Appendix A

Overview of TIMSS Procedures: Science Achievement Results for Seventh- and Eighth-Grade Students

History

TIMSS 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 curricular areas such as 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. Since the subjects of mathematics and science are related in many respects, the third studies were conducted together as an integrated effort.[1]

The number of participating countries and the inclusion of both mathematics and science resulted in TIMSS becoming the largest, most complex IEA study to date and the largest international study of educational achievement ever undertaken. Traditionally, IEA studies have systematically worked toward gaining more in-depth understanding of how various factors contribute to the overall outcomes of schooling. Particular emphasis has been given to refining our understanding of students' opportunity to learn as this opportunity becomes successively defined and implemented by curricular and instructional practices. In an effort to extend what had been learned from previous studies and provide contextual and explanatory information, the magnitude of TIMSS expanded beyond the already substantial task of measuring achievement in two subject areas to also include a thorough investigation of curriculum and how it is delivered in classrooms around the world.

The Components of TIMSS

Continuing the approach of previous IEA studies, TIMSS addressed three conceptual levels of curriculum. The **intended curriculum** is composed of the mathematics and science instructional and learning goals as defined at the system level. The **implemented curriculum** is the mathematics and science curriculum as interpreted by teachers and made available to students. The **attained curriculum** is the mathematics and science content that students have learned and their attitudes towards these subjects. To aid in meaningful interpretation and comparison of results, TIMSS

[1] Because a substantial amount of time has elapsed since earlier IEA studies in mathematics and science, curriculum and testing methods in these two subjects have undergone many changes. Since TIMSS has devoted considerable energy toward reflecting the most current educational and measurement practices, changes in items and methods as well as differences in the populations tested make comparisons of TIMSS results with those of previous studies very difficult. The focus of TIMSS is not on measuring achievement trends, but rather on providing up-to-date information about the current quality of education in mathematics and science.

also collected extensive information about the social and cultural contexts for learning, many of which are related to variation among different educational systems.

Even though slightly fewer countries completed all the steps necessary to have their data included in this report, nearly 50 countries participated in one or more of the various components of the TIMSS data collection effort, including the curriculum analysis. To gather information about the intended curriculum, mathematics and science specialists within each participating country worked section-by-section through curriculum guides, textbooks, and other curricular materials to categorize aspects of these materials in accordance with detailed specifications derived from the TIMSS mathematics and science curriculum frameworks.[2] Initial results from this component of TIMSS can be found in two companion volumes: *Many Visions, Many Aims: A Cross-National Investigation of Curricular Intention in School Mathematics* and *Many Visions, Many Aims: A Cross-National Investigation of Curricular Intentions in School Science*. [3] This component of TIMSS is conducted by researchers at Michigan State University.

To measure the attained curriculum, TIMSS tested more than half a million students in mathematics and science at five grade levels. TIMSS included testing at three separate populations:

Population 1. Students enrolled in the two adjacent grades that contained the largest proportion of 9-year-old students at the time of testing – third- and fourth-grade students in most countries.

Population 2. Students enrolled in the two adjacent grades that contained the largest proportion of 13-year-old students at the time of testing – seventh- and eighth-grade students in most countries.

Population 3. Students in their final year of secondary education. As an additional option, countries could test two special subgroups of these students:

1) Students taking advanced courses in mathematics,
2) Students taking physics.

Countries participating in the study were required to administer tests to the students in the two grades at Population 2, but could choose whether or not to participate at the other levels. In about half of the countries at Populations 1 and 2, subsets of the upper-grade students who completed the written tests also participated in a performance assessment. In the performance assessment, students engaged in a number of hands-on mathematics and science activities. The students designed experiments, tested

[2] Robitaille, D.F., McKnight, C., Schmidt, W., Britton, E., Raizen, S., and Nicol, C. (1993). *TIMSS Monograph No. 1: Curriculum Frameworks for Mathematics and Science*. Vancouver, B.C.: Pacific Educational Press.

[3] Schmidt, W.H., McKnight, C.C., Valverde, G. A., Houang, R.T., and Wiley, D. E. (in press). *Many Visions, Many Aims: A Cross-National Investigation of Curricular Intentions in School Mathematics*. Dordrecht, The Netherlands: Kluwer Academic Publishers. Schmidt, W.H., Raizen, S.A., Britton, E.D., Bianchi, L.J., and Wolfe, R.G., (in press). *Many Visions, Many Aims: A Cross-National Investigation of Curricular Intentions in School Science*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

hypotheses, and recorded their findings. For example, in one task, students were asked to design and conduct a controlled experiment to measure the effect of water temperature on the rate at which tablets dissolve, requiring organization and interpretation of data to draw conclusions and explain results. Figure A.1 shows the countries that participated in the various components of TIMSS achievement testing.

TIMSS also administered a broad array of questionnaires to collect data about how the curriculum is implemented in classrooms, including the instructional practices used to deliver it. The questionnaires also were used to collect information about the social and cultural contexts for learning. Questionnaires were administered at the **country level** about decision-making and organizational features within their educational systems. The **students** who were tested answered questions pertaining to their attitudes towards mathematics and science, 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, textbook usage, 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. In addition, a volume was compiled that presents descriptions of the educational systems of the participating countries.[4]

With its enormous array of data, TIMSS has numerous possibilities for policy-related research, focused studies related to students' understandings of mathematics and science subtopics and processes, and integrated analyses linking the various components of TIMSS. The initial round of reports is only the beginning of a number of research efforts and publications aimed at increasing our understanding of how mathematics and science education functions across countries, investigating what impacts student performance, and helping to improve mathematics and science education.

[4] Robitaille D.F. (in press). *National Contexts for Mathematics and Science Education: An Encyclopedia of the Education Systems Participating in TIMSS.* Vancouver, B.C.: Pacific Educational Press.

