Chapter 1

Overview of TIMSS 2003

Michael O. Martin and Ina V.S. Mullis

1.1 Introduction

Since pioneering cross-national studies of educational achievement with the First International Mathematics Study (FIMS) in 1964, the International Association for the Evaluation of Educational Achievement (IEA) has conducted almost 20 studies of student achievement in the curricular areas of mathematics, science, language, civics, and reading. The Third International Mathematics and Science Study (TIMSS) in 1994-1995 was the largest and most complex IEA study ever conducted, including both mathematics and science at third and fourth grades, seventh and eighth grades, and the final year of secondary school.

In 1999, TIMSS (now renamed the Trends in International Mathematics and Science Study) again assessed eighth-grade students in both mathematics and science to measure trends in student achievement since 1995. Also, 1999 represented four years since the first TIMSS, and the population of students originally assessed as fourth-graders had advanced to the eighth grade. Thus, TIMSS 1999 also provided information about whether the relative performance of these students had changed in the intervening years.

TIMSS 2003, the third data collection in the TIMSS cycle of studies, was administered at the eighth and fourth grades. For countries that participated in previous assessments, TIMSS 2003 provides three-cycle trends at the eighth grade (1995, 1999, 2003) and data over two points in time at the fourth grade (1995 and 2003). In countries new to the study, the 2003 results can help policy makers and practitioners assess their comparative standing and gauge the rigor and effectiveness of their mathematics and science programs.

This volume describes the technical aspects of TIMSS 2003 and summarizes the main activities involved in the development of the data collection instruments, the data collection itself, and the analysis and reporting of the data.

1.2 Participants in TIMSS 2003

Exhibit 1.1 lists all the countries that have participated in TIMSS in 1995, 1999, or 2003 at fourth or eighth grade. In all, 67 countries have participated in TIMSS at one time or another. Of the 49 countries that participated in TIMSS 2003, 48 participated at the eighth grade and 26 at the fourth grade. Yemen participated at the fourth but not the eighth grade. The exhibit shows that at the eighth grade 23 countries also participated in TIMSS 1995 and TIMSS 1999. For these participants, trend data across three points in time are available. Eleven countries participated in TIMSS 2003 and TIMSS 1999 only, while three countries participated in TIMSS 2003 and TIMSS 1995. These countries have trend data for two points in time. Of the 12 new countries participating in the study, 11 participated at eighth grade and 2 at the fourth grade. Of the 26 countries participating in TIMSS 2003 at the fourth grade, 16 also participated in 1995, providing data at two points in time.

Following the success of the TIMSS 1999 benchmarking initiative in the United States,[1] in which 13 states and 14 school districts or district consortia administered the TIMSS assessment and compared their students' achievement to student achievement world wide, TIMSS 2003 included an international benchmarking program, whereby regions of countries could participate in the study to compare to international standards. TIMSS 2003 included four benchmarking participants at the eighth grade: the Basque Country of Spain, the U.S. state of Indiana, and the Canadian provinces of Ontario and Quebec. Indiana, Ontario, and Quebec participated also at the fourth grade. Having also participated in 1999, Indiana has data at two points in time at eighth grade. Ontario and Quebec participated also in 1995 and 1999, and so have trend data across three points in time at both grade levels.

1.3 Student Populations

TIMSS 2003 had as its intended target population all students at the end of their eighth and fourth years of formal schooling in the participating countries. However, for comparability with previous TIMSS assessments, the formal definition for the eighth-grade population specified all 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. This grade level was intended to represent eight years of schooling, counting from the first year of primary or elementary schooling, and was indeed the eighth grade in most countries. Similarly, for the fourth-grade population, the formal definition specified all students enrolled in the upper of the two adjacent grades that contained the largest proportion of 9-year-olds. This grade level was intended to represent

1 See Mullis, Martin, Gonzalez, O'Connor, Chrostowski, Gregory, Garden, and Smith (2001) for the results of the benchmarking in mathematics and Martin, Mullis, Gonzalez, O'Connor, Chrostowski, Gregory, Smith, and Garden (2001) for the results in science.

four years of schooling, counting from the first year of primary or elementary schooling, and was the fourth grade in most countries.

