.9 |
| Idaho | 1847 | 501 | 102 | 6.1 | 7.2 |
| Illinois | 4781 | 510 | 98 | 6.3 | 7.1 |
| Indiana | 2046 | 518 | 95 | 5.5 | 6.3 |
| Maryland | 3317 | 504 | 99 | 5.5 | 6.4 |
| Massachusetts | 2353 | 521 | 102 | 6.0 | 6.3 |
| Michigan | 2623 | 517 | 101 | 6.7 | 6.8 |
| Missouri | 1979 | 500 | 96 | 4.4 | 5.0 |
| North Carolina | 3089 | 502 | 101 | 5.4 | 5.8 |
| Oregon | 1889 | 516 | 97 | 6.2 | 7.0 |
| Pennsylvania | 3236 | 510 | 99 | 7.2 | 8.6 |
| South Carolina | 2011 | 507 | 105 | 6.3 | 7.5 |
| Texas | 1996 | 527 | 111 | 9.5 | 10.2 |
| **Districts and Consortia** | | | | | |
| Academy School Dist. #20, CO | 1233 | 527 | 98 | 2.4 | 4.1 |
| Chicago Public Schools, IL | 1132 | 472 | 93 | 6.9 | 7.2 |
| Delaware Science Coalition, DE | 1268 | 493 | 107 | 9.3 | 9.7 |
| First in the World Consort., IL | 750 | 558 | 96 | 5.2 | 7.3 |
| Fremont/Lincoln/WestSide PS, NE | 1093 | 496 | 106 | 8.9 | 10.8 |
| Guilford County, NC | 1018 | 520 | 105 | 8.3 | 10.1 |
| Jersey City Public Schools, NJ | 1004 | 488 | 109 | 8.5 | 9.6 |
| Miami-Dade County PS, FL | 1229 | 445 | 105 | 8.0 | 9.0 |
| Michigan Invitational Group, MI | 903 | 538 | 104 | 6.5 | 6.9 |
| Montgomery County, MD | 1155 | 541 | 110 | 4.1 | 4.8 |
| Naperville Sch. Dist. #203, IL | 1212 | 559 | 90 | 2.7 | 4.9 |
| Project SMART Consortium, OH | 1096 | 534 | 97 | 7.7 | 8.6 |
| Rochester City Sch. Dist., NY | 966 | 465 | 96 | 4.9 | 6.2 |
| SW Math/Sci. Collaborative, PA | 1538 | 518 | 99 | 5.9 | 6.5 |

a. Average across the five plausible values.
b. Includes error due to sampling and imputation.

**Exhibit 11.5 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 11.5 (continued) Summary Statistics and Standard Errors for Measurement Proficiency**

| Participants | Sample Size | Mean Proficiency[a] | Standard Deviation[a] | Jackknife Sampling Error | Overall Standard Error[b] |
|---|---|---|---|---|---|
| **States** | | | | | |
| Connecticut | 2023 | 493 | 89 | 8.0 | 8.3 |
| Idaho | 1847 | 482 | 87 | 6.9 | 8.1 |
| Illinois | 4781 | 491 | 86 | 6.1 | 6.3 |
| Indiana | 2046 | 489 | 88 | 6.3 | 6.8 |
| Maryland | 3317 | 482 | 89 | 5.5 | 5.9 |
| Massachusetts | 2353 | 491 | 89 | 6.0 | 7.0 |
| Michigan | 2623 | 494 | 85 | 6.7 | 7.4 |
| Missouri | 1979 | 474 | 86 | 5.5 | 6.3 |
| North Carolina | 3089 | 472 | 94 | 7.0 | 7.5 |
| Oregon | 1889 | 500 | 90 | 6.0 | 6.3 |
| Pennsylvania | 3236 | 489 | 88 | 5.7 | 6.0 |
| South Carolina | 2011 | 475 | 95 | 6.7 | 7.1 |
| Texas | 1996 | 489 | 99 | 9.1 | 9.1 |
| **Districts and Consortia** | | | | | |
| Academy School Dist. #20, CO | 1233 | 507 | 85 | 2.3 | 3.5 |
| Chicago Public Schools, IL | 1132 | 439 | 90 | 7.7 | 8.1 |
| Delaware Science Coalition, DE | 1268 | 459 | 98 | 8.4 | 8.7 |
| First in the World Consort., IL | 750 | 535 | 90 | 5.0 | 5.8 |
| Fremont/Lincoln/WestSide PS, NE | 1093 | 474 | 98 | 8.3 | 8.7 |
| Guilford County, NC | 1018 | 487 | 93 | 6.5 | 7.1 |
| Jersey City Public Schools, NJ | 1004 | 450 | 105 | 8.9 | 9.1 |
| Miami-Dade County PS, FL | 1229 | 407 | 104 | 7.4 | 8.9 |
| Michigan Invitational Group, MI | 903 | 516 | 89 | 5.1 | 5.8 |
| Montgomery County, MD | 1155 | 516 | 92 | 3.7 | 4.3 |
| Naperville Sch. Dist. #203, IL | 1212 | 549 | 80 | 3.0 | 3.4 |
| Project SMART Consortium, OH | 1096 | 498 | 91 | 7.5 | 7.8 |
| Rochester City Sch. Dist., NY | 966 | 417 | 98 | 5.2 | 6.2 |
| SW Math/Sci. Collaborative, PA | 1538 | 495 | 90 | 6.7 | 7.0 |

a. Average across the five plausible values.
b. Includes error due to sampling and imputation.

**Exhibit 11.6 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 11.6 (continued) Summary Statistics and Standard Errors for Algebra Proficiency**

| Participants | Sample Size | Mean Proficiency[a] | Standard Deviation[a] | Jackknife Sampling Error | Overall Standard Error[b] |
|---|---|---|---|---|---|
| **States** | | | | | |
| Connecticut | 2023 | 513 | 84 | 8.0 | 8.2 |
| Idaho | 1847 | 500 | 83 | 6.5 | 7.3 |
| Illinois | 4781 | 513 | 84 | 5.5 | 5.7 |
| Indiana | 2046 | 515 | 78 | 6.4 | 6.5 |
| Maryland | 3317 | 499 | 89 | 5.6 | 6.4 |
| Massachusetts | 2353 | 521 | 84 | 5.5 | 5.6 |
| Michigan | 2623 | 520 | 82 | 5.9 | 6.0 |
| Missouri | 1979 | 494 | 81 | 4.6 | 4.9 |
| North Carolina | 3089 | 510 | 80 | 5.4 | 6.1 |
| Oregon | 1889 | 515 | 87 | 5.7 | 6.2 |
| Pennsylvania | 3236 | 511 | 86 | 5.6 | 6.1 |
| South Carolina | 2011 | 511 | 92 | 5.6 | 6.2 |
| Texas | 1996 | 514 | 89 | 7.9 | 8.5 |
| **Districts and Consortia** | | | | | |
| Academy School Dist. #20, CO | 1233 | 532 | 79 | 1.8 | 3.3 |
| Chicago Public Schools, IL | 1132 | 474 | 80 | 6.0 | 6.5 |
| Delaware Science Coalition, DE | 1268 | 497 | 90 | 8.1 | 8.3 |
| First in the World Consort., IL | 750 | 561 | 79 | 5.3 | 5.8 |
| Fremont/Lincoln/WestSide PS, NE | 1093 | 495 | 87 | 6.5 | 6.9 |
| Guilford County, NC | 1018 | 524 | 82 | 6.3 | 6.5 |
| Jersey City Public Schools, NJ | 1004 | 496 | 86 | 7.2 | 7.4 |
| Miami-Dade County PS, FL | 1229 | 452 | 97 | 6.2 | 7.3 |
| Michigan Invitational Group, MI | 903 | 533 | 79 | 6.6 | 7.1 |
| Montgomery County, MD | 1155 | 540 | 93 | 3.4 | 4.7 |
| Naperville Sch. Dist. #203, IL | 1212 | 563 | 77 | 2.8 | 4.0 |
| Project SMART Consortium, OH | 1096 | 521 | 79 | 6.8 | 7.6 |
| Rochester City Sch. Dist., NY | 966 | 466 | 87 | 6.3 | 7.1 |
| SW Math/Sci. Collaborative, PA | 1538 | 519 | 83 | 7.6 | 8.5 |

a. Average across the five plausible values.
b. Includes error due to sampling and imputation.

**Exhibit 11.7 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 11.7 (continued) Summary Statistics and Standard Errors for Fractions and Number Sense Proficiency**

| Participants | Sample Size | Mean Proficiency[a] | Standard Deviation[a] | Jackknife Sampling Error | Overall Standard Error[b] |
|---|---|---|---|---|---|
| **States** | | | | | |
| Connecticut | 2023 | 522 | 84 | 7.5 | 7.9 |
| Idaho | 1847 | 505 | 81 | 6.6 | 6.9 |
| Illinois | 4781 | 516 | 83 | 6.2 | 6.2 |
| Indiana | 2046 | 526 | 79 | 7.4 | 7.6 |
| Maryland | 3317 | 501 | 87 | 5.8 | 5.9 |
| Massachusetts | 2353 | 521 | 85 | 5.9 | 5.9 |
| Michigan | 2623 | 525 | 80 | 6.9 | 7.2 |
| Missouri | 1979 | 497 | 78 | 4.2 | 4.8 |
| North Carolina | 3089 | 497 | 86 | 6.9 | 7.0 |
| Oregon | 1889 | 521 | 78 | 5.9 | 6.2 |
| Pennsylvania | 3236 | 517 | 80 | 5.3 | 5.3 |
| South Carolina | 2011 | 509 | 88 | 6.4 | 7.0 |
| Texas | 1996 | 527 | 87 | 8.6 | 8.9 |
| **Districts and Consortia** | | | | | |
| Academy School Dist. #20, CO | 1233 | 534 | 70 | 1.4 | 2.8 |
| Chicago Public Schools, IL | 1132 | 474 | 79 | 5.9 | 6.1 |
| Delaware Science Coalition, DE | 1268 | 487 | 91 | 7.9 | 8.3 |
| First in the World Consort., IL | 750 | 561 | 77 | 4.6 | 4.9 |
| Fremont/Lincoln/WestSide PS, NE | 1093 | 498 | 89 | 6.3 | 6.4 |
| Guilford County, NC | 1018 | 513 | 87 | 7.0 | 7.3 |
| Jersey City Public Schools, NJ | 1004 | 483 | 84 | 6.8 | 7.3 |
| Miami-Dade County PS, FL | 1229 | 434 | 96 | 8.1 | 9.0 |
| Michigan Invitational Group, MI | 903 | 535 | 71 | 4.4 | 5.1 |
| Montgomery County, MD | 1155 | 540 | 83 | 3.0 | 5.1 |
| Naperville Sch. Dist. #203, IL | 1212 | 569 | 69 | 2.8 | 3.9 |
| Project SMART Consortium, OH | 1096 | 527 | 79 | 7.6 | 7.9 |
| Rochester City Sch. Dist., NY | 966 | 458 | 83 | 5.6 | 5.7 |
| SW Math/Sci. Collaborative, PA | 1538 | 524 | 80 | 6.4 | 6.6 |

a. Average across the five plausible values.
b. Includes error due to sampling and imputation.

**Exhibit 11.8 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 11.8 (continued) Summary Statistics and Standard Errors for Science Proficiency**

| Participants | Sample Size | Mean Proficiency[a] | Standard Deviation[a] | Jackknife Sampling Error | Overall Standard Error[b] |
|---|---|---|---|---|---|
| **States** | | | | | |
| Connecticut | 2023 | 529 | 91 | 10.4 | 10.4 |
| Idaho | 1847 | 526 | 85 | 6.5 | 6.6 |
| Illinois | 4781 | 521 | 89 | 6.4 | 6.5 |
| Indiana | 2046 | 534 | 86 | 6.7 | 7.0 |
| Maryland | 3317 | 506 | 95 | 7.2 | 7.7 |
| Massachusetts | 2353 | 533 | 89 | 7.1 | 7.4 |
| Michigan | 2623 | 544 | 94 | 8.4 | 8.6 |
| Missouri | 1979 | 523 | 89 | 6.1 | 6.5 |
| North Carolina | 3089 | 508 | 90 | 6.2 | 6.5 |
| Oregon | 1889 | 536 | 91 | 5.7 | 6.1 |
| Pennsylvania | 3236 | 529 | 87 | 6.3 | 6.5 |
| South Carolina | 2011 | 511 | 95 | 6.7 | 6.7 |
| Texas | 1996 | 509 | 104 | 10.4 | 10.4 |
| **Districts and Consortia** | | | | | |
| Academy School Dist. #20, CO | 1233 | 559 | 77 | 1.7 | 2.1 |
| Chicago Public Schools, IL | 1132 | 449 | 90 | 9.4 | 9.5 |
| Delaware Science Coalition, DE | 1268 | 500 | 94 | 8.3 | 8.4 |
| First in the World Consort., IL | 750 | 565 | 78 | 4.0 | 5.3 |
| Fremont/Lincoln/WestSide PS, NE | 1093 | 511 | 91 | 4.8 | 5.8 |
| Guilford County, NC | 1018 | 534 | 93 | 7.0 | 7.1 |
| Jersey City Public Schools, NJ | 1004 | 440 | 96 | 9.6 | 9.8 |
| Miami-Dade County PS, FL | 1229 | 426 | 106 | 10.9 | 10.9 |
| Michigan Invitational Group, MI | 903 | 563 | 82 | 5.7 | 6.2 |
| Montgomery County, MD | 1155 | 531 | 92 | 3.5 | 4.3 |
| Naperville Sch. Dist. #203, IL | 1212 | 584 | 76 | 3.6 | 4.1 |
| Project SMART Consortium, OH | 1096 | 539 | 86 | 8.3 | 8.4 |
| Rochester City Sch. Dist., NY | 966 | 452 | 89 | 7.2 | 7.4 |
| SW Math/Sci. Collaborative, PA | 1538 | 543 | 85 | 7.3 | 7.4 |

a. Average across the five plausible values.
b. Includes error due to sampling and imputation.

**Exhibit 11.9 Summary Statistics and Standard Errors for Life Science Proficiency**

| 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 11.9 (continued) Summary Statistics and Standard Errors for Life Science Proficiency**

| Participants | Sample Size | Mean Proficiency[a] | Standard Deviation[a] | Jackknife Sampling Error | Overall Standard Error[b] |
|---|---|---|---|---|---|
| **States** | | | | | |
| Connecticut | 2023 | 533 | 96 | 9.5 | 9.6 |
| Idaho | 1847 | 531 | 98 | 5.7 | 5.7 |
| Illinois | 4781 | 525 | 94 | 6.1 | 6.8 |
| Indiana | 2046 | 539 | 95 | 6.6 | 8.4 |
| Maryland | 3317 | 510 | 99 | 6.1 | 6.8 |
| Massachusetts | 2353 | 531 | 97 | 5.7 | 6.4 |
| Michigan | 2623 | 541 | 100 | 7.6 | 7.6 |
| Missouri | 1979 | 525 | 96 | 5.4 | 6.1 |
| North Carolina | 3089 | 513 | 95 | 5.0 | 5.7 |
| Oregon | 1889 | 541 | 100 | 4.6 | 5.6 |
| Pennsylvania | 3236 | 530 | 97 | 6.9 | 7.6 |
| South Carolina | 2011 | 518 | 99 | 5.6 | 5.7 |
| Texas | 1996 | 513 | 108 | 9.3 | 9.4 |
| **Districts and Consortia** | | | | | |
| Academy School Dist. #20, CO | 1233 | 559 | 93 | 2.2 | 4.6 |
| Chicago Public Schools, IL | 1132 | 471 | 95 | 9.9 | 10.8 |
| Delaware Science Coalition, DE | 1268 | 507 | 101 | 6.9 | 7.5 |
| First in the World Consort., IL | 750 | 567 | 89 | 4.2 | 4.5 |
| Fremont/Lincoln/WestSide PS, NE | 1093 | 524 | 96 | 5.4 | 5.7 |
| Guilford County, NC | 1018 | 532 | 97 | 6.7 | 7.6 |
| Jersey City Public Schools, NJ | 1004 | 457 | 100 | 8.1 | 8.6 |
| Miami-Dade County PS, FL | 1229 | 445 | 109 | 11.6 | 12.7 |
| Michigan Invitational Group, MI | 903 | 558 | 92 | 5.5 | 7.5 |
| Montgomery County, MD | 1155 | 530 | 103 | 4.1 | 5.0 |
| Naperville Sch. Dist. #203, IL | 1212 | 573 | 92 | 3.1 | 3.4 |
| Project SMART Consortium, OH | 1096 | 540 | 95 | 7.7 | 8.3 |
| Rochester City Sch. Dist., NY | 966 | 476 | 100 | 7.9 | 8.7 |
| SW Math/Sci. Collaborative, PA | 1538 | 544 | 98 | 8.5 | 8.6 |

a. Average across the five plausible values.
b. Includes error due to sampling and imputation.

**Exhibit 11.10 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 11.10 (continued) Summary Statistics and Standard Errors for Earth Science Proficiency**

| Participants | Sample Size | Mean Proficiency[a] | Standard Deviation[a] | Jackknife Sampling Error | Overall Standard Error[b] |
|---|---|---|---|---|---|
| **States** | | | | | |
| Connecticut | 2023 | 508 | 93 | 6.1 | 6.5 |
| Idaho | 1847 | 513 | 96 | 5.5 | 6.6 |
| Illinois | 4781 | 505 | 95 | 5.1 | 7.2 |
| Indiana | 2046 | 515 | 92 | 5.8 | 6.3 |
| Maryland | 3317 | 495 | 94 | 4.7 | 6.1 |
| Massachusetts | 2353 | 516 | 95 | 6.6 | 7.6 |
| Michigan | 2623 | 526 | 101 | 7.3 | 7.9 |
| Missouri | 1979 | 511 | 98 | 4.4 | 5.8 |
| North Carolina | 3089 | 500 | 92 | 5.2 | 7.0 |
| Oregon | 1889 | 528 | 97 | 4.8 | 6.7 |
| Pennsylvania | 3236 | 515 | 92 | 5.8 | 6.6 |
| South Carolina | 2011 | 514 | 99 | 6.2 | 6.5 |
| Texas | 1996 | 503 | 99 | 8.0 | 9.4 |
| **Districts and Consortia** | | | | | |
| Academy School Dist. #20, CO | 1233 | 535 | 91 | 2.4 | 3.9 |
| Chicago Public Schools, IL | 1132 | 456 | 86 | 2.7 | 4.1 |
| Delaware Science Coalition, DE | 1268 | 500 | 94 | 7.0 | 7.2 |
| First in the World Consort., IL | 750 | 539 | 94 | 3.6 | 3.8 |
| Fremont/Lincoln/WestSide PS, NE | 1093 | 497 | 91 | 4.0 | 4.6 |
| Guilford County, NC | 1018 | 519 | 95 | 6.3 | 8.0 |
| Jersey City Public Schools, NJ | 1004 | 447 | 85 | 6.1 | 9.3 |
| Miami-Dade County PS, FL | 1229 | 446 | 94 | 8.1 | 9.0 |
| Michigan Invitational Group, MI | 903 | 546 | 94 | 4.4 | 6.5 |
| Montgomery County, MD | 1155 | 518 | 96 | 3.6 | 5.9 |
| Naperville Sch. Dist. #203, IL | 1212 | 554 | 96 | 4.3 | 5.6 |
| Project SMART Consortium, OH | 1096 | 531 | 102 | 7.1 | 7.8 |
| Rochester City Sch. Dist., NY | 966 | 461 | 91 | 4.7 | 5.1 |
| SW Math/Sci. Collaborative, PA | 1538 | 528 | 98 | 6.3 | 6.6 |

a. Average across the five plausible values.
b. Includes error due to sampling and imputation.