Figure A.1

Countries Participating in Additional Components of TIMSS Testing

Country	Population 1		Population 2		Population 3		
	Written Test	Performance Assessment	Written Test	Performance Assessment	Mathematics & Science Literacy	Advanced Mathematics	Physics
Argentina			●				
Australia	●	●	●	●	●	●	●
Austria	●		●		●	●	●
Belgium (Fl)			●				
Belgium (Fr)			●				
Bulgaria			●				
Canada	●	●	●	●	●	●	●
Colombia			●	●			
Cyprus	●	●	●	●	●	●	●
Czech Republic	●	●	●	●	●	●	●
Denmark			●	●	●	●	●
England	●		●	●			
France			●		●	●	●
Germany			●		●	●	●
Greece	●		●		●	●	●
Hong Kong	●	●	●	●			
Hungary	●		●		●		
Iceland	●		●		●		
Indonesia	●		●				
Iran, Islamic Rep.	●	●	●	●			
Ireland	●		●				
Israel	●	●	●	●	●	●	●
Italy	●		●		●		
Japan	●		●				●
Korea	●		●				
Kuwait	●		●				
Latvia	●		●				●
Lithuania			●		●	●	
Mexico	●		●		●	●	●
Netherlands	●		●		●		
New Zealand	●	●	●	●	●		
Norway	●		●	●	●		●
Philippines			●				
Portugal	●	●	●	●			
Romania			●	●			
Russian Federation			●		●	●	●
Scotland	●		●	●			
Singapore	●		●	●			
Slovak Republic			●				
Slovenia	●	●	●	●	●	●	●
South Africa			●		●		
Spain			●	●			
Sweden			●	●	●	●	●
Switzerland			●	●	●	●	●
Thailand	●		●				
United States	●	●	●	●	●	●	●

DEVELOPING THE TIMSS SCIENCE TEST

The TIMSS curriculum framework underlying the science tests at all three populations was developed by groups of science educators with input from the TIMSS National Research Coordinators (NRCs). As shown in Figure A.2, the science curriculum framework contains three dimensions or aspects. The **content** aspect represents the subject matter content of school science. 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 science. The **perspectives** aspect focuses on the development of students' attitudes, interest, and motivations in science.[5]

Working within the science curriculum framework, science test specifications were developed for Population 2 that included items representing a wide range of science topics and eliciting a range of skills from the students. The tests were developed through an international consensus involving input from experts in science and measurement specialists. The TIMSS Subject Matter Advisory Committee, including distinguished scholars from 10 countries, ensured that the test reflected current thinking and priorities in the sciences. The items underwent an iterative development and review process, with one of the pilot testing efforts involving 43 countries. Every effort was made to help ensure that the tests represented the curricula of the participating countries and that the items did not exhibit any bias towards or against particular countries, including modifying specifications in accordance with data from the curriculum analysis component, obtaining ratings of the items by subject-matter specialists within the participating countries, and conducting thorough statistical item analysis of data collected in the pilot testing. The final forms of the test were endorsed by the NRCs of the participating countries.[6] In addition, countries had an opportunity to match the content of the test to their curricula at the seventh and eighth grades. They identified items measuring topics not covered in their intended curriculum. The information from this Test-Curriculum Matching Analysis indicates that omitting such items has little effect on the overall pattern of results (see Appendix B).

Table A.1 presents the five content areas included in the Population 2 science test and the numbers of items and score points in each category. Distributions also are included for the five performance categories derived from the performance expectations aspect of the curriculum framework. Approximately one-fourth of the items were in the free-response format, requiring students to generate and write their own answers. Designed to represent approximately one-third of students' response time, some free-response questions asked for short answers while others required extended

[5] 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, B.C.: Pacific Educational Press.

[6] For a full discussion of the TIMSS test development effort, please see: Garden, R.A. and Orpwood, G. (1996). "TIMSS Test Development" in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study Technical Report, Volume I*. Chestnut Hill, MA: Boston College; and Garden, R.A. (1996). "Development of the TIMSS Achievement Items" in D.F. Robitaille and R.A. Garden (eds.), *TIMSS Monograph No.2: Research Questions and Study Design*. Vancouver, B.C.: Pacific Educational Press.

Figure A.2

The Three Aspects and Major Categories of the Science Framework

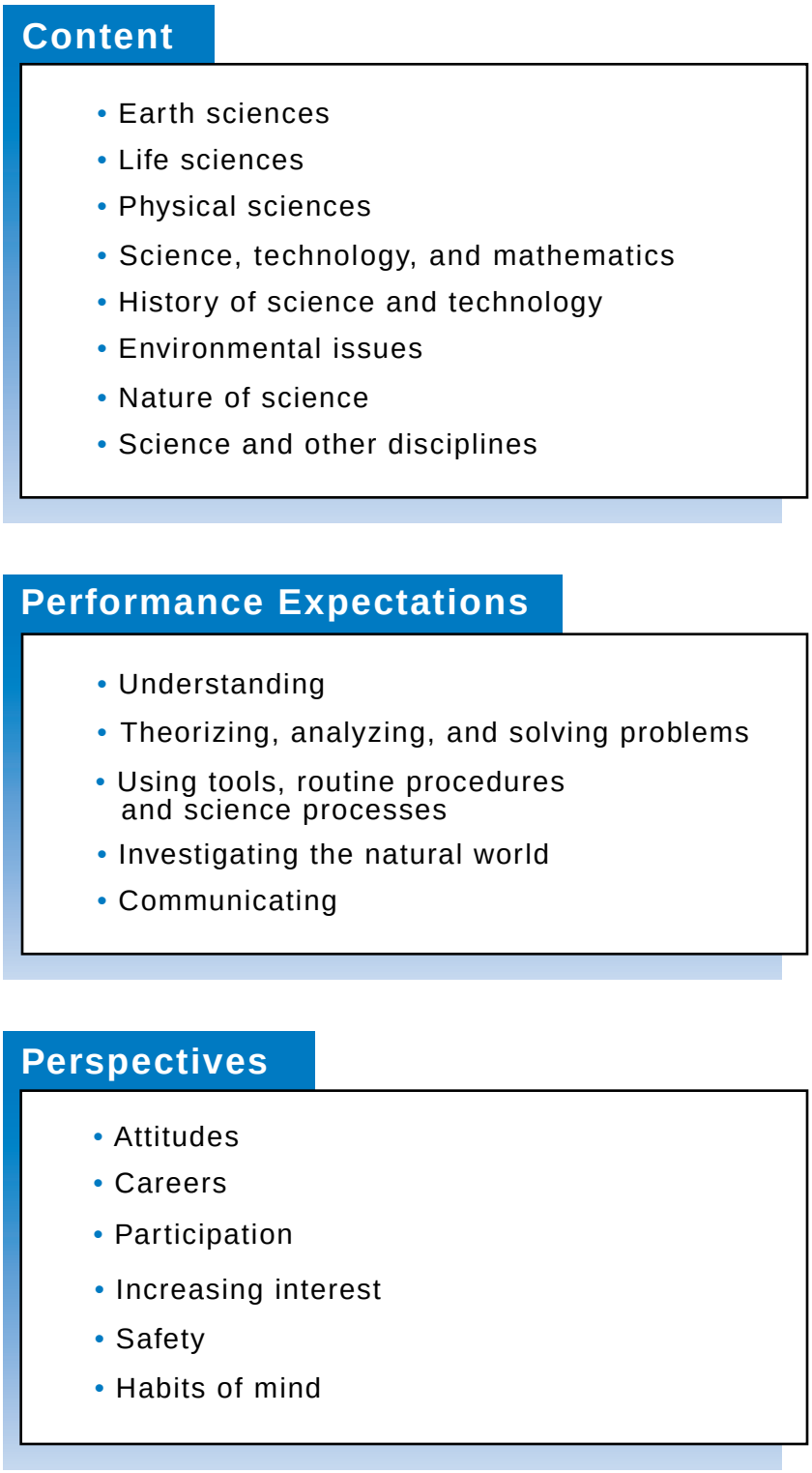

Table A.1

Distribution of Science Items by Content Reporting Category and Performance Category - Population 2