1.4 Assessment Dates

TIMSS 2003 was administered near the end of the school year in each country. In countries in the Southern Hemisphere (where the school year typically ends in November or December) the assessment was conducted in October or November 2002. In the Northern Hemisphere, the school year typically ends in June; so in these countries the assessment was conducted in April, May, or June 2003.

1.5 Study Management and Organization

TIMSS 2003 was conducted under the auspices of the IEA. The study was directed by Michael O. Martin and Ina V.S. Mullis of the TIMSS & PIRLS International Study Center at Boston College, Lynch School of Education, where they also direct IEA's Progress in International Reading Literacy Study (PIRLS). The International Study Center was responsible for the design, development, and implementation of the study – including developing the assessment framework, assessment instruments, and survey procedures; ensuring quality in data collection; and analyzing and reporting the study results. Staff at the International Study Center worked closely with the organizations responsible for particular aspects of the study, the representatives of participating countries, and the TIMSS advisory committees.

In the IEA Secretariat, Hans Wagemaker, Executive Director, was responsible for overseeing fundraising and country participation. The IEA Secretariat also managed the ambitious translation verification effort conducted for the field test and main assessment and recruited international quality control monitors in each country. The IEA Data Processing Center was responsible for processing and verifying the data from the participating countries and for constructing the international database. Working closely with the Data Processing Center, Statistics Canada was responsible for collecting and evaluating the sampling documentation from each country and for calculating the sampling weights. Educational Testing Service in Princeton, New Jersey provided consultation on psychometric issues as well as technical support and software for scaling the achievement data. The Project Management Team, comprising the study directors and representatives from the International Study Center, IEA, Statistics Canada, and Educational Testing Service, met regularly throughout the study to discuss the study's progress, procedures, and schedule.

Exhibit 1.1 **Countries Participating in TIMSS 2003, 1999, and 1995**

Countries	Grade 8			Grade 4	
	2003	1999	1995	2003	1995
Argentina*	●	●			
Armenia	●			●	
Australia	●	●	●	●	●
Austria			●		●
Bahrain	●				
Belgium (Flemish)	●	●	●	●	
Belgium (French)			●		
Botswana	●				
Bulgaria	●	●	●		
Canada		●	●		●
Chile	●	●			
Chinese Taipei	●	●		●	
Colombia			●		
Cyprus	●	●	●	●	●
Czech Republic		●	●		●
Denmark			●		
Egypt	●				
England	●	●	●	●	●
Estonia	●				
Finland		●			
France			●		
Germany			●		
Ghana	●				
Greece			●		●
Hong Kong, SAR	●	●	●	●	●
Hungary	●	●	●	●	●
Iceland			●		●
Indonesia	●	●			
Iran, Islamic Rep. of	●	●	●	●	●
Ireland			●		●
Israel	●	●	●		●
Italy	●	●	●	●	●
Japan	●	●	●	●	●
Jordan	●	●			
Korea, Rep. of	●	●	●		●
Kuwait			●		●
Latvia	●	●	●	●	●
Lebanon	●				
Lithuania	●	●	●	●	
Macedonia, Rep. of	●	●			
Malaysia	●	●			
Moldova, Rep. of	●	●		●	

Exhibit 1.1 **Countries Participating in TIMSS 2003, 1999, and 1995** *(...Continued)*

