Appendix A

Overview of TIMSS Procedures for Assessing Science

History

TIMSS 2003 is the latest in 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 almost 20 studies of cross-national achievement in the curricular areas of mathematics, science, language, civics, and reading.

In particular, TIMSS 2003 continues a rich tradition of studies designed to improve teaching and learning in mathematics and science. IEA conducted the pioneering First International Science Study (FISS) in 1970-71 and the Second International Science Study (SISS) in 1983-84. The First and Second International Mathematics Studies (FIMS and SIMS) were conducted in 1964 and 1980-82, respectively. 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. TIMSS 2007 will again assess mathematics and science achievement at fourth and eighth grades, providing previously participating countries an opportunity to extend their trend lines and new countries an opportunity to join a valuable and exciting endeavor.

Participants in TIMSS

Exhibit A.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.

Inspired by the very successful 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 provided an international benchmarking program, whereby regions or localities 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 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.

Exhibit A.1: Countries Participating in TIMSS 2003, 1999, and 1995

Countries	Grade 8			Grade 4	
	2003	1999	1995	2003	1995
[1] 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	●	●		●	
Morocco	●	●		●	
Netherlands	●	●	●	●	●
New Zealand	●	●	●	●	●
Norway	●		●	●	●
Palestinian Nat'l Auth.	●				
Philippines	●	●		●	
Portugal			●		●

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

1 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.

Exhibit A.1: Countries Participating in TIMSS 2003, 1999, and 1995

Countries	Grade 8			Grade 4	
	2003	1999	1995	2003	1995
Romania	●	●	●		
Russian Federation	●	●	●	●	
Saudi Arabia	●				
Scotland	●		●	●	●
Serbia	●				
Singapore	●	●	●	●	●
Slovak Republic	●	●	●		
Slovenia	●	●	●	●	●
South Africa	●	●	●		
Spain			●		
Sweden	●		●		
Switzerland			●		
[2] Syrian Arab Republic	●				
Thailand		●	●		●
Tunisia	●	●		●	
Turkey		●			
United States	●	●	●	●	●
[2] Yemen				●	
Benchmarking Participants					
[2] Basque Country, Spain	●				
Indiana State, US	●	●		●	
[3] Ontario Province, Can.	●	●	●	●	●
[3] Quebec Province, Can.	●	●	●	●	●

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

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

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

Developing the TIMSS 2003 Science Assessment

The development of the TIMSS 2003 science assessment was a collaborative process spanning a two-and-a-half-year period and involving science educators and development specialists from all over the world.[2] 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.[3]

As shown in Exhibit A.2, the science assessment framework for TIMSS 2003 is framed by two organizing dimensions or aspects, a content domain and a cognitive domain. The content domains – life science, chemistry, physics, earth science, and environmental science at the eighth grade and life science, physical science, and earth science at the fourth grade – define the specific science subject matter covered by the assessment. The three cognitive domains – factual knowledge, conceptual understanding, and reasoning and analysis – define the sets of behaviors expected of students as they engage with the science content.

Developing the TIMSS assessments for 2003 was a cooperative venture involving all of the National Research Coordinators (NRCs) during the entire process. Although about half of the items in the 1999 eighth-grade assessment had been kept secure and were available for use in 2003 to measure trends from 1995 and 1999, the ambitious goals for curriculum coverage and innovative problem solving tasks specified in the *Frameworks and Specifications* necessitated a tremendous item development effort.

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, an item development "Task Force" consisting of the science coordinator and two experienced science item writers reviewed general

2 For a full discussion of the TIMSS 2003 test development effort, please see Smith Neidorf, T.A. and Garden, R.A. (2004), "Developing the TIMSS 2003 Mathematics and Science Assessment and Scoring Guides" in M.O. Martin, I.V.S. Mullis and S.J. Chrostowski (eds.), *TIMSS 2003 Technical Report*, Chestnut Hill, MA: Boston College.

3 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.

For the TIMSS frameworks used in 1995 and 1999, see Robitaillle, D.F., McKnight, C.C., Schmidt, W.H., Britton, E.D., Raisen, S.A., and Nicol, C. (1993), *TIMSS Monograph No. 1: Curriculum Frameworks for Mathematics and Science*, Vancouver, BC: Pacific Educational Press.

Exhibit A.2: **The Content and the Cognitive Domains of the Science Framework**

Content Domain

Grade 8

- Life Science
- Chemistry
- Physics
- Earth Science
- Environmental Science

Grade 4

- Life Science
- Physical Science
- Earth Science

Cognitive Domain

- Factual Knowledge
- Conceptual Understanding
- Reasoning and Analysis

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

item-writing guidelines for multiple-choice and constructed-response items and provided specific training in writing science items in accordance with the *TIMSS Assessment Frameworks and Specifications 2003*. In the weeks that followed, more than 2,000 items and scoring guides were drafted and reviewed by the task force. The items were further reviewed by the Science and Mathematics Item Review Committee, a group of internationally prominent mathematics and science educators nominated by participating countries to advise on subject-matter issues in the assessment. Committee members also contributed enormously to the quality of the assessment by helping 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. The NRCs had several opportunities to review the items and scoring criteria. The resulting TIMSS 2003 science tests contained 189 items at the eighth grade and 152 items at the fourth grade.

Exhibit A.3 presents the number and percentage of items, the number of multiple-choice and constructed-response items, and the number of score points in each of the science content domains for eighth and fourth grades. Comparable information is presented for the three cognitive domains. About 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 (see later section on scoring). The total number of score points available for analysis thus somewhat exceeds the number of items (211 and 168

score points for eighth- and fourth-grades, respectively). Less than half of the students' testing time (48% at eighth grade and 46% at fourth grade) was allocated to constructed-response items.

To ensure reliable measurement of trends over time, the TIMSS 2003 assessment included items that had been used in the 1995 and 1999 assessments as well as items developed for the first time in 2003. Exhibit A.4 shows the distribution of score points across content domains for both trend items and items used for the first time. Of the 211 score points available in the entire 2003 science assessment, 24 came from items used also in 1995, 52 from items used also in 1999, and 135 from items used for the first time in 2003. At fourth grade, 33 score points came from 1995 items, and the remaining 135 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, indicates that omitting such items has little effect on the overall pattern of results.

Exhibit A.3: **Distribution of Science Items by Content Domain and Cognitive Domain**

Content Domain	Percentage of Items	Total Number of Items	Number of Multiple-Choice Items	Number of Constructed-Response Items[1]	Number of Score Points[2]
Life Science	29	54	29	25	65
Chemistry	16	31	20	11	34
Physics	24	46	28	18	49
Earth Science	16	31	22	9	33
Environmental Science	14	27	10	17	30
Total	100	189	109	80	211

Cognitive Domain	Percentage of Items	Total Number of Items	Number of Multiple-Choice Items	Number of Constructed-Response Items[1]	Number of Score Points[2]
Factual Knowledge	30	57	50	7	59
Conceptual Understanding	39	73	42	31	80
Reasoning and Analysis	31	59	17	42	72
Total	100	189	109	80	211

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

1 Constructed-response items include both short-answer and extended-response types. Because results are rounded to the nearest whole number, some totals may appear inconsistent.