**Exhibit 11.11 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 11.11 (continued) Summary Statistics and Standard Errors for Physics Proficiency**

| Participants | Sample Size | Mean Proficiency[a] | Standard Deviation[a] | Jackknife Sampling Error | Overall Standard Error[b] |
|---|---|---|---|---|---|
| **States** | | | | | |
| Connecticut | 2023 | 508 | 93 | 7.6 | 8.0 |
| Idaho | 1847 | 507 | 90 | 5.7 | 7.3 |
| Illinois | 4781 | 506 | 94 | 5.5 | 6.4 |
| Indiana | 2046 | 509 | 90 | 5.5 | 6.4 |
| Maryland | 3317 | 487 | 96 | 6.2 | 7.3 |
| Massachusetts | 2353 | 510 | 90 | 5.0 | 5.8 |
| Michigan | 2623 | 524 | 97 | 6.7 | 6.8 |
| Missouri | 1979 | 506 | 90 | 4.4 | 5.6 |
| North Carolina | 3089 | 487 | 92 | 5.5 | 6.7 |
| Oregon | 1889 | 513 | 96 | 5.6 | 6.9 |
| Pennsylvania | 3236 | 503 | 94 | 5.2 | 6.5 |
| South Carolina | 2011 | 488 | 95 | 5.6 | 6.8 |
| Texas | 1996 | 492 | 97 | 7.6 | 7.9 |
| **Districts and Consortia** | | | | | |
| Academy School Dist. #20, CO | 1233 | 533 | 86 | 2.6 | 5.8 |
| Chicago Public Schools, IL | 1132 | 453 | 94 | 7.2 | 7.6 |
| Delaware Science Coalition, DE | 1268 | 484 | 92 | 6.6 | 7.5 |
| First in the World Consort., IL | 750 | 538 | 90 | 4.8 | 5.7 |
| Fremont/Lincoln/WestSide PS, NE | 1093 | 490 | 99 | 4.8 | 5.2 |
| Guilford County, NC | 1018 | 510 | 94 | 6.4 | 7.5 |
| Jersey City Public Schools, NJ | 1004 | 451 | 98 | 8.0 | 8.2 |
| Miami-Dade County PS, FL | 1229 | 440 | 102 | 8.6 | 9.5 |
| Michigan Invitational Group, MI | 903 | 536 | 96 | 5.6 | 7.1 |
| Montgomery County, MD | 1155 | 514 | 93 | 3.3 | 4.0 |
| Naperville Sch. Dist. #203, IL | 1212 | 557 | 91 | 3.8 | 4.5 |
| Project SMART Consortium, OH | 1096 | 516 | 90 | 6.8 | 7.0 |
| Rochester City Sch. Dist., NY | 966 | 452 | 93 | 5.4 | 6.5 |
| SW Math/Sci. Collaborative, PA | 1538 | 516 | 90 | 6.0 | 7.2 |

a. Average across the five plausible values.
b. Includes error due to sampling and imputation.

## Exhibit 11.12 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 11.12 (continued) Summary Statistics and Standard Errors for Chemistry Proficiency**

| Participants | Sample Size | Mean Proficiency[a] | Standard Deviation[a] | Jackknife Sampling Error | Overall Standard Error[b] |
|---|---|---|---|---|---|
| **States** | | | | | |
| Connecticut | 2023 | 521 | 107 | 8.6 | 9.1 |
| Idaho | 1847 | 518 | 103 | 6.5 | 8.0 |
| Illinois | 4781 | 508 | 104 | 6.9 | 7.1 |
| Indiana | 2046 | 524 | 100 | 5.6 | 7.4 |
| Maryland | 3317 | 498 | 105 | 5.1 | 6.9 |
| Massachusetts | 2353 | 522 | 108 | 7.6 | 7.8 |
| Michigan | 2623 | 537 | 105 | 7.1 | 7.2 |
| Missouri | 1979 | 513 | 108 | 6.3 | 7.1 |
| North Carolina | 3089 | 498 | 104 | 6.1 | 7.8 |
| Oregon | 1889 | 527 | 100 | 4.5 | 7.0 |
| Pennsylvania | 3236 | 516 | 100 | 5.9 | 8.8 |
| South Carolina | 2011 | 502 | 107 | 5.9 | 8.1 |
| Texas | 1996 | 497 | 119 | 10.0 | 10.5 |
| **Districts and Consortia** | | | | | |
| Academy School Dist. #20, CO | 1233 | 551 | 98 | 2.7 | 5.8 |
| Chicago Public Schools, IL | 1132 | 441 | 115 | 10.1 | 10.4 |
| Delaware Science Coalition, DE | 1268 | 495 | 97 | 5.9 | 8.4 |
| First in the World Consort., IL | 750 | 548 | 108 | 5.6 | 6.6 |
| Fremont/Lincoln/WestSide PS, NE | 1093 | 513 | 107 | 4.8 | 6.2 |
| Guilford County, NC | 1018 | 518 | 114 | 7.6 | 8.6 |
| Jersey City Public Schools, NJ | 1004 | 428 | 113 | 7.7 | 8.4 |
| Miami-Dade County PS, FL | 1229 | 436 | 115 | 9.7 | 10.5 |
| Michigan Invitational Group, MI | 903 | 554 | 106 | 8.3 | 9.4 |
| Montgomery County, MD | 1155 | 519 | 102 | 3.3 | 4.2 |
| Naperville Sch. Dist. #203, IL | 1212 | 558 | 93 | 3.2 | 4.5 |
| Project SMART Consortium, OH | 1096 | 534 | 101 | 6.5 | 8.6 |
| Rochester City Sch. Dist., NY | 966 | 453 | 100 | 6.3 | 7.3 |
| SW Math/Sci. Collaborative, PA | 1538 | 537 | 96 | 7.1 | 7.8 |

a. Average across the five plausible values.
b. Includes error due to sampling and imputation.

**Exhibit 11.13 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 11.13 (continued) Summary Statistics and Standard Errors for Scientific Inquiry and the Nature of Science Proficiency**

| Participants | Sample Size | Mean Proficiency[a] | Standard Deviation[a] | Jackknife Sampling Error | Overall Standard Error[b] |
|---|---|---|---|---|---|
| **States** | | | | | |
| Connecticut | 2023 | 533 | 97 | 5.6 | 7.3 |
| Idaho | 1847 | 513 | 100 | 6.1 | 7.1 |
| Illinois | 4781 | 532 | 91 | 6.1 | 8.3 |
| Indiana | 2046 | 527 | 97 | 4.4 | 5.0 |
| Maryland | 3317 | 524 | 93 | 4.8 | 5.4 |
| Massachusetts | 2353 | 542 | 89 | 4.5 | 4.7 |
| Michigan | 2623 | 538 | 98 | 6.3 | 6.8 |
| Missouri | 1979 | 515 | 97 | 3.6 | 4.1 |
| North Carolina | 3089 | 516 | 89 | 4.8 | 5.1 |
| Oregon | 1889 | 525 | 96 | 5.0 | 6.0 |
| Pennsylvania | 3236 | 531 | 88 | 4.3 | 5.4 |
| South Carolina | 2011 | 521 | 101 | 5.2 | 6.7 |
| Texas | 1996 | 514 | 98 | 6.8 | 7.6 |
| **Districts and Consortia** | | | | | |
| Academy School Dist. #20, CO | 1233 | 541 | 94 | 2.4 | 5.1 |
| Chicago Public Schools, IL | 1132 | 491 | 117 | 6.6 | 8.1 |
| Delaware Science Coalition, DE | 1268 | 501 | 109 | 7.1 | 7.3 |
| First in the World Consort., IL | 750 | 574 | 108 | 6.7 | 8.8 |
| Fremont/Lincoln/WestSide PS, NE | 1093 | 511 | 109 | 8.4 | 8.4 |
| Guilford County, NC | 1018 | 533 | 98 | 5.8 | 6.8 |
| Jersey City Public Schools, NJ | 1004 | 492 | 116 | 6.7 | 9.8 |
| Miami-Dade County PS, FL | 1229 | 462 | 118 | 9.0 | 9.4 |
| Michigan Invitational Group, MI | 903 | 545 | 99 | 4.4 | 5.1 |
| Montgomery County, MD | 1155 | 542 | 108 | 3.7 | 4.4 |
| Naperville Sch. Dist. #203, IL | 1212 | 581 | 86 | 2.7 | 3.8 |
| Project SMART Consortium, OH | 1096 | 527 | 101 | 7.0 | 8.7 |
| Rochester City Sch. Dist., NY | 966 | 476 | 115 | 5.1 | 7.9 |
| SW Math/Sci. Collaborative, PA | 1538 | 541 | 94 | 4.7 | 5.9 |

a. Average across the five plausible values.
b. Includes error due to sampling and imputation.

**Exhibit 11.14 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.

**Exhibit 11.14 (continued) Summary Statistics and Standard Errors for Environment and Resources Issues Proficiency**

| Participants | Sample Size | Mean Proficiency[a] | Standard Deviation[a] | Jackknife Sampling Error | Overall Standard Error[b] |
|---|---|---|---|---|---|
| **States** | | | | | |
| Connecticut | 2023 | 515 | 106 | 7.1 | 7.5 |
| Idaho | 1847 | 522 | 102 | 6.0 | 7.1 |
| Illinois | 4781 | 513 | 103 | 4.9 | 6.8 |
| Indiana | 2046 | 527 | 107 | 6.6 | 7.1 |
| Maryland | 3317 | 505 | 107 | 5.8 | 6.4 |
| Massachusetts | 2353 | 522 | 102 | 5.7 | 8.1 |
| Michigan | 2623 | 529 | 105 | 6.0 | 7.5 |
| Missouri | 1979 | 514 | 105 | 6.1 | 7.2 |
| North Carolina | 3089 | 505 | 104 | 5.6 | 7.2 |
| Oregon | 1889 | 520 | 103 | 4.6 | 6.5 |
| Pennsylvania | 3236 | 522 | 105 | 6.7 | 8.3 |
| South Carolina | 2011 | 505 | 108 | 5.1 | 9.1 |
| Texas | 1996 | 502 | 114 | 9.1 | 9.6 |
| **Districts and Consortia** | | | | | |
| Academy School Dist. #20, CO | 1233 | 540 | 100 | 2.6 | 5.7 |
| Chicago Public Schools, IL | 1132 | 442 | 115 | 8.5 | 9.8 |
| Delaware Science Coalition, DE | 1268 | 494 | 119 | 6.8 | 7.3 |
| First in the World Consort., IL | 750 | 549 | 103 | 3.5 | 5.9 |
| Fremont/Lincoln/WestSide PS, NE | 1093 | 508 | 106 | 4.7 | 5.2 |
| Guilford County, NC | 1018 | 531 | 117 | 6.5 | 9.3 |
| Jersey City Public Schools, NJ | 1004 | 451 | 124 | 8.6 | 10.1 |
| Miami-Dade County PS, FL | 1229 | 426 | 123 | 11.3 | 11.9 |
| Michigan Invitational Group, MI | 903 | 550 | 122 | 5.7 | 8.0 |
| Montgomery County, MD | 1155 | 517 | 110 | 4.3 | 6.4 |
| Naperville Sch. Dist. #203, IL | 1212 | 566 | 102 | 2.6 | 6.9 |
| Project SMART Consortium, OH | 1096 | 525 | 114 | 7.3 | 7.8 |
| Rochester City Sch. Dist., NY | 966 | 438 | 118 | 7.5 | 9.6 |
| SW Math/Sci. Collaborative, PA | 1538 | 528 | 100 | 5.5 | 6.8 |

a. Average across the five plausible values.
b. Includes error due to sampling and imputation.

## References

Gonzalez, E.G., & Foy, P. (2000). Estimation of sampling variance. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS 1999 technical report* (pp. 203-224). Chestnut Hill, MA: Boston College.

Gonzalez, E.G., & Miles, J.A. (Eds.). (2001). *TIMSS 1999 international database user guide*. Chestnut Hill, MA: Boston College.

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*. 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.

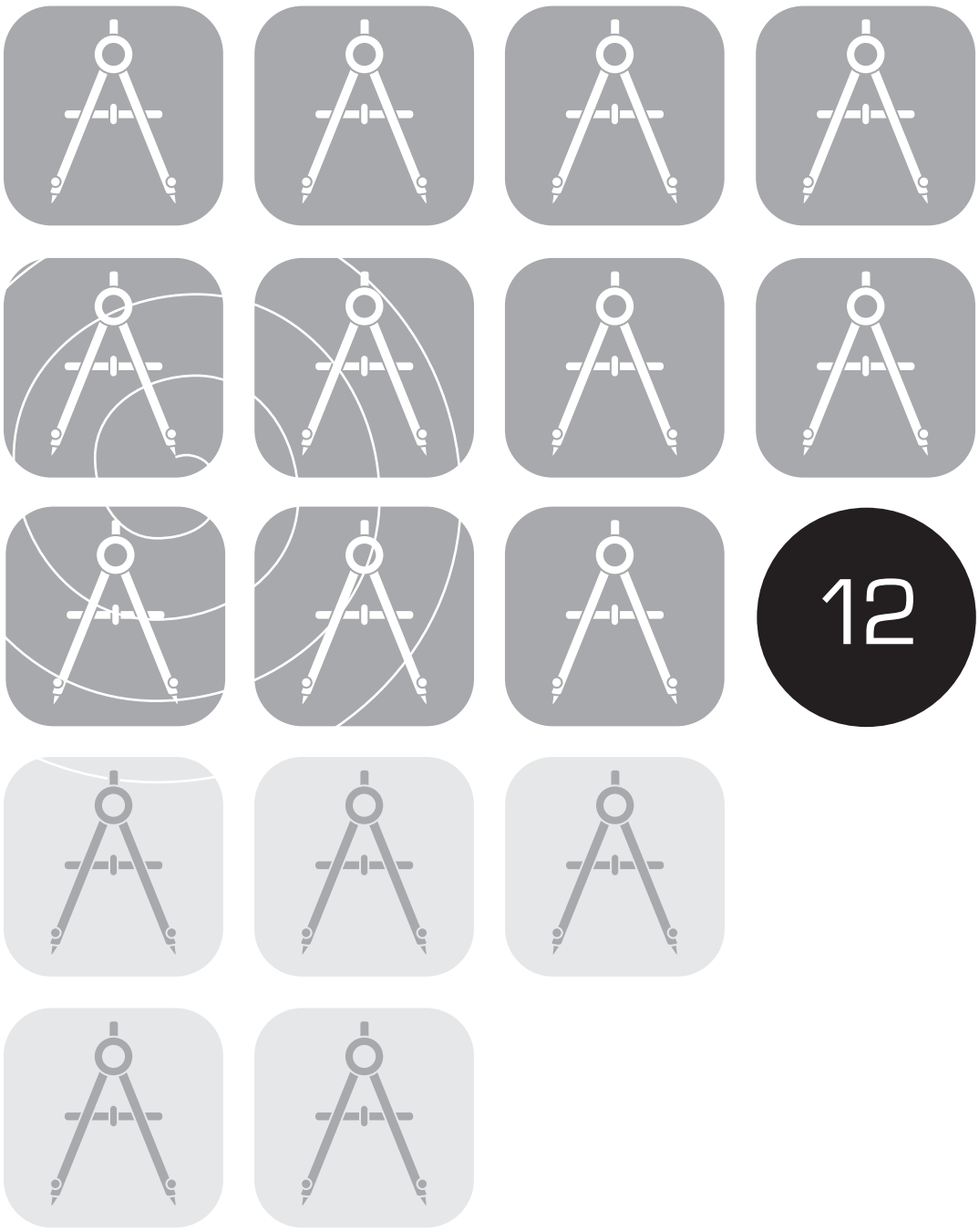
Mislevy, R.J., Beaton, A.E., Kaplan, B., & Sheenan, K.M. (1992). Estimating Population Characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133-161.

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*. 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.

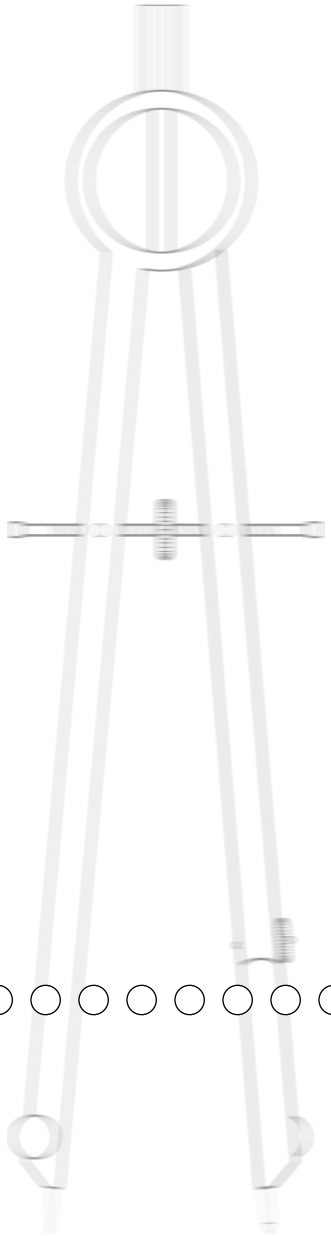
Westat, Inc. (1997). *A user's guide to WesVarPC*. Rockville, MD: Westat, Inc.