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]
Earth Science	16	22	17	5	24
Life Science	30	40	31	9	44
Physics	30	40	28	12	42
Chemistry	14	19	15	4	21
Environmental Issues and the Nature of Science	10	14	11	3	15

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]
Understanding Simple Information	40	55	53	2	55
Understanding Complex Information	29	39	29	10	41
Theorizing, Analyzing, and Solving Problems	21	28	9	19	36
Using Tools, Routine Procedures, and Science Processes	6	8	8	0	8
Investigating the Natural World	4	5	3	2	6

[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 constructed-response items were evaluated for partial credit with a fully correct answer awarded up to three points. In addition, some items had two parts. Thus, the number of score points exceeds the number of items in the test.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

responses where students needed to show their work or provide 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 or even three points (see later section on scoring). This, in addition to the fact that several items had two parts, means that the total number of score points available for analysis somewhat exceeds the number of items included in the test.

The TIMSS instruments were prepared in English and translated into 30 additional languages. In addition, it sometimes was necessary to adapt the international versions for cultural purposes, including the 11 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 and using two or more independent translations, (3) consultation with subject-matter experts regarding cultural adaptations to ensure that the meaning and difficulty of items did not change, (4) verification of the quality of the translations by professional translators from an independent translation company, (5) corrections by the national centers in accordance with the suggestions made, (6) verification that corrections were implemented, and (7) a series of statistical checks after the testing to detect items that did not perform comparably across countries.[7]

[7] More details about the translation verification procedures can be found in Mullis, I.V.S., Kelly, D.L., and Haley, K. (1996). "Translation Verification Procedures" in M.O. Martin and I.V.S. Mullis (eds.), *Third International Mathematics and Science Study: Quality Assurance in Data Collection*. Chestnut Hill, MA: Boston College; and Maxwell, B. (1996). "Translation and Cultural Adaptation of the TIMSS Instruments" in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study: Technical Report, Volume I*. Chestnut Hill, MA: Boston College.

TIMSS Test Design

Not all of the students in Population 2 responded to all of the science 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. The TIMSS Population 2 test consisted of eight booklets, with each booklet requiring 90 minutes of student response time. In accordance with the design, the mathematics and science items were assembled into 26 different clusters (labeled A through Z). Eight of the clusters were designed to take students 12 minutes to complete; 10 of the clusters, 22 minutes; and 8 clusters, 10 minutes. In all, the design provided a total of 396 unique 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.[8]

Sample Implementation and Participation Rates

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 available and on the quality 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. This documentation was used by the International Study Center in consultation with Statistics Canada, the sampling referee, and the Technical Advisory Committee, to evaluate the quality of the samples.

In a few situations where it was not possible to implement TIMSS testing for the entire internationally desired definition of Population 2 (all students in the two adjacent grades with the greatest proportion of 13-year-olds), countries were permitted to define a national desired population that did not include part of the internationally desired population. Table A.2 shows any differences in coverage between the international and national desired populations. Most participants achieved 100% coverage (36 out of 42). The countries with less than 100% coverage are annotated in tables in this report. In some instances, countries, as a matter of practicality, needed to define their tested population according to the structure of school systems, but in Germany and Switzerland, parts of the country were simply unwilling to take part

[8] The design is fully documented in Adams, R. and Gonzalez, E. (1996). "Design of the TIMSS Achievement Instruments" in D.F. Robitaille and R.A. Garden (eds.), *TIMSS Monograph No. 2: Research Questions and Study Design*. Vancouver, B.C.: Pacific Education Press and 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.

Table A.2

Coverage of TIMSS Target Population

The International Desired Population is defined as follows:
Population 2 - All students enrolled in the two adjacent grades with the largest proportion of 13-year-old students at the time of testing.

Country	International Desired Population		National Desired Population		
	Coverage	Notes on Coverage	School-Level Exclusions	Within-Sample Exclusions	Overall Exclusions
Australia	100%		0.2%	0.7%	0.8%
Austria	100%		2.9%	0.2%	3.1%
Belgium (Fl)	100%		3.8%	0.0%	3.8%
Belgium (Fr)	100%		4.5%	0.0%	4.5%
Bulgaria	100%		0.6%	0.0%	0.6%
Canada	100%		2.4%	2.1%	4.5%
Colombia	100%		3.8%	0.0%	3.8%
Cyprus	100%		0.0%	0.0%	0.0%
Czech Republic	100%		4.9%	0.0%	4.9%
Denmark	100%		0.0%	0.0%	0.0%
[2] England	100%		8.4%	2.9%	11.3%
France	100%		2.0%	0.0%	2.0%
[1] Germany	88%	15 of 16 regions*	8.8%	0.9%	9.7%
Greece	100%		1.5%	1.3%	2.8%
Hong Kong	100%		2.0%	0.0%	2.0%
Hungary	100%		3.8%	0.0%	3.8%
Iceland	100%		1.7%	2.9%	4.5%
Iran, Islamic Rep.	100%		0.3%	0.0%	0.3%
Ireland	100%		0.0%	0.4%	0.4%
[1] Israel	74%	Hebrew Public Education System	3.1%	0.0%	3.1%
Japan	100%		0.6%	0.0%	0.6%
Korea	100%		2.2%	1.6%	3.8%
Kuwait	100%		0.0%	0.0%	0.0%
[1] Latvia (LSS)	51%	Latvian-speaking schools	2.9%	0.0%	2.9%
[1] Lithuania	84%	Lithuanian-speaking schools	6.6%	0.0%	6.6%
Netherlands	100%		1.2%	0.0%	1.2%
New Zealand	100%		1.3%	0.4%	1.7%
Norway	100%		0.3%	1.9%	2.2%
Philippines	91%	2 provinces and autonomous regions excluded	6.5%	0.0%	6.5%
Portugal	100%		0.0%	0.3%	0.3%
Romania	100%		2.8%	0.0%	2.8%
Russian Federation	100%		6.1%	0.2%	6.3%
Scotland	100%		0.3%	1.9%	2.2%
Singapore	100%		4.6%	0.0%	4.6%
Slovak Republic	100%		7.4%	0.1%	7.4%
Slovenia	100%		2.4%	0.2%	2.6%
South Africa	100%		9.6%	0.0%	9.6%
Spain	100%		6.0%	2.7%	8.7%
Sweden	100%		0.0%	0.9%	0.9%
[1] Switzerland	86%	22 of 26 cantons	4.4%	0.8%	5.3%
Thailand	100%		6.2%	0.0%	6.2%
United States	100%		0.4%	1.7%	2.1%