2 In scoring the tests, correct answers to most items were worth one point. However, responses to some constructed-response items were evaluated for partial credit with a fully correct answer awarded two points. Thus, the number of score points exceeds the number of items in the test.

Exhibit A.3: **Distribution of Science Items by Content Domain and Cognitive Domain**

Content Domain	Percentage of Items	Total Number of Items	Number of Multiple-Choice Items	Number of Constructed-Response Items[1]	Number of Score Points[2]
Life Science	43	65	41	24	72
Physical Science	35	53	29	24	59
Earth Science	22	34	21	13	37
Total	100	152	91	61	168

Cognitive Domain	Percentage of Items	Total Number of Items	Number of Multiple-Choice Items	Number of Constructed-Response Items[1]	Number of Score Points[2]
Factual Knowledge	35	54	41	13	59
Conceptual Understanding	42	64	38	26	70
Reasoning and Analysis	23	34	12	22	39
Total	100	152	91	61	168

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

1 Constructed-response items include both short-answer and extended-response types.

2 In scoring the tests, correct answers to most items were worth one point. However, responses to some constructed-response items were evaluated for partial credit with a fully correct answer awarded two points. Thus, the number of score points exceeds the number of items in the test.

Exhibit A.4: **Distribution of Score Points in TIMSS 2003 from Each Assessment Year by Science Content Domain**

Grade 8

Content Domain	From 1995	From 1999	New in 2003	Total
Life Science	6	12	47	65
Chemistry	4	11	19	34
Physics	5	17	27	49
Earth Science	6	6	21	33
Environmental Science	3	6	21	30
Total	24	52	135	211

Grade 4

Content Domain	From 1995	From 1999	New in 2003	Total
Life Science	12	N/A	60	72
Physical Science	9	N/A	50	59
Earth Science	12	N/A	25	37
Total	33	N/A	135	168

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

TIMSS 2003 Assessment Design

Not all of the students in the TIMSS assessment responded to all of the science 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.

Exhibit A.5 summarizes the TIMSS 2003 assessment design, presenting both the matrix-sampling item blocks for mathematics and science and the item block-to-booklet assignment plan. According to the design, the 313 mathematics and science items at fourth grade and 383 items at eighth grade are 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 contains either mathematics items only or science items only. This general block design is the same for both grades, although the planned assessment time per block is 12 minutes for fourth grade and 15 minutes for eighth grade. At the eighth grade, six blocks in each subject (blocks 01 – 06) contain secure items from 1995 and 1999 to measure trends and eight blocks (07 – 14) contain new items developed for TIMSS 2003. Since fourth grade was not included in the 1999 assessment, trend items from 1995 only were available, and these were placed in the first three blocks. The remaining 11 blocks contain items new in 2003.

In the TIMSS 2003 design, the 28 blocks of items are distributed across 12 student booklets, as shown in Exhibit A.5. Each booklet consists of six blocks of items. To enable linking between booklets, each block appears in two, three, or four different booklets. The assessment time for individual students is 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 are 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 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.

Exhibit A.5: TIMSS 2003 Assessment Design

TIMSS 2003 Item Blocks for Matrix-Sampling

Source of Items	Mathematics Blocks	Science Blocks
Trend Items (TIMSS 1995 or 1999)	M01	S01
Trend Items (TIMSS 1995 or 1999)	M02	S02
Trend Items (TIMSS 1995 or 1999)	M03	S03
Trend Items (TIMSS 1999)	M04	S04
Trend Items (TIMSS 1999)	M05	S05
Trend Items (TIMSS 1999)	M06	S06
New Replacement Items (TIMSS 2003)	M07	S07
New Replacement Items (TIMSS 2003)	M08	S08
New Replacement Items (TIMSS 2003)	M09	S09
New Replacement Items (TIMSS 2003)	M10	S10
New Replacement Items (TIMSS 2003)	M11	S11
New Replacement Items (TIMSS 2003)	M12	S12
New Replacement Items (TIMSS 2003)	M13	S13
New Replacement Items (TIMSS 2003)	M14	S14

Booklet Design for TIMSS 2003

Student Booklet	Part I			Part II		
Booklet 1	M01	M02	S06	S07	M05	M07
Booklet 2	M02	M03	S05	S08	M06	M08
Booklet 3	M03	M04	S04	S09	M13	M11
Booklet 4	M04	M05	S03	S10	M14	M12
Booklet 5	M05	M06	S02	S11	M09	M13
Booklet 6	M06	M01	S01	S12	M10	M14
Booklet 7	S01	S02	M06	M07	S05	S07
Booklet 8	S02	S03	M05	M08	S06	S08
Booklet 9	S03	S04	M04	M09	S13	S11
Booklet 10	S04	S05	M03	M10	S14	S12
Booklet 11	S05	S06	M02	M11	S09	S13
Booklet 12	S06	S01	M01	M12	S10	S14

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

Background Questionnaires

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. This work began with articulating the information to be collected in the TIMSS 2003 framework and continued with extensive field testing.[4]

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.

4 For more information, see Chrostowski, S.J. (2004), "Developing the TIMSS 2003 Background Questionnaires" in M.O. Martin, I.V.S. Mullis, and S.J. Chrostowski (eds.), *TIMSS 2003 Technical Report*, Chestnut Hill, MA: Boston College.

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.[5]

5 More details about the translation verification procedures can be found in Chrostowski, S.J. and Malak, B. (2004), "Translation and Cultural Adaptation of the TIMSS 2003 Instruments" in M.O. Martin, I.V.S. Mullis, and S.J. Chrostowski (eds.), *TIMSS 2003 Technical Report*, Chestnut Hill, MA: Boston College.

Population Definition and Sampling

Since it is a curriculum-based study, 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 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, and for fourth grade, all students enrolled in the upper of the two adjacent grades that contained the largest proportion of 9-year-olds. These correspond to the eighth and fourth grades in practically every country.[6]

The selection of valid and efficient samples is crucial to the quality and success of an international comparative study such as TIMSS. The accuracy of the survey results depends on the quality of sampling information and that of the sampling activities themselves. For TIMSS, NRCs worked on all phases of sampling with the TIMSS sampling experts from Statistics Canada and the IEA Data Processing Center (DPC). NRCs received training in how to select the school and student samples and in the use of the sampling software. In consultation with the TIMSS sampling referee (Keith Rust, Westat, Inc.), the TIMSS sampling experts reviewed the national sampling plans, sampling data, sampling frames, and sample execution. The sampling documentation was used by the TIMSS & PIRLS International Study Center, in consultation with the sampling experts and the sampling referee, to evaluate the quality of the samples.

