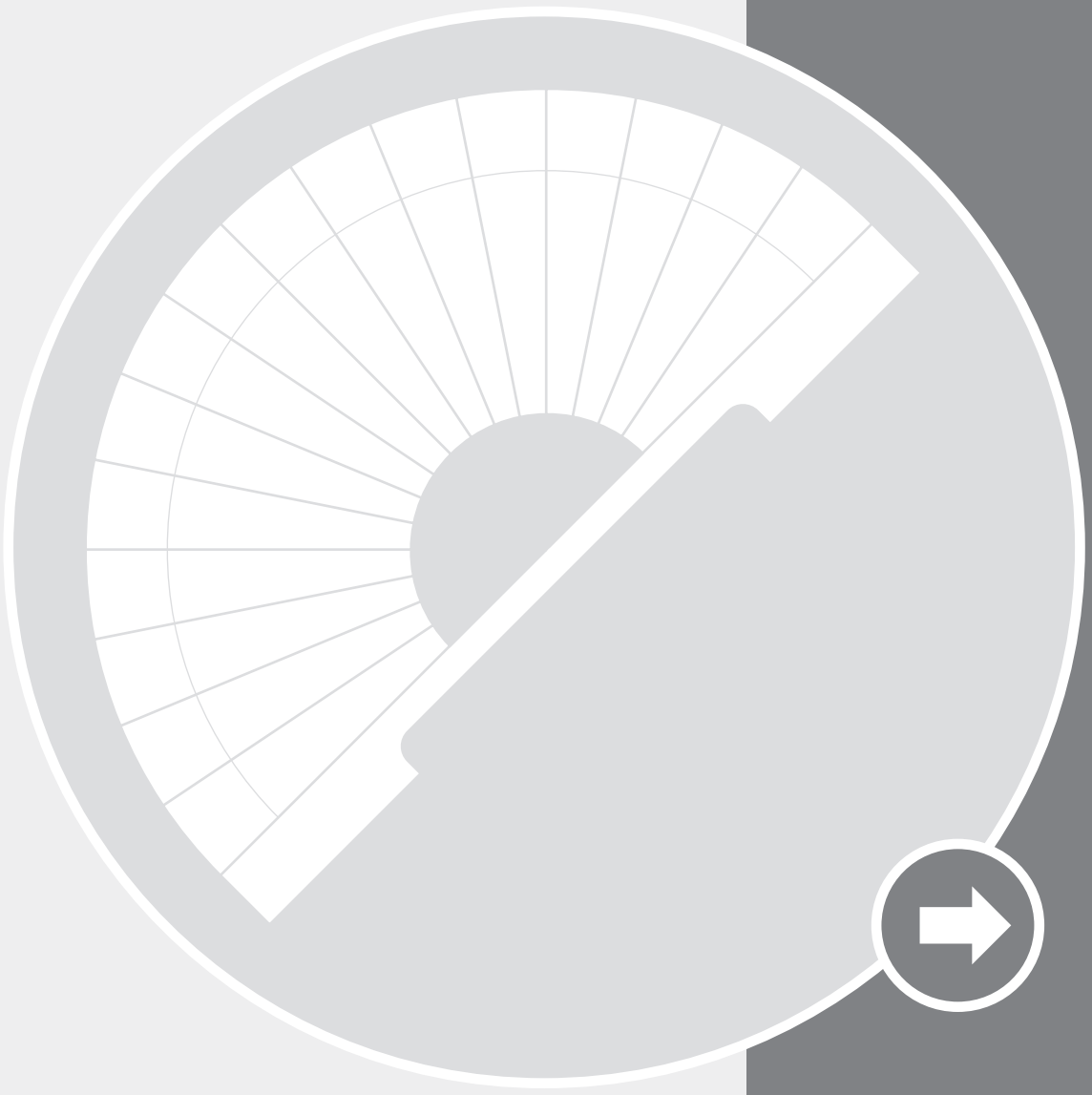
APPENDIX A

A

Overview of TIMSS Procedures: Mathematics Achievement



A

History

TIMSS 1999 represents the continuation of a long series of studies conducted by the International Association for the Evaluation of Educational Achievement (IEA). Since its inception in 1959, the IEA has conducted more than 15 studies of cross-national achievement in the curricular areas of mathematics, science, language, civics, and reading. IEA conducted its First International Science Study (FISS) in 1970-71 and the Second International Science Study (SISS) in 1983-84. The First and Second International Mathematics Studies (FIMS and SIMS) were conducted in 1964 and 1980-82, respectively. The Third International Mathematics and Science Study (TIMSS), conducted in 1994-1995, was the largest and most complex IEA study to date, and included both mathematics and science at third and fourth grades, seventh and eighth grades, and the final year of secondary school.

In 1999, TIMSS again assessed eighth-grade students in both mathematics and science to measure trends in student achievement since 1995. TIMSS 1999 was also known as TIMSS-Repeat, or TIMSS-R.

Participants in TIMSS

Of the 42 countries that participated in TIMSS[1] at the eighth grade in 1995, 26 availed themselves of the opportunity to measure changes in the achievement of their students by also taking part in 1999 (see Exhibit A.1). Twelve additional countries participated in 1999, for a total of 38 countries. Of those taking part in 1999, 19 had also participated in 1995 at the fourth grade.[2] Since fourth-grade students in 1995 were in eighth grade in 1999, these countries can compare the eighth-grade performance of this cohort of students with their performance at the fourth grade, as well as with the eighth-grade performance of students in other countries.