Countries	Grade 8			Grade 4	
	2003	1999	1995	2003	1995
Morocco	●	●		●	
Netherlands	●	●	●	●	●
NewZealand	●	●	●	●	●
Norway	●		●	●	●
Palestinian Nat'l Auth.	●				
Philippines	●	●		●	
Portugal			●		●
Romania	●	●	●		
Russian Federation	●	●	●	●	
SaudiArabia	●				
Scotland	●		●	●	●
Serbia	●				
Singapore	●	●	●	●	●
Slovak Republic	●	●	●		
Slovenia	●	●	●	●	●
South Africa	●	●	●		
Spain			●		
Sweden	●		●		
Switzerland			●		
Syrian Arab Republic**	●				
Thailand		●	●		●
Tunisia	●	●		●	
Turkey		●			
United States	●	●	●	●	●
Yemen**				●	
Benchmarking Participants					
BasqueCountry, Spain	●				
IndianaState, US	●	●		●	
OntarioProvince, Can.***	●	●	●	●	●
QuebecProvince, Can.***	●	●	●	●	●

* Argentina administered the TIMSS 2003 data collection one year late, and did not score and process its data in time for inclusion in this report.

**Because the characteristics of their samples are not completely known, achievement data for Syrian Arab Republic and Yemen are presented in Appendix F of the International reports.

***Ontario and Quebec participated in TIMSS 1999 and 1995 as part of Canada.

Each participating country appointed a National Research Coordinator (NRC) and a national center responsible for all aspects of TIMSS 2003 within that country. The TIMSS & PIRLS International Study Center organized meetings of the NRCs several times a year to review study materials and procedures, and to provide training in student sampling, constructed-response item scoring, and data entry and database construction.

The TIMSS & PIRLS International Study Center was supported in its work by a number of advisory committees. The International Expert Panel in Mathematics and Science played a crucial role in developing the TIMSS 2003 frameworks and specifications for the assessment. The Mathematics and Science Item Development Task Forces coordinated the work of the National Research Coordinators in developing and reviewing the mathematics and science achievement items. The Science and Mathematics Item Review Committee reviewed and revised successive drafts of the achievement items and was an integral part of the scale anchoring process. The Questionnaire Item Review Committee revised the TIMSS context questionnaires for the 2003 assessment.

1.6 The TIMSS 2003 Assessment Frameworks

The development of the TIMSS 2003 assessment was a collaborative process spanning a two-and-a-half-year period and involving mathematics and science educators and development specialists from all over the world. Central to this effort was a major updating and revision of the existing TIMSS assessment frameworks to address changes during the last decade in curricula and the way science is taught. The resulting publication entitled *TIMSS Assessment Frameworks and Specifications 2003* serves as the basis of TIMSS 2003 and beyond (Mullis, Martin, Smith, Garden, Gregory, Gonzalez, Chrostowski, and O'Connor, 2003).

As shown in Exhibit 1.2, the mathematics and science assessment frameworks for TIMSS 2003 are framed by two organizing dimensions or aspects, a content domain and a cognitive domain. There are five content domains in mathematics (number, algebra, measurement, geometry, and data) and five in science (life science, chemistry, physics, earth science, and environmental science) that define the specific mathematics and science subject matter covered by the assessment. The cognitive domains, four in mathematics (knowing facts and procedures, using concepts, solving routine problems, and reasoning) and three in science (factual knowledge, conceptual understanding, and reasoning and analysis) define the sets of behaviors expected of students as they engage with the mathematics and science content.

Exhibit 1.2 **The Content and the Cognitive Domains of the Mathematics and Science Framework**

Mathematics		Science	
Content Domain		**Content Domain**	
Grade 8	Number	Grade 8	Life Science
	Algebra		Chemistry
	Measurement		Physics
	Geometry		Earth Science
	Data		Environmental Science
Grade 4	Number	Grade 4**	Life Science
	Patterns and Relationships*		Physical Science
	Measurement		Earth Science
	Geometry		
	Data		
Cognitive Domain		**Cognitive Domain**	
	Knowing Facts and Procedures		Factual Knowledge
	Using Concepts		Conceptual Understanding
	Solving Routine Problems		Reasoning and Analysis
	Reasoning		

* At fourth grade, the algebra content domain is called patterns and relationships.
**At the fourth grade, there are only three content areas in science, namely life science, physical science, and earth science.