In a few situations where it was not possible to test the entire internationally desired population (all students enrolled in the upper of the two adjacent grades that contained the largest proportion of 13-year-old or 9-year-old students at the time of testing), countries were permitted to define a national desired population that excluded part of the internationally desired population. Exhibit A.6 shows any differences in coverage between the international and national desired populations for eighth and fourth grades. Almost all participants at the

6 The sample design for TIMSS is described in detail in Foy, P., and Joncas, M. (2004), "TIMSS 2003 Sampling Design" in M.O. Martin, I.V.S. Mullis and S.J. Chrostowski (eds.), *TIMSS 2003 Technical Report*, Chestnut Hill, MA: Boston College.

eighth grade achieved 100 percent coverage (47 out of 51), with Indonesia, Lithuania, Morocco, and Serbia the exceptions. Consequently, the results for these countries are annotated in exhibits in this report. At fourth grade, only Lithuania of the 29 participants had less than 100 percent coverage.

Within the desired population, countries could define a population that excluded a small percentage (less than five percent) of certain kinds of schools or students that would be very difficult or resource-intensive to test (e.g., schools for students with special needs or schools that were very small or located in extremely rural areas). Countries excluding more than 10 percent of their population are annotated in the exhibits in this report. Exhibit A.6 shows that only three countries exceeded the 10 percent limit at eighth grade (Israel, Macedonia, and Syria) and no fourth-grade participant did so.

Within countries, TIMSS used a two-stage sample design, in which the first stage involved selecting about 150 public and private schools in each country. Within each school, countries were to use random procedures to select one eighth-grade mathematics class (for eighth-grade participants) and one fourth-grade classroom (fourth-grade participants). All of the students in the sampled class were to participate in the TIMSS testing. This approach was designed to yield a representative sample of at least 4,000 students per country at each grade level. Typically, between 1,200 and 2,000 students responded to each achievement item in each country, depending on the booklets in which the items appeared.

Exhibits A.7 and A.8 present achieved sample sizes for schools and students, respectively, for participating countries. Exhibit A.9 shows the participation rates for schools, students, and overall, both with and without the use of replacement schools. Most countries achieved the minimum acceptable participation rates – 85 percent of both the schools and students, or a combined rate (the product of school and student participation) of 75 percent – although Hong Kong SAR, the Netherlands, and Scotland did so only after including replacement

Exhibit A.6: Coverage of TIMSS 2003 Target Population

SCIENCE Grade 8

Countries	International Desired Population		National Desired Population		
	Coverage	Notes on Coverage	School-Level Exclusions	Within-Sample Exclusions	Overall Exclusions
Armenia	100%		2.9%	0.0%	2.9%
Australia	100%		0.4%	0.9%	1.3%
Bahrain	100%		0.0%	0.0%	0.0%
Belgium (Flemish)	100%		3.1%	0.1%	3.2%
Botswana	100%		0.8%	2.2%	3.0%
Bulgaria	100%		0.5%	0.0%	0.5%
Chile	100%		1.6%	0.7%	2.2%
Chinese Taipei	100%		0.2%	4.6%	4.8%
Cyprus	100%		1.1%	1.5%	2.5%
Egypt	100%		3.4%	0.0%	3.4%
England	100%		2.1%	0.0%	2.1%
Estonia	100%		2.6%	0.8%	3.4%
Ghana	100%		0.9%	0.0%	0.9%
Hong Kong, SAR	100%		3.3%	0.1%	3.4%
Hungary	100%		5.5%	3.2%	8.5%
Indonesia	80%	Non-islamic schools	0.1%	0.3%	0.4%
Iran, Islamic Rep. of	100%		5.5%	1.1%	6.5%
Israel	100%		15.2%	8.6%	22.5%
Italy	100%		0.0%	3.6%	3.6%
Japan	100%		0.5%	0.1%	0.6%
Jordan	100%		0.5%	0.8%	1.3%
Korea, Rep. of	100%		1.5%	3.4%	4.9%
Latvia	100%		3.6%	0.1%	3.7%
Lebanon	100%		1.4%	0.0%	1.4%
Lithuania	89%	Students taught in Lithuanian	1.4%	1.2%	2.6%
Macedonia, Rep. of	100%		12.5%	0.0%	12.5%
Malaysia	100%		4.0%	0.0%	4.0%
Moldova, Rep. of	100%		0.7%	0.5%	1.2%
Morocco	69%	All students but Souss Massa Draa, Casablanca, Gharb-Chrarda	1.5%	0.0%	1.5%
Netherlands	100%		3.0%	0.0%	3.0%
New Zealand	100%		1.7%	2.7%	4.4%
Norway	100%		0.9%	1.5%	2.3%
Palestinian Nat'l Auth.	100%		0.2%	0.3%	0.5%
Philippines	100%		1.5%	0.0%	1.5%
Romania	100%		0.4%	0.1%	0.5%
Russian Federation	100%		1.7%	3.9%	5.5%
Saudi Arabia	100%		0.3%	0.2%	0.5%
Scotland	100%		0.0%	0.0%	0.0%
Serbia	81%	Serbia without Kosovo	2.4%	0.6%	2.9%
Singapore	100%		0.0%	0.0%	0.0%
Slovak Republic	100%		5.0%	0.0%	5.0%
Slovenia	100%		1.3%	0.1%	1.4%
South Africa	100%		0.6%	0.0%	0.6%
Sweden	100%		0.3%	2.5%	2.8%
Syrian Arab Republic	100%		18.7%	0.0%	18.8%
Tunisia	100%		1.8%	0.0%	1.8%
United States	100%		0.0%	4.9%	4.9%
Benchmarking Participants					
Basque Region, Spain	100%		2.1%	3.8%	5.8%
Indiana State, US	100%		0.0%	7.8%	7.8%
Ontario Province, Can.	100%		1.0%	5.0%	6.0%
Quebec Province, Can.	100%		1.4%	3.5%	4.8%

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

Exhibit A.6: Coverage of TIMSS 2003 Target Population

Countries	International Desired Population		National Desired Population		
	Coverage	Notes on Coverage	School-Level Exclusions	Within-Sample Exclusions	Overall Exclusions
Armenia	100%		2.9%	0.0%	2.9%
Australia	100%		1.2%	1.6%	2.7%
Belgium (Flemish)	100%		5.9%	0.4%	6.3%
Chinese Taipei	100%		0.3%	2.8%	3.1%
Cyprus	100%		1.5%	1.4%	2.9%
England	100%		1.9%	0.0%	1.9%
Hong Kong, SAR	100%		3.7%	0.1%	3.8%
Hungary	100%		4.4%	3.9%	8.1%
Iran, Islamic Rep. of	100%		3.6%	2.1%	5.7%
Italy	100%		0.1%	4.1%	4.2%
Japan	100%		0.4%	0.3%	0.8%
Latvia	100%		4.3%	0.1%	4.4%
Lithuania	92%	Students taught in Lithuanian	2.1%	2.6%	4.6%
Moldova, Rep. of	100%		2.0%	1.6%	3.6%
Morocco	100%		2.2%	0.0%	2.2%
Netherlands	100%		4.1%	1.1%	5.2%
New Zealand	100%		1.5%	2.5%	4.0%
Norway	100%		1.7%	2.7%	4.4%
Philippines	100%		3.8%	0.7%	4.5%
Russian Federation	100%		2.2%	4.7%	6.8%
Scotland	100%		1.5%	0.0%	1.5%
Singapore	100%		0.0%	0.0%	0.0%
Slovenia	100%		0.8%	0.5%	1.3%
Tunisia	100%		0.9%	0.0%	0.9%
United States	100%		0.0%	5.1%	5.1%
Yemen	100%		0.6%	8.9%	9.5%
Benchmarking Participants					
Indiana State, US	100%		0.0%	7.2%	7.2%
Ontario Province, Can.	100%		1.3%	3.5%	4.8%
Quebec Province, Can.	100%		2.7%	0.9%	3.6%