[1]National Desired Population does not cover all of International Desired Population. Because coverage falls below 65%, Latvia is annotated LSS for Latvian Speaking Schools only.
[2]National Defined Population covers less than 90 percent of National Desired Population.
* One region (Baden-Wuerttemberg) did not participate.
SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

in TIMSS. Because coverage fell below 65% for Latvia, the Latvian results have been labeled "Latvia (LSS)," for Latvian Speaking Schools, throughout the report.

Within the desired population, countries could define a population that excluded a small percent (less than 10%) 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 remote areas). Table A.2 also shows that the degree of such exclusions was small. Only England exceeded the 10% limit, and this is annotated in the tables in this report.

Countries were required to test the two adjacent grades with the greatest proportion of 13-year-olds. Table A.3 presents, for each country, the percentage of 13-year-olds in the lower grade tested, the percentage in the upper grade, and the percentage in both the upper and lower grades combined.

Within countries, TIMSS used a two-stage sample design at Population 2, where the first stage involved selecting 150 public and private schools within each country. Within each school, the basic approach required countries to use random procedures to select one mathematics class at the eighth grade and one at the seventh grade (or the corresponding upper and lower grades in that country). All of the students in those two classes were to participate in the TIMSS testing. This approach was designed to yield a representative sample of 7,500 students per country, with approximately 3,750 students at each grade.[9] Typically, between 450 and 3,750 students responded to each item at each grade level, depending on the booklets in which the items were located.

Countries were required to obtain a participation rate of at least 85% of both the schools and the students, or a combined rate (the product of school and student participation) of 75%. Tables A.4 through A.8 present the participation rates and achieved sample sizes for the eighth and seventh grades.

[9] The sample design for TIMSS is described in detail in Foy, P., Rust, K. and Schleicher, A. (1996). "TIMSS Sample 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.

Table A.3

Coverage of 13-Year-Old Students

Country	Percent of 13-Year-Olds in Lower Grade (Seventh Grade*)	Percent of 13-Year-Olds in Upper Grade (Eighth Grade*)	Percent of 13-Year-Olds in Both Grades
Australia	64	28	92
Austria	62	27	89
Belgium (Fl)	46	49	94
Belgium (Fr)	41	46	87
Bulgaria	58	37	95
Canada	48	43	91
Colombia	30	15	45
Cyprus	28	70	98
Czech Republic	73	17	90
Denmark	35	64	98
England	57	42	99
France	44	35	78
Germany	71	2	73
Greece	11	85	96
Hong Kong	44	46	90
Hungary	65	24	89
Iceland	16	83	100
Iran, Islamic Rep.	47	25	72
Ireland	69	17	86
Israel	–	–	–
Japan	91	9	100
Korea	70	28	98
Kuwait	–	–	–
Latvia (LSS)	60	26	86
Lithuania	64	26	90
Netherlands	59	31	90
New Zealand	52	47	99
Norway	43	57	100
Philippines	–	–	–
Portugal	44	32	76
Romania	67	9	76
Russian Federation	50	44	95
Scotland	24	75	99
Singapore	82	15	97
Slovak Republic	73	22	95
Slovenia	65	2	67
South Africa	36	20	55
Spain	46	39	85
Sweden	45	54	99
Switzerland	48	44	92
Thailand	58	20	78
United States	58	33	91

*Seventh and eighth grades in most countries; see Table 2 for more information about the grades tested in each country.
A dash (–) indicates data are unavailable. Israel and Kuwait did not test the lower (seventh) grade.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

Table A.4

School Participation Rates and Sample Sizes - Upper Grade (Eighth Grade*)

Country	School Participation Before Replacement (Weighted Percentage)	School Participation After Replacement (Weighted Percentage)	Number of Schools in Original Sample	Number of Eligible Schools in Original Sample	Number of Schools in Original Sample That Participated	Number of Replacement Schools That Participated	Total Number of Schools That Participated
Australia	75	77	214	214	158	3	161
Austria	41	84	159	159	62	62	124
Belgium (Fl)	61	94	150	150	92	49	141
Belgium (Fr)	57	79	150	150	85	34	119
Bulgaria	72	74	167	167	111	4	115
Canada	90	91	413	388	363	1	364
Colombia	91	93	150	150	136	4	140
Cyprus	100	100	55	55	55	0	55
Czech Republic	96	100	150	149	143	6	149
Denmark	93	93	158	157	144	0	144
England	56	85	150	144	80	41	121
France	86	86	151	151	127	0	127
Germany	72	93	153	150	102	32	134
Greece	87	87	180	180	156	0	156
Hong Kong	82	82	105	104	85	0	85
Hungary	100	100	150	150	150	0	150
Iceland	98	98	161	132	129	0	129
Iran, Islamic Rep.	100	100	192	191	191	0	191
Ireland	84	89	150	149	125	7	132
Israel	45	46	100	100	45	1	46
Japan	92	95	158	158	146	5	151
Korea	100	100	150	150	150	0	150
Kuwait	100	100	69	69	69	0	69
Latvia (LSS)	83	83	170	169	140	1	141
Lithuania	96	96	151	151	145	0	145
Netherlands	24	63	150	150	36	59	95
New Zealand	91	99	150	150	137	12	149
Norway	91	97	150	150	136	10	146
Philippines	96 **	97 **	200	200	192	1	193
Portugal	95	95	150	150	142	0	142
Romania	94	94	176	176	163	0	163
Russian Federation	97	100	175	175	170	4	174
Scotland	79	83	153	153	119	8	127
Singapore	100	100	137	137	137	0	137
Slovak Republic	91	97	150	150	136	9	145
Slovenia	81	81	150	150	121	0	121
South Africa	60	64	180	180	107	7	114
Spain	96	100	155	154	147	6	153
Sweden	97	97	120	120	116	0	116
Switzerland	93	95	259	258	247	3	250
Thailand	99	99	150	150	147	0	147
United States	77	85	220	217	169	14	183