1.7 Developing the TIMSS 2003 Assessment

Given TIMSS' ambitious goals for curriculum coverage and innovative problem solving tasks, as specified in the Frameworks and Specifications, the development of the assessment items required a tremendous cooperative effort, crucially dependent on the contribution of the National Research Coordinators (NRCs) during the entire process. To maximize the effectiveness of the contributions from national centers, the TIMSS & PIRLS International Study Center developed a detailed item-writing manual and conducted a workshop for countries that wished to provide items for the international item pool. At this workshop, two item development "Task Forces" reviewed general item-writing guidelines for multiple-choice and constructed-response items and provided specific training in writing mathematics and science items in accordance with the *TIMSS Assessment Frameworks and Specifications 2003*. The mathematics task force consisted of the mathematics coordinator and two experienced mathematics item writers, and similarly the science task force comprised the science coordinator and two experienced science item writers.

More than 2,000 items and scoring guides were drafted, and reviewed by the task forces. The items were further reviewed by the Science and Mathematics Item Review Committee, a group of internationally prominent math-

ematics and science educators nominated by participating countries to advise on subject-matter issues in the assessment. Committee members also helped to develop tasks and items to assess problem solving and scientific inquiry.

Participating countries field-tested the items with representative samples of students, and all of the potential new items were again reviewed by the Science and Mathematics Item Review Committee as well as by NRCs. The resulting TIMSS 2003 eighth-grade assessment contained 383 items, 194 in mathematics and 189 in science. The fourth grade assessment contained 313 items, 161 in mathematics and 152 in science.

Between one-third and two-fifths of the items at each grade level were in constructed-response format, requiring students to generate and write their own answers. Some constructed-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 used a multiple-choice format. In scoring the items, correct answers to most questions were worth one point. However, responses to some constructed-response questions (particularly those requiring extended responses) were evaluated for partial credit, with a fully correct answer being awarded two points. The total number of score points available for analysis thus somewhat exceeds the number of items.

Not all of the items in the TIMSS 2003 assessment were newly developed for 2003. To ensure reliable measurement of trends over time, the assessment included also items that had been used in the 1995 and 1999 assessments. For example, of the 426 score points available in the entire 2003 mathematics and science assessment, 47 came from items used also in 1995, 102 from items used also in 1999, and 267 from items used for the first time in 2003. At fourth grade, 70 score points came from 1995 items, and the remaining 267 from new 2003 items.

Every effort was made to ensure that the tests represented the curricula of the participating countries and that the items exhibited no bias toward or against particular countries. The final forms of the test were endorsed by the NRCs of the participating countries. In addition, countries had an opportunity to match the content of the test to their curriculum. They identified items measuring topics not covered in their intended curriculum. The information from this Test-Curriculum Matching Analysis, provided in Appendix C of the International Reports, indicates that omitting such items has little effect on the overall pattern of results.

1.8 TIMSS 2003 Assessment Design

With the large number of mathematics and science items, it was not possible for every student to respond to all items. To ensure broad subject-matter coverage without overburdening individual students, TIMSS 2003, as in the 1995 and 1999 assessments, used a matrix-sampling technique that assigns each assessment item to one of a set of item blocks, and then assembles student test booklets by combining the item blocks according to a balanced design. Each student takes one booklet containing both mathematics and science items. Thus, the same students participated in both the mathematics and science testing.

In the TIMSS 2003 assessment design, the 313 fourth-grade mathematics and science items and the 383 eighth-grade items were divided among 28 item blocks at each grade, 14 mathematics blocks labeled M01 through M14, and 14 science blocks labeled S01 through S14. Each block contained either mathematics items only or science items only. This general block design was the same for both grades, although the planned assessment time per block was 12 minutes for fourth grade and 15 minutes for eighth grade.