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

Exhibit A.7: School Sample Sizes

Countries	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
Armenia	150	150	149	0	149
Australia	230	226	186	21	207
Bahrain	67	67	67	0	67
Belgium (Flemish)	150	150	122	26	148
Botswana	152	150	146	0	146
Bulgaria	170	169	163	1	164
Chile	195	195	191	4	195
Chinese Taipei	150	150	150	0	150
Cyprus	59	59	59	0	59
Egypt	217	217	215	2	217
England	160	160	62	25	87
Estonia	154	152	151	0	151
Ghana	150	150	150	0	150
Hong Kong, SAR	150	150	112	13	125
Hungary	160	157	154	1	155
Indonesia	150	150	148	2	150
Iran, Islamic Rep. of	188	181	181	0	181
Israel	150	147	143	3	146
Italy	172	171	164	7	171
Japan	150	150	146	0	146
Jordan	150	140	140	0	140
Korea, Rep. of	151	150	149	0	149
Latvia	150	149	137	3	140
Lebanon	160	160	148	4	152
Lithuania	150	150	137	6	143
Macedonia, Rep. of	150	150	142	7	149
Malaysia	150	150	150	0	150
Moldova, Rep. of	150	149	147	2	149
Morocco	227	165	131	0	131
Netherlands	150	150	118	12	130
New Zealand	175	174	149	20	169
Norway	150	150	138	0	138
Palestinian Nat'l Auth.	150	145	145	0	145
Philippines	160	160	132	5	137
Romania	150	149	148	0	148
Russian Federation	216	216	214	0	214
Saudi Arabia	160	160	154	1	155
Scotland	150	150	115	13	128
Serbia	150	150	149	0	149
Singapore	164	164	164	0	164
Slovak Republic	180	179	170	9	179
Slovenia	177	177	169	5	174
South Africa	265	265	241	14	255
Sweden	160	160	155	4	159
Syrian Arab Republic	150	150	121	13	134
Tunisia	150	150	150	0	150
United States	301	296	211	21	232
Benchmarking Participants					
Basque Region, Spain	120	120	119	1	120
Indiana State, US	56	56	54	0	54
Ontario Province, Can.	200	196	171	15	186
Quebec Province, Can.	199	185	173	2	175

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

Exhibit A.7: School Sample Sizes

Countries	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
Armenia	150	150	148	0	148
Australia	230	227	178	26	204
Belgium (Flemish)	150	150	133	16	149
Chinese Taipei	150	150	150	0	150
Cyprus	150	150	150	0	150
England	150	150	79	44	123
Hong Kong, SAR	150	150	116	16	132
Hungary	160	159	156	1	157
Iran, Islamic Rep. of	176	171	171	0	171
Italy	172	171	165	6	171
Japan	150	150	150	0	150
Latvia	150	149	137	3	140
Lithuania	160	160	147	6	153
Moldova, Rep. of	153	151	147	4	151
Morocco	227	225	197	0	197
Netherlands	150	149	77	53	130
New Zealand	228	228	194	26	220
Norway	150	150	134	5	139
Philippines	160	160	122	13	135
Russian Federation	206	205	204	1	205
Scotland	150	150	94	31	125
Singapore	182	182	182	0	182
Slovenia	177	177	169	5	174
Tunisia	150	150	150	0	150
United States	310	300	212	36	248
Yemen	150	150	150	0	150
Benchmarking Participants					
Indiana State, US	56	56	56	0	56
Ontario Province, Can.	200	196	179	10	189
Quebec Province, Can.	198	194	192	1	193

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

Exhibit A.8: Student Sample Sizes

SCIENCE Grade 8

Countries	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 Eligible Students	Number of Students Absent	Number of Students Assessed
Armenia	90%	6388	56	0	6332	606	5726
Australia	93%	5286	60	16	5210	419	4791
Bahrain	98%	4351	64	0	4287	88	4199
Belgium (Flemish)	97%	5161	19	7	5135	165	4970
Botswana	98%	5388	70	70	5248	98	5150
Bulgaria	96%	4489	167	0	4322	205	4117
Chile	99%	6528	15	39	6474	97	6377
Chinese Taipei	99%	5525	54	37	5434	55	5379
Cyprus	96%	4314	79	66	4169	167	4002
Egypt	97%	7259	0	0	7259	164	7095
England	86%	3360	34	0	3326	496	2830
Estonia	96%	4242	28	5	4209	169	4040
Ghana	93%	5690	189	0	5501	401	5100
Hong Kong, SAR	97%	5204	33	4	5167	195	4972
Hungary	95%	3506	7	34	3465	163	3302
Indonesia	99%	5884	61	0	5823	61	5762
Iran, Islamic Rep. of	98%	5215	118	52	5045	103	4942
Israel	95%	4880	2	319	4559	241	4318
Italy	97%	4628	35	173	4420	142	4278
Japan	96%	5121	51	5	5065	209	4856
Jordan	96%	4871	176	41	4654	165	4489
Korea, Rep. of	99%	5451	18	50	5383	74	5309
Latvia	89%	4146	23	5	4118	488	3630
Lebanon	96%	4030	64	0	3966	152	3814
Lithuania	89%	6619	58	955	5606	642	4964
Macedonia, Rep. of	97%	4028	0	0	4028	135	3893
Malaysia	98%	5464	46	0	5418	104	5314
Moldova, Rep. of	96%	4262	58	0	4204	171	4033
Morocco	91%	3243	25	0	3218	275	2943
Netherlands	94%	3283	2	0	3281	216	3065
New Zealand	93%	4343	170	65	4108	307	3801
Norway	92%	4569	24	61	4484	351	4133
Palestinian Nat'l Auth.	99%	5543	117	14	5412	55	5357
Philippines	96%	7498	288	0	7210	293	6917
Romania	98%	4249	53	4	4192	88	4104
Russian Federation	97%	4926	50	62	4814	147	4667
Saudi Arabia	97%	4553	115	5	4433	138	4295
Scotland	89%	3962	24	0	3938	422	3516
Serbia	96%	4514	52	2	4460	164	4296
Singapore	97%	6236	5	0	6231	213	6018
Slovak Republic	95%	4428	16	0	4412	197	4215
Slovenia	93%	3883	19	2	3862	284	3578
South Africa	92%	9905	320	0	9585	633	8952
Sweden	89%	4941	58	93	4790	534	4256
Syrian Arab Republic	98%	5001	0	1	5000	105	4895
Tunisia	98%	5106	74	0	5032	101	4931
United States	94%	9891	90	279	9522	610	8912
Benchmarking Participants							
Basque Region, Spain	98%	2736	41	113	2582	68	2514
Indiana State, US	97%	2402	43	107	2252	64	2188
Ontario Province, Can.	95%	4693	59	208	4426	209	4217
Quebec Province, Can.	92%	4919	78	46	4795	384	4411