1 Results for 41 countries are reported in the 1995 international reports. Italy also completed the 1995 testing, but too late to be included in the international reports. It is counted as a 1995 country in this report and included in all trend exhibits in the 1999 international reports. Unweighted data for the Philippines were reported in an appendix to the international reports in 1995. These data were not included in trend exhibits in the 1999 international reports.

2 Two of the 19 countries with fourth-grade data from 1995 (Israel and Thailand) did not satisfy guidelines for sampling procedures at the classroom level and were not included in the comparisons for fourth and eighth grade.

Exhibit A.1 Countries Participating in TIMSS 1999 and 1995

	TIMSS 1999	TIMSS 1995 (Grade 8)	TIMSS 1995 (Grade 4)
Australia	●	●	●
Austria		●	●
Belgium (Flemish)	●	●	
Belgium (French)		●	
Bulgaria	●	●	
Canada	●	●	●
Chile	●		
Chinese Taipei	●		
Colombia		●	
Cyprus	●	●	●
Czech Republic	●	●	●
Denmark		●	
England	●	●	●
Finland	●		
France		●	
Germany		●	
Greece		●	●
Hong Kong, SAR	●	●	●
Hungary	●	●	●
Iceland		●	●
Indonesia	●		
Iran, Islamic Republic	●	●	●
Ireland		●	●
Israel	●	●	●
Italy	●	●	●
Japan	●	●	●
Jordan	●		
Korea, Republic of	●	●	●
Kuwait		●	●
Latvia	●	●	●
Lithuania	●	●	
Macedonia, Republic of	●		
Malaysia	●		
Moldova	●		
Morocco	●		
Netherlands	●	●	●
New Zealand	●	●	●
Norway		●	●
Philippines	●		
Portugal		●	●
Romania	●	●	
Russian Federation	●	●	
Scotland		●	●
Singapore	●	●	●
Slovak Republic	●	●	
Slovenia	●	●	●
South Africa	●	●	
Spain		●	
Sweden		●	
Switzerland		●	
Thailand	●	●	●
Tunisia	●		
Turkey	●		
United States	●	●	●

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1998-1999.

Developing the TIMSS 1999 Mathematics Test

The TIMSS curriculum framework underlying the mathematics tests was developed for TIMSS in 1995 by groups of mathematics educators with input from the TIMSS National Research Coordinators (NRCs). As shown in Exhibit A.2, the mathematics curriculum framework contains three dimensions or aspects. The *content* aspect represents the subject matter content of school mathematics. The *performance expectations* aspect describes, in a non-hierarchical way, the many kinds of performances or behaviors that might be expected of students in school mathematics. The *perspectives* aspect focuses on the development of students' attitudes, interest, and motivation in mathematics. Because the frameworks were developed to include content, performance expectations, and perspectives for the entire span of curricula from the beginning of schooling through the completion of secondary school, some aspects may not be reflected in the eighth-grade TIMSS assessment.[3] Working within the framework, mathematics test specifications for TIMSS in 1995 were developed that included items representing a wide range of mathematics topics and eliciting a range of skills from the students. The 1995 tests were developed through an international consensus involving input from experts in mathematics and measurement specialists, ensuring they reflected current thinking and priorities in the mathematics.

About one-third of the items in the 1995 assessment were kept secure to measure trends over time; the remaining items were released for public use. An essential part of the development of the 1999 assessment, therefore, was to replace the released items with items of similar content, format, and difficulty. With the assistance of the Science and Mathematics Item Replacement Committee, a group of internationally prominent mathematics and science educators nominated by participating countries to advise on subject-matter issues in the assessment, over 300 mathematics and science items were developed as potential replacements. After an extensive process of review and field testing, 114 items were selected for use as replacements in the 1999 mathematics assessment.

Exhibit A.3 presents the five content areas included in the 1999 mathematics test and the numbers of items and score points in each area. Distributions are also included for the five performance categories derived from the performance expectations aspect of the curriculum framework. Exhibit A.4 shows how the trend and replacement items were distributed across these content areas and performance categories. About one-fourth of the items were in the free-response format, requiring students to generate and write their own answers. Designed to take about

3 The complete TIMSS curriculum frameworks can be found in Robitaille, D.F., et al. (1993), *TIMSS Monograph No.1: Curriculum Frameworks for Mathematics and Science*, Vancouver, BC: Pacific Educational Press.

one-third of students' test time, some free-response questions asked for short answers while others required extended responses with students showing their work or providing explanations for their answers. The remaining questions used a multiple-choice format. In scoring the tests, correct answers to most questions were worth one point. Consistent with the approach of allotting students longer response time for the constructed-response questions than for multiple-choice questions, however, responses to some of these questions (particularly those requiring extended responses) were evaluated for partial credit, with a fully correct answer being awarded two points (see later section on scoring). The total number of score points available for analysis thus somewhat exceeds the number of items.

Every effort was made to help ensure that the tests represented the curricula of the participating countries and that the items exhibited no bias towards or against particular countries. The final forms of the test were endorsed by the NRCs of the participating countries.[4] 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.

4 For a full discussion of the TIMSS 1999 test development effort, please see Garden, R.A. and Smith, T.A. (2000), "TIMSS Test Development" in M.O. Martin, K.D. Gregory and S.E. Stemler (eds.), *TIMSS 1999 Technical Report*, Chestnut Hill, MA: Boston College.