There were 12 student booklets at each grade level, with six blocks of items in each booklet. To enable linking between booklets, each block appears in two, three, or four different booklets. The assessment time for individual students was 72 minutes at fourth grade (six 12-minute blocks) and 90 minutes at eighth grade (six 15-minute blocks), which is comparable to that in the 1995 and 1999 assessments. The booklets were organized into two three-block sessions (Parts I and II), with a break between the parts.

The 2003 assessment was the first TIMSS assessment in which calculators were permitted, and so it was important that the design allow students to use calculators when working on the new 2003 items. However, because calculators were not permitted in TIMSS 1995 or 1999, the 2003 design also had to ensure that students did not use calculators when working on trend items from these assessments. The solution was to place the blocks containing trend items (blocks M01 – M06 and S01 – S06) in Part I of the test booklets, to be completed without calculators before the break. After the break, calculators were allowed for the new items (blocks M07 – M14 and S07 – S14). To provide a more balanced design, however, and have information about differences with calculator access, two mathematics trend blocks (M05 and M06) and two science trend blocks (S05 and S06) also were placed in Part II of one booklet each. Note that calculators were allowed only at the eighth grade, and not at the fourth grade.

1.9 Background Questionnaires

By gathering information about students' educational experiences together with their mathematics and science achievement on the TIMSS assessment, it is possible to identify factors or combinations of factors related to high achievement. As in previous assessments, TIMSS in 2003 administered a broad array of questionnaires to collect data on the educational context for student achievement. For TIMSS 2003, a concerted effort was made to streamline and upgrade the questionnaires. The TIMSS 2003 contextual framework (Mullis, et al., 2003) articulated the goals of the questionnaire data collection and laid the foundation for the questionnaire development work.

Across the two grades and two subjects, TIMSS 2003 involved 11 questionnaires. *National Research Coordinators* completed four questionnaires. With the assistance of their curriculum experts, they provided detailed information on the organization, emphasis, and content coverage of the mathematics and science curriculum at fourth and eighth grades. The *fourth- and eighth-grade students* who were tested answered questions pertaining to their attitudes 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 the topics in the curriculum frameworks, instructional practices, professional training and education, and their views on mathematics and science. Separate questionnaires for mathematics and science teachers were administered at the eighth grade, while to reflect the fact that most younger students are taught all subjects by the same teacher, a single questionnaire was used at the fourth grade. The principals or heads of *schools* at the fourth and eighth grades responded to questions about school staffing and resources, school safety, mathematics and science course offerings, and teacher support.

1.10 Translation and Verification

The TIMSS data collection instruments were prepared in English and translated into 34 languages. Of the 49 countries and four benchmarking participants, 17 collected data in two languages and one country, Egypt, in three languages – Arabic, English, and French. In addition to translation, it sometimes was necessary to modify the international versions for cultural reasons, even in the countries that tested wholly or partly in English. This process represented an enormous effort for the national centers, with many checks along the way. The translation effort included (1) developing explicit guidelines for translation and cultural adaptation; (2) translation of the instruments by the national centers in accordance with the guidelines, using two or more independent translations; (3) consultation with subject-matter experts on

cultural adaptations to ensure that the meaning and difficulty of items did not change; (4) verification of translation quality by professional translators from an independent translation company; (5) corrections by the national centers in accordance with the suggestions made; (6) verification by the International Study Center that corrections were made; and (7) a series of statistical checks after the testing to detect items that did not perform comparably across countries.

1.11 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. These manuals covered procedures for test security, standardized scripts to regulate directions and timing, rules for answering students' questions, and steps to ensure that identification on the test booklets and questionnaires corresponded to the information on the forms used to track students.

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

In all, 50 quality control monitors drawn from the 49 countries and four Benchmarking participants participated in the training. Where necessary, quality control monitors who attended the training session were permitted to recruit other monitors to assist them in covering the territory and meeting the testing timetable. All together, the international quality control monitors and those trained by them observed 1,147 testing sessions (755 for grade 8 and 392 for grade 4), and conducted interviews with the National Research Coordinator in each of the participating countries.