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

Exhibit A.8: Student Sample Sizes

Countries	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 Eligible Students	Number of Students Absent	Number of Students Assessed
Armenia	91%	6275	57	0	6218	544	5674
Australia	94%	4675	69	39	4567	246	4321
Belgium (Flemish)	98%	4866	17	20	4829	117	4712
Chinese Taipei	99%	4793	11	88	4694	33	4661
Cyprus	97%	4536	27	60	4449	121	4328
England	93%	3917	45	0	3872	287	3585
Hong Kong, SAR	95%	4901	23	4	4874	266	4608
Hungary	94%	3603	11	67	3525	206	3319
Iran, Islamic Rep. of	98%	4587	83	80	4424	72	4352
Italy	97%	4641	23	185	4433	151	4282
Japan	97%	4690	16	16	4658	123	4535
Latvia	94%	3980	16	4	3960	273	3687
Lithuania	92%	5701	35	852	4814	392	4422
Moldova, Rep. of	97%	4162	46	0	4116	135	3981
Morocco	93%	4546	0	0	4546	282	4264
Netherlands	96%	3080	0	30	3050	113	2937
New Zealand	95%	4785	145	107	4533	225	4308
Norway	95%	4706	22	107	4577	235	4342
Philippines	95%	5225	40	31	5154	582	4572
Russian Federation	97%	4229	54	66	4109	146	3963
Scotland	92%	4283	34	0	4249	313	3936
Singapore	98%	6851	16	0	6835	167	6668
Slovenia	92%	3410	13	17	3380	254	3126
Tunisia	99%	4408	23	0	4385	51	4334
United States	95%	10795	49	429	10317	488	9829
Yemen	93%	4550	0	0	4550	345	4205
Benchmarking Participants							
Indiana State, US	98%	2472	44	151	2277	44	2233
Ontario Province, Can.	96%	4813	91	158	4564	202	4362
Quebec Province, Can.	91%	4864	51	73	4740	390	4350

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

Exhibit A.9: Participation Rates (Weighted)

Countries	School Participation		Class Participation	Student Participation	Overall Participation	
	Before Replacement	After Replacement			Before Replacement	After Replacement
Armenia	99%	99%	99%	90%	89%	89%
Australia	81%	90%	100%	93%	75%	83%
Bahrain	100%	100%	100%	98%	98%	98%
Belgium (Flemish)	82%	99%	98%	97%	77%	94%
Botswana	98%	98%	100%	98%	96%	96%
Bulgaria	97%	97%	99%	96%	92%	92%
Chile	98%	100%	100%	99%	97%	99%
Chinese Taipei	100%	100%	100%	99%	99%	99%
Cyprus	100%	100%	100%	96%	96%	96%
Egypt	99%	100%	100%	97%	97%	97%
England	40%	54%	99%	86%	34%	46%
Estonia	99%	99%	100%	96%	95%	95%
Ghana	100%	100%	100%	93%	93%	93%
Hong Kong, SAR	74%	83%	99%	97%	72%	80%
Hungary	98%	99%	100%	95%	94%	94%
Indonesia	98%	100%	100%	99%	97%	99%
Iran, Islamic Rep. of	100%	100%	100%	98%	98%	98%
Israel	98%	99%	100%	95%	93%	94%
Italy	96%	100%	100%	97%	93%	97%
Japan	97%	97%	100%	96%	93%	93%
Jordan	100%	100%	100%	96%	96%	96%
Korea, Rep. of	99%	99%	100%	99%	98%	98%
Latvia	92%	94%	100%	89%	81%	83%
Lebanon	93%	95%	100%	96%	89%	91%
Lithuania	92%	95%	100%	89%	81%	84%
Macedonia, Rep. of	94%	99%	100%	97%	91%	96%
Malaysia	100%	100%	100%	98%	98%	98%
Moldova, Rep. of	99%	100%	100%	96%	95%	96%
Morocco	79%	79%	100%	91%	71%	71%
Netherlands	79%	87%	100%	94%	74%	81%
New Zealand	86%	97%	100%	93%	80%	90%
Norway	92%	92%	100%	92%	85%	85%
Palestinian Nat'l Auth.	100%	100%	100%	99%	99%	99%
Philippines	81%	86%	100%	96%	78%	82%
Romania	99%	99%	100%	98%	98%	98%
Russian Federation	99%	99%	100%	97%	96%	96%
Saudi Arabia	95%	97%	100%	97%	93%	94%
Scotland	76%	85%	100%	89%	68%	76%
Serbia	99%	99%	100%	96%	96%	96%
Singapore	100%	100%	100%	97%	97%	97%
Slovak Republic	96%	100%	100%	95%	91%	95%
Slovenia	94%	99%	100%	93%	87%	91%
South Africa	89%	96%	100%	92%	82%	88%
Sweden	97%	99%	99%	89%	85%	87%
Syrian Arab Republic	81%	89%	100%	98%	79%	87%
Tunisia	100%	100%	100%	98%	98%	98%
United States	71%	78%	99%	94%	66%	73%
Benchmarking Participants						
Basque Region, Spain	100%	100%	100%	98%	97%	98%
Indiana State, US	97%	97%	100%	97%	94%	94%
Ontario Province, Can.	84%	93%	100%	95%	80%	89%
Quebec Province, Can.	91%	93%	100%	92%	84%	85%

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

Exhibit A.9: Participation Rates (Weighted)