*Eighth grade in most countries; see Table 2 for more information about the grades tested in each country.
**Participation rates for the Philippines are unweighted.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

Table A.5

Student Participation Rates and Sample Sizes - Upper Grade (Eighth Grade*)

Country	Within School Student Participation (Weighted Percentage)	Number of Sampled Students in Participating Schools	Number of Students Withdrawn from Class/School	Number of Students Excluded	Number of Students Eligible	Number of Students Absent	Total Number of Students Assessed
Australia	92	8027	63	61	7903	650	7253
Austria	95	2969	14	4	2951	178	2773
Belgium (Fl)	97	2979	1	0	2978	84	2894
Belgium (Fr)	91	2824	0	1	2823	232	2591
Bulgaria	86	2300	0	0	2300	327	1973
Canada	93	9240	134	206	8900	538	8362
Colombia	94	2843	6	0	2837	188	2649
Cyprus	97	3045	15	0	3030	107	2923
Czech Republic	92	3608	6	0	3602	275	3327
Denmark	93	2487	0	0	2487	190	2297
England	91	2015	37	60	1918	142	1776
France	95	3141	0	0	3141	143	2998
Germany	87	3318	0	35	3283	413	2870
Greece	97	4154	27	23	4104	114	3990
Hong Kong	98	3415	12	0	3403	64	3339
Hungary	87	3339	0	0	3339	427	2912
Iceland	90	2025	10	65	1950	177	1773
Iran, Islamic Rep.	98	3770	20	0	3750	56	3694
Ireland	91	3411	28	10	3373	297	3076
Israel	98	1453	6	0	1447	32	1415
Japan	95	5441	0	0	5441	300	5141
Korea	95	2998	31	0	2967	47	2920
Kuwait	83	1980	3	0	1977	322	1655
Latvia (LSS)	90	2705	19	0	2686	277	2409
Lithuania	87	2915	2	0	2913	388	2525
Netherlands	95	2112	14	1	2097	110	1987
New Zealand	94	4038	121	12	3905	222	3683
Norway	96	3482	26	49	3407	140	3267
Philippines	91 **	6586	93	0	6493	492	6001
Portugal	97	3589	70	13	3506	115	3391
Romania	96	3899	0	0	3899	174	3725
Russian Federation	95	4311	42	10	4259	237	4022
Scotland	88	3289	0	46	3243	380	2863
Singapore	95	4910	18	0	4892	248	4644
Slovak Republic	95	3718	5	3	3710	209	3501
Slovenia	95	2869	15	8	2846	138	2708
South Africa	97	4793	0	0	4793	302	4491
Spain	95	4198	27	102	4069	214	3855
Sweden	93	4483	71	28	4384	309	4075
Switzerland	98	4989	16	24	4949	94	4855
Thailand	100	5850	0	0	5850	0	5850
United States	92	8026	104	108	7814	727	7087

*Eighth grade in most countries; see Table 2 for more information about the grades tested in each country.
**Participation rates for the Philippines are unweighted.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

Table A.6

School Participation Rates and Sample Sizes - Lower Grade (Seventh Grade*)

Country	School Participation Before Replacement (Weighted Percentage)	School Participation After Replacement (Weighted Percentage)	Number of Schools in Original Sample	Number of Eligible Schools in Original Sample	Number of Schools in Original Sample That Participated	Number of Replacement Schools That Participated	Total Number of Schools That Participated
Australia	75	76	214	213	156	3	159
Austria	43	86	159	159	63	62	125
Belgium (Fl)	61	93	150	150	91	49	140
Belgium (Fr)	57	80	150	150	85	35	120
Bulgaria	75	77	150	150	101	3	104
Canada	90	90	413	390	366	1	367
Colombia	91	93	150	150	136	4	140
Cyprus	100	100	55	55	55	0	55
Czech Republic	96	100	150	150	144	6	150
Denmark	88	88	158	154	137	0	137
England	57	85	150	145	81	41	122
France	87	87	151	151	126	0	126
Germany	70	90	153	153	101	31	132
Greece	87	87	180	180	156	0	156
Hong Kong	83	83	105	104	86	0	86
Hungary	99	99	150	150	149	0	149
Iceland	97	97	161	149	144	0	144
Iran, Islamic Rep.	100	100	192	192	192	0	192
Ireland	82	87	150	148	122	7	129
Israel	–	–	–	–	–	–	–
Japan	92	95	158	158	146	5	151
Korea	100	100	150	150	150	0	150
Kuwait	–	–	–	–	–	–	–
Latvia (LSS)	83	84	170	169	141	1	142
Lithuania	96	96	151	151	145	0	145
Netherlands	23	61	150	150	34	58	92
New Zealand	90	99	150	150	135	13	148
Norway	84	96	150	147	124	17	141
Philippines	97 **	97 **	200	200	194	0	194
Portugal	94	94	150	150	141	0	141
Romania	94	94	176	175	162	0	162
Russian Federation	97	100	175	175	170	4	174
Scotland	79	85	153	153	120	9	129
Singapore	100	100	137	137	137	0	137
Slovak Republic	91	97	150	150	136	9	145
Slovenia	81	81	150	150	122	0	122
South Africa	83	85	161	161	133	4	137
Spain	96	100	155	154	147	6	153
Sweden	96	96	160	160	154	0	154
Switzerland	90	94	217	217	200	6	206
Thailand	99	99	150	150	146	0	146
United States	77	84	220	214	165	14	179

*Seventh grade in most countries; see Table 2 for more information about the grades tested in each country.
**Participation rates for the Philippines are unweighted.
A dash (–) indicates data are unavailable. Israel and Kuwait did not test the lower grade.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

Table A.7

Student Participation Rates and Sample Sizes - Lower Grade (Seventh Grade*)