The results of the interviews indicate that, in general, NRCs had prepared well for data collection and, despite the heavy demands of the schedule and shortages of resources, were able to conduct the data collection efficiently and professionally. Similarly, the TIMSS tests appeared to have

been administered in compliance with international procedures, including the activities before the testing session, those during testing, and the school-level activities related to receiving, distributing, and returning material from the national centers.

1.12 Scoring the Constructed-Response Items

Because a large proportion of the assessment time was devoted to constructed-response items, TIMSS needed to develop procedures for reliably evaluating student responses within and across countries. Scoring used two-digit codes with rubrics specific to each item. The first digit designates the correctness level of the response. The second digit, combined with the first, represents a diagnostic code identifying specific types of approaches, strategies, or common errors and misconceptions. Although not used in this report, analyses of responses based on the second digit should provide insight into ways to help students better understand science 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 implement them, together with example student responses for the various rubric categories. These guides, along with training packets containing extensive examples of student responses for practice in applying the rubrics, were used as a basis for intensive training in scoring the constructed-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.

To gather and document empirical information about agreement among scorers in each country, TIMSS arranged to have systematic samples of at least 100 student responses to each item scored independently by two readers. The results showed a high degree of agreement for both the correctness score (the first digit) and for the two-digit diagnostic score. At the eighth grade, the percentage of exact agreement between scorers averaged 99 and 97 percent for the correctness score in mathematics and science, respectively, and 97 and 92 percent for the diagnostic score. At fourth grade, the figures were 99 and 96 percent for the mathematics and science correctness score and 97 and 92 percent for the diagnostic score. The TIMSS data from the reliability studies indicate that scoring procedures were robust for the mathematics and science items, especially for the correctness score used for the analyses in the International reports.

TIMSS 2003 also took steps to show that those constructed-response items from 1999 that were used in 2003 were scored in the same way in both

assessments. In anticipation of this, countries that participated in TIMSS 1999 sent samples of scored student booklets from the 1999 eighth-grade data collection to the IEA Data Processing Center, where they were digitally scanned and stored in presentation software for later use. As a check on scoring consistency from 1999 to 2003, staff members working in each country on scoring the 2003 eighth-grade data were asked also to score these 1999 responses using the DPC software. The items from 1995 that were used in TIMSS 2003 all were in multiple-choice format, and therefore scoring reliability was not an issue. There was a high degree of scoring consistency, with 92 percent exact agreement, on average, internationally, in mathematics and 98 percent in science between the scores awarded in 1999 and those given by the 2003 scorers. There was somewhat less agreement at the diagnostic score level, with 93 percent exact agreement, on average, in mathematics and 81 percent in science.

To monitor the consistency with which the scoring rubrics were applied across countries, TIMSS collected from the Southern-Hemisphere countries that administered TIMSS in English a sample of 150 student responses to 41 constructed-response mathematics and science questions. This set of student responses was then sent to each Northern-Hemisphere country having scorers proficient in English and scored independently by one or if possible two of these scorers. All 150 responses to each of the 41 items were scored by 37 scorers from the countries that participated. Agreement across countries was defined in terms of the percentage of these scores that were in exact agreement. The results showed that scorer reliability across countries was high, particularly in mathematics, with the percent exact agreement averaging 96 percent across the mathematics items and 87 percent across the science items for the correctness score and 92 percent and 76 percent across mathematics and science items, respectively, for the diagnostic score.

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.

Throughout the process, the TIMSS 2003 data were checked and double-checked by the IEA Data Processing Center, the International Study Center, and the national centers. The national centers were contacted regularly and given multiple opportunities to review the data for their countries. In conjunction with the IEA Data Processing Center, the International Study Center reviewed item statistics for each cognitive item in each country to identify poorly performing items. In general, the items exhibited very good psychometric properties in all countries. In the few instances where there were poor item statistics (negative point-biserials for the key, large item-by-country interactions, and statistics indicating lack of fit with the model), these were a result of translation, adaptation, or printing errors.