Countries	School Participation		Class Participation	Student Participation	Overall Participation	
	Before Replacement	After Replacement			Before Replacement	After Replacement
Armenia	99%	99%	100%	91%	90%	90%
Australia	78%	90%	100%	94%	73%	85%
Belgium (Flemish)	89%	99%	100%	98%	87%	97%
Chinese Taipei	100%	100%	100%	99%	99%	99%
Cyprus	100%	100%	100%	97%	97%	97%
England	54%	82%	100%	93%	50%	76%
Hong Kong, SAR	77%	88%	99%	95%	73%	83%
Hungary	98%	99%	100%	94%	92%	93%
Iran, Islamic Rep. of	100%	100%	100%	98%	98%	98%
Italy	97%	100%	100%	97%	93%	97%
Japan	100%	100%	100%	97%	97%	97%
Latvia	91%	94%	100%	94%	85%	88%
Lithuania	92%	96%	99%	92%	84%	87%
Moldova, Rep. of	97%	100%	100%	97%	94%	97%
Morocco	87%	87%	100%	93%	81%	81%
Netherlands	52%	87%	100%	96%	50%	84%
New Zealand	87%	98%	100%	95%	82%	93%
Norway	89%	93%	100%	95%	85%	88%
Philippines	78%	85%	100%	95%	75%	81%
Russian Federation	99%	100%	100%	97%	96%	97%
Scotland	64%	83%	100%	92%	59%	77%
Singapore	100%	100%	100%	98%	98%	98%
Slovenia	95%	99%	100%	92%	87%	91%
Tunisia	100%	100%	100%	99%	99%	99%
United States	70%	82%	99%	95%	66%	78%
Yemen	100%	100%	100%	93%	93%	93%
Benchmarking Participants						
Indiana State, US	100%	100%	100%	98%	98%	98%
Ontario Province, Can.	89%	94%	100%	96%	85%	90%
Quebec Province, Can.	99%	100%	100%	91%	90%	91%

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

schools. The United States and Morocco had overall participation rates after including replacement schools of just below 75 percent (73% and 71%, respectively), and were annotated accordingly. Despite extraordinary efforts to secure full participation, England's participation fell below the minimum requirement of 50 percent, and so their results were annotated and placed below a line in exhibits showing achievement. Because of scheduling difficulties, Korea was unable to test its eighth-grade students in May 2003 as planned. Instead, the students were tested in September 2003, when they had moved into the ninth grade. The results for Korea are annotated accordingly in exhibits in this report.

At fourth grade, all participants achieved the minimum acceptable participation rates, although Australia, England, Hong Kong SAR, the Netherlands, Scotland, and the United States did so only after including replacement schools.

Whereas countries achieved a high degree of compliance with sampling guidelines in 2003, occasionally countries' data were omitted from exhibits dealing with trends from earlier assessments because of comparability issues. Because of differences in population coverage, 1999 eighth-grade data for Australia, Morocco, and Slovenia and fourth-grade data for Italy are not shown in this report. Israel, Italy, and South Africa, experienced difficulties with sampling at the classroom level in 1995; consequently their eighth-grade data from that assessment are not shown in this report.

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.[7]

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 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.[8] 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),[9] and conducted interviews with the National Research Coordinator in each of the participating countries.[10]

The results of the interviews indicate that, in general, NRCs had prepared well for data collection and, despite the heavy demands

7 Data collection procedures for TIMSS is described in detail in Barth, J., Gonzalez, E.J., and Neuschmidt, O. (2004), "TIMSS 2003 Survey Operations Procedures" in M.O. Martin, I.V.S. Mullis and S.J. Chrostowski (eds.), *TIMSS 2003 Technical Report*, Chestnut Hill, MA: Boston College.

8 Iran and Israel were the only countries whose quality control monitors were not trained; Ontario and Quebec shared the same quality control monitor.

9 Operational constraints prevented quality control monitor visits in five testing sessions in Japan.

10 Steps taken to ensure high-quality data collection in TIMSS are described in detail in Gonzalez, E.J. and Diaconu, D. (2004), "Quality Assurance in the TIMSS 2003 Data Collection" in M.O. Martin, I.V.S. Mullis, and S.J. Chrostowski (eds.), *TIMSS 2003 Technical Report*, Chestnut Hill, MA: Boston College.

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.

Scoring the Constructed-Response Items

Because 40 to 50 percent of the test 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 the within-country agreement among scorers, TIMSS arranged to have systematic samples of at least 100 student responses to each item scored independently by two readers. Exhibit A.10 shows the average and range of the within-country exact percent of agreement between scorers on the

constructed-response items in the science test for the TIMSS participants. The exhibit shows agreement for both the correctness score (the first digit) and for the two-digit diagnostic score. A high percentage of exact agreement was observed, with an overall average of 97 percent for correctness score and 92 percent for diagnostic score at the eighth grade and 96 and 92 percent, respectively at the fourth grade. The TIMSS data from the reliability studies indicate that scoring procedures were robust for the science items, especially for the correctness score used for the analyses in this report.

The double scoring of a sample of the student test booklets provided a measure of the consistency within each country with which constructed-response questions were scored. 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. As shown in Exhibit A.11, there was a very high degree of scoring consistency, with 92 percent exact agreement, on average, internationally, between the scores awarded in 1999 and those given by the 2003 scorers. There was somewhat less agreement at the diagnostic score level, with 81 percent exact agreement, on average.

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 21 constructed-response science questions. This set of 3,150 student responses was then sent to each Northern-Hemisphere

Exhibit A.10: TIMSS 2003 Within-Country Scoring Reliability for the Constructed-Response Science Items

SCIENCE Grade 8

Countries	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Armenia	98	92	100	97	90	100
Australia	99	94	100	97	89	100
Bahrain	98	94	100	95	85	100
Belgium (Flemish)	97	89	100	93	83	100
Botswana	95	74	100	87	74	97
Bulgaria	91	72	99	84	64	99
Chile	97	91	100	94	89	99
Chinese Taipei	99	97	100	98	86	100
Cyprus	96	87	100	91	80	99
Egypt	100	98	100	100	97	100
England	98	92	100	96	85	100
Estonia	99	97	100	98	88	100
Ghana	98	93	100	93	83	99
Hong Kong, SAR	99	97	100	97	92	100
Hungary	96	87	100	92	83	100
Indonesia	96	87	100	86	68	99
Iran, Islamic Rep. of	98	87	100	95	84	100
Israel	95	89	100	84	66	98
Italy	98	91	100	96	90	100
Japan	97	81	100	93	80	100
Jordan	99	97	100	96	91	100
Korea, Rep. of	98	84	100	95	74	100
Latvia	94	78	100	87	50	100
Lebanon	100	98	100	99	95	100
Lithuania	90	69	100	82	58	100
Macedonia, Rep. of	99	96	100	97	92	100
Malaysia	99	98	100	99	97	100
Moldova, Rep. of	100	99	100	100	99	100
Morocco	94	86	100	86	69	95
Netherlands	90	70	100	84	61	100
New Zealand	98	92	100	93	84	100
Norway	95	83	100	91	80	100
Palestinian Nat'l Auth.	95	82	100	87	69	99
Philippines	98	89	100	94	83	99
Romania	99	96	100	98	94	100
Russian Federation	99	92	100	98	91	100
Saudi Arabia	97	87	100	91	68	99
Scotland	97	89	100	94	85	100
Serbia	99	94	100	98	92	100
Singapore	100	99	100	99	98	100
Slovak Republic	99	95	100	97	89	100
Slovenia	90	70	100	81	61	100
South Africa	99	94	100	96	88	99
Sweden	92	76	100	85	68	99
Tunisia	98	90	100	94	73	100
United States	92	72	100	83	68	99
International Avg.	97	88	100	92	80	99
Benchmarking Participants						
Basque Country, Spain	96	87	100	92	79	100
Indiana State, US	94	82	100	87	67	100
Ontario Province, Can.	91	77	100	83	62	98
Quebec Province, Can.	92	80	100	84	66	100