Country	Within School Student Participation (Weighted Percentage)	Number of Sampled Students in Participating Schools	Number of Students Withdrawn from Class/School	Number of Students Excluded	Number of Students Eligible	Number of Students Absent	Total Number of Students Assessed
Australia	93	6067	26	21	6020	421	5599
Austria	95	3196	22	5	3169	156	3013
Belgium (Fl)	97	2857	3	0	2854	86	2768
Belgium (Fr)	95	2418	0	1	2417	125	2292
Bulgaria	87	2080	0	0	2080	282	1798
Canada	95	8962	89	248	8625	406	8219
Colombia	93	2840	2	0	2838	183	2655
Cyprus	98	3028	17	0	3011	82	2929
Czech Republic	92	3641	11	0	3630	285	3345
Denmark	86	2408	0	0	2408	335	2073
England	92	2031	31	67	1933	130	1803
France	95	3164	0	0	3164	148	3016
Germany	87	3388	0	37	3351	458	2893
Greece	97	4166	30	78	4058	127	3931
Hong Kong	98	3507	11	0	3496	83	3413
Hungary	94	3266	0	0	3266	200	3066
Iceland	92	2243	11	72	2160	203	1957
Iran, Islamic Rep.	99	3789	18	0	3771	36	3735
Ireland	91	3480	23	17	3440	313	3127
Israel	–	–	–	–	–	–	–
Japan	96	5337	0	0	5337	207	5130
Korea	94	2996	51	0	2945	38	2907
Kuwait	–	–	–	–	–	–	–
Latvia (LSS)	91	2853	7	0	2846	279	2567
Lithuania	89	2852	3	0	2849	318	2531
Netherlands	95	2220	23	0	2197	100	2097
New Zealand	95	3471	98	17	3356	172	3184
Norway	96	2629	8	53	2568	99	2469
Philippines	93 **	6283	29	1	6253	401	5852
Portugal	96	3594	80	4	3510	148	3362
Romania	95	3938	0	0	3938	192	3746
Russian Federation	96	4408	39	11	4358	220	4138
Scotland	90	3313	0	81	3232	319	2913
Singapore	98	3744	19	0	3725	84	3641
Slovak Republic	95	3797	10	3	3784	184	3600
Slovenia	95	3058	12	4	3042	144	2898
South Africa	96	5532	0	0	5532	231	5301
Spain	95	4087	38	116	3933	192	3741
Sweden	95	3055	27	36	2992	161	2831
Switzerland	99	4199	14	44	4141	56	4085
Thailand	100	5845	0	0	5845	0	5845
United States	94	4295	42	85	4168	282	3886

*Seventh grade in most countries; see Table 2 for more information about the grades tested in each country.
**Participation rates for the Philippines are unweighted.
A dash (–) indicates data are unavailable. Israel and Kuwait did not test the lower grade.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

Table A.8

Overall Participation Rates
Upper and Lower Grades (Eighth and Seventh Grades*)

Country	Upper Grade		Lower Grade	
	Overall Participation Before Replacement (Weighted Percentage)	Overall Participation After Replacement (Weighted Percentage)	Overall Participation Before Replacement (Weighted Percentage)	Overall Participation After Replacement (Weighted Percentage)
Australia	69	70	69	71
Austria	39	80	41	82
Belgium (Fl)	59	91	59	91
Belgium (Fr)	52	72	54	76
Bulgaria	62	63	65	67
Canada	84	84	86	86
Colombia	85	87	84	86
Cyprus	97	97	98	98
Czech Republic	89	92	88	92
Denmark	86	86	76	76
England	51	77	52	78
France	82	82	82	82
Germany	63	81	61	78
Greece	84	84	84	84
Hong Kong	81	81	81	81
Hungary	87	87	93	93
Iceland	88	88	89	89
Iran, Islamic Rep.	98	98	99	99
Ireland	76	81	75	79
Israel	44	45	–	–
Japan	87	90	88	91
Korea	95	95	94	94
Kuwait	83	83	–	–
Latvia (LSS)	75	75	75	76
Lithuania	83	83	86	86
Netherlands	23	60	22	58
New Zealand	86	94	85	94
Norway	87	93	81	92
Philippines	87**	88**	90**	90**
Portugal	92	92	90	90
Romania	89	89	89	89
Russian Federation	93	95	93	95
Scotland	69	73	71	76
Singapore	95	95	98	98
Slovak Republic	86	91	86	92
Slovenia	77	77	77	77
South Africa	58	62	79	82
Spain	91	94	91	95
Sweden	90	90	91	91
Switzerland	92	94	89	93
Thailand	99	99	99	99
United States	71	78	72	79

*Seventh and eighth grades in most countries; see Table 2 for information about the grades tested in each country.
** Participation rates for the Philippines are unweighted.
A dash (–) indicates data are unavailable. Israel and Kuwait did not test the lower grade.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

Indicating Compliance with Sampling Guidelines in the Report

Figure A.3 shows how countries have been grouped in tables reporting achievement results. Countries that achieved acceptable participation rates – 85% of both the schools and students, or a combined rate (the product of school and student participation) of 75% – with or without replacement schools, and that complied with the TIMSS guidelines for grade selection and classroom sampling are shown in the first panel of Figure A.3. Countries that met the guidelines only after including replacement schools are annotated. These countries (25 at the eighth grade and 27 at the seventh grade) appear in the tables in Chapters 1, 2, and 3 ordered by achievement.

Countries not reaching at least 50% school participation without the use of replacement schools, or that failed to reach the sampling participation standard even with the inclusion of replacement schools, are shown in the second panel of Figure A.3. These countries are presented in a separate section of the achievement tables in Chapters 1, 2, and 3 in alphabetical order, and are shown in tables in Chapters 4 and 5 in italics.

To provide a better curricular match, four countries (i.e., Colombia, Germany, Romania, and Slovenia) elected to test their seventh- and eighth-grade students even though that meant not testing the two grades with the most 13-year-olds and led to their students being somewhat older than in the other countries. These countries are also presented in a separate section of the achievement tables in Chapters 1, 2, and 3 in alphabetical order, and are shown in tables in Chapters 4 and 5 in italics.

For a variety of reasons, three countries (Denmark, Greece, and Thailand) did not comply with the guidelines for sampling classrooms. Their results are also presented in a separate section of the achievement tables in Chapters 1, 2, and 3 in alphabetical order, and are italicized in tables in Chapters 4 and 5. At the eighth grade, Israel, Kuwait, and South Africa also had difficulty complying with the classroom selection guidelines, but in addition had other difficulties (Kuwait tested a single grade with relatively few 13-year-olds; Israel and South Africa had low sampling participation rates), and so these countries are also presented in separate sections in tables in Chapters 1, 2, and 3, and are italicized in tables in Chapters 4 and 5. At the seventh grade, South Africa had a better sampling participation rate, and is presented in the same section of tables as Denmark, Greece and Thailand. Israel and Kuwait did not test at the seventh grade.