Exhibit A.2 The Three Aspects and Major Categories of the Mathematics Frameworks

TIMSS 1999 8th grade Mathematics

Content	Performance Expectations	Perspectives
Numbers	Knowing	Attitudes
Measurement	Using Routine Procedures	Careers
Geometry	Investigating and Problem Solving	Participation
Proportionality	Mathematical Reasoning	Increasing Interest
Functions, Relations, and Equations	Communicating	Habits of Mind
Data Representation		
Probability and Statistics		
Elementary Analysis, Validation, and Structure		

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1998-1999.

Exhibit A.3 Distribution of Mathematics Items by Content Reporting Category and Performance Category

TIMSS 1999 8th grade Mathematics

Content Category	Percentage of Items	Total Number of Items	Number of Multiple-Choice Items	Number of Free-Response Items[1]	Number of Score Points[2]
Fractions and Number Sense	38	61	47	14	62
Measurement	15	24	15	9	26
Data Representation, Analysis and Probability	13	21	19	2	22
Geometry	13	21	20	1	21
Algebra	22	35	24	11	38
Total	100	162	125	37	169

Performance Category	Percentage of Items	Total Number of Items	Number of Multiple-Choice Items	Number of Free-Response Items[1]	Number of Score Points[2]
Knowing	19	30	28	2	30
Using Routine Procedures	23	38	28	10	39
Using Complex Procedures	24	39	34	5	40
Investigating and Solving Problems	31	51	34	17	53
Communicating and Reasoning	2	4	1	3	7
Total	100	162	125	37	169

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1998-1999.

1 Free-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 free-response items were evaluated for partial credit with a fully correct answer awarded up to two points. Thus, the number of score points exceeds the number of items in the test.

Exhibit A.4 Distribution of Mathematics Trend and Replacement Items by Content Reporting Category and Performance Category

Content Category	Trend Items		Replacement Items		Total Items	
	Number of Items	Number of Points	Number of Items	Number of Points	Number of Items	Number of Points
Fractions and Number Sense	17	17	44	45	61	62
Measurement	6	6	18	20	24	26
Data Representation, Analysis and Probability	8	8	13	14	21	22
Geometry	6	6	15	15	21	21
Algebra	11	11	24	27	35	38
Total	48	48	114	121	162	169

Performance Category	Trend Items		Replacement Items		Total Items	
	Number of Items	Number of Points	Number of Items	Number of Points	Number of Items	Number of Points
Knowing	16	16	14	14	30	30
Using Routine Procedures	8	8	30	31	38	39
Using Complex Procedures	13	13	26	27	39	40
Investigating and Solving Problems	11	11	40	42	51	53
Communicating and Reasoning	–	–	4	7	4	7
Total	48	48	114	121	162	169

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1998-1999.

TIMSS Test Design

Not all of the students in the TIMSS assessment responded to all of the mathematics items. To ensure broad subject-matter coverage without overburdening individual students, TIMSS used a rotated design that included both the mathematics and science items. Thus, the same students participated in both the mathematics and science testing. As in 1995, the 1999 assessment consisted of eight booklets, each requiring 90 minutes of response time. Each participating student was assigned one booklet only. In accordance with the design, the mathematics and science items were assembled into 26 clusters (labeled A through Z). The secure trend items were in clusters A through H, and items replacing the released 1995 items in clusters I through Z. Eight of the clusters were designed to take 12 minutes to complete; 10 of the clusters, 22 minutes; and 8 clusters, 10 minutes. In all, the design provided 396 testing minutes, 198 for mathematics and 198 for science. Cluster A was a core cluster assigned to all booklets. The remaining clusters were assigned to the booklets in accordance with the rotated design so that representative samples of students responded to each cluster.[5]

Background Questionnaires

TIMSS in 1999 administered a broad array of questionnaires to collect data on the educational context for student achievement and to measure trends since 1995. *National Research Coordinators*, with the assistance of their curriculum experts, provided detailed information on the organization, emphases, and content coverage of the mathematics and science curriculum. The *students* who were tested answered questions pertaining to their 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. The heads of *schools* responded to questions about school staffing and resources, mathematics and science course offerings, and teacher support.

5 The 1999 TIMSS test design is identical to the design for 1995, which is fully documented in Adams, R. and Gonzalez, E. (1996), "TIMSS Test Design" in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study Technical Report, Volume I*, Chestnut Hill, MA: Boston College.

Translation and Verification

The TIMSS instruments were prepared in English and translated into 33 languages, with 10 of the 38 countries collecting data in two languages. In addition, it sometimes was necessary to modify the international versions for cultural reasons, even in the nine countries that tested in English. This process represented an enormous effort for the national centers, with many checks along the way. The translation effort included (1) developing explicit guidelines for translation and cultural adaptation; (2) translation of the instruments by the national centers in accordance with the guidelines, using two or more independent 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.[6]

Population Definition and Sampling

TIMSS in 1995 had as its target population students enrolled in the two adjacent grades that contained the largest proportion of 13-year-old students at the time of testing, which were seventh- and eighth-grade students in most countries. TIMSS in 1999 used the same definition to identify the target grades, but assessed students in the upper of the two grades only, which was the eighth grade in most countries.[7]

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 staff from Statistics Canada. 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.), staff from Statistics Canada reviewed the national sampling plans, sampling data, sampling frames, and sample execution. The sampling documentation was used by the International Study Center, in consultation with Statistics Canada and the sampling referee, to evaluate the quality of the samples.

6 More details about the translation verification procedures can be found in O'Connor, K., and Malak, B. (2000), "Translation and Cultural Adaptation of the TIMSS Instruments" in M.O. Martin, K.D. Gregory and S.E. Stemler (eds.), *TIMSS 1999 Technical Report*, Chestnut Hill, MA: Boston College.