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

Exhibit A.10: TIMSS 2003 Within-Country Scoring Reliability for the Constructed-Response Science Items

Countries	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Armenia	99	97	100	97	91	100
Australia	99	94	100	98	91	100
Belgium (Flemish)	99	89	100	95	86	100
Chinese Taipei	98	89	100	96	89	100
Cyprus	94	76	100	89	75	99
England	98	87	100	96	86	100
Hong Kong, SAR	99	97	100	97	89	100
Hungary	95	80	100	91	78	100
Iran, Islamic Rep. of	96	85	100	93	83	99
Italy	94	77	100	90	77	100
Japan	97	86	100	94	83	100
Latvia	96	82	100	92	71	99
Lithuania	93	81	100	86	50	99
Moldova, Rep. of	100	100	100	100	100	100
Morocco	97	93	100	92	78	99
Netherlands	91	71	99	84	70	99
New Zealand	97	86	100	92	83	99
Norway	97	85	100	93	84	100
Philippines	97	89	100	91	77	99
Russian Federation	99	98	100	99	96	100
Scotland	98	90	100	96	85	100
Singapore	100	99	100	99	97	100
Slovenia	91	74	100	85	69	100
Tunisia	93	79	100	82	68	96
United States	93	70	100	86	68	99
International Avg.	96	85	100	92	80	99
Benchmarking Participants						
Indiana State, US	95	76	100	92	62	100
Ontario Province, Can.	95	80	100	90	75	100
Quebec Province, Can.	95	81	100	89	72	99

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

Exhibit A.11: TIMSS 2003 Trend Scoring Reliability (1999–2003) for the Constructed-Response Science Items

Countries	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Australia	93	75	100	81	56	100
Belgium (Flemish)	92	79	100	83	68	100
Bulgaria	96	87	100	83	45	100
Chile	91	80	100	77	47	100
Chinese Taipei	92	70	100	80	38	100
Cyprus	90	70	99	79	50	99
Hong Kong, SAR	89	74	100	80	58	100
Hungary	92	74	100	84	64	100
Indonesia	90	63	100	75	41	97
Iran, Islamic Rep.	92	68	100	82	55	99
Israel	93	80	100	81	46	100
Italy	94	86	100	88	73	100
Japan	92	72	100	84	62	100
Jordan	96	90	100	87	76	99
Korea, Rep. of	93	77	100	85	56	100
Latvia	79	36	100	65	21	98
Lithuania	86	66	100	74	40	100
Macedonia, Rep. of	99	89	100	98	80	100
Malaysia	92	80	100	74	35	100
New Zealand	94	87	99	79	52	98
Philippines	90	44	100	76	32	100
Romania	96	91	100	90	73	100
Russian Federation	93	80	100	79	55	99
Singapore	97	93	100	88	61	100
Slovak Republic	89	73	100	76	56	100
Slovenia	94	71	100	90	72	100
South Africa	93	71	100	79	19	100
United States	94	83	100	84	70	100
International Avg.	92	75	100	81	54	100
Benchmarking Participants						
Ontario Province, Can.	91	76	100	81	60	100
Quebec Province, Can.	91	76	100	81	60	100

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

Exhibit A.12: TIMSS 2003 Cross-Country Scoring Reliability for the Constructed-Response Science Items

Item Label	Total Valid Comparisons	Exact Percent Agreement	
		Correctness Score Agreement	Diagnostic Score Agreement
S032202	99900	83	73
S022283	99900	93	86
S022154	99900	83	70
S022191	99900	94	83
S022088A	99900	83	72
S022088B	99900	76	61
S022286	99900	91	77
S032625A	99900	97	94
S032625B	99900	92	72
S032120A	99900	78	61
S032120B	99900	87	69
S032063	99900	81	73
S032306	99900	88	83
S032640	99900	89	79
S032272	99900	95	88
S032650A	99900	90	84
S032650B	99900	87	80
S032056	99900	88	74
S032369	99900	80	71
S032565	99900	90	78
S032516	99900	84	74
Average Percent Agreement		87	76

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

country having scorers proficient in English and scored independently by one or if possible two of these scorers. Each of the responses was scored by 37 scorers from the countries that participated. Making all possible comparisons among scorers gave 666 comparisons for each student response to each item, and 99,900 total comparisons when aggregated across all 150 student responses to that item. Agreement across countries was defined in terms of the percentage of these comparisons that were in exact agreement. Exhibit A.12 shows that scorer reliability across countries was high, with the percent exact agreement averaging 87 percent across the 21 items for the correctness score and 76 percent for the diagnostic score.

Test Reliability

Exhibit A.13 displays the mathematics test reliability coefficient for each country. This coefficient is the median Cronbach's alpha reliability across the 12 test booklets. At both grade levels, median reliabilities generally were high, with an international median (the median of the reliability coefficients for all countries) of 0.84 at both grades. Despite the generally high reliabilities, there were some countries with median reliabilities below 0.80, namely Bahrain, Botswana, Ghana, Indonesia, Morocco, Saudi Arabia, Syria, and Tunisia at the eighth grade and Belgium (Flemish), Hong Kong SAR, Morocco, and the Netherlands at the fourth grade.

Exhibit A.13: **Cronbach's Alpha Reliability Coefficient – TIMSS 2003 Science Test**

Countries	Reliability Coefficient[1]	
	Grade 8	Grade 4
Armenia	0.81	0.84
Australia	0.86	0.85
Bahrain	0.78	
Belgium (Flemish)	0.84	0.77
Botswana	0.72	
Bulgaria	0.88	
Chile	0.82	
Chinese Taipei	0.89	0.80
Cyprus	0.81	0.81
Egypt	0.85	
England	0.88	0.86
Estonia	0.84	
Ghana	0.63	
Hong Kong, SAR	0.83	0.76
Hungary	0.88	0.84
Indonesia	0.79	
Iran, Islamic Rep. of	0.81	0.81
Israel	0.87	
Italy	0.85	0.85
Japan	0.86	0.82
Jordan	0.87	
Korea, Rep. of	0.87	
Latvia	0.83	0.80
Lebanon	0.81	
Lithuania	0.85	0.80
Macedonia, Rep. of	0.85	
Malaysia	0.83	
Moldova, Rep. of	0.82	0.87
Morocco	0.70	0.74
Netherlands	0.85	0.75
New Zealand	0.87	0.87
Norway	0.83	0.84
Palestinian Nat'l Auth.	0.83	
Philippines	0.81	0.86
Romania	0.89	
Russian Federation	0.86	0.86
Saudi Arabia	0.71	
Scotland	0.85	0.85
Serbia	0.86	
Singapore	0.91	0.87
Slovak Republic	0.87	
Slovenia	0.84	0.83
South Africa	0.84	
Sweden	0.86	
Syrian Arab Republic	0.77	
Tunisia	0.67	0.81
United States	0.88	0.85
Yemen		0.80
International Median	0.84	0.84
Benchmarking Participants		
Basque Country, Spain	0.81	
Indiana State, US	0.85	0.82
Ontario Province, Can.	0.84	0.84
Quebec Province, Can.	0.82	0.81