Because the Philippines was unable to document clearly the school sampling procedures used, its results are not presented in the main body of the report. A small set of results for the Philippines can be found in Appendix C.

Figure A.3

Countries Grouped for Reporting of Achievement According to Their Compliance with Guidelines for Sample Implementation and Participation Rates

	Eighth Grade	Seventh Grade
Countries satisfying guidelines for sample participation rates, grade selection and sampling procedures	† Belgium (Fl) Canada Cyprus Czech Republic †2 England France Hong Kong Hungary Iceland Iran, Islamic Rep. Ireland Japan Korea 1 Latvia 1 Lithuania New Zealand Norway Portugal Russian Federation Singapore Slovak Republic Spain Sweden 1 Switzerland † United States	† Belgium (Fr) † Belgium (Fl) Canada Cyprus Czech Republic †2 England France Hong Kong Hungary Iceland Iran, Islamic Rep. Ireland Japan Korea 1 Latvia (LSS) 1 Lithuania New Zealand Norway Portugal Russian Federation † Scotland Singapore Slovak Republic Spain Sweden 1 Switzerland † United States
Countries not satisfying guidelines for sample participation	Australia Austria Belgium (Fr) Bulgaria Netherlands Scotland	Australia Austria Bulgaria Netherlands
Countries not meeting age/grade specifications (high percentage of older students)	Colombia †1 Germany Romania Slovenia	Colombia †1 Germany Romania Slovenia
Countries with unapproved sampling procedures at the classroom level	Denmark Greece Thailand	Denmark Greece 1 South Africa Thailand
Countries with unapproved sampling procedures at classroom level and not meeting other guidelines	1 Israel Kuwait South Africa	
Countries with unapproved sampling procedures at school level	3 Philippines	3 Philippines

†Met guidelines for sample participation rates only after replacement schools were included.

1National Desired Population does not cover all of International Desired Population (see Table 1). Because coverage falls below 65%, Latvia is annotated LSS for Latvian Speaking Schools only.

2National Defined Population covers less than 90 percent of National Desired Population (see Table 1).

3TIMSS was unable to compute sampling weights for the Philippines. Selected unweighted achievement results for the Philippines are presented in Appendix C.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

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 developed 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. The test administrator 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 as part of the NRC's report documenting procedures used in the study. In addition, the International Study Center considered it essential to establish some method to monitor compliance with standardized procedures. NRCs were asked to nominate a person, such as a retired school teacher, to serve as quality control monitor for their countries, and in almost all cases, the International Study Center adopted the NRCs' first suggestion. The International Study Center developed manuals for the quality control 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.

The quality control monitors interviewed the NRCs about data collection plans and procedures. They also selected a sample of approximately 10 schools to visit where they observed testing sessions and interviewed school coordinators.[10] Quality control monitors observed test administrations and interviewed school coordinators in 37 countries, and interviewed school coordinators or test administrators in 3 additional countries.

The results of the interviews indicate that, in general, NRCs had prepared well for data collection and, despite the heavy demands of the schedule and shortages of resources, were in a position to conduct the data collection in an efficient and professional manner. Similarly, the TIMSS tests appeared to have been administered in compliance with international procedures, including the activities preliminary to the testing session, the activities during the testing sessions, and the school-level activities related to receiving, distributing, and returning materials from the national centers.

[10] The results of the interviews and observations by the quality control monitors are presented in Martin, M.O., Hoyle, C.D., and Gregory, K.D. (1996). "Monitoring the TIMSS Data Collection" and "Observing the TIMSS Test Administration" both in M.O. Martin and I.V.S. Mullis (eds.), *Third International Mathematics and Science Study: Quality Assurance in Data Collection*. Chestnut Hill, MA: Boston College.

Scoring the Free-Response Items

Because approximately 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 utilized two-digit codes with rubrics specific to each item. Development of the rubrics was led by the Norwegian TIMSS national center. The first digit designates the correctness level of the response. The second digit, combined with the first digit, represents a diagnostic code used to identify specific types of approaches, strategies, or common errors and misconceptions. Although not specifically 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 meet the goal of implementing reliable scoring procedures based on the TIMSS rubrics, the International Study Center prepared guides containing the rubrics and explanations of how to implement them together with example student responses for the various rubric categories. These guides, together with more examples of student responses for practice in applying the rubrics, were used as a basis for an ambitious series of regional training sessions. The training sessions were designed to assist representatives of national centers who would then be responsible for training personnel in their respective countries to apply the two-digit codes reliably.[11]

To gather and document empirical information about the within-country agreement among scorers, TIMSS developed a procedure whereby systematic subsamples of approximately 10% of the students' responses were to be coded independently by two different readers. To provide information about the cross-country agreement among scorers, TIMSS conducted a special study at Population 2, where 39 scorers from 21 of the participating countries evaluated common sets of students' responses to more than half of the free-response items.

Table A.9 shows the average and range of the within-country exact percent of agreement between scorers on the free-response items in the Population 2 science test for 26 countries. Unfortunately, lack of resources precluded several countries from providing this information. A high percent of exact agreement was observed, with averages across the items for the correctness score ranging from 88% to 100% and an overall average of 95% across the 26 countries.

The cross-country coding reliability study involved 350 students' responses for each of 14 mathematics and 17 science items, totaling 10,850 responses in all. The responses were random samples from the within-country reliability samples from seven English-test countries: Australia, Canada, England, Ireland, New Zealand, Singapore, and

[11] The procedures used in the training sessions are documented in Mullis, I.V.S., Garden, R.A., and Jones, C.A. (1996). "Training for Scoring the TIMSS Free-Response Items" in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study Technical Report, Volume I.* Chestnut Hill, MA: Boston College.