7 The sample design for TIMSS is described in detail in Foy, P., and Joncas, M. (2000), "TIMSS Sample Design" in M.O. Martin, K.D. Gregory and S.E. Stemler (eds.), *TIMSS 1999 Technical Report*, Chestnut Hill, MA: Boston College.

In a few situations where it was not possible to test the entire internationally desired population (all students in the upper of the two adjacent grades with the greatest proportion of 13-year-olds), countries were permitted to define a national desired population that excluded part of the internationally desired population. Exhibit A.5 shows any differences in coverage between the international and national desired populations. Almost all participants achieved 100 percent coverage (36 out of 38), with Lithuania and Latvia the exceptions. Consequently, the results for Lithuania are annotated in exhibits in this report, and because coverage fell below 65 percent for Latvia, the Latvian results have been labeled "Latvia (LSS)," for Latvian-Speaking Schools. Although achieving 100 percent coverage of their populations in 1999, both Italy and Israel had less than complete coverage in 1995 – Italy because four of its provinces did not take part, and Israel because it did not test the Arabic-speaking population. Comparisons between 1995 and 1999 for these countries are based on the subsets of the 1999 populations that were comparable to the populations tested in 1995.

Within the desired population, countries could define a population that excluded a small percentage (less than 10 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). Exhibit A.5 also shows that the degree of such exclusions was small. Only Israel exceeded the 10 percent limit, and this is annotated in the exhibits in this report.

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 mathematics class at the eighth grade. All of the students in that class were to participate in the TIMSS testing. This approach was designed to yield a representative sample of about 3,750 students per country. Typically, between 450 and 3,750 students responded to each achievement item in each country, depending on the booklets in which the items appeared.

Exhibits A.6 and A.7 present achieved sample sizes for schools and students, respectively, for participating countries. Exhibit A.8 shows the participation rates for schools, students, and overall, both with and without the use of replacement schools. All 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 Belgium (Flemish), England, Hong Kong, and the Netherlands did so only after including replacement schools.

Because of scheduling difficulties, Lithuania was unable to test its eighth-grade students in May 1999 as planned. Instead, the students were tested in September 1999, when they had moved into the ninth grade. The results for Lithuania are annotated accordingly in exhibits in this report.

Whereas all countries achieved a high degree of compliance with sampling guidelines in 1999, three of them (Israel, South Africa, and Thailand) had experienced difficulties with sampling at the classroom level in 1995. Accordingly, results for these three countries are reported in a separate panel of the exhibits in these reports that deal with trends from 1995.

Exhibit A.5 Coverage of TIMSS 1999 Target Population

TIMSS 1999 8th grade Mathematics

	International Desired Population		National Desired Population		
	Coverage	Notes on Coverage	School-Level Exclusions	Within-Sample Exclusions	Overall Exclusions
Australia	100%		1%	1%	2%
Belgium (Flemish)	100%		1%	0%	1%
Bulgaria	100%		5%	0%	5%
Canada	100%		4%	2%	6%
Chile	100%		3%	0%	3%
Chinese Taipei	100%		1%	1%	2%
Cyprus	100%		0%	1%	1%
Czech Republic	100%		5%	0%	5%
England	100%		2%	3%	5%
Finland	100%		3%	0%	4%
Hong Kong, SAR	100%		1%	0%	1%
Hungary	100%		4%	0%	4%
Indonesia	100%		0%	0%	0%
Iran, Islamic Rep. of	100%		4%	0%	4%
Israel	100%		8%	8%	16%
Italy	100%		4%	2%	7%
Japan	100%		1%	0%	1%
Jordan	100%		2%	1%	3%
Korea, Rep. of	100%		2%	2%	4%
Latvia (LSS)	61%	Latvian-speaking students only	4%	0%	4%
Lithuania	87%	Lithuanian-speaking students only	5%	0%	5%
Macedonia, Rep. of	100%		1%	0%	1%
Malaysia	100%		5%	0%	5%
Moldova	100%		2%	0%	2%
Morocco	100%		1%	0%	1%
Netherlands	100%		1%	0%	1%
New Zealand	100%		2%	1%	2%
Philippines	100%		3%	0%	3%
Romania	100%		4%	0%	4%
Russian Federation	100%		1%	1%	2%
Singapore	100%		0%	0%	0%
Slovak Republic	100%		7%	0%	7%
Slovenia	100%		3%	0%	3%
South Africa	100%		2%	0%	2%
Thailand	100%		3%	0%	3%
Tunisia	100%		0%	0%	0%
Turkey	100%		2%	0%	2%
United States	100%		0%	4%	4%

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1998-1999.

Exhibit A.6 School Sample Sizes

	Number of Schools in Original Sample	Number of Eligible Schools in Original Sample	Number of Schools in Original Sample That Participated	Number of Replacement Schools That Participated	Total Number of Schools That Participated
Australia	184	182	152	18	170
Belgium (Flemish)	150	150	106	29	135
Bulgaria	172	169	163	0	163
Canada	410	398	376	9	385
Chile	186	185	181	4	185
Chinese Taipei	150	150	150	0	150
Cyprus	61	61	61	0	61
Czech Republic	150	142	136	6	142
England	150	150	76	52	128
Finland	160	160	155	4	159
Hong Kong, SAR	180	180	135	2	137
Hungary	150	150	147	0	147
Indonesia	150	150	132	18	150
Iran, Islamic Rep. of	170	170	164	6	170
Israel	150	139	137	2	139
Italy	180	180	170	10	180
Japan	150	150	140	0	140
Jordan	150	147	146	1	147
Korea, Rep. of	150	150	150	0	150
Latvia (LSS)	150	148	143	2	145
Lithuania	150	150	150	0	150
Macedonia, Rep. of	150	150	149	0	149
Malaysia	150	150	148	2	150
Moldova	150	150	145	5	150
Morocco	174	174	172	1	173
Netherlands	150	148	86	40	126
New Zealand	156	156	145	7	152
Philippines	150	150	148	2	150
Romania	150	150	147	0	147
Russian Federation	190	190	186	3	189
Singapore	145	145	145	0	145
Slovak Republic	150	150	143	2	145
Slovenia	150	150	147	2	149
South Africa	225	219	183	11	194
Thailand	150	150	143	7	150
Tunisia	150	149	126	23	149
Turkey	204	204	202	2	204
United States	250	246	202	19	221