SOURCE: IEA's Trends in International Mathematics and Science Study (TIMSS) 2003

1 The reliability coefficient for each country is the median Cronbach's alpha reliability across the 12 test booklets.

Data Processing

To ensure the availability of comparable, high-quality data for analysis, TIMSS took rigorous quality control steps to create the international database.[11] 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. On the fourth-grade science test, two items were deleted for all countries. In addition, 10 countries had one or more items deleted (in most cases, one or two). Usually the poor statistics (negative point-biserials for the key, large item-by-country interactions, and statistics indicating lack of fit with the model) were a result of translation, adaptation, or printing deviations. At eighth grade, no science items were deleted for all countries, but 16 countries had one or more items deleted (mostly one or two).

11 These steps are detailed in Barth, J., Carstens, R., and Neuschmidt, O. (2004), "Creating and Checking the TIMSS 2003 Database" in M.O. Martin, I.V.S. Mullis, and S.J. Chrostowski (eds.), *TIMSS 2003 Technical Report*, Chestnut Hill, MA: Boston College.

IRT Scaling and Data Analysis

The general approach to reporting the TIMSS achievement data was based primarily on item response theory (IRT) scaling methods.[12] The 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. The methodology used in TIMSS includes refinements that enable reliable scores to be produced even though individual students responded to relatively small subsets of the total science item pool. Achievement scales were produced for each of the science content areas (life science, chemistry, physics, earth science, and environmental science at the eighth grade and life science, physical science, and earth science at the fourth grade), as well as for science overall.

The IRT methodology was preferred 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.

As shown in Exhibit A.5, TIMSS has a complicated booklet design, with blocks of items appearing in different positions in different booklets. For example, the items in block M1 appear as the first block in Booklet 1, as the second block in Booklet 6, and as the third block in Booklet 12. This allows the booklets to be linked together efficiently, but also to monitor and counterbalance any position effect. In TIMSS 2003, the counterbalanced booklet design made it possible to detect an unexpectedly strong position effect in the data as the item statistics for each country were reviewed. More specifically, this position

12 For a detailed description of the TIMSS scaling, see Gonzalez, E.J., Galia, J., and Li, I. (2004), "Scaling Methods and Procedures for the TIMSS 2003 Mathematics and Science Scales" in M.O. Martin, I.V.S. Mullis, and S.J. Chrostowski (eds.), *TIMSS 2003 Technical Report*, Chestnut Hill, MA: Boston College.

effect occurred because some students in all countries did not reach all the items in the third block position, which was the end of the first half of each booklet before the break. The same effect was evident for the sixth block position, which was the last block in the booklets. The IRT scaling addressed this problem by treating items in the third and sixth block positions as if they were unique, even though they also appeared in other positions. For example, the mathematics items in block M1 from Booklet 1 (the first position) and from Booklet 6 (second position) were considered to be the same items for scaling and reporting purposes, but those in Booklet 12 (the third position) were scaled as items that were different and unique.

The TIMSS science achievement scale was designed to provide a reliable measure of student achievement spanning 1995, 1999, and 2003. The metric of the scale was established originally with the 1995 assessment. When all countries participating in 1995 at the eighth grade are treated equally, the TIMSS scale average over those countries is 500 and the standard deviation is 100. The same applies 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,[13] 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

13 Because the 1995 student data had already been linked to the 1999 student data, it was not necessary to include the 1995 data in the 1999-2003 calibration.

students tested in 2003 but not 1995 were assigned scores on the basis of the scale.

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 student's 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 the score estimation process.

In addition to the scales for science overall, IRT scales also were created for each of the science content areas for the 2003 data. However, insufficient common items were used in 1995 and 1999 to establish reliable IRT content area scales for trend purposes. The trend exhibits presented in Chapter 3 were based on the average percentage of students responding correctly to the common items in each content area.

Estimating Sampling Error

Because the statistics presented in this report 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.[14] 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.

14 Procedures for computing jackknifed standard errors are presented in Gonzalez, E.J., Galia, J., Arora, A., Erberber, E., and Diaconu, D. (2004), "Reporting Student Achievement in Mathematics and Science" in M.O. Martin, I.V.S. Mullis, and S.J. Chrotowski (eds.), *TIMSS 2003 Technical Report*, Chestnut Hill, MA: Boston College.

Assessing Statistical Significance

This report makes extensive use of statistical hypothesis-testing to provide a basis for evaluating the significance of differences in percentages and in average achievement scores. Each separate test follows the usual convention of holding to 0.05 the probability that reported differences could be due to sampling variability alone. There is one important difference in the way TIMSS 2003 reports significance tests compared with the practice in 1995 and 1999. In the previous assessments, significance tests in exhibits where the results of many tests are reported simultaneously were based on a Bonferroni procedure for multiple comparisons. The Bonferroni procedure was not used in TIMSS 2003. The procedure takes into account the number of comparisons being made, which is a function of the number of countries participating. Since this varies from assessment to assessment, the Bonferroni procedure makes it difficult to compare results from one assessment to the next. However, users of the reports should be aware that, following the logic of statistical hypothesis testing, on average, about five percent of statistical tests will be significant by chance alone.

Setting International Benchmarks of Student Achievement

In order to provide meaningful descriptions of what performance on the TIMSS science scale could mean in terms of the science that students know and can do, TIMSS identified four points on the scale for use as international benchmarks. Selected to represent the range of performance shown by students internationally, the advanced benchmark is 625, the high benchmark is 550, the intermediate benchmark is 475, and the low benchmark is 400. Although the fourth- and eighth-grade scales are different, the same benchmark points are used at both grades.

To interpret the TIMSS scale scores and analyze achievement at the international benchmarks, TIMSS conducted a scale anchoring analysis to describe achievement of students at those four points on

the scale. 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.[15]

15 The scale-anchoring procedure is described fully in Gonzalez, E.J., Galia, J., Arora, A., Erberber, E., and Diaconu, D. (2004), "Reporting Student Achievement in Mathematics and Science" in M.O. Martin, I.V.S. Mullis, and S.J. Chrotowski (eds.), *TIMSS 2003 Technical Report*, Chestnut Hill, MA: Boston College. An application of the procedure to the 1995 TIMSS data may be found in Kelly, D.L., Mullis, I.V.S., and 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.