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




# Item Analysis and Review for TIMSS 1999 Benchmarking

Ina V.S. Mullis
Michael O. Martin



# 12 Item Analysis and Review for TIMSS 1999 Benchmarking[1]

Ina V.S. Mullis
Michael O. Martin

## 12.1 Overview

In order to assess the psychometric properties of the TIMSS 1999 achievement items before proceeding with Item Response Theory (IRT) scaling,[2] 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 peculiar item behavior. If an item was exceptionally easy or difficult for a particular country, or had unusually low discriminating power, this sometimes suggested a translation or printing problem. For the few such items 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.

To ensure that TIMSS 1999 Benchmarking study met the same exacting standards as TIMSS 1999, all items were subjected to the same review process at the state and district level. This chapter describes the TIMSS 1999 item analysis and review carried out for the Benchmarking study.

## 12.2 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 countries as well as for the 13 states and 14 districts or consortia. Where countries tested in more than one languages, statistics were computed separately for each language group. For each item, the basic item-analysis display presents the number of students that responded in each Benchmarking entity, the difficulty level (i.e., the percentage of students that answered the item correctly), and the discrimination index (i.e., the point-biserial correlation between success

○○○

1. This chapter is based on Mullis & Martin (2000) from the international technical report for TIMSS 1999 (Martin, Gregory, & Stemler, 2000).
2. The TIMSS IRT scaling is described in chapter 13 of this volume.

on the item and total score).[3] For multiple-choice items (see Exhibits 12.1 and 12.2 for examples), 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 Exhibits 12.3 and 12.4 for examples), the display 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 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 (i.e., the mean of the item difficulties across countries) and the international item discrimination (i.e., the mean of the item discriminations).

○○○

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

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 | RDFF | 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.

Third International Mathematics and Science Study - 1999 Main Survey May 17, 2000 47
International Item Statistics (Unweighted) - Review Version
For Internal Review Only: DO NOT CITE OR CIRCULATE

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

| State/District | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Academy School Dist. | 465 | 62.6 | 0.51 | 6.2 | 7.5 | 62.6 | 12.7 | 10.3 | 0.0 | 0.6 | 0.0 | -0.21 | -0.17 | 0.51 | -0.32 | -0.11 | 0.00 | -0.12 | 0.03 | __E_F_ |
| Chicago Public School | 422 | 45.5 | 0.42 | 7.6 | 11.6 | 45.5 | 14.5 | 20.6 | 0.0 | 0.2 | 0.0 | -0.11 | -0.12 | 0.42 | -0.22 | -0.15 | 0.00 | -0.07 | -0.07 | __E_F_ |
| Connecticut | 745 | 59.2 | 0.46 | 5.1 | 7.5 | 59.2 | 16.4 | 11.8 | 0.0 | 0.0 | 0.5 | -0.13 | -0.20 | 0.46 | -0.27 | -0.15 | 0.00 | 0.00 | -0.09 | __E_F_ |
| Delaware Science Coal | 477 | 54.9 | 0.50 | 5.7 | 9.6 | 54.9 | 17.6 | 11.9 | 0.0 | 0.2 | 0.4 | -0.12 | -0.20 | 0.50 | -0.26 | -0.16 | 0.00 | -0.06 | -0.15 | __E_F_ |
| First in the World Co | 474 | 70.0 | 0.50 | 4.2 | 7.8 | 67.3 | 10.3 | 10.3 | 0.0 | 0.0 | 0.0 | -0.19 | -0.23 | 0.49 | -0.22 | -0.21 | 0.00 | 0.00 | 0.02 | __E_F_ |
| Fremont/Lincoln/WestS | 395 | 56.2 | 0.48 | 5.8 | 10.1 | 56.2 | 15.4 | 11.9 | 0.0 | 0.5 | 0.3 | -0.15 | -0.24 | 0.48 | -0.26 | -0.08 | 0.00 | -0.12 | -0.06 | __E_F_ |
| Guilford County, NC | 385 | 62.3 | 0.40 | 5.5 | 11.2 | 62.3 | 11.2 | 9.9 | 0.0 | 0.0 | 0.0 | -0.18 | -0.16 | 0.40 | -0.14 | -0.19 | 0.00 | 0.00 | -0.20 | __E_F_ |
| Idaho | 694 | 55.0 | 0.49 | 4.6 | 11.2 | 55.0 | 14.8 | 14.3 | 0.0 | 0.0 | 0.0 | -0.17 | -0.19 | 0.49 | -0.30 | -0.13 | 0.00 | 0.00 | -0.08 | __E_F_ |
| Illinois | 628 | 60.8 | 0.51 | 5.1 | 10.2 | 57.0 | 13.7 | 13.9 | 0.0 | 0.2 | 0.3 | -0.14 | -0.25 | 0.51 | -0.27 | -0.15 | 0.00 | -0.07 | 0.11 | __E_F_ |
| Indiana | 763 | 57.3 | 0.48 | 3.7 | 11.7 | 57.0 | 12.2 | 15.2 | 0.1 | 0.1 | 0.1 | -0.18 | -0.16 | 0.48 | -0.23 | -0.20 | -0.05 | -0.01 | -0.04 | __E_F_ |
| Jersey City Public Sc | 364 | 58.5 | 0.50 | 5.8 | 10.4 | 58.5 | 11.5 | 12.9 | 0.0 | 0.8 | 0.0 | -0.17 | -0.21 | 0.50 | -0.30 | -0.13 | 0.00 | -0.07 | -0.57 | __E_F_ |
| Maryland | 811 | 58.7 | 0.45 | 8.1 | 11.6 | 52.2 | 14.1 | 13.3 | 0.0 | 0.7 | 0.2 | -0.15 | -0.16 | 0.45 | -0.25 | -0.11 | 0.00 | -0.07 | -0.09 | __E_F_ |
| Massachusetts | 875 | 63.4 | 0.50 | 5.1 | 8.1 | 63.4 | 11.3 | 11.5 | 0.0 | 0.5 | 0.1 | -0.19 | -0.21 | 0.50 | -0.25 | -0.18 | 0.00 | -0.08 | -0.33 | __E_F_ |
| Miami-Dade County PS, | 464 | 43.3 | 0.36 | 8.6 | 13.1 | 43.3 | 17.9 | 16.4 | 0.0 | 0.6 | 0.0 | -0.12 | -0.14 | 0.36 | -0.14 | -0.11 | 0.00 | -0.01 | -0.29 | __E_F_ |
| Michigan | 1000 | 58.4 | 0.47 | 5.5 | 8.7 | 58.6 | 14.2 | 12.7 | 0.0 | 0.3 | 0.0 | -0.16 | -0.17 | 0.47 | -0.24 | -0.18 | 0.00 | -0.05 | -0.10 | __E_F_ |
| Michigan Invitational | 339 | 64.0 | 0.48 | 3.2 | 10.3 | 64.0 | 12.4 | 9.7 | 0.3 | 0.0 | 0.0 | -0.11 | -0.24 | 0.48 | -0.24 | -0.18 | -0.11 | 0.00 | -0.01 | __E_F_ |
| Missouri | 725 | 54.9 | 0.42 | 6.8 | 8.7 | 54.9 | 15.2 | 14.3 | 0.0 | 0.1 | 0.0 | -0.10 | -0.19 | 0.41 | -0.20 | -0.15 | 0.00 | -0.08 | -0.15 | __E_F_ |
| Montgomery County, MD | 427 | 71.2 | 0.51 | 2.8 | 6.8 | 71.2 | 9.6 | 9.1 | 0.0 | 0.5 | 0.5 | -0.13 | -0.20 | 0.51 | -0.30 | -0.21 | 0.00 | -0.04 | -0.26 | __E_F_ |
| Naperville Sch. Dist. | 466 | 74.2 | 0.44 | 3.6 | 5.6 | 74.2 | 7.9 | 8.6 | 0.0 | 0.0 | 0.0 | -0.13 | -0.17 | 0.44 | -0.23 | -0.23 | 0.00 | 0.00 | -0.07 | __E_F_ |
| North Carolina | 767 | 57.5 | 0.47 | 7.2 | 9.5 | 55.1 | 13.3 | 14.1 | 0.0 | 0.8 | 0.0 | -0.12 | -0.20 | 0.47 | -0.22 | -0.17 | 0.00 | -0.08 | -0.20 | __E_F_ |
| Oregon | 694 | 60.8 | 0.51 | 4.3 | 10.4 | 60.8 | 13.5 | 10.5 | 0.0 | 0.4 | 0.3 | -0.14 | -0.21 | 0.51 | -0.27 | -0.19 | 0.00 | -0.05 | -0.16 | __E_F_ |
| Pennsylvania | 641 | 61.2 | 0.51 | 5.5 | 10.0 | 60.2 | 10.9 | 13.3 | 0.0 | 0.2 | 0.2 | -0.22 | -0.20 | 0.51 | -0.20 | -0.22 | 0.00 | -0.06 | -0.19 | __E_F_ |
| Project SMART Consort | 408 | 62.7 | 0.56 | 3.7 | 9.3 | 62.7 | 11.0 | 13.2 | 0.0 | 0.0 | 0.0 | -0.21 | -0.27 | 0.56 | -0.29 | -0.18 | 0.00 | 0.00 | -0.15 | __E_F_ |
| Rochester City Sch. D | 349 | 37.2 | 0.45 | 9.7 | 12.6 | 37.2 | 18.6 | 20.1 | 0.0 | 1.7 | 1.4 | -0.16 | -0.18 | 0.45 | -0.13 | -0.10 | 0.00 | -0.15 | 0.14 | __E_F_ |
| SW Math/Sci. Collabor | 580 | 62.2 | 0.48 | 5.3 | 9.1 | 62.2 | 11.6 | 11.0 | 0.0 | 0.7 | 0.2 | -0.20 | -0.23 | 0.48 | -0.25 | -0.11 | 0.00 | -0.08 | -0.28 | __E_F_ |
| South Carolina | 754 | 54.1 | 0.48 | 6.4 | 9.2 | 54.1 | 14.2 | 15.6 | 0.0 | 0.5 | 0.1 | -0.18 | -0.16 | 0.48 | -0.24 | -0.16 | 0.00 | -0.06 | 0.01 | __E_F_ |
| Texas | 748 | 68.7 | 0.51 | 4.0 | 7.6 | 68.7 | 9.0 | 10.3 | 0.0 | 0.4 | 0.0 | -0.14 | -0.23 | 0.51 | -0.29 | -0.20 | 0.00 | -0.03 | -0.58 | __E_F_ |
| United States | 3353 | 56.4 | 0.45 | 4.9 | 9.9 | 56.4 | 13.5 | 14.9 | 0.0 | 0.4 | 0.1 | -0.13 | -0.18 | 0.45 | -0.24 | -0.17 | -0.01 | -0.03 | -0.14 | __E_F_ |

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.

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 )

| 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 | Reliability Cases | Reliability 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 (Englis | 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 R | 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 (Hebrew | 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. o | 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 (Er | 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.

Third International Mathematics and Science Study - 1999 Main Survey — May 17, 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 )

| State/District | 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 | Reliability Cases | Score | Code | Flags |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Academy School Dist. | 157 | 40.1 | 0.37 | 51.6 | 40.1 | . | . | 8.3 | 0.0 | -0.33 | 0.37 | . | . | -0.06 | 1.03 | . | . | . | __E___ |
| Chicago Public School | 141 | 39.7 | 0.37 | 53.2 | 39.7 | . | . | 7.1 | 0.0 | -0.26 | 0.37 | . | . | -0.21 | 0.09 | . | . | . | __E___ |
| Connecticut | 233 | 33.5 | 0.37 | 57.1 | 33.5 | . | . | 9.4 | 2.5 | -0.27 | 0.37 | . | . | -0.19 | 1.11 | . | . | . | __E___ |
| Delaware Science Coal | 162 | 20.4 | 0.35 | 65.4 | 20.4 | . | . | 14.2 | 0.0 | -0.16 | 0.34 | . | . | -0.18 | 1.57 | . | . | . | __E___ |
| First in the World Co | 150 | 40.7 | 0.45 | 57.3 | 40.7 | . | . | 2.0 | 0.0 | -0.45 | 0.45 | . | . | 0.02 | 1.28 | . | . | . | __E___ |
| Fremont/Lincoln/WestS | 142 | 35.2 | 0.42 | 55.6 | 35.2 | . | . | 9.2 | 0.0 | -0.33 | 0.41 | . | . | -0.12 | 0.83 | . | . | . | __E___ |
| Guilford County, NC | 129 | 34.1 | 0.40 | 58.9 | 34.1 | . | . | 7.0 | 0.0 | -0.41 | 0.40 | . | . | 0.05 | 1.33 | . | . | . | __E___ |
| Idaho | 224 | 41.5 | 0.45 | 50.4 | 41.5 | . | . | 8.0 | 0.0 | -0.30 | 0.45 | . | . | -0.27 | 0.49 | . | . | . | __E___ |
| Illinois | 207 | 41.1 | 0.48 | 50.2 | 41.1 | . | . | 8.7 | 0.0 | -0.34 | 0.47 | . | . | -0.22 | 0.95 | . | . | . | __E___ |
| Indiana | 265 | 46.8 | 0.42 | 49.4 | 46.8 | . | . | 3.8 | 0.0 | -0.36 | 0.42 | . | . | -0.17 | 0.51 | . | . | . | __E___ |
| Jersey City Public Sc | 124 | 16.9 | 0.37 | 71.8 | 16.9 | . | . | 11.3 | 0.8 | -0.29 | 0.37 | . | . | -0.02 | 1.94 | . | . | . | __E___ |
| Maryland | 257 | 28.8 | 0.42 | 59.9 | 28.8 | . | . | 11.3 | 0.8 | -0.31 | 0.42 | . | . | -0.12 | 1.30 | . | . | . | __E___ |
| Massachusetts | 296 | 37.2 | 0.47 | 52.7 | 37.2 | . | . | 10.1 | 0.0 | -0.33 | 0.47 | . | . | -0.21 | 1.05 | . | . | . | __E___ |
| Miami-Dade County PS, | 139 | 18.0 | 0.59 | 66.2 | 18.0 | . | . | 15.8 | 0.0 | -0.26 | 0.59 | . | . | -0.29 | 1.22 | . | . | . | __E___ |
| Michigan | 337 | 28.8 | 0.50 | 61.4 | 28.8 | . | . | 9.8 | 0.9 | -0.29 | 0.50 | . | . | -0.23 | 1.44 | . | . | . | __E___ |
| Michigan Invitational | 116 | 27.6 | 0.36 | 66.4 | 27.6 | . | . | 6.0 | 0.0 | -0.29 | 0.35 | . | . | -0.08 | 1.65 | . | . | . | __E___ |
| Missouri | 252 | 25.0 | 0.34 | 65.9 | 25.0 | . | . | 9.1 | 0.0 | -0.20 | 0.34 | . | . | -0.17 | 1.38 | . | . | . | __E___ |
| Montgomery County, MD | 140 | 40.0 | 0.47 | 52.9 | 40.0 | . | . | 7.1 | 0.0 | -0.39 | 0.47 | . | . | -0.14 | 1.12 | . | . | . | __E___ |
| Naperville Sch. Dist. | 152 | 44.7 | 0.41 | 50.7 | 44.7 | . | . | 4.6 | 0.0 | -0.41 | 0.41 | . | . | 0.00 | 1.62 | . | . | . | __E___ |
| North Carolina | 269 | 27.9 | 0.42 | 65.8 | 27.9 | . | . | 6.3 | 0.0 | -0.29 | 0.42 | . | . | -0.20 | 1.28 | . | . | . | __E___ |
| Oregon | 232 | 31.9 | 0.34 | 56.5 | 31.9 | . | . | 11.6 | 0.0 | -0.18 | 0.33 | . | . | -0.21 | 1.37 | . | . | . | __E___ |
| Pennsylvania | 218 | 37.6 | 0.41 | 56.0 | 37.6 | . | . | 6.4 | 0.0 | -0.32 | 0.41 | . | . | -0.16 | 0.78 | . | . | . | __E___ |
| Project SMART Consort | 132 | 51.5 | 0.40 | 43.9 | 51.5 | . | . | 4.5 | 0.0 | -0.28 | 0.39 | . | . | -0.29 | 0.50 | . | . | . | __E___ |
| Rochester City Sch. D | 124 | 28.2 | 0.49 | 58.1 | 28.2 | . | . | 13.7 | 1.6 | -0.29 | 0.49 | . | . | -0.20 | 0.75 | . | . | . | __E___ |
| SW Math/Sci. Collabor | 192 | 29.2 | 0.39 | 63.0 | 29.2 | . | . | 7.8 | 0.0 | -0.33 | 0.39 | . | . | -0.07 | 1.44 | . | . | . | __E___ |
| South Carolina | 242 | 46.3 | 0.54 | 49.2 | 46.3 | . | . | 4.5 | 0.0 | -0.49 | 0.54 | . | . | -0.11 | 0.45 | . | . | . | __E___ |
| Texas | 247 | 41.7 | 0.51 | 53.4 | 41.7 | . | . | 4.9 | 0.0 | -0.45 | 0.50 | . | . | -0.11 | 0.80 | . | . | . | __E___ |
| United States | 1128 | 37.4 | 0.52 | 55.3 | 37.4 | . | . | 7.3 | 0.1 | -0.41 | 0.52 | . | . | -0.18 | 0.77 | 118 | 99.2 | 97.5 | __E___ |

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 12.1 through 12.4 contain 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.[4]

**Diff:** The item difficulty is 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, items that were not reached were treated as not administered.

**Disc:** The item discrimination is the correlation between a correct answer to the item and the total score on all of the items in the subject area in the test booklet.[5] 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 (Exhibits 12.1 and 12.2), 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 (Exhibits 12.3 and 12.4), these are the percentages 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 is 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.