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1998-1999.

Exhibit A.7 Student Sample Sizes

	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
Australia	90%	4600	96	53	4451	419	4032
Belgium (Flemish)	97%	5387	12	0	5375	116	5259
Bulgaria	96%	3461	63	0	3398	126	3272
Canada	96%	9490	84	245	9161	391	8770
Chile	96%	6283	119	18	6146	239	5907
Chinese Taipei	99%	5889	30	42	5817	45	5772
Cyprus	97%	3296	38	32	3226	110	3116
Czech Republic	96%	3640	24	0	3616	163	3453
England	90%	3400	27	115	3258	298	2960
Finland	96%	3060	17	13	3030	110	2920
Hong Kong, SAR	98%	5310	18	1	5291	112	5179
Hungary	95%	3350	0	0	3350	167	3183
Indonesia	97%	6162	106	1	6055	207	5848
Iran, Islamic Rep. of	98%	5497	104	0	5393	92	5301
Israel	94%	4670	29	187	4454	259	4195
Italy	97%	3531	23	86	3422	94	3328
Japan	95%	4996	15	12	4969	224	4745
Jordan	99%	5300	130	42	5128	76	5052
Korea, Rep. of	100%	6285	29	128	6128	14	6114
Latvia (LSS)	93%	3128	16	4	3108	235	2873
Lithuania	89%	2668	0	0	2668	307	2361
Macedonia, Rep. of	98%	4096	0	0	4096	73	4023
Malaysia	99%	5713	98	0	5615	38	5577
Moldova	98%	3824	23	0	3801	90	3711
Morocco	92%	5841	42	0	5799	397	5402
Netherlands	95%	3099	12	0	3087	125	2962
New Zealand	94%	3966	96	22	3848	235	3613
Philippines	92%	7591	461	0	7130	529	6601
Romania	98%	3514	36	0	3478	53	3425
Russian Federation	97%	4557	48	34	4475	143	4332
Singapore	98%	5100	37	0	5063	97	4966
Slovak Republic	98%	3695	149	0	3546	49	3497
Slovenia	95%	3287	0	4	3283	174	3109
South Africa	93%	9071	256	0	8815	669	8146
Thailand	99%	5831	59	0	5772	40	5732
Tunisia	98%	5189	45	0	5144	93	5051
Turkey	99%	7972	49	0	7923	82	7841
United States	94%	9981	115	142	9724	652	9072

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1998-1999.

Exhibit A.8 Overall Participation Rates

	School Participation		Student Participation	Overall Participation	
	Before Replacement	After Replacement		Before Replacement	After Replacement
Australia	83%	93%	90%	75%	84%
Belgium (Flemish)	72%	89%	97%	70%	87%
Bulgaria	97%	97%	96%	93%	93%
Canada	92%	95%	96%	88%	92%
Chile	98%	100%	96%	94%	96%
Chinese Taipei	100%	100%	99%	99%	99%
Cyprus	100%	100%	97%	97%	97%
Czech Republic	94%	100%	96%	90%	96%
England	49%	85%	90%	45%	77%
Finland	97%	100%	96%	93%	96%
Hong Kong, SAR	75%	76%	98%	74%	75%
Hungary	98%	98%	95%	93%	93%
Indonesia	84%	100%	97%	81%	97%
Iran, Islamic Rep. of	96%	100%	98%	95%	98%
Israel	98%	100%	94%	93%	94%
Italy	94%	100%	97%	91%	97%
Japan	93%	93%	95%	89%	89%
Jordan	99%	100%	99%	98%	99%
Korea, Rep. of	100%	100%	100%	100%	100%
Latvia (LSS)	96%	98%	93%	89%	91%
Lithuania	100%	100%	89%	89%	89%
Macedonia, Rep. of	99%	99%	98%	98%	98%
Malaysia	99%	100%	99%	98%	99%
Moldova	96%	100%	98%	94%	98%
Morocco	99%	99%	92%	91%	92%
Netherlands	62%	85%	95%	59%	81%
New Zealand	93%	97%	94%	87%	91%
Philippines	98%	100%	92%	91%	92%
Romania	98%	98%	98%	97%	97%
Russian Federation	98%	100%	97%	95%	97%
Singapore	100%	100%	98%	98%	98%
Slovak Republic	95%	96%	98%	93%	94%
Slovenia	98%	99%	95%	93%	94%
South Africa	85%	91%	93%	79%	84%
Thailand	93%	100%	99%	93%	99%
Tunisia	84%	100%	98%	82%	98%
Turkey	99%	100%	99%	98%	99%
United States	83%	90%	94%	78%	85%

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1998-1999.