1.14 Scaling the TIMSS Achievement Data

Deriving reliable student achievement scores from a large-scale assessment measuring trends over time like TIMSS poses a difficult challenge. Firstly, because of the ambitious coverage goals of TIMSS 2003, there was not enough testing time for a student to complete the entire assessment, and so a matrix-sampling design was adopted whereby each student's test booklet contained just a part of the assessment. Although this solved the problem of administering the assessment, it complicated the calculation of student achievement scores, since not all students took the same set of items, and the items that students did take were not all equally difficult. Secondly, in measuring trends over time (1995, 1999, 2003, and so on), it was not possible for TIMSS to keep reusing the same mathematics and science achievement items. In order to keep the assessment at the cutting edge of mathematics and science education, it was necessary to replace older items with new material at each cycle. In addition, TIMSS has a policy of publishing a large proportion of the items used in each assessment so that educators, policy makers, and the public may have a good understanding of the mathematics and science addressed by the assessment. Accordingly, the composition of the assessment evolves at each assessment cycle, as items are published and used for illustrative purposes and new items are developed to replace the published items. This further complicated the calculation of student achievement scores.

To meet the challenge of estimating student achievement, TIMSS relies primarily on item response theory (IRT) scaling methods. With IRT scaling, students' scores do not depend on taking the same set of items, and so this methodology is particularly useful when different blocks of items and different

samples of students have to be linked. This being the case, IRT methodology was preferred by TIMSS for developing comparable estimates of performance for all students, since students answered different test items depending upon which of the 12 test booklets they received. The IRT analysis provides a common scale on which performance can be compared across countries. In addition to providing a basis for estimating mean achievement, scale scores permit estimates of how students within countries vary and provide information on percentiles of performance.

In TIMSS 2003, the mathematics and science 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 that he or she took in a way that takes into account the difficulty and discriminating power of each item. As with any method of scaling student achievement, measurement is most reliable when a student responds to a large number of items, and is less reliable when the number of items is small. In the matrix-sampling approach adopted by TIMSS, with each student responding to a limited number of items, and given TIMSS' ambitious reporting goals – scales for two subjects (mathematics and science) and for five content domains in each subject – each student may respond to just a few items related to a particular scale.

To improve reliability, the TIMSS scaling methodology draws on information about students' background characteristics as well as their responses to the achievement items. This approach, known as "conditioning," enables reliable scores to be produced even though individual students responded to relatively small subsets of the total mathematics or science item pool. Rather than estimating student scores directly, TIMSS combines information about item characteristics, student responses to the items that they took, and student background information to estimate student achievement distributions. Having determined the overall achievement distribution, TIMSS estimates each student's achievement conditional on the student's responses to the items that they took and the student's background characteristics. To account for error in this imputation process, TIMSS draws five such estimates, or "plausible values," for each student on each of the scales, and incorporates the variability between the five estimates in the standard error of any statistics reported.

The TIMSS mathematics and science achievement scales were designed to provide reliable measures of student achievement spanning 1995, 1999, and 2003. The metric of the scale was established originally with the 1995 assessment. Treating equally all the countries that participated in 1995 at the

eighth grade, the TIMSS scale average over those countries was set at 500 and the standard deviation at 100. The same applied for the fourth-grade assessment. Since the countries varied 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. To preserve the metric of the original 1995 scale, the 1999 eighth-grade assessment was scaled using students from the countries that participated in both 1995 and 1999. Then students from the countries that tested in 1999 but not 1995 were assigned scores on the basis of the scale.

At the eighth grade, TIMSS developed the 2003 scale in the same way as in 1999, preserving the metric first with students from countries that participated in both 1999 and 2003, and then assigning scores on the basis of the scale to students tested in 2003 but not the earlier assessment. At fourth grade, because there was no assessment in 1999, the 2003 and 1995 data were linked directly together using students from countries that participated in both assessments, and the students tested in 2003 but not 1995 were assigned scores on the basis of the scale.