○○○

4. 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.
5. 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 (i.e., 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 (zero, one, two, or three) 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 (i.e., 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 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
- Differences 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

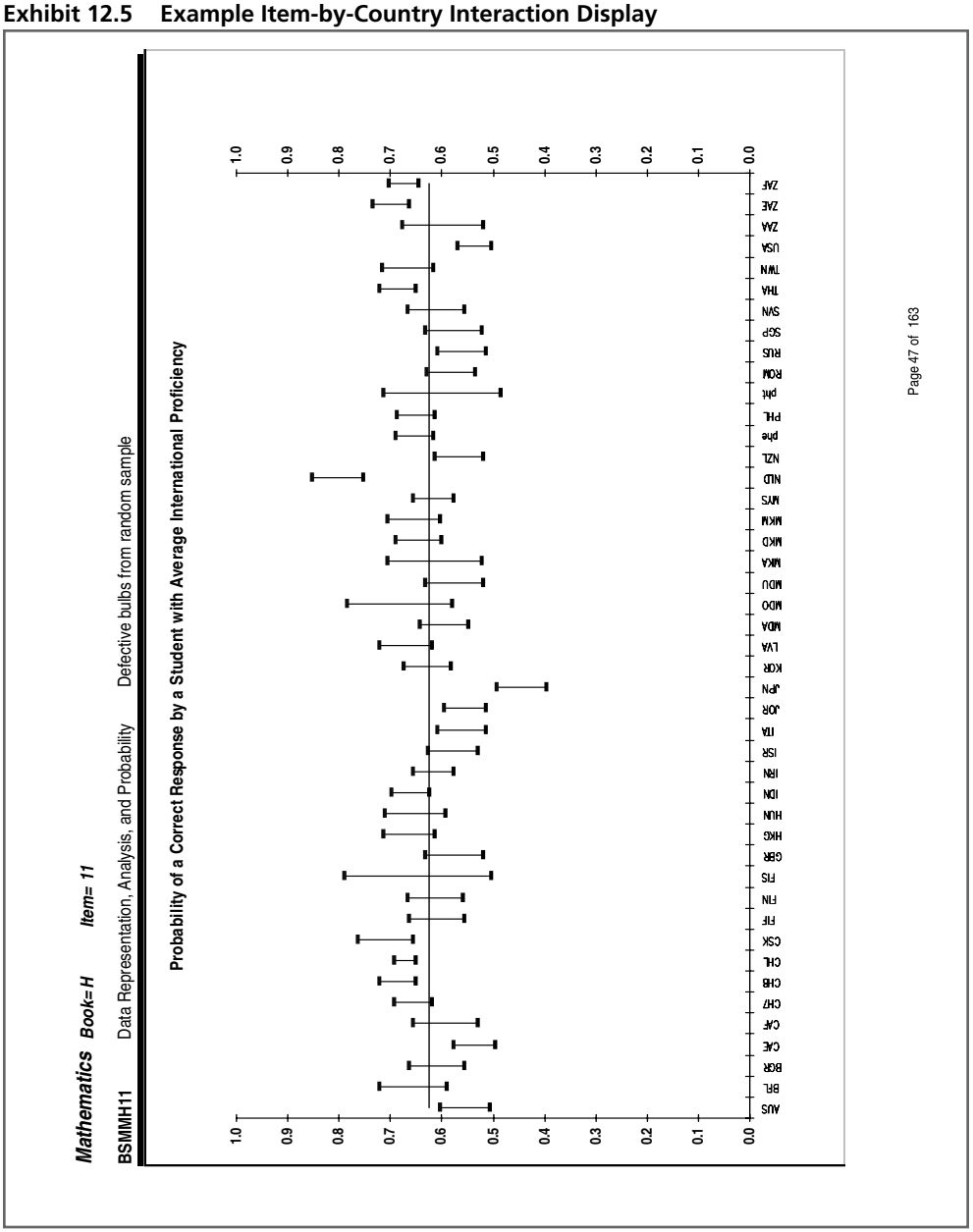
Although not all of these conditions necessarily indicate a problem, the flags are a useful way to draw attention to potential sources of concern. The IEA Data Processing Center also produced information about the inter-rater agreement for the free-response items.

### 12.2.1 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. When this does not occur (i.e., when a high-scoring group has low perfor-

mance on an item on which other groups are doing well) there is said to be an item-by-group interaction. Since large item-by-group interactions can indicate an item that is flawed in some way, the item review also included this aspect of item performance. Item-by-country interactions for the U.S. were assessed using the TIMSS 1999 national sample.

**Exhibit 12.5 Example Item-by-Country Interaction Display**



To help examine item-by-group interactions, the International Study Center produced a graphical display for each item showing the average probability across all groups 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 group (see Exhibit 12.5 for an example). The probability for each group 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 group $i$; $SE_{RDIFFik}$ is the standard error of the difficulty of item $k$ in group $i$; and $Z_b$ is the critical value from the $Z$ distribution, corrected for multiple comparisons using the Bonferroni procedure.

## 12.3 Item Checking Procedures

Before 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.

When the item statistics indicated a problem with an item, the documentation from the translation verification[6] was used as an aid in checking the test booklets and contacting national research coordinators. If a problem could be detected by the

○○○

6. See chapter 5 for a description of the translation and verification of the TIMSS data-collection instruments.

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 translation or cultural issues, however, the national research coordinator was consulted before deciding how the item should be treated. Appendix D of the TIMSS 1999 Technical Report (Martin, Gregory, & Stemler, 2000) provides a list of deleted items as well as a list of recodes made to free-response items. No items were deleted for the United States, nor for any of the states and districts involved in the Benchmarking study.[7]

## 12.4 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. No items were found to be problematic specifically for the states and districts involved in TIMSS 1999 Benchmarking.

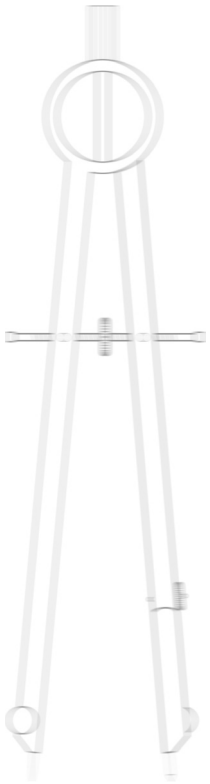
○○○

7. Note, however, that a single mathematics item, M02, was discovered to have an incorrectly drawn figure in the international version. This item was subsequently deleted from all countries, including the United States and the Benchmarking jurisdictions.

## References

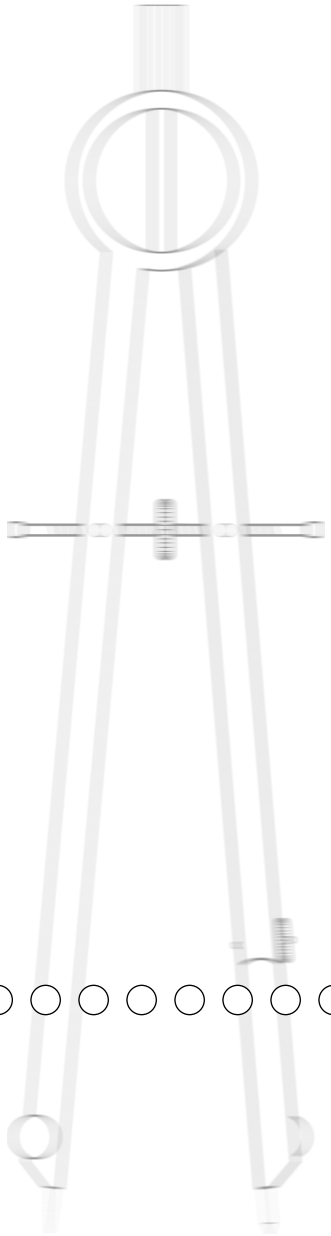
Martin, M.O., Gregory, K.D., & Stemler, S.E. (Eds.). (2000). *TIMSS 1999 technical report.* Chestnut Hill, MA: Boston College.

Mullis, I.V.S., & Martin, M.O. (2000). Item Analysis and Review. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS 1999 technical report* (pp. 225-236). Chestnut Hill, MA: Boston College.




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

Kentaro Yamamoto
Edward Kulick



# 13 Scaling Methodology and Procedures for the TIMSS Mathematics and Science Scales

Kentaro Yamamoto
Edward Kulick

## 13.1 Overview

The TIMSS achievement test design made use of matrix sampling techniques to divide the assessment item pool so that each sampled student responded 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 relied on a sophisticated form of psychometric scaling known as item response theory (IRT) scaling to combine the student responses in a way that provided accurate estimates of achievement. The TIMSS IRT scaling used multiple imputations or "plausible values" 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.

The TIMSS 1999 Benchmarking study used the same scaling and imputation methodology as the TIMSS 1999 International component. This chapter summarizes that methodology; further details can be found in the TIMSS 1999 Technical Report (see Yamamoto & Kulick, 2000).

## 13.2 TIMSS 1999 Benchmarking Scaling Methodology

Three distinct scaling models, depending on item type and scoring procedure, were used in analyzing the Benchmarking assessment data. Each is a *latent variable* model that describes the probability of a specific response to an 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, scored as correct or incorrect.

○○○

1. The TIMSS 1999 achievement test design is described in chapter 2.

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).

### 13.2.1 Three- and Two-Parameter IRT Models for Dichotomous Items

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

$$Pi1(x_i = 1|\theta_k, a_i, b_i, c_i) = c_i + \frac{(1 - c_i)}{1.0 + \exp(-1.7\, a_i(\theta_k - b_i))} \quad (1)$$

where

- $x_i$ is the response to item $i$, 1 if correct and 0 if incorrect;
- $\theta_k$ is the proficiency of a person on a scale $k$;
- $a_i$ 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.

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

$$P_{i0} \equiv P(x_i = 0|\theta_k, a_i, b_i, c_i) = 1 - P_{i1}(\theta_k) \quad . \quad (2)$$

The two-parameter logistic (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.

In scaling the Benchmarking data, the three- and two-parameter models were used in preference to the one-parameter Rasch model, 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, and the 3PL model has a parameter that can be used to model guessing behavior among low-ability students.

Modeling item response functions as accurately as possible by using 2PL and 3PL models also reduces errors due to model mis-specification. 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 generally preferable to use models that characterize the item response function as well as possible.

### 13.2.2 The IRT Model for Polytomous Items

Free-response items requiring an extended response were scored for partial credit, with zero, one, and two 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 $\theta_k$ 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;

$\theta_k$ is the proficiency of a person on scale $k$;

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

$b_i$ is the location parameter of item $i$, characterizing its difficulty;

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

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

(4)

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

## 13.3 Item Parameter Estimation

For all of the IRT models, there is a linear indeterminacy between the values of item parameters and proficiency parameters; that is, 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.

IRT modeling relies on a number of assumptions, the most important being conditional independence. Under this assumption, item response probabilities depend only on $\theta_\kappa$ (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. The joint probability of a particular response pattern $x$ across a set of $n$ items is then given by:

(5)
$$P(x|\theta_k, \textit{item parameters}) = \prod_{i=1}^{n} \prod_{l=0}^{m_i - 1} P_{il}(\theta_k)^{u_{il}}$$

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

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 Benchmarking 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 $\theta_\kappa$ was induced from student responses to the calibrated items. This likelihood function for the proficiency $\theta_k$ is called the posterior distribution of the $\theta$s for each respondent.

### 13.3.1 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 13.1 and 13.2). 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 13.1 TIMSS 1999 Grade 8 Science Assessment Example Item Response Function—Dichotomous Item**

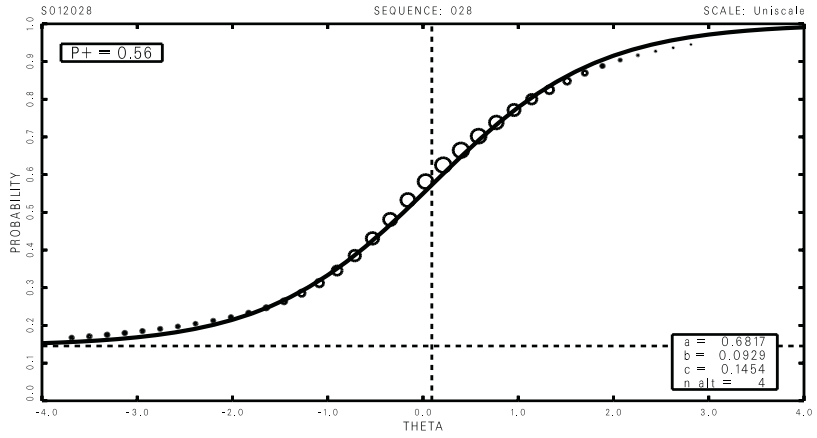


**Exhibit 13.2 TIMSS 1999 Grade 8 Science Assessment Example Item Response Function—Polytomous Item**

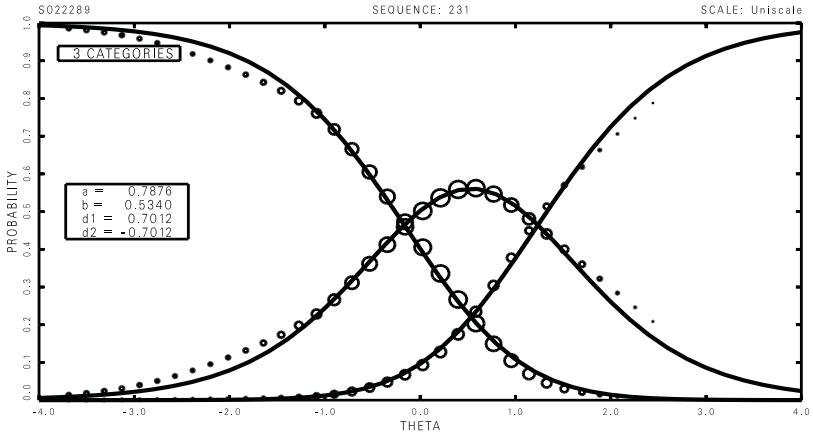



Exhibit 13.1 shows a plot of the empirical and theoretical item response functions for a dichotomous item. The horizontal axis represents the proficiency scale, and the vertical axis represents the probability of a correct response. The solid curve is the theoretical curve based on the estimated item 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 13.2 shows a plot of the empirical and theoretical item response functions for a polytomous item. Again, 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 13.1. For items where the model fits the data well, the empirical and theoretical curves are close together.

## 13.4 Scaling Mathematics and Science Domains and Content Areas

In order to estimate student proficiency scores 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 distribu-

tion. This was found to be a more reliable way to compare proficiency across countries than to make a scale for each content area, such as algebra, geometry, etc., and then form a composite measure of mathematics by combining the content area scales.

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.

### 13.4.1 Omitted and Not-Reached Responses.

Apart from data that by design were not administered to a student, missing data could also occur when 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 as not having 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 1999 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.

### 13.4.2 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 $\theta$ in such tests is negligible, the distribution of $\theta$ or the joint distribution of $\theta$ with other variables can be approximated using individual $\theta$s.

For the distribution of proficiencies in large populations, 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 $\theta$ 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).[2]

The following is a brief overview of the plausible-values approach. Let $\underline{y}$ represent the responses of all sampled students to background questions or background data of sampled students collected from other sources, and let $\underline{\theta}$ represent the proficiency of interest. If $\underline{\theta}$ were known for all sampled students, it would be possible to compute a statistic $t(\underline{\theta}, \underline{y})$ - 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, $\underline{\theta}$ values are not known even for sampled respondents. The solution to this problem is to follow Rubin (1987) by considering $\underline{\theta}$ as "missing data" and approximate $t(\underline{\theta}, \underline{y})$ by its expectation given $(\underline{x}, \underline{y})$, the data that actually were observed, as follows:

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

○○○

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

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 sampled student *j*. 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.[3] The value of $\theta$ 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 process several times so that the uncertainty 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.

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 conditioning model is correctly specified, plausible values will provide consistent estimates of population proficiency, 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 $P(\underline{\theta}_j|(\underline{x}_j, \underline{y}_j, \Gamma, \Sigma))$, where $\Gamma$ is a matrix of regression coefficients for the background variables, and $\Sigma$ is a common variance matrix for residuals. Using standard rules of probability, the conditional probability of proficiency can be represented as

**(8)** $$P(\underline{\theta}_j|\underline{x}_j, \underline{y}_j, \Gamma, \Sigma) \propto P(\underline{x}_j|\underline{\theta}_j, \underline{y}_j, \Gamma, \Sigma)P(\underline{\theta}_j|\underline{y}_j, \Gamma, \Sigma) = P(\underline{x}_j|\underline{\theta}_j)P(\underline{\theta}_j|\underline{y}_j, \Gamma, \Sigma)$$

where $\underline{\theta}_j$ is a vector of scale values, $P(\underline{x}_j|\underline{\theta}_j)$ is the product over the scales of the independent likelihoods induced by responses to items within each scale, and $P(\underline{\theta}_j|(\underline{y}_j, \Gamma, \Sigma))$ is the multivariate joint density of proficiencies of the scales, conditional on the observed

○○○

3. U.S. National Assessment of Educational Progress (NAEP), U.S. National Adult Literacy Survey (NALS), the International Adult Literacy and Life Skills Survey (IALLS).
4. For further discussion, see Mislevy, Beaton, Kaplan, & Sheehan (1992).

value $\underline{y}_j$ 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.

### 13.4.3 Conditioning

A multivariate normal distribution was assumed for $P(\underline{\theta}_j, \underline{y}_j, \Gamma, \Sigma)$, 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 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 $\underline{y}^c$. The following model is then fit to the data:

$$\theta = \Gamma' \underline{y}^c + \varepsilon \quad (9)$$

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

In order to be strictly correct for all functions Γ of $\underline{\theta}$, it is necessary that $p(\underline{\theta}|\underline{y})$ be correctly specified for all background variables in the survey. In Benchmarking, 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 $\underline{\theta}$ for these variables is nearly optimal. Estimates of functions Γ involving background variables not conditioned in this manner are subject to estimation error due to mis-specification. The nature of these errors is 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, $\underline{\theta}$, and variance Σ, of the posterior distribution in equation (7). 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.

### 13.4.4 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 $\Gamma$ for all sampled. First, a value of $\Gamma$ is drawn from a normal approximation to $P(\Gamma, \Sigma|(\underline{x}_j, \underline{y}_j))$ that fixes $\Sigma$ at the value $\hat{\Sigma}$ (Thomas, 1993). Second, conditional on the generated value of $\Gamma$ (and the fixed value of $\Sigma=\hat{\Sigma}$), the mean, $\underline{\theta}$, and variance, $\Sigma_j^P$, 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 $\underline{\theta}$ and variance $\Sigma_j^P$. These three steps are repeated five times, producing five imputations of $\underline{\theta}$ for each sampled respondent.