Data Collection

Each participating country was responsible for carrying out all aspects of the data collection, using standardized procedures developed for the study. Training manuals were created for school coordinators and test administrators that explained procedures for receipt and distribution of materials as well as for the activities related to the testing sessions. These manuals covered procedures for test security, standardized scripts to regulate directions and timing, rules for answering students' questions, and steps to ensure that identification on the test booklets and questionnaires corresponded to the information on the forms used to track students.

Each country was responsible for conducting quality control procedures and describing this effort in the NRC's report documenting procedures used in the study. In addition, the International Study Center considered it essential to monitor compliance with standardized procedures. NRCs were asked to nominate one or more persons unconnected with their national center, such as retired school teachers, to serve as quality control monitors for their countries. The International Study Center developed manuals for the monitors and briefed them in two-day training sessions about TIMSS, the responsibilities of the national centers in conducting the study, and their own roles and responsibilities. In all, 71 quality control monitors participated in this training.

The quality control monitors interviewed the NRCs about data collection plans and procedures. They also visited a sample of 15 schools where they observed testing sessions and interviewed school coordinators.[8] Quality control monitors interviewed school coordinators in all 38 countries, and observed a total of 550 testing sessions.

The results of the interviews indicate that, in general, NRCs had prepared well for data collection and, despite the heavy demands of the schedule and shortages of resources, were able to conduct the data collection efficiently and professionally. Similarly, the TIMSS tests appeared to have been administered in compliance with international procedures, including the activities before the testing session, those during testing, and the school-level activities related to receiving, distributing, and returning material from the national centers.

8 Steps taken to ensure high-quality data collection in TIMSS are described in detail in O'Connor, K., and Stemler, S. (2000), "Quality Control in the TIMSS Data Collection" in M.O. Martin, K.D. Gregory and S.E. Stemler (eds.), *TIMSS 1999 Technical Report*, Chestnut Hill, MA: Boston College.

Scoring the Free-Response Items

Because about one-third of the written test time was devoted to free-response items, TIMSS needed to develop procedures for reliably evaluating student responses within and across countries. Scoring used two-digit codes with rubrics specific to each item. The first digit designates the correctness level of the response. The second digit, combined with the first, represents a diagnostic code identifying specific types of approaches, strategies, or common errors and misconceptions. Although not used in this report, analyses of responses based on the second digit should provide insight into ways to help students better understand mathematics concepts and problem-solving approaches. Because of the burden of maintaining scoring consistency across time, no free-response items were used to measure trends from 1995 to 1999. However, samples of student responses from each country to selected items in 1999 have been scanned using advanced imaging technology in preparation for studying trends to 2003 and beyond.

To ensure reliable scoring procedures based on the TIMSS rubrics, the International Study Center prepared detailed guides containing the rubrics and explanations of how to 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 free-response items. The training sessions were designed to help representatives of national centers who would then be responsible for training personnel in their countries to apply the two-digit codes reliably.

To gather and document empirical information about the within-country agreement among scorers, TIMSS arranged to have systematic subsamples of at least 100 students' responses to each item coded independently by two readers. Exhibit A.9 shows the average and range of the within-country exact percent of agreement between scorers on the free-response items in the mathematics test for 37 of the 38 countries. A high percentage of exact agreement was observed, with an overall average of 99 percent across the 37 countries. The TIMSS data from the reliability studies indicate that scoring procedures were robust for the mathematics items, especially for the correctness score used for the analyses in this report.

Exhibit A.9 TIMSS 1999 Within-Country Free-Response Scoring Reliability Data for Mathematics Items

TIMSS 1999 8th grade Mathematics

	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	98	94	100	95	80	100
Belgium (Flemish)	99	92	100	98	91	100
Bulgaria	99	94	100	96	73	100
Canada	98	88	100	94	80	99
Chile	99	94	100	97	88	100
Chinese Taipei	100	98	100	99	93	100
Cyprus	–	–	–	–	–	–
Czech Republic	97	81	100	92	63	99
England	99	96	100	97	87	100
Finland	99	97	100	97	90	100
Hong Kong, SAR	98	84	100	95	80	100
Hungary	98	87	100	96	76	100
Indonesia	99	92	100	94	79	100
Iran, Islamic Rep.	99	93	100	94	74	100
Israel	98	92	100	95	81	100
Italy	99	95	100	97	89	100
Japan	99	90	100	96	88	100
Jordan	99	96	100	96	89	100
Korea, Rep. of	98	88	100	96	73	100
Latvia (LSS)	99	96	100	96	79	100
Lithuania	99	90	100	98	88	100
Macedonia, Rep. of	99	97	100	98	95	100
Malaysia	100	98	100	99	97	100
Moldova	97	92	100	94	86	99
Morocco	97	84	100	88	65	99
Netherlands	99	85	100	94	79	100
New Zealand	99	95	100	95	88	100
Philippines	99	97	100	95	84	100
Romania	99	96	100	97	92	100
Russian Federation	100	98	100	98	92	100
Singapore	99	94	100	97	87	100
Slovak Republic	99	97	100	99	96	100
Slovenia	100	99	100	96	83	100
South Africa	99	93	100	96	85	99
Thailand	100	100	100	100	100	100
Tunisia	98	92	100	96	88	100
Turkey	100	97	100	99	97	100
United States	99	96	100	97	89	100
International Avg.	99	93	100	96	85	100

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1998-1999.

A dash (–) indicates data are not available.

Test Reliability

A.10 

Exhibit A.10 displays the mathematics test reliability coefficient for each country. This coefficient is the median KR-20 reliability across the eight test booklets. Median reliabilities ranged from 0.76 in the Philippines to 0.94 in Chinese Taipei. The international median, 0.89, is the median of the reliability coefficients for all countries.