In addition to the scales for mathematics and science overall, TIMSS created IRT scales for each of the mathematics and science content domains for the 2003 data. These included number, algebra, measurement, geometry, and data in mathematics; and life science, chemistry, physics, earth science, and environmental science in science.[2] However, insufficient common items were used in 1995 and 1999 to establish reliable IRT content area scales for trend purposes.

1.15 Data Analysis and Reporting

The TIMSS 2003 International Mathematics Report (Mullis, Martin, Gonzalez, and Chrostowski, 2004) and the TIMSS 2003 International Science Report (Martin, Mullis, Gonzalez, and Chrostowski, 2004) summarize fourth- and eighth- grade students' mathematics and science achievement, respectively, in each participating country. The reports present trend results from 1995 and 1999 at the eighth grade, as well as from 1995 for the fourth grade. Average achievement is reported separately for girls and for boys.

To provide additional information about mathematics and science achievement among high- and low-achieving students, TIMSS reported the percentage of students in each country performing at each of four international benchmarks of student achievement. Selected to represent the range of performance shown by students internationally, the advanced benchmark was 625, the high benchmark was 550, the intermediate benchmark was 475,

2 At the fourth grade, scales were constructed only for life science, physical science, and earth science.

and the low benchmark was 400. Although the fourth- and eighth-grade scales are different, the same benchmark points were used at both grades. To enhance this reporting approach, TIMSS conducted a scale anchoring analysis to describe achievement of students at those four points on the scales. Scale anchoring is a way of describing students' performance at different points on a scale in terms of what they know and can do. It involves a statistical component, in which items that discriminate between successive points on the scale are identified, and a judgmental component, in which subject-matter experts examine the items and generalize to students' knowledge and understandings. Complementing this approach further, the TIMSS 2003 International Reports present examples of mathematics and science items that anchor at each of the benchmarks, and display student performance in each country on the example items.

TIMSS 2003 collected a wide array of information about the homes, schools, classrooms, and teachers of the participating students, as well as about the mathematics and science curriculum in each country. The TIMSS 2003 International Reports summarize much of this information, combining data into composite indices showing an association with achievement where appropriate. In particular, student mathematics and science achievement is described in relation to characteristics of the home, curriculum coverage, classroom instruction, and school environment.

Because the statistics presented in the international reports are estimates of national performance based on samples of students, rather than the values that could be calculated if every student in every country had answered every question, it is important to have measures of the degree of uncertainty of the estimates. The jackknife procedure was used to estimate the standard error associated with each statistic presented in this report. The jackknife standard errors also include an error component due to variation among the five plausible values generated for each student. The use of confidence intervals, based on the standard errors, provides a way to make inferences about the population means and proportions in a manner that reflects the uncertainty associated with the sample estimates. An estimated sample statistic plus or minus two standard errors represents a 95 percent confidence interval for the corresponding population result.

References

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., O'Connor, K.M., Chrostowski, S.J., Gregory, K.D., Garden, R.A., and Smith, T.A. (2001), *Mathematics Benchmarking Report TIMSS 1999 – Eighth Grade: Achievement for U.S. States and Districts in an International Context*. Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Smith, T.A., Garden, R.A., Gregory, K.D., Gonzalez, E.J., Chrostowski, S.J, and O'Connor, K.M. (2003), *TIMSS Assessment Frameworks and Specifications 2003 (2nd Edition)*, Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., and Chrostowski, S.J. (2004), *TIMSS 2003 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. 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., and Garden, R.A. (2001), *Science Benchmarking Report TIMSS 1999 – Eighth Grade: Achievement for U.S. States and Districts in an International Context*. Chestnut Hill, MA: Boston College.

Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., and Chrostowski, S.J. (2004), *TIMSS 2003 International Science Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Chestnut Hill, MA: Boston College.