For respondents with an insufficient number of responses, the $\Gamma$ and $\Sigma$ 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 (7) 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 $\theta$. 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 $T_u$s:

(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' $\theta$s are not known precisely, but only indirectly through $x$ and $y$.

### 13.4.5 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 1999 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. Yet, retaining this component of uncertainty requires that additional analytic procedures be used to estimate respondents' proficiencies, as follows.

If $\underline{\theta}$ values were observed for sampled respondents, the statistic $(t\text{-}T)/U^{1/2}$ would follow a $t$-distribution with $d$ degrees of freedom. Then the incomplete-data statistic $(t^*\text{-}T)/(Var(t^*))^{1/2}$ 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 $\underline{\theta}$ values:

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

Here $B_M$ 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 $(T\text{-}t^*)V^{1}\ (T\text{-}t^*)'$ is approximately $F$ distributed with degrees of freedom equal to $k$ and $\nu$, with $\nu$ defined as above but with a matrix generalization of $f_M$

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

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 student 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 student background variables.

## 13.5 Implementing the TIMSS Benchmarking Scaling Procedures

This section provides a synopsis of the IRT scaling and plausible-value methodology applied to the TIMSS 1999 data. Three major tasks were completed, as follows.

### 13.5.1 Rescaling of the TIMSS 1995 Data

TIMSS 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 TIMSS 1999. Since a major goal of TIMSS 1999 was to measure trends 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.

The rescaling of the TIMSS 1995 data was conducted according to the method described in the TIMSS 1999 Technical Report (Yamamoto & Kulick, 2000). The scale was set so the distribution of eighth-grade proficiency scores in 1995 had a mean of 500 and a standard deviation of 100 for both the mathematics and science scales (Gonzalez, 1997). Setting the scale metric in

this way produces slightly different means and standard deviations than in the original TIMSS 1995 results. Comparison of the original and rescaled 1995 proficiency scores is not appropriate because of this difference in the scale metric.

### 13.5.2 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.

The linking of the 1995 and 1999 scales was done at the mathematics and science domain levels only, since there were not enough common items to enable reliable linking within each content area.

### 13.5.3 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 two scales were not linked, but rather each was established independently.

For TIMSS 1999, the international mean for mathematics was 487 and the international mean for science was 488. The international mean for each content area was set to be equal to the subject area international mean.

### 13.5.4 Proficiency Scores for Benchmarking Students

Benchmarking plausible values for each student were generated using item statistics obtained from the international study. Consequently, the benchmarking plausible values are directly comparable to those obtained in the international study. For each student, five plausible values 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. Similarly, plausible values were generated for each student in each of the six science content areas (earth science; life science; physics; chemistry; scientific inquiry; and the nature of science) and science overall.

## 13.6 Summary

IRT was used to model the TIMSS achievement data. TIMSS used two- and three-parameter IRT models, and plausible-value technology to reanalyze the 1995 achievement data and analyze the 1999 achievement data. Plausible-value methodology was used to generate proficiency estimates for each subject and each content area.

## References

Adams, R.J., Wu, M.L., & Macaskill, G. (1997). Scaling methodology and procedures for the mathematics and science scales. In M.O. Martin & D. L. Kelly (Eds.), *TIMSS technical report volume II: Implementation and analysis* (pp. 111-145). 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* (pp. 397-479). 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.* (Department of Education Research Report No. 87-1). Twente, The Netherlands: 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* (pp. 147-174). 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 No. HB-91-1040-EX). Groningen, The Netherlands: University of Groningen, Psychological Institute.

Johnson, E.G., & Rust, K.F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics, 17*(2), 175-190.

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 data*. 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.* Reading, 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). Princeton, NJ: Educational Testing Service.

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.

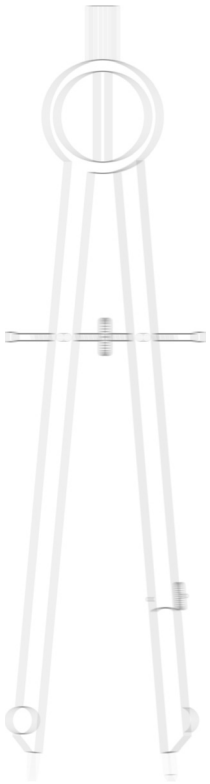
Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics, 2*, 309-22.

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). Princeton, NJ: Educational Testing Service.

Van Der Linden, W.J., & Hambleton, R. (1996). *Handbook of modern item response theory.* New York. Springer-Verlag.

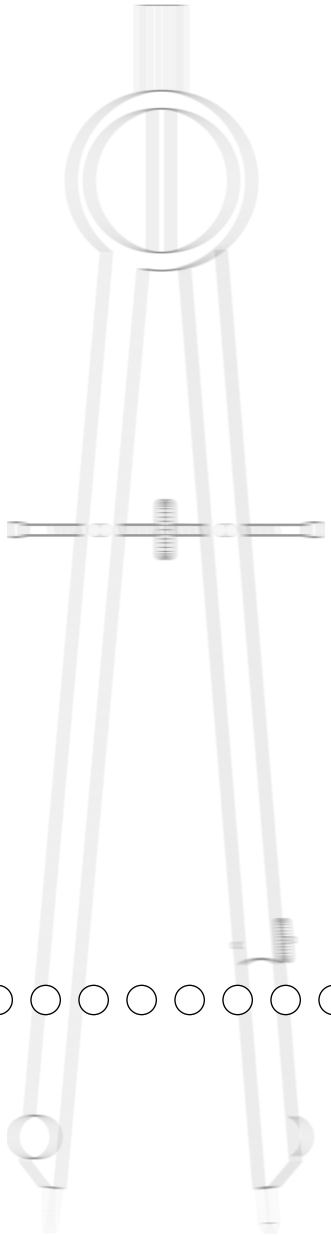
Yamamoto, K., & Kulick, E. (2000). Scaling methodology and procedures for the TIMSS mathematics and science scales. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS 1999 technical report* (pp. 237-263). Chestnut Hill, MA: Boston College.

Zwinderman, A.H. (1991). Logistic regression Rasch models. *Psychometrika*, *56*, 589-600.




# Describing TIMSS 1999 International Benchmarks of Student Achievement

Kelvin D. Gregory
Ina V. S. Mullis



# 14 Describing TIMSS 1999 International Benchmarks of Student Achievement[1]

Kelvin D. Gregory
Ina V. S. Mullis

## 14.1 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.[2] 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 portrait, illustrated by example items, of what students scoring at the anchor points are likely to know and be able to do.

The theoretical underpinnings of scale anchoring and decisions related to the application of scale anchoring to the TIMSS data can be found in Kelly (1999). This chapter is derived from chapter three of Kelly's work and describes how the TIMSS 1999 International Benchmarks were developed. These benchmarks are used in TIMSS 1999 Benchmarking Reports.

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.

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, & Martin (2000); those for science are presented in Smith, Martin, Mullis, & Kelly (2000).

○○○

1. This chapter was mainly reproduced from Gregory & Mullis (2000) in the international technical report for TIMSS 1999 (Martin, Gregory, & Stemler, 2000).
2. The international benchmarks - top 10%, upper quarter, median, and lower quarter - correspond to the 90th, 75th, 50th, and 25th percentiles, respectively, of the international distribution of student achievement in mathematics and science. The international benchmarks should not be confused with the TIMSS Benchmarking study, in which states and school districts compared or "benchmarked" their school systems against high-performing countries around the world.

## 14.2 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 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

### 14.2.1 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 an anchor point, and by fewer students at the next lower point.

TIMSS 1999, like TIMSS 1995, based the scale-anchoring descriptions on the international benchmarks, the $25^{th}$, $50^{th}$, $75^{th}$ and $90^{th}$ percentiles. These percentiles were labelled the lower quarter, median, upper quarter, and top 10% international benchmarks, respectively. The international percentiles were computed using the combined data from the countries that participated. Exhibit 14.1 shows the scale scores representing the international benchmarks for mathematics and science, respectively.

**Exhibit 14.1 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 five scale points was used. The ranges around the international percentiles and the number of observations within each range are shown in Exhibit 14.2.

**Exhibit 14.2 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 |

## 14.3 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 fewer 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
- Fewer 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
- Fewer 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
- Fewer 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
- Fewer 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
- Fewer 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
- Fewer than 50% of students at the 75th percentile answered the item correctly

Items answered correctly by at least 60% to 65% of the students regardless of the performance of students at the next lower point were identified to further supplement the item pool. Items that anchored, almost anchored, and met the 60% to 65% criterion were placed into three mutually exclusive categories. Each of these items helped to inform the descriptions of student achievement at the anchor levels.

## 14.4 Computing the Item Percent Correct at Each Level

The percentage of students scoring in the range around each anchor point and who answered a given 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 were 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 zero, one, or two. Two variables were created:

$v_1$ = 1 if the student received a one or two, and

$v_1$ = 0 otherwise, and

$v_2$ = 1 if the student received a two and

$v_2$ = 0 otherwise.

The percentage of students receiving a 1 on $v_1$ and of those receiving a 1 on $v_2$ was computed. This yielded the percentage of students receiving at least one point and a percentage of students receiving 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.

## 14.5 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 14.3 and 14.4 present the number of these items at each anchor point. Altogether, six mathematics items met the anchoring criteria at the 25$^{th}$ percentile, 36 did so for the 50$^{th}$ percentile, 73 for the 75$^{th}$ percentile, and 43 for the 90$^{th}$ percentile. Eleven items were too difficult for the 90$^{th}$ percentile. In science, 15 items met one of the criteria for anchoring at the 25$^{th}$ percentile, 33 for the 50$^{th}$ percentile, 39 for the 75$^{th}$ percentile, and 41 for the 90$^{th}$ percentile. Twenty-eight items were too difficult to anchor at the 90$^{th}$ 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. Despite not meeting the 65%/50% criteria, these were still items that students scoring at the anchor points had a high probability of answering correctly.

**Exhibit 14.3 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 14.4 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 |

## 14.6 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 14.5 and 14.6 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 90$^{th}$ percentile.

In mathematics, each of the five reporting categories had the most items anchoring at the 75$^{th}$ percentile. Fractions and number sense, data representation, analysis and probability, and algebra had at least one item anchoring at the 25$^{th}$ 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. The categories of environmental and resource issues, and scientific inquiry and the nature of science had no items that anchored at the 25$^{th}$ percentile, but it should be remembered that they contained the fewest items.

**Exhibit 14.5 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 14.6 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 |

## 14.7 The Anchoring Expert Panels

Two panels of experts 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 14.7 and 14.8, respectively. The members had extensive experience in their subject areas and a thorough knowledge of the TIMSS curriculum frameworks and achievement tests.

**Exhibit 14.7 Mathematics Scale Anchoring Panel Members**

| | |
|---|---|
| Lillie Albert<br>Boston College<br>United States | Anica Aleksova<br>Pedagosiki Zawod na Makedonija<br>Republic of Macedonia |
| Kiril Bankov<br>University of Sofia<br>Bulgaria | Jau-D Chen<br>Taiwan Normal University<br>Taiwan |
| John Dossey<br>Consultant<br>United States | Barbara Japelj<br>Educational Research Institute<br>Slovenia |
| Mary Lindquist<br>National Council of Teachers of Mathematics<br>United States | David Robitaille<br>University of British Columbia<br>Canada |
| Graham Ruddock<br>National Foundation for Education Research<br>England | Hanako Senuma<br>National Institute for Educational Research<br>Japan |
| Pauline Vos<br>University of Twente<br>Netherlands | |

**Exhibit 14.8 Science Scale Anchoring Panel Members**

| | |
|---|---|
| Audrey Champagne<br>State University of New York<br>United States | Galina Kovalyova<br>Center for Evaluating the Quality of Education<br>Russian Federation |
| Jan Lokan<br>Australian Council for Educational Research<br>Australia | Jana Paleckova<br>Institute for Information on Education<br>Czech Republic |
| Senta Raizen<br>National Center for Improving Science Education<br>United States | Vivien Talisayon<br>Institute of Science and<br>Mathematics Education Development<br>University of the Philippines |
| Hong Kim Tan<br>Ministry of Education Research and Evaluation<br>Singapore | |

## 14.8 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' were assigned 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% to 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. These drafts were then edited and revised as necessary, and the panelists reviewed and approved the item descriptions, anchor point descriptions, and example items for use in the TIMSS 1999 International Reports.

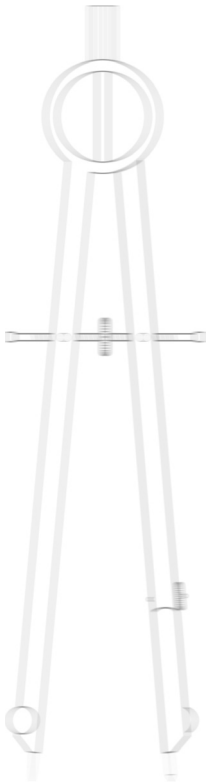
## References

Gregory, K., & Mullis, I.V.S. (2000). Describing international benchmarks of student achievement. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS 1999 technical report* (pp. 267-278). Chestnut Hill, MA: Boston College.

Kelly, D. L. (1999). *Interpreting the Third International Mathematics and Science Study (TIMSS) achievement scales using scale anchoring*. Unpublished doctoral dissertation, Chestnut Hill, MA: Boston College.

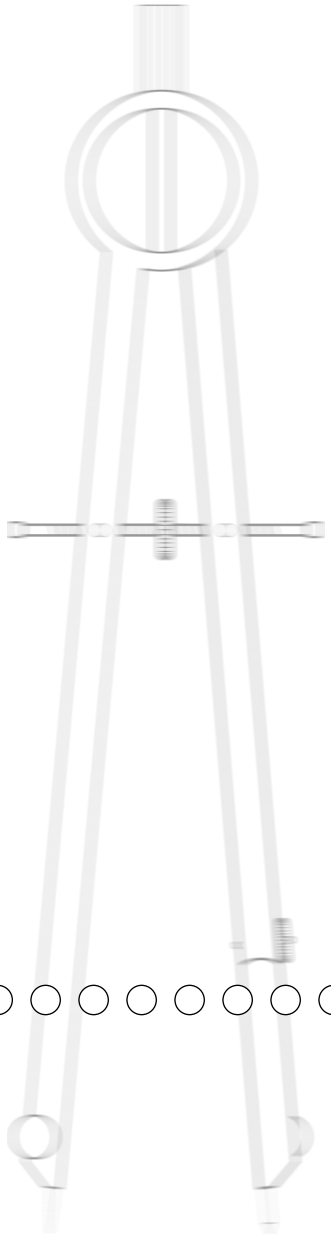
Kelly, D.L., Mullis, I.V.S., & Martin, M.O. (2000). *Profiles of student achievement in mathematics at the TIMSS international benchmarks: U.S. performance and standards in an international context*. Chestnut Hill, MA: Boston College.

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 for TIMSS 1999 Benchmarking

Eugenio J. Gonzalez
Kelvin D. Gregory



# 15 Reporting Student Achievement in Mathematics and Science for TIMSS 1999 Benchmarking[1]

Eugenio J. Gonzalez
Kelvin D. Gregory

## 15.1 Overview

As described in earlier chapters, the Benchmarking study makes extensive use of imputed student proficiency scores to report achievement in mathematics and science, both in the subjects overall and in the separate content areas. This chapter describes the procedures followed in computing the major statistics used to summarize achievement in the TIMSS 1999 Benchmarking Reports (Mullis et al., 2001; Martin et al., 2001), including average scores based on plausible values, Bonferroni adjustments for multiple comparisons, international benchmarks of achievement, and profiles of relative performance in subject-matter areas.

## 15.2 Computing Average Student Achievement

The item response theory (IRT) scaling procedure described in chapter 13 yields five imputed scores or plausible values in mathematics and science and in each of their content areas for each student. Average mathematics or science scores for countries or Benchmarking jurisdictions were computed by first taking the mean for each of the five plausible values, and then taking the mean of the five plausible-value means, as follows: The 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 or jurisdiction mean for plausible value $l$

$pv_{lj}$ is the $l^{th}$ plausible value for the $j^{th}$ student

○○○

1. This chapter is based on Gonzalez & Gregory (2000) from the TIMSS 1999 international technical report (Martin, Gregory, & Stemler, 2000).

$W^{i,j}$ is the weight associated with the $j^{th}$ student in class $i$, described in chapters 5 and 6

$N$ is the number of students in the sample.

The country or jurisdiction average is the mean of the five plausible value means.

The international average for mathematics and science was computed by taking the mean of the country means for each of the five plausible values and averaging across these five international means, as follows: The international average for each plausible value was computed as the average of that 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 these five international means. The international averages were based on all TIMSS 1999 countries. Data from Benchmarking jurisdictions were not included in the computation of international averages.

## 15.3 Achievement Differences Across Benchmarking Jurisdictions

The TIMSS 1999 Benchmarking Reports aim to provide fair and accurate comparisons of student achievement across the participating jurisdictions. Most of the exhibits summarize achievement using 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 jurisdictions, standard errors were used to assess the statistical significance of the difference between the summary statistics.

The charts presented in the TIMSS 1999 Benchmarking Reports provide comparisons of average performance of a jurisdiction with that of the TIMSS 1999 countries as well as with other participating jurisdictions. The significance tests reported in these

charts include a Bonferroni adjustment for multiple comparisons. The Bonferroni adjustment is necessary because the probability of finding a difference that is an artifact of chance greatly increases as the number of simultaneous comparisons increases.

### 15.3.1 Bonferroni Adjustments in TIMSS

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 $\alpha = .05$ level (i.e., with 95% confidence) was tested, then it would be expected that 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 $\alpha$. Conversely, the probability of not making such an error is $1 - \alpha$, which in the case of a single test is .95. When $\alpha = .05$, 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 avoiding a Type I error in any of the three is the product of the individual probabilities, which is $(1 - \alpha)(1 - \alpha)(1 - \alpha)$. With $\alpha = .05$, the overall probability of avoiding a Type I error is only .873. Stated differently, the probability of committing a Type I error rises from .05 for one comparisons to .127 with three comparisons, which is considerably less than the probability for a single test. As the number of tests increases, the probability of making a Type I error increases rapidly.

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 of $\alpha$ 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 ($\alpha$)of a statistical test is adjusted by establishing the number of comparisons that are

planned and then looking up the appropriate quantile from the normal distribution. In choosing the 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. This risks an error of a different kind, however, that of concluding that a difference in sample means is not significant when in fact there is a difference in the population means (i.e., Type II error).