Data Processing

To ensure the availability of comparable, high-quality data for analysis, TIMSS took rigorous quality control steps to create the international database.[9] TIMSS prepared manuals and software for countries to use in entering their data, so that the information would be in a standardized international format before being forwarded to the IEA Data Processing Center in Hamburg for creation of the international database. Upon arrival at the Data Processing Center, the data underwent an exhaustive cleaning process. This involved several iterative steps and procedures designed to identify, document, and correct deviations from the international instruments, file structures, and coding schemes. The process also emphasized consistency of information within national data sets and appropriate linking among the many student, teacher, and school data files.

Throughout the process, the data were checked and double-checked by the IEA Data Processing Center, the International Study Center, and the national centers. The national centers were contacted regularly and given multiple opportunities to review the data for their countries. In conjunction with the IEA Data Processing Center, the International Study Center reviewed item statistics for each cognitive item in each country to identify poorly performing items. On the mathematics test, one item was deleted for all countries. In addition, 14 countries had one or more items deleted (in most cases, one). 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.

9 These steps are detailed in Hastedt, D., and Gonzalez, E. (2000), "Data Management and Database Construction" in M.O. Martin, K.D. Gregory and S.E. Stemler (eds.), *TIMSS 1999 Technical Report*, Chestnut Hill, MA: Boston College.

Exhibit A.10 Cronbach's Alpha Reliability Coefficient – TIMSS 1999 Mathematics Test

	Reliability Coefficient[1]
Australia	0.90
Belgium (Flemish)	0.89
Bulgaria	0.90
Canada	0.88
Chile	0.83
Chinese Taipei	0.94
Cyprus	0.87
Czech Republic	0.90
England	0.90
Finland	0.86
Hong Kong, SAR	0.89
Hungary	0.91
Indonesia	0.87
Iran, Islamic Rep.	0.83
Israel	0.90
Italy	0.89
Japan	0.91
Jordan	0.89
Korea, Rep. of	0.91
Latvia (LSS)	0.89
Lithuania	0.89
Macedonia, Rep. of	0.88
Malaysia	0.90
Moldova	0.88
Morocco	0.69
Netherlands	0.89
New Zealand	0.91
Philippines	0.76
Romania	0.90
Russian Federation	0.91
Singapore	0.90
Slovak Republic	0.89
Slovenia	0.90
South Africa	0.77
Thailand	0.87
Tunisia	0.79
Turkey	0.86
United States	0.90
International Median	0.89

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1998-1999.

1 The reliability coefficient for each country is the median KR-20 reliability across the eight test booklets.

IRT Scaling and Data Analysis

The general approach to reporting the TIMSS achievement data was based primarily on item response theory (IRT) scaling methods.[10] The mathematics results were summarized using a family of 2-parameter and 3-parameter IRT models for dichotomously-scored items (right or wrong), and generalized partial credit models for items with 0, 1, or 2 available score points. The IRT scaling method produces a score by averaging the responses of each student to the items 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 mathematics item pool. Achievement scales were produced for each of the five mathematics content areas (fractions and number sense, measurement, data representation, analysis, and probability, geometry, and algebra), as well as for mathematics overall.

The IRT methodology was preferred for developing comparable estimates of performance for all students, since students answered different test items depending upon which of the eight test booklets they received. 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. To provide a reliable measure of student achievement in both 1999 and 1995, the overall mathematics scale was calibrated using students from the countries that participated in both years. When all countries participating in 1995 at the eighth grade are treated equally, the TIMSS scale average over those countries is 500 and the standard deviation is 100. Since the countries 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. When the metric of the scale had been established, students from the countries that tested in 1999 but not 1995 were assigned scores on the basis of the new scale.

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

[10] For a detailed description of the TIMSS scaling, see Yamamoto, K., and Kulick, E. (2000), "Scaling Methods and Procedures for the TIMSS Mathematics and Science Scales" in M.O. Martin, K.D. Gregory and S.E. Stemler (eds.), *TIMSS 1999 Technical Report*, Chestnut Hill, MA: Boston College.

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.

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.[11] The jackknife standard errors also include an error component due to variation between 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.

Making Multiple Comparisons

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. However, in exhibits where statistical significance tests are reported, the results of many tests are reported simultaneously, usually at least one for each country in the exhibit. The significance tests in these exhibits are based on a Bonferroni procedure for multiple comparisons that hold to 0.05 the probability of erroneously declaring a statistic (mean or percentage)

[11] Procedures for computing jackknifed standard errors are presented in Gonzalez, E. and Foy, P. (2000), "Estimation of Sampling Variance" in M.O. Martin, K.D. Gregory and S.E. Stemler (eds.), *TIMSS 1999 Technical Report*, Chestnut Hill, MA: Boston College.

for one country to be different from that for another country. In the multiple comparison charts (Exhibit 1.2 and those in Appendix B), the Bonferroni procedure adjusts for the number of countries in the chart, minus one. In exhibits where a country statistic is compared to the international average, the adjustment is for the number of countries.[12]

Setting International Benchmarks of Student Achievement

International benchmarks of student achievement were computed at each grade level for both mathematics and science. The benchmarks are points in the weighted international distribution of achievement scores that separate the 10 percent of students located on top of the distribution, the top 25 percent of students, the top 50 percent, and the bottom 25 percent. The percentage of students in each country meeting or exceeding the international benchmarks is reported. The benchmarks correspond to the 90th, 75th, 50th, and 25th percentiles of the international distribution of achievement. When computing these percentiles, each country contributed as many students to the distribution as there were students in the target population in the country. That is, each country's contribution to setting the international benchmarks was proportional to the estimated population enrolled at the eighth grade.

In order 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.[13]

12 The application of the Bonferroni procedures is described in Gonzalez, E., and Gregory, K. (2000), "Reporting Student Achievement in Mathematics and Science" in M.O. Martin, K.D. Gregory and S.E. Stemler (eds.), *TIMSS 1999 Technical Report*, Chestnut Hill, MA: Boston College.

13 The scale-anchoring procedure is described fully in Gregory, K., and Mullis, I. (2000), "Describing International Benchmarks of Student Achievement" in M.O. Martin, K.D. Gregory and S.E. Stemler (eds.), *TIMSS 1999 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.

Mathematics Curriculum Questionnaire

In an effort to collect information about the content of the intended curriculum in mathematics, TIMSS asked National Research Coordinators to complete a questionnaire about the structure, organization, and content coverage of their national curricula. NRCs reviewed 56 mathematics topics and reported the percentage of their eighth-grade students for which each topic was intended in their curriculum. Although most topic descriptions were used without modification, there were occasions when NRCs found it necessary to expand on or qualify the topic description to describe their situation accurately. These country-specific adaptations to the mathematics curriculum questionnaire are presented in Exhibit A.11.

Exhibit A.11 Country-Specific Variations in Mathematics Topics in the Curriculum Questionnaire

	Topic	Response	Comments
Bulgaria	Geometry: Congruence and similarity	All or almost all of the students (at least 90%)	Similarity not included in curriculum through grade 8.
Czech Republic	Measurement: Volume of other solids (e.g., pyramids, cylinders, cones, spheres)	All or almost all of the students (at least 90%)	Volume of pyramids, cones, & spheres not included in curriculum through grade 8.
	Geometry: Congruence and similarity	All or almost all of the students (at least 90%)	Similarity not included in curriculum through grade 8.
Finland	Fractions and Number Sense: Concepts of ratio and proportion; ratio and proportion problems	Not included in curriculum through grade 8	Concepts of ratio and proportion included in curriculum through grade 8.
	Geometry: Symmetry and transformations (reflection and rotation)	Not included in curriculum through grade 8	Symmetry included in curriculum through grade 8.
	Algebra: Representing situations algebraically; formulas	All or almost all of the students (at least 90%)	Formulas not included in curriculum through grade 8.
Israel	Fractions and Number Sense: Whole numbers - including place values, factorization and operations (+, -, x, ÷)	All or almost all of the students (at least 90%)	Factorization not included in curriculum through grade 8.
	Fractions and Number Sense: Computations with common fractions	All or almost all of the students (at least 90%)	Division with common fractions not included in curriculum through grade 8.
	Fractions and Number Sense: Computations with decimal fractions	All or almost all of the students (at least 90%)	Division with decimal fractions is not included in curriculum through grade 8.
	Measurement: Estimates of measurement; accuracy of measurement	Only the most advanced students (10% or less)	Accuracy of measurement is not included in curriculum through grade 8.
	Geometry: Simple two dimensional geometry - angles on a straight line, parallel lines, triangles and quadrilaterals	About half of the students	Quadrilaterals not included in curriculum through grade 8.
	Geometry: Congruence and similarity	All or almost all of the students (at least 90%)	Similarity not included in curriculum through grade 8.
Japan	Fractions and Number Sense: Prime factors, highest common factor, lowest common multiple, rules for divisibility	Not included in curriculum through grade 8	Highest common factor and lowest common multiple included in curriculum through grade 8.
Korea, Rep. of	Fractions and Number Sense: Number lines	All or almost all of the students (at least 90%)	Whole number and integer number lines included in curriculum through grade 8. The real number line is taught in grade 9.
	Geometry: Cartesian coordinates of points in a plane	Not included in curriculum through grade 8	Linear function and its graph is included in curriculum through grade 8.
Morocco	Geometry: Symmetry and transformations (reflection and rotation)	All or almost all of the students (at least 90%)	Transformations (reflection & rotation) not included in curriculum through grade 8.
Netherlands	Geometry: Congruence and similarity	Not included in curriculum through grade 8	Symmetry taught to all or almost all of the students.
New Zealand	Fractions and Number Sense: Computations with common fractions	All or almost all of the students (at least 90%)	Division with common fractions is not included in curriculum through grade 8.
	Fractions and Number Sense: Square roots (of perfect squares less than 144), small integer exponents	All or almost all of the students (at least 90%)	Small integer exponents taught to about half of the students.
	Algebra: Representing situations algebraically; formulas	About half of the students	Formulas not included in curriculum through grade 8.
	Algebra: Using the graph of a relationship to interpolate/extrapolate	All or almost all of the students (at least 90%)	Using the graph of a relationship to extrapolate not included in curriculum through grade 8.
Russian Federation	Measurement: Perimeter and area of simple shapes - triangles, rectangles, and circles	About half of the students	Perimeter and area of rectangles and circles included in curriculum through grade 8.
	Geometry: Congruence and similarity	About half of the students	Congruence included in curriculum through grade 8.
South Africa	Measurement: Volume of other solids (e.g., pyramids, cylinders, cones, spheres)	All or almost all of the students (at least 90%)	Volume of pyramids, cones, & spheres not included in curriculum through grade 8.
Tunisia	Geometry: Symmetry and transformations (reflection and rotation)	All or almost all of the students (at least 90%)	Rotation not included in curriculum through grade 8.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1998-1999.