Most users of the multiple comparison exhibits in the international reports 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 therefore adopted for the international reports. 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.

In the multiple comparison exhibits of the TIMSS 1999 Benchmarking Reports (Martin et al., 2001; Mullis et al., 2001), it was decided to keep the same Bonferroni correction as in the international reports so that between-country significance tests in both sets of reports would have the same results. This decision was taken despite the fact that Benchmarking exhibits that included all 38 TIMSS countries as well as the 27 Benchmarking participants had more comparisons (65) than exhibits in the international reports, which involved just the 38 countries. Consequently, exhibits with all 65 comparisons, which are confined to the first chapter in each Benchmarking report, present significance tests that are slightly less conservative than they would otherwise be.

### 15.3.2 Standard Error of the Difference

Mean proficiencies were considered significantly different if the absolute difference between them, divided by the standard error of the difference, was greater than the Bonferroni-adjusted critical value. For differences between countries or Benchmarking

jurisdictions, 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. Exhibits 15.1 and 15.2 show the means and standard errors for mathematics and science used in the calculation of statistical significance for countries and Benchmarking jurisdictions, respectively.

**Exhibit 15.1 Means and Standard Errors for Multiple-Comparisons Exhibits-Countries**

| Country | Math | | Science | |
|---|---|---|---|---|
| | Mean | S.E. | Mean | SE |
| United States | 501.633 | 3.971 | 514.915 | 4.553 |
| 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 |

**Exhibit 15.2 Means and Standard Errors for Multiple-Comparisons Exhibits - States and Districts**

| States | Math | | Science | |
|---|---|---|---|---|
| | Mean | S.E. | Mean | SE |
| Connecticut | 512.389 | 9.075 | 529.485 | 10.436 |
| Idaho | 494.886 | 7.385 | 526.368 | 6.585 |
| Illinois | 509.478 | 6.730 | 520.515 | 6.546 |
| Indiana | 514.626 | 7.186 | 534.202 | 6.973 |
| Maryland | 494.610 | 6.245 | 506.110 | 7.689 |
| Massachusetts | 513.469 | 5.938 | 533.194 | 7.363 |
| Michigan | 516.630 | 7.452 | 544.142 | 8.624 |
| Missouri | 489.731 | 5.314 | 522.826 | 6.486 |
| North Carolina | 495.218 | 7.026 | 507.792 | 6.544 |
| Oregon | 514.110 | 5.953 | 536.094 | 6.051 |
| Pennsylvania | 507.452 | 6.299 | 528.951 | 6.475 |
| South Carolina | 501.610 | 7.393 | 510.958 | 6.693 |
| Texas | 516.445 | 9.066 | 508.698 | 10.427 |
| **Districts and Consortia** | **Math** | | **Science** | |
| | Mean | S.E. | Mean | SE |
| Academy School Dist. #20, CO | 528.464 | 1.828 | 558.742 | 2.116 |
| Chicago Public Schools, IL | 462.500 | 6.102 | 449.447 | 9.505 |
| Delaware Science Coalition, DE | 479.483 | 8.928 | 500.446 | 8.379 |
| First in the World Consort., IL | 559.633 | 5.775 | 565.461 | 5.255 |
| Fremont/Lincoln/WestSide PS, NE | 488.142 | 8.215 | 511.302 | 5.780 |
| Guilford County, NC | 513.565 | 7.705 | 533.780 | 7.063 |
| Jersey City Public Schools, NJ | 474.814 | 8.610 | 439.666 | 9.756 |
| Miami-Dade County PS, FL | 421.330 | 9.449 | 425.956 | 10.937 |
| Michigan Invitational Group, MI" | 531.748 | 5.815 | 563.495 | 6.246 |
| Montgomery County, MD | 537.370 | 3.548 | 531.480 | 4.252 |
| Naperville Sch. Dist. #203, IL | 569.172 | 2.835 | 583.727 | 4.092 |
| Project SMART Consortium, OH | 520.593 | 7.507 | 539.223 | 8.370 |
| Rochester City Sch. Dist., NY | 444.404 | 6.462 | 451.669 | 7.372 |
| SW Math/Sci. Collaborative, PA | 516.719 | 7.547 | 543.249 | 7.429 |

## 15.4 Comparing Achievement with the International Mean

Many of the data exhibits in the TIMSS 1999 International Reports show countries' and jurisdictions; mean achievement compared with the international mean. Since this resulted in 38 simultaneous comparisons, the critical value was adjusted to 3.2125 using the Dunn-Bonferroni procedure. In the Benchmarking Reports, the corresponding exhibits contained 40 comparisons (27 Benchmarking participants and 13 selected countries), but for consistency with the international reports, the critical value for 38 comparisons was used in Benchmarking exhibits also.

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

$se_j^2$ is the sampling standard error for country *j*

$se_k^2$ is the sampling standard error for country *k*.

The imputation component of the standard error was computed by taking the square root of the imputation variance calculated as follows

$$se_{i\_dif\_j} = \sqrt{\frac{6}{5} Var(d_{1...} d_{l...} d_5)}$$

where $d_l$ is the difference between the international mean and the jurisdiction mean for plausible value *l*.

Finally, the standard error of the difference was calculated as:

$$se_{dif\_j} = \sqrt{se^2_{i\_dif\_j} + se^2_{s\_dif\_j}}.$$

## 15.5 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 benchmark as described in chapter 14. The top 10% benchmark was defined as the 90$^{th}$ 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. Comparing the percentage of students in Benchmarking jurisdictions that reached the achievement levels defined by these international benchmarks was a very useful way of describing student performance at various points of the ability distribution.

### 15.5.1 Establishing the International Benchmarks of Achievement

In computing 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 15.3 shows the contribution of each country to the estimation of the international benchmarks.

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 fewer than 1% in Tunisia, the Philippines, South Africa, and Morocco.

**Exhibit 15.3 Estimated Enrollment at the Eighth Grade**

| 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 |

Because of the imputation technology used to derive the student achievement 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 by taking into account the sampling design as well as the variance due to imputation. The international benchmarks are presented in Exhibit 15.4 and 15.5 for mathematics and science, respectively.

**Exhibit 15.4 International Benchmarks of Achievement for Eighth Grade—Mathematics**

| 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 15.5 International Benchmarks of Achievement for Eighth Grade—Science**

| 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 |

## Exhibit 15.6 Percentages of Students Reaching TIMSS 1999 International Benchmarks of Mathematics Achievement

| States, Districts and Consortia | Top 10% | Upper Quarter | Median | Lower Quarter |
|---|---|---|---|---|
| Connecticut | 11 (2.5) | 31 (3.9) | 67 (4.4) | 91 (1.9) |
| Idaho | 5 (1.1) | 24 (2.9) | 61 (3.5) | 88 (2.2) |
| Illinois | 10 (1.6) | 29 (2.9) | 65 (3.3) | 92 (1.5) |
| Indiana † | 9 (1.9) | 30 (3.9) | 69 (3.6) | 94 (1.2) |
| Maryland | 8 (1.4) | 27 (2.5) | 57 (3.2) | 87 (2.0) |
| Massachusetts | 10 (1.6) | 31 (2.6) | 68 (3.0) | 92 (1.6) |
| Michigan | 10 (2.0) | 33 (3.7) | 70 (3.3) | 92 (1.7) |
| Missouri | 4 (0.9) | 20 (2.4) | 58 (2.9) | 89 (1.5) |
| North Carolina | 7 (1.6) | 25 (3.1) | 57 (3.3) | 88 (2.0) |
| Oregon | 10 (1.8) | 32 (2.8) | 69 (2.8) | 91 (1.4) |
| *Pennsylvania* | 9 (1.3) | 28 (2.6) | 65 (3.0) | 91 (1.8) |
| South Carolina | 10 (2.0) | 30 (3.2) | 60 (3.5) | 88 (1.8) |
| *Texas* | 13 (2.2) | 37 (3.8) | 66 (4.3) | 90 (2.1) |
| **States, Districts and Consortia** | **Top 10%** | **Upper Quarter** | **Median** | **Lower Quarter** |
| Academy School Dist. #20, CO | 12 (0.8) | 38 (1.5) | 75 (1.5) | 95 (0.7) |
| Chicago Public Schools, IL | 2 (0.9) | 12 (1.7) | 41 (4.3) | 81 (2.5) |
| Delaware Science Coalition, DE | 5 (1.8) | 22 (4.1) | 51 (4.5) | 83 (2.4) |
| First in the World Consort., IL | 22 (3.2) | 56 (3.3) | 87 (2.1) | 98 (0.6) |
| Fremont/Lincoln/WestSide PS, NE | 6 (2.3) | 23 (4.1) | 58 (4.0) | 84 (2.7) |
| Guilford County, NC 1 | 10 (2.2) | 33 (3.5) | 66 (4.1) | 91 (1.6) |
| Jersey City Public Schools, NJ | 6 (1.9) | 17 (3.4) | 48 (3.9) | 82 (2.9) |
| Miami-Dade County PS, FL | 2 (0.9) | 9 (2.4) | 29 (3.6) | 61 (3.5) |
| Michigan Invitational Group, MI | 12 (2.4) | 39 (3.4) | 77 (3.0) | 96 (1.3) |
| Montgomery County, MD 1 | 17 (2.2) | 45 (1.8) | 77 (1.4) | 95 (1.1) |
| Naperville Sch. Dist. #203, IL | 24 (1.7) | 59 (2.2) | 91 (1.1) | 99 (0.4) |
| Project SMART Consortium, OH | 11 (2.9) | 34 (4.7) | 70 (3.1) | 95 (1.0) |
| Rochester City Sch. Dist., NY | 2 (0.9) | 9 (2.5) | 32 (3.2) | 73 (2.9) |
| SW Math/Sci. Collaborative, PA | 11 (2.7) | 32 (3.9) | 68 (3.1) | 93 (1.6) |

| | |
|---|---|
| Top 10% Benchmark (90th Percentile) | 616 |
| Upper Quarter Benchmark (75th Percentile) | 555 |
| Median Benchmark (50th Percentile) | 479 |
| Lower Quarter Benchmark (25th Percentile) | 396 |

States in *italics* did not fully satisfy guidelines for sample participation rates (see Appendix A for details).

† Met guidelines for sample participation rates only after replacement schools were included (see Exhibit A.6).

1 National Defined Population covers less than 90 percent of National Desired Population (see Exhibit A.3).

( ) Standard errors appear in parentheses. Because results are rounded to the nearest whole number, some totals may appear inconsistent.

## Exhibit 15.7 Percentages of Students Reaching TIMSS 1999 International Benchmarks of Science Achievement

| States | Top 10% | Upper Quarter | Median | Lower Quarter |
|---|---|---|---|---|
| Connecticut | 17 (3.0) | 39 (4.4) | 69 (4.6) | 90 (2.5) |
| Idaho | 13 (1.8) | 37 (3.2) | 70 (3.3) | 91 (1.8) |
| Illinois | 14 (1.9) | 36 (3.0) | 66 (3.0) | 88 (1.5) |
| Indiana † | 18 (2.5) | 41 (3.6) | 72 (2.8) | 92 (1.4) |
| Maryland | 12 (1.3) | 31 (3.0) | 59 (3.5) | 84 (2.5) |
| Massachusetts | 17 (2.4) | 40 (3.0) | 71 (3.4) | 92 (1.7) |
| Michigan | 22 (2.6) | 47 (3.6) | 75 (3.4) | 91 (2.2) |
| Missouri | 14 (2.3) | 36 (3.0) | 67 (2.8) | 89 (1.8) |
| North Carolina | 11 (1.4) | 30 (2.9) | 60 (3.4) | 85 (2.1) |
| Oregon | 19 (2.3) | 43 (2.7) | 73 (2.6) | 91 (1.9) |
| *Pennsylvania* | 15 (1.5) | 38 (2.5) | 70 (3.2) | 91 (1.6) |
| South Carolina | 13 (1.8) | 34 (2.7) | 60 (3.4) | 85 (1.7) |
| *Texas* | 15 (2.1) | 35 (3.6) | 61 (4.5) | 83 (3.3) |
| **Districts and Consortia** | **Top 10%** | **Upper Quarter** | **Median** | **Lower Quarter** |
| Academy School Dist. #20, CO | 23 (1.6) | 52 (1.5) | 84 (1.2) | 97 (0.6) |
| Chicago Public Schools, IL | 3 (1.1) | 11 (2.4) | 34 (3.9) | 67 (3.8) |
| Delaware Science Coalition, DE | 10 (1.8) | 29 (4.0) | 56 (4.2) | 83 (2.1) |
| First in the World Consort., IL | 27 (3.7) | 54 (3.6) | 85 (2.0) | 97 (0.9) |
| Fremont/Lincoln/WestSide PS, NE | 11 (1.7) | 32 (3.1) | 63 (3.2) | 86 (2.1) |
| Guilford County, NC 1 | 19 (2.5) | 43 (3.6) | 69 (3.5) | 90 (2.0) |
| Jersey City Public Schools, NJ | 3 (1.5) | 11 (3.1) | 31 (3.6) | 64 (3.5) |
| Miami-Dade County PS, FL | 4 (1.4) | 10 (2.4) | 28 (3.0) | 58 (3.7) |
| Michigan Invitational Group, MI | 25 (3.1) | 54 (3.0) | 84 (2.1) | 96 (1.1) |
| Montgomery County, MD 1 | 17 (1.1) | 40 (2.5) | 70 (2.3) | 91 (1.3) |
| Naperville Sch. Dist. #203, IL | 33 (2.5) | 64 (2.2) | 90 (1.2) | 98 (0.6) |
| Project SMART Consortium, OH | 19 (3.6) | 43 (5.0) | 73 (3.3) | 93 (1.1) |
| Rochester City Sch. Dist., NY | 3 (1.3) | 12 (2.5) | 33 (3.7) | 68 (3.0) |
| SW Math/Sci. Collaborative, PA | 19 (3.1) | 45 (3.6) | 75 (3.5) | 94 (1.7) |

| | |
|---|---|
| Top 10% Benchmark (90th Percentile) | 616 |
| Upper Quarter Benchmark (75th Percentile) | 558 |
| Median Benchmark (50th Percentile) | 488 |
| Lower Quarter Benchmark (25th Percentile) | 410 |

States in *italics* did not fully satisfy guidelines for sample participation rates (see Appendix A for details).

† Met guidelines for sample participation rates only after replacement schools were included (see Exhibit A.6).

1 National Defined Population covers less than 90 percent of National Desired Population (see Exhibit A.3).

( ) Standard errors appear in parentheses. Because results are rounded to the nearest whole number, some totals may appear inconsistent.

### 15.5.2 Reporting Student Achievement at the International Benchmarks

To compare student performance at the international benchmarks, TIMSS computed the percentage of students in each Benchmarking jurisdiction reaching each international benchmark. These percentages and their standard errors are presented in Exhibit 15.6 for mathematics and in Exhibit 15.7 for science.

## 15.6 Reporting Gender Differences

TIMSS reported gender differences in student achievement in mathematics and science overall, as well as in content areas. Gender differences in countries and Benchmarking jurisdictions were presented in an exhibit showing mean achievement for males and females, the differences between them, and 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 two samples 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 11.

## 15.7 Relative Performance by Content Areas

In addition to performance in mathematics and science overall, it was of interest to see how Benchmarking participants and 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. The average across content area scores was computed for each jurisdiction, and then performance in each content area was shown 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 vector to form a matrix called $R_{ks}$, where a row contained the average proficiency score for jurisdiction $k$ on scale $s$ for a specific subject. This $R_{ks}$ matrix had also a "zero$^{th}$" row and column. The elements in $r_{k0}$ contained the average of the elements on the $k^{th}$

row of the $R_{ks}$ matrix. These were the jurisdiction averages across the content areas. The elements in $r_{0s}$ contained the average of the elements of the $s^{th}$ column of the $R_{ks}$ matrix. These were the content area averages across all jurisdictions. The element $r_{00}$ contained 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 jurisdiction $k$ in content area $s$. A value of zero for an element $i_{ks}$ indicates a level of performance for jurisdiction $k$ in content area $s$ that would be expected given its performance in other content areas and its performance relative to other jurisdictions on that content area. A negative value for an element $i_{ks}$ indicates a performance for jurisdiction $k$ on content area $s$ lower than would be expected on the basis of the jurisdiction's overall performance. A positive value for an element $i_{ks}$ indicates a better than expected performance for jurisdiction $k$ in the content areas. This procedure was applied to each of the five plausible values and the results were averaged.

To construct confidence intervals, the standard error for each content area in each jurisdiction 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 jurisdiction replicates for each content area in the test. For each jurisdiction and each content area, 75 replicates were created.[2] 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 jurisdiction $k$ in content area s for a specific subject, for the $h^{th}$ international set of replicates. Each of these $R^h_{ks}$ matrices has also a zero$^{th}$ row and

○○○

2. In countries and jurisdictions where there were fewer than 75 jackknife zones, 75 replicates were also created by assigning the overall mean to as many replicates as were necessary to have 75.

column. The elements in $r^h_{k0}$ contain the average of the elements on the $k^{th}$ row of the $R^h_{ks}$ matrix. These are the jurisdiction averages across the content areas. The elements in $r^h_{0s}$ contain the average of the elements of the $s^{th}$ 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 jackknife repeated replication (JRR) standard error is then given by the formula

$$jse_{r_{ks}} = \sqrt{\sum_h (i_{ks} - i^h_{ks})^2}.$$

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 $\alpha = .05$ 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.

## 15.8 Percent Correct for Individual Items

To portray student achievement as fully as possible, the TIMSS 1999 Benchmarking Reports present many examples of the items used in the TIMSS 1999 tests, together with the percentage of students in each jurisdiction responding correctly to the item. These percentages were 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 percent 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 percent 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.

## References

Dunn, O.J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association, 56*, 52-64.

Gonzalez, E.G., & Gregory, K.D. (2000). Reporting student achievement in mathematics and science. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS 1999 technical report.* Chestnut Hill, MA: Boston College.

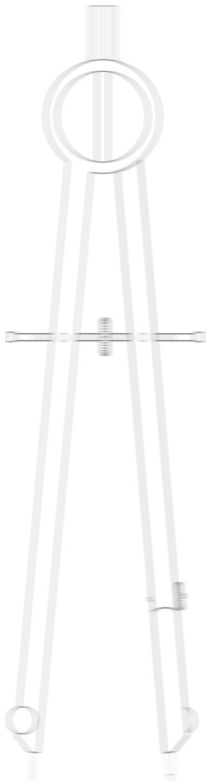
Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., O'Connor, K.M., Chrostowski, S.J., Gregory, K.D., Smith, T.A., & Garden, R.A. (2001). *Science benchmarking report: TIMSS 1999—Eighth grade.* Chestnut Hill, MA: Boston College.

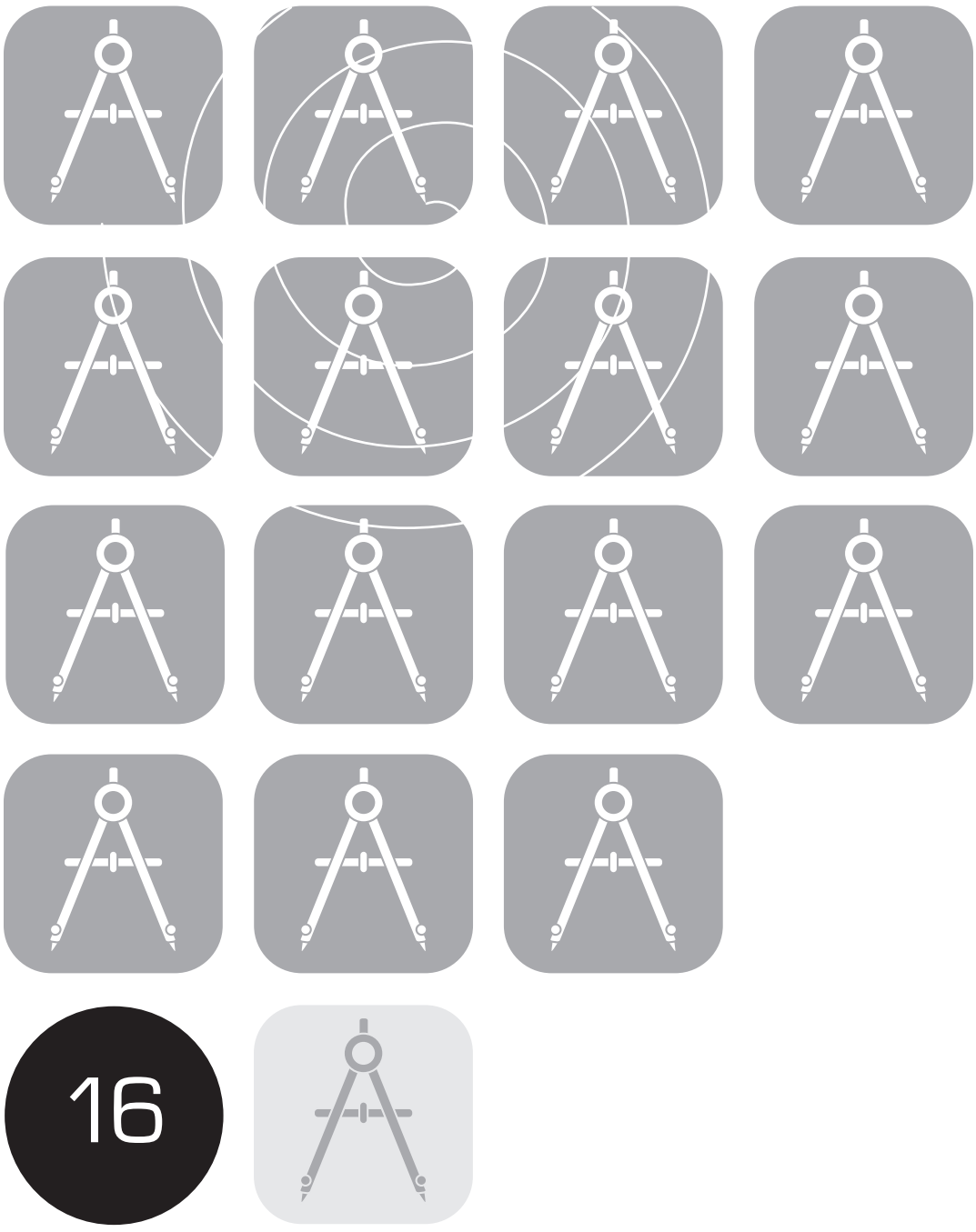
Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., Gregory, K.D., Smith, T.A., Chrostowski, S.J., Garden, R.A., & O'Connor, K.M. (2000). *TIMSS 1999 international science report.* Chestnut Hill, MA: Boston College.

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.

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.

Winer, B.J., Brown, D.R., & Michels, K.M. (1991). *Statistical principles in experimental design.* New York: McGraw Hill.

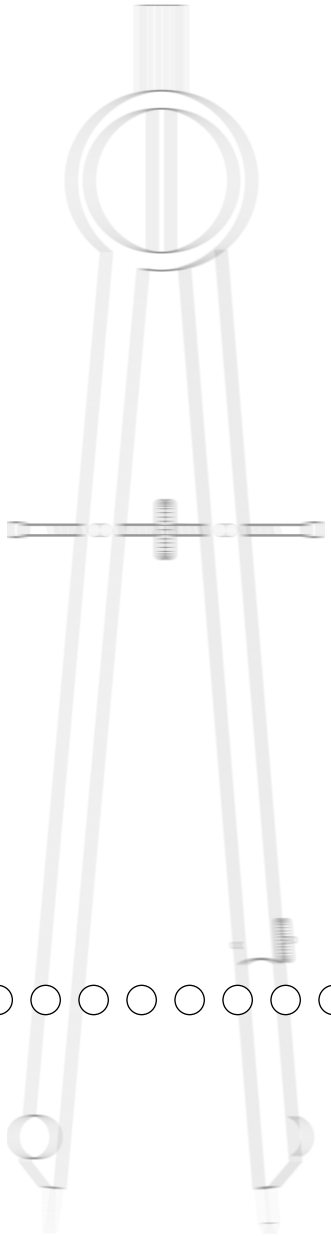




16

# Reporting Questionnaire Data for TIMSS 1999 Benchmarking

Teresa A. Smith



# 16 Reporting Questionnaire Data for TIMSS 1999 Benchmarking[1]

Teresa A. Smith

## 16.1 Overview

This chapter documents the analysis and reporting procedures used for the background questionnaire data in producing the TIMSS 1999 Benchmarking Reports. In particular, it describes the consensus process used to develop the report outlines and prototype exhibits; the development and computation of indices based on student, teacher, and school background variables; special considerations in reporting the questionnaire data; and the handling of non-response issues.

## 16.2 Background Questionnaires

As described in chapter 3, TIMSS 1999 used four types of background questionnaires to gather information at various levels of the educational system.

1. Curriculum Questionnaires addressing issues of curriculum design and curricular emphasis in mathematics and science were completed by national research coordinators (NRCs) in the TIMSS countries and by the appropriate authority in the Benchmarking jurisdictions.
2. A School Questionnaire providing 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. Student Questionnaires provided information about students' 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 (i.e., biology, chemistry, earth science, and physics). The general science version was used in the United States and in the Benchmarking jurisdictions.

○○○

1. This chapter is based on Smith (2000) from the international technical report for TIMSS 1999 (Martin, Gregory, & Stemler, 2000).

## 16.3 Reporting the TIMSS 1999 Benchmarking Data

The TIMSS 1999 Benchmarking results were reported in separate mathematics and science volumes (Mullis et al., 2001; Martin et al., 2001). These reports each 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. Each report included a number of summary indices based on some of the background data. These were presented to highlight issues related to good educational practice.

The TIMSS 1999 Benchmarking Reports present the questionnaire data of the participating jurisdictions along with those of a selected number of countries. The countries included for the purposes of comparison are the United States as well as a dozen European and Asian countries of interest. These countries include several high-performing European countries, countries that are major economic trading partners of the United States, and top-scoring Asian countries[2].

### 16.3.1 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, students were placed in a "high," "medium," or "low" category, with the high level being set to correspond 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 (i.e., 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 having at least one parent who had finished university. Students at the low level

○○○

2. See chapter 2 of Martin et al. (2001) or Mullis et al. (2001) for further details.

reported having 25 or fewer books in the home, not all three educational aids, and some secondary or less education as the highest level for either parent. Students with all other response combinations were assigned to the middle category.

The 17 indices computed for the TIMSS 1999 Benchmarking Reports are listed in Exhibit 16.1, which gives the name of the index; the label used to identify it in the TIMSS 1999 Benchmarking Reports and database; the mathematics or science exhibit in the TIMSS 1999 Benchmarking Report where the index data were reported; and the method used to compute the index.

**Exhibit 16.1 Summary Indices from Background Data in the TIMSS 1999 Benchmarking Report**

| Name of Index | Label | Exhibit[a] | Analysis Method |
|---|---|---|---|
| Index of Home Educational Resources | HER | 4.1 (M)<br>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 | 4.6 (M)<br>4.6 (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<br>SCS-E<br>SCS-B<br>SCS-P<br>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. |

**Exhibit 16.1 (continued) Summary Indices from Background Data in the TIMSS 1999 Benchmarking Report**

| 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<br>PATS-E<br>PATS-B<br>PATS-P<br>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 R3.1) 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 | 6.11 (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 | 6.11 (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. |

**Exhibit 16.1 (continued) Summary Indices from Background Data in the TIMSS 1999 Benchmarking Report**

| Name of Index | Label | Exhibit[a] | Analysis Method |
|---|---|---|---|
| Index of Emphasis on Conducting Experiments in Science Classes* | ECES-G<br>ECES-E<br>ECES-B<br>ECES-P<br>ECES-C | 6.13 (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 | 6.13 (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 | 6.15 (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 | 6.16 (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 | 7.2 (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 | 7.2 (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. |

**Exhibit 16.1 (continued) Summary Indices from Background Data in the TIMSS 1999 Benchmarking Report**

| Name of Index | Label | Exhibit[a] | Analysis Method |
|---|---|---|---|
| Index of Good School and Class Attendance | SCA | 7.5 (M)<br>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 Benchmarking report where data based on the index were presented. An (M) indicates mathematics report; (S) indicates science report.

* In the U.S. and the Benchmarking jurisdictions only the (G) indices were computed. In countries where science is taught as separate subjects, 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 percentage of students at each level, together with their average mathematics or science achievement. In addition, the percentage at the high level was shown graphically, with the Benchmarking jurisdictions and comparison countries ranked. For some of the sciences indices, the results were presented in separate panels for each science subject.

## 16.4 Development of the Benchmarking Reports

TIMSS 1999 was designed to investigate student learning of mathematics and science and the way in which aspects of the education systems relate to the learning opportunities and experiences of individual students. The TIMSS Benchmarking Study was designed to allow participants to compare, or benchmark, not just the mathematics and science achievement of their students with that of students in high-achieving TIMSS countries, but also key attributes of their education systems, including:

- 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

The goal of the Benchmarking reports was to present as much descriptive information about the contexts for learning as possible without overburdening the reader. Indices based on variables from the TIMSS 1999 background questionnaires were proposed to summarize information.

Analyses began with the preparation of prototype exhibits. This required a careful review of the questionnaires, detailed documentation of the variables and response categories, 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 reviewed by the coordinators from the Benchmarking jurisdictions in September 2000. Consensus was reached on the reporting priorities: which indices and variables should be reported, how much information should be included, and which trend tables to present. Coordinators reviewed the final data exhibits and draft text in January 2001. The Mathematics and Science Benchmarking reports were published in April 2001 (Martin et al., 2001; Mullis et al., 2001).

## 16.5 Reporting Student Background Data

Reporting the data from the Student Questionnaire was fairly straightforward. Most of the tables in the TIMSS 1999 Benchmarking 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 reports include some information separately by gender. For gender-based exhibits, the percentages of boys and girls in each category were displayed, along with the statistical significance of the gender difference.

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 separate-subjects version. The

○○○

3. See chapter 11 for a description of the jackknife methodology.

exhibits showing results for questions that differed in the two versions have separate sections for countries that administered each one. The United States and the Benchmarking participants used the general science version.

In the exhibits based on questions 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. Countries appear in the appropriate panels. Countries where earth science or chemistry was not applicable for the eighth grade were excluded from these panels. Also, in some countries combined courses were taught, such as physical science (physics/chemistry) or natural science (biology/earth science). 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 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 16.2 lists the countries and jurisdictions administering the two versions and indicates which science subjects were taught in each. 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 16.2 Countries that Administered the Two 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 | ● | | | | |
| [d] United States | ● | | | | |

a Chinese Taipei: separate sciences are taught starting in 7th grade, with biology in 7th grade and physics/chemistry in 8th grade. 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 8th grade.

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 The Netherlands: students were asked questions about integrated physics/chemistry; data for questions pertaining to physics/chemistry were reported in the physics panel.

d All Benchmarking jurisdictions in the U.S. also administered the general version.

## 16.6 Reporting Teacher Background Data

In the eighth grade, mathematics and science are generally taught by different teachers. Accordingly, there was one 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, all science teachers of the students in the sampled classes completed questionnaires. 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 Section B for each class taught. Thus, the information about instruction was tied directly to the students tested and the 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 or Benchmarking jurisdictions. Rather, they represent teachers of the representative samples of students assessed. It is important to note that in the TIMSS 1999 Benchmarking 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 given to 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 TIMSS 1999 Mathematics Benchmarking Report, and those collected from science teachers in the TIMSS 1999 Science Benchmarking Report. Where possible and relevant, the average achievement of students was reported for each teacher

response category exhibit to show the relationship with achievement. For indices computed from teacher data, percentages of students and average achievement are displayed at the high, medium, and low level of the index.

The data obtained from the science teachers were displayed in two ways. The general science data were presented together for all science teachers in each country or jurisdiction. Data 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 science teachers and for the separate-subject 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 16.2.[4] 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 those 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

○○○

4. Not all countries listed in Exhibit 16.2 had questionnaire data reported in the Benchmarking reports. Data from only 13 selected countries were reported.

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).

## 16.7 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 or Benchmarking jurisdiction and for the international average of all TIMSS countries. In a few instances, average numerical values for open-ended questions were computed across students (e.g., instructional time, hours the principal spends on different activities).

## 16.8 Reporting Curriculum Questionnaire Data

One chapter in each of the TIMSS 1999 Benchmarking 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

○○○

5. See chapter 5 for a description of the sample design for the TIMSS 1999 countries and chapter 6 for a description of the sample design for the TIMSS 1999 Benchmarking participants.

curriculum was introduced and its current status; methods 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 percentages 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 or Benchmarking jurisdiction. Detailed information on the percentage of students intended to be taught each 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 Benchmarking reports present both teachers' reports of the topics actually taught (i.e., the implemented curriculum) and reports from Benchmarking Coordinators 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.

## 16.9 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 Benchmarking, 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. For index variables, however, cases were coded as missing 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 1999 Benchmarking 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 or jurisdictions 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 Benchmarking 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 TIMSS 1999 Benchmarking reports.[6]

- For a country or jurisdiction where student, teacher or school responses were available for 70% to 84% of the students, "r" appears next to the data.
- Where student, teacher, or school responses were available for 50% to 69% of the students, "s" appears next to the data for that country or jurisdiction.
- When student, teacher or school responses were available for less than 50% of the students, "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 (~).

○○○

6. Since the information from the Curriculum Questionnaires was obtained at the national or jurisdiction level, no non-response flags were necessary in exhibits based on these data.

- When data were unavailable for all respondents in a country or jurisdiction, dashes (–) were used in place of data in all of the affected columns.[7]

## 16.10 Summary

This chapter presented how TIMSS reported and analyzed the background data collected for the Benchmarking Study from students, teachers, schools, and Benchmarking jurisdiction coordinators. It documented how summary indices were created, as well as the consensus approach used in developing the TIMSS 1999 Benchmarking reports.

○○○

7. A dash usually indicates that a background question was not administered, 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 science subjects not included in each country.

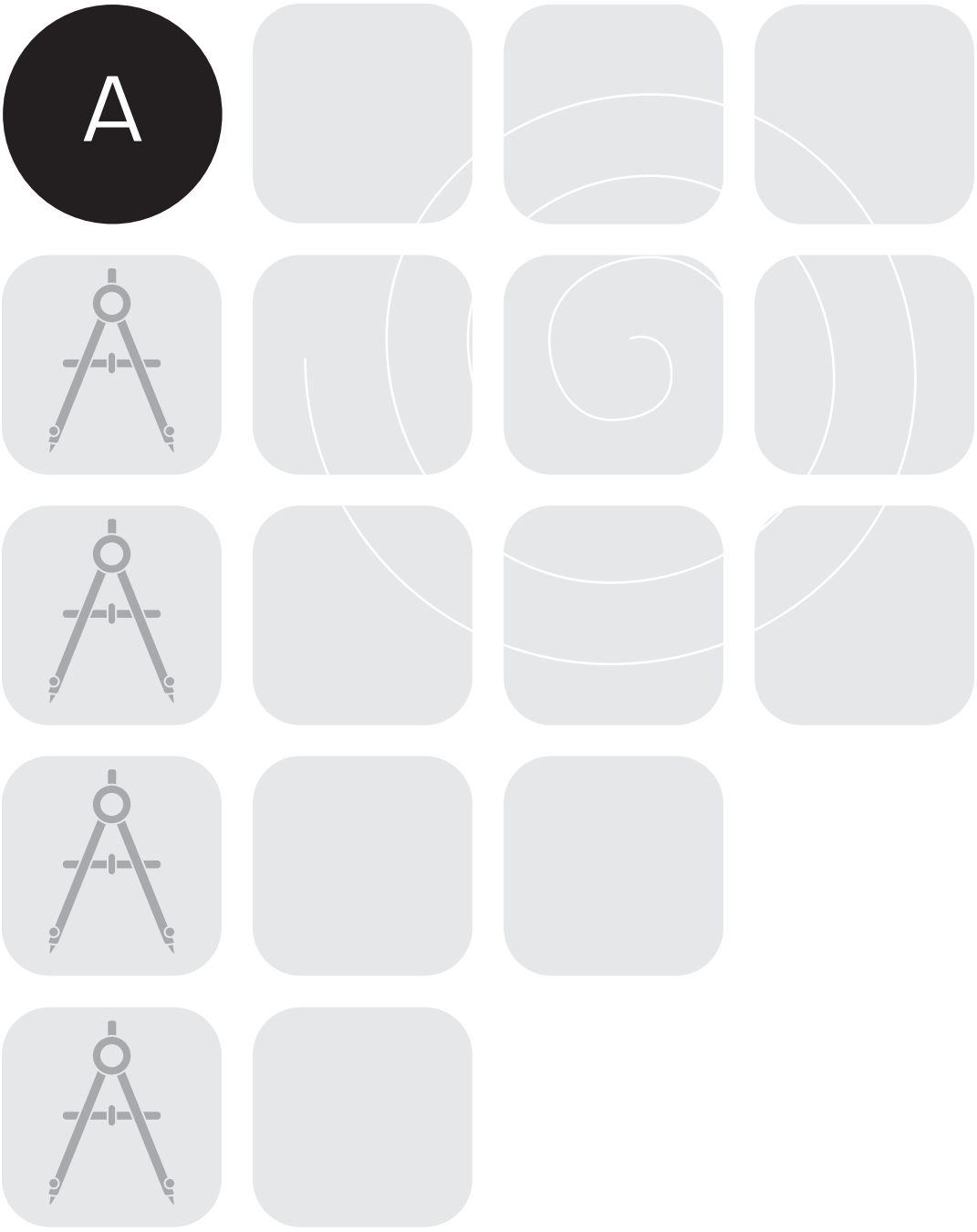
## References

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.

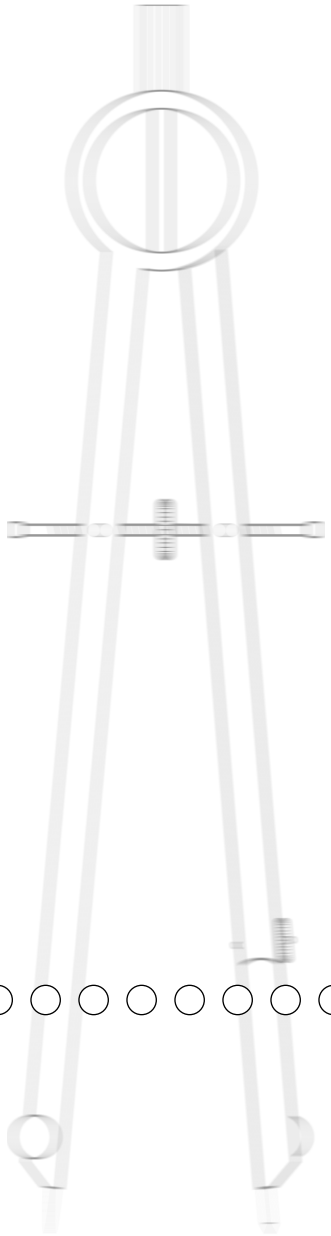
Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., Gregory, K.D., Garden, R.A., O'Connor, K.M., Chrostowski, S.J., & Smith, T.A. (2000). *TIMSS 1999 international mathematics report.* Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., O'Connor, K.M., Chrostowski, S.J., & Smith, T.A. (2000). TIMSS 1999 international mathematics report. Chestnut Hill, MA: Boston College.

Smith, T. A. (2000). Reporting Questionnaire Data. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS 1999 technical report.* Chestnut Hill, MA: Boston College.




# Appendix A: Acknowledgements



# Acknowledgements

TIMSS 1999 and the TIMSS Benchmarking Study were collaborative efforts among hundreds of individuals around the world. Staff from the national research centers in each participating country and from each Benchmarking jurisdiction, 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 the projects. They 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 the studies 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 and the Benchmarking Study also acknowledge 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 (NCES) in the U.S. Department of Education, the U.S. National Science Foundation (NSF), the World Bank, and participating countries. Funding for the overall design, administration, data management, and quality assurance activities of TIMSS Benchmarking was provided by NCES, NSF, and the Office of Educational Research and Improvement (OERI) in the U.S. Department of Education. Valena Plisko, Eugene Owen, and Patrick Gonzales of NCES; Janice Earle, Larry Suter, and Elizabeth VanderPutten of NSF; Carol Sue Fromboluti and Jill Edwards Staton of OERI, and Maggie McNeely formerly of OERI each played a crucial role in making TIMSS 1999 and the Benchmarking Study possible and for ensuring the quality of the studies. Each participating country was responsible for funding local project costs and implementing TIMSS 1999 in accordance with the international procedures. Each Benchmarking participant contracted directly with Boston College to fund data-collection activities in its own jurisdiction.

## 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 in the Lynch School of Education at Boston College. 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 in Amsterdam, Hans Wagemaker, Executive Director, was responsible for overseeing fundraising and country participation. The IEA Secretariat

also coordinated translation verification and recruiting of international quality control monitors. The data were processed centrally by the IEA Data Processing Center in Hamburg. Statistics Canada in Ottawa was responsible for collecting and evaluating the sampling documentation from each country and for calculating the sampling weights. Educational Testing Service (ETS) in Princeton, New Jersey, conducted the scaling of the achievement data.

For the Benchmarking Study, Westat in Rockville, Maryland, was responsible for sampling, data collection activities, and preliminary data processing. National Computer Systems (NCS) in Iowa City, Iowa, conducted the scoring for Benchmarking jurisdictions along with the national scoring effort. All data were processed in accordance with international standards at the IEA Data Processing Center. Scaling of the achievement data was conducted by Educational Testing Service.

### 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. González, Director of Operations and Data Analysis
Kathleen M. O'Connor, TIMSS Benchmarking Coordinator
Kelvin D. Gregory, TIMSS Study Coordinator
Teresa A. Smith, TIMSS Science Coordinator
Robert Garden, TIMSS Mathematics Coordinator
Dana L. Kelly, PIRLS Study Coordinator
Steven Chrostowski, Research Associate
Ce Shen, Research Associate (former)
Julie Miles, Research Associate
Steven Stemler, Research Associate
Ann Kennedy, Research Associate
Maria José Ramirez, Research Assistant
Joseph Galia, Statistician/Programmer
Lana Seliger, Statistician/Programmer (former)
Andrea Pastelis, Database Manager
Kieran Brosnan, Technology Support Specialist
Christine Conley, Publications Design Manager
José Nieto, Publications Manager
Tom Hoffmann, Internet Communications Manager
Betty Hugh, Data Graphics Specialist
Susan Messner, Data Graphics Specialist
Mario Pita, 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
Rita Holmes, 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 Service

Kentaro Yamamoto, Principal Research Scientist
Ed Kulick, Manager, Research Data Analysis

### Westat

Nancy Caldwell, Vice President and Associate Director, Survey Operations Group
Keith Rust, Vice President and Associate Director, Statistical Group
Stephen Roey, Senior Systems Analyst

### Project Management Team

Michael Martin, International Study Center
Ina Mullis, International Study Center
Eugenio González, International Study Center
Hans Wagemaker, IEA Secretariat
Dirk Hastedt, IEA Data Processing Center
Pierre Foy, Statistics Canada
Kentaro Yamamoto, Educational Testing Service
Eugene Johnson, American Institutes for Research

### Sampling Referees

Pierre Foy, Statistics Canada – TIMSS 1999 Benchmarking
Keith Rust, Westat – TIMSS 1999 International

## Benchmarking Participants

Individuals from each Benchmarking jurisdiction were instrumental in conducting the TIMSS Benchmarking Study in their state, district, or consortium. They were responsible for obtaining funding for the project; obtaining cooperation of sampled schools, classes, and students; responding to curriculum questionnaires; reviewing data; contributing to the development of the Benchmarking reports; and coordinating activities with the International Study Center. Jurisdictions would like to acknowledge the following people for their extensive contributions.

### States

#### Connecticut

Patricia Brandt
Connecticut Department of Education
165 Capital Avenue
Hartford CT 06145-2219

Abigail L. Hughes
Connecticut Department of Education
165 Capital Avenue
Hartford CT 06145-2219

Douglas Rindone
Connecticut Department of Education
165 Capital Avenue
Hartford CT 06145-2219

Theodore S. Sergi
Connecticut Department of Education
165 Capital Avenue
Hartford CT 06145-2219

#### Idaho

Tom Farley
Idaho Department of Education
P.O. Box 83720
Boise ID 83720-0027

Susan Harrington
Idaho Department of Education
P.O. Box 83720
Boise ID 83720-0027

Sally Tiel
Idaho Department of Education
P.O. Box 83720
Boise ID 83720-0027

#### Illinois

Mervin Brennan
Illinois State Board of Education
100 North First Street
Springfield IL 62777

Carmen Chapman
Illinois State Board of Education
100 North First Street
Springfield IL 62777

Megan Forness
Illinois State Board of Education
Assessment E216
100 North First Street
Springfield IL 62777

Andy Metcalf
Illinois State Board of Education
100 North First Street
Springfield IL 62777

Pam Stanko
Illinois State Board of Education
100 North First Street
Springfield IL 62777

**Indiana**

Larry Grau
Office of the Governor
State House
200 West Washington Street, Room 206
Indianapolis IN 46204-2797

Dwayne James
Indiana Department of Education
Room 229, State House
Indianapolis IN 46204

Stan Jones
Commissioner for Higher Education
101 West Ohio Street - Suite 550
Indianapolis IN 46204

Cheryl Orr
Indiana's Education Roundtable
101 West Ohio Street - Suite 550
Indianapolis IN 46204

Suellen Reed
Superintendent of Public Instruction
Room 229, State House
Indianapolis IN 46204-2797

Cynthia Roach
Indiana Department of Education
Division of Assessment
Room 229, State House
Indianapolis IN 46204-2797

**Maryland**

Diane Householder
Maryland State Department of Education
200 West Baltimore Street
Baltimore MD 21201-2595

Mark Moody
Maryland State Department of Education
200 West Baltimore Street
Baltimore MD 21201-2595

Kathy Rosenberg
Maryland State Department of Education
200 West Baltimore Street
Baltimore MD 21201-2595

**Massachusetts**

Jeffrey Nellhaus
Massachusetts Department of Education
350 Main Street
Malden MA 02148-5023

Sheldon Rothman
Massachusetts Department of Education
350 Main Street
Malden MA 02148-5023

Kit Viator
Massachusetts Department of Education
350 Main Street
Malden MA 02148-5023

Lori Wright
Massachusetts Department of Education
350 Main Street
Malden MA 02148-5023

**Michigan**

Charles Allan
Michigan Department of Education
Curriculum Development Program
P.O. Box 30008
Lansing MI 48909

**Missouri**

James Friedebach
Missouri Department of Education
205 Jefferson
P.O. Box 480
Jefferson City MO 65102-0480

**North Carolina**

Louis Fabrizio
North Carolina Department of Public Instruction
301 North Wilmington Street
Raleigh NC 27601-2825

Tammy Howard
North Carolina Department of Public Instruction
301 North Wilmington Street
Raleigh NC 27601-2825

#### Oregon

Joanne Flint
Oregon Department of Education
255 Capital Street NE
Salem OR 97310-0203

Wayne Neuberger
Oregon Department of Education
255 Capital Street NE
Salem OR 97310-0203

#### Pennsylvania

R. Jay Gift
Pennsylvania Department of Education
333 Market Street, 8th Floor
Harrisburg PA 17126-0333

Frank Marburger
Pennsylvania Department of Education
333 Market Street, 8th Floor
Harrisburg PA 17126-0333

Lee Plempel
Pennsylvania Department of Education
333 Market Street, 8th Floor
Harrisburg PA 17126-0333

Charlie Wayne
Pennsylvania Department of Education
333 Market Street, 8th Floor
Harrisburg PA 17126-0333

#### South Carolina

Karen Horne
South Carolina Department of Education
1429 Senate Street
Columbia SC 29201

Susan Agruso
South Carolina Department of Education
1429 Senate Street
Columbia SC 29201

Lane Peeler
South Carolina Department of Education
611-B Rutledge Building
1429 Senate Street
Columbia SC 29201

Paul Sandifer
South Carolina Department of Education
607 Rutledge Building
1429 Senate Street
Columbia SC 29201

Teri Siskind
South Carolina Department of Education
607 Rutledge Building
1429 Senate Street
Columbia SC 29201

#### Texas

Chris Castillo Comer
Texas Education Agency
1701 North Congress Avenue
Austin TX 78701

Ed Miller
Texas Education Agency
1701 North Congress Avenue
Austin TX 78701-1494

Phyllis Stolp
Texas Education Agency
1700 North Congress Avenue
Austin TX 78701

### Districts and Consortia

#### Academy School District #20

Wendy Crist
Academy School District #20
7610 North Union Boulevard
Colorado Springs CO 80920

Alisabeth Hohn
Academy School District #20
7610 North Union Boulevard
Colorado Springs CO 80920

**Chicago Public Schools**
Gery Chico
Chicago Public Schools
125 South Clark Street
Chicago IL 60603

Richard Daley
City Hall
121 North LaSalle Street
Chicago IL 60603

Joseph Hahn
Chicago Public Schools
125 South Clark Street 11th Floor
Chicago IL 60603

Phil Hansen
Chicago Public Schools
125 South Clark Street
Chicago IL 60603

Paul Vallas
Chicago Public Schools
125 South Clark Street
Chicago IL 60603

Melanie Wojtulewicz
Chicago Public Schools
1326 West 14th Place
Room 315A
Chicago IL 60608

**Delaware Science Coalition**
Gail Ames
Delaware Science Coalition
2916 Duncan Road
Wilmington DE 19808

John Collette
Delaware Science Coalition
309 Brockton Road
Wilmington DE 19803

Julie Cwikla Banks
University of Delaware
305 M Willard Hall
Newark DE 19716

Valerie Maxwell
Appoquinimink School District
118 South Sixth Street
Odessa DE 19730

**First in the World Consortium**
Elaine Aumiller
North Central Regional Education Lab (NCREL)
1120 East Diehl Road, Suite 200
Naperville IL 60563

Blase Masini
North Central Regional Education Lab (NCREL)
1120 East Diehl Road, Suite 200
Naperville IL 60563

Paul Kimmelman
1306 Hidden Lake Drive
Buffalo Grove IL 60089

David Kroeze
First in the World Consortium
Northbrook School District #27
1250 Sanders Road
Northbrook IL 60062

**Fremont/Lincoln/WestSide Public Schools**
James Findley
Westside Public Schools
909 South 76th Street
Omaha NE 68114-4599

Marilyn Moore
Lincoln Public Schools
Box 82889
Lincoln NE 68501-2889

Stephen Sexton
Fremont Public Schools
957 North Pierce Street
Fremont NE 68025

Terry Snyder
Fremont Public Schools
957 North Pierce Street
Fremont NE 68025

**Guilford County Schools**
Lynne Johnson
Guilford County Schools
120 Franklin Boulevard
Greensboro NC 27401

Diane Spencer
Guilford County Schools
120 Franklin Boulevard
Greensboro NC 27401

Sadie Bryant Woods
Guilford County Schools
134 Franklin Boulevard
Greensboro NC 27401

**Jersey City Public Schools**
Richard DiPatri
Jersey City Public Schools
State District Superintendent
346 Claremont Avenue
Jersey City NJ 07305

Adele Macula
Jersey City Board of Education
346 Claremont Avenue
Jersey City NJ 07305

Aldo Sanchez-Abreu
Jersey City Board of Education
346 Claremont Avenue
Jersey City NJ 07305

Patsy Wang-Iverson
Mid-Atlantic Eisenhower Consortium
Research for Better Schools
444 North Third Street
Philadelphia PA 19123

**Miami-Dade County Public Schools**
Joseph Burke
Miami-Dade County Public Schools
1500 Biscayne Boulevard
Room 327T
Miami FL 33132

Gisela Feild
Miami-Dade County Public Schools
1500 Biscayne Boulevard
Suite 225
Miami FL 33132

Joseph Mathos
Miami-Dade County Public Schools
1450 Northeast 2nd Avenue #931
Miami FL 33132

Vilma Rubiera
Miami-Dade County Public Schools
1500 Biscayne Boulevard
Suite 225
Miami FL 33132

Alex Shneyderman
Miami-Dade County Public Schools
1500 Biscayne Boulevard
Suite 225
Miami FL 33132

Constance Thornton
Miami-Dade County Public Schools
1500 Biscayne Boulevard
Suite 327
Miami FL 33132

**Michigan Invitational Group**
Robert Dunn
Michigan Invitational Group
Michigan Department of Education
658 Grat Strasse
Manchester MI 48158

**Montgomery County Public Schools**
Marlaine Hartzman
Montgomery County Public Schools
850 Hungerford Drive, Room 11
Rockville MD 20850

John Larson
Montgomery County Public Schools
850 Hungerford Drive, Room 11
Rockville MD 20850

**Naperville Community School District 203**

Russ Bryan
Naperville Community School District 203
203 West Hillside Road
Naperville IL 60540

Lenore Johnson
Naperville Community School District 203
203 West Hillside Road
Naperville IL 60540

Jack Hinterlong
Naperville Community School District 203
203 Hillside Road
Naperville IL 60540

Donald E. Weber, Ed.D
Naperville Community School District 203
203 West Hillside Road
Naperville IL 60540

Jodi Wirt
Naperville Community School District 203
203 Hillside Road
Naperville IL 60540

**Project SMART Consortium**

Dennis Kowalski
Strongville City School
13200 Pearl Road
Stongsville OH 44136

Terry Krivak
c/o Ohio Aerospace Institute
22800 Cedar Point Road
Cleveland OH 44142

Anne Mikesell
Ohio Department of Education
25 South Front Street, 5th Floor
Columbus OH 43215

Linda Williams
Mentor Exempted Village
6451 Center Street
Mentor OH 44060

Paul R. Williams
Project SMART Consortium
Beachwood City School District
24601 Fairmount Boulevard
Beachwood OH 44122

**Rochester City School District**

Ann Pinnella Brown
Rochester City School District
131 West Broad Street
Rochester NY 14614

Cecilia Golden
Rochester City School District
131 West Broad Street
Rochester NY 14614

Corinthia Sims
Rochester City School District
131 West Broad Street
Rochester NY 14614

**Southwest Pennsylvania Math and Science Collaborative**

Nancy Bunt
2650 Regional Enterprise Tower
425 Sixth Avenue
Pittsburgh PA 15219

Marcia Seeley
2650 Regional Enterprise Tower
425 Sixth Avenue
Pittsburgh PA 15219

Lou Tamler
2650 Regional Enterprise Tower
425 Sixth Avenue
Pittsburgh PA 15219

Cynthia A. Tananis
University of Pittsburgh
5P26 WWPH School of Education
Pittsburgh PA 15260

## 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.Didaktiek
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
Hong Kong, SAR

**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

**Jordan**
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)
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

**United States**
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 the 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

#### 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

**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
Susan Messner
Mario Pita

**Cover Design** Christine Conley