Table A.9

TIMSS Within-Country Free-Response Coding Reliability Data for Population 2 Science Items*

Country	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	91	69	99	78	48	97
Belgium (Fl)	100	95	100	98	82	100
Bulgaria	91	63	100	81	50	100
Canada	92	76	100	80	59	99
Colombia	97	83	100	91	73	100
Czech Republic	96	87	100	90	61	100
England	97	90	100	91	65	100
France	99	95	100	97	89	100
Germany	94	81	100	84	66	100
Hong Kong	94	72	100	87	56	100
Iceland	95	74	100	83	22	98
Iran, Islamic Rep.	88	67	100	73	33	99
Ireland	95	87	100	89	69	100
Japan	100	96	100	98	87	100
Netherlands	92	75	100	79	17	100
New Zealand	97	90	100	90	63	100
Norway	95	87	100	91	71	100
Portugal	96	88	100	91	75	100
Russian Federation	96	87	100	91	73	100
Scotland	89	73	99	74	52	96
Singapore	98	92	100	95	86	100
Slovak Republic	92	62	100	81	43	100
Spain	95	85	100	88	73	98
Sweden	94	80	100	83	54	99
Switzerland	98	93	100	93	85	99
United States	97	90	100	89	74	100
AVERAGE	95	82	100	87	63	99

*Based on 33 science items, including 4 multiple-part items.
Note: Percent agreement was computed separately for each part, and each part was treated as a separate item in computing averages and ranges.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

the United States. The responses were presented to the scorers according to a rotated design whereby each response was coded by 7 to 18 different scorers. This design resulted in a large number of comparisons between coders, approximately 10,000 or more for each item.

Table A.10 presents the percent of exact agreement for the 17 science items and the scorers involved in the international study. For comparison purposes, it also shows the average and range of the percent of exact agreement for each of the items within the 26 countries submitting data about their scoring reliability. The percent of exact agreement for each science item was fairly high on the correctness score agreement. Most measures fell between 80% and 99%, although measures for three items were between 72% and 78%. In general, the average international correctness score agreement for the science items was not as high as the within-country agreement (86% as opposed to 94%), but results are acceptable, and to be expected given the nature of the science items and the nature of the international coding reliability study. 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.[12]

[12] Details about the reliability studies can be found in Mullis, I.V.S. and Smith, T.A. (1996). "Quality Control Steps for Free-Response Scoring" in M.O. Martin and I.V.S. Mullis (eds.), *Third International Mathematics and Science Study: Quality Assurance in Data Collection.* Chestnut Hill, MA: Boston College.

Table A.10

Percent Exact Agreement for Coding of Science Items for International and Within-Country Reliability Studies

Item Label	Total Valid Comparisons in International Study	Correctness Score Agreement				Diagnostic Code Agreement			
		International Study	Within-Country Study			International Study	Within-Country Study		
			Average	Min	Max		Average	Min	Max
O10	9078	99	99	95	100	98	97	80	100
O17	46035	94	97	77	100	74	86	64	100
Q18	9150	93	96	81	100	85	91	54	100
K19	12600	93	95	83	100	67	80	52	99
P03	46050	92	97	88	100	78	88	58	100
K10	46050	91	96	90	100	79	91	79	99
[1] W01A	9150	90	95	83	100	71	87	67	99
[1] W01B	9150	89	95	87	100	77	89	74	98
R04	45930	89	96	90	100	70	84	65	98
P06	46050	88	93	74	100	74	87	64	100
O14	9150	88	96	86	100	83	91	65	100
R05	9122	86	95	86	100	72	87	61	100
O16	45930	86	95	81	100	59	80	53	96
Q17	46034	82	93	74	100	66	87	65	100
P05	9150	80	93	82	100	59	82	47	100
W02	46050	78	92	75	100	70	89	69	99
Q12	12600	75	91	74	100	51	78	55	100
R03	9129	72	90	70	100	50	82	59	100
AVERAGE SCIENCE ITEMS		86	94	81	100	70	86	62	99

[1]Two-part items; each part is analyzed separately.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

Test Reliability

Table A.11 displays the science test reliability coefficient for each country for the lower and upper grades (usually seventh and eighth grades). This coefficient is the median KR-20 reliability across the eight test booklets. Median reliabilities in the lower grade ranged from 0.83 in the United States and the Philippines to 0.68 in Portugal and in the upper grade from 0.84 in Australia, Bulgaria, and the Philippines to 0.69 in Kuwait. The international median, shown in the last row of the table, is the median of the reliability coefficients for all countries. These international medians are 0.77 for the lower grade and 0.78 for the upper grade.

Data Processing

To ensure the availability of comparable, high quality data for analysis, TIMSS engaged in a rigorous set of quality control steps to create the international database.[13] TIMSS prepared manuals and software for countries to use in entering their data so 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 IEA Data Processing Center, the data from each country underwent an exhaustive cleaning process. The data-cleaning process involved several iterative steps and procedures designed to identify, document, and correct deviations from the international instruments, file structures, and coding schemes. This 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 Australian Council for Educational Research (ACER), the International Study Center conducted a review of items statistics for each of the cognitive items in each of the countries to identify poorly performing items. Twenty-one 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.

[13] These steps are detailed in Jungclaus, H. and Bruneforth, M. (1996). "Data Consistency Checking Across Countries" in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study Technical Report, Volume I.* Chestnut Hill, MA: Boston College.

Table A.11

Cronbach's Alpha Reliability Coefficients[1] - TIMSS Science Test Lower and Upper Grades (Seventh and Eighth Grades*)

Country	Lower Grade	Upper Grade
Australia	0.81	0.84
Austria	0.80	0.81
Belgium (Fl)	0.68	0.78
Belgium (Fr)	0.72	0.79
Bulgaria	0.81	0.84
Canada	0.79	0.78
Colombia	0.69	0.72
Cyprus	0.74	0.79
Czech Republic	0.75	0.78
Denmark	0.77	0.77
England	0.82	0.83
France	0.71	0.73
Germany	0.80	0.82
Greece	0.78	0.77
Hong Kong	0.78	0.78
Hungary	0.80	0.79
Iceland	0.74	0.75
Iran, Islamic Rep.	0.71	0.71
Ireland	0.78	0.82
Israel	–	0.83
Japan	0.76	0.79
Korea	0.79	0.79
Kuwait	–	0.69
Latvia (LSS)	0.74	0.76
Lithuania	0.75	0.75
Netherlands	0.74	0.76
New Zealand	0.80	0.82
Norway	0.77	0.78
Philippines	0.83	0.84
Portugal	0.68	0.75
Romania	0.81	0.82
Russian Federation	0.79	0.79
Scotland	0.79	0.82
Singapore	0.81	0.77
Slovak Republic	0.77	0.81
Slovenia	0.77	0.78
South Africa	0.78	0.82
Spain	0.75	0.73
Sweden	0.76	0.77
Switzerland	0.74	0.78
Thailand	0.70	0.72
United States	0.83	0.83
International Median	0.77	0.78

*Seventh and eighth grade in most countries; see Table 2 for more information about the grades tested in each country. Israel and Kuwait did not test the lower grade.

[1]The reliability coefficient for each country is the median KR-20 reliability across the eight test booklets. The international median is the median of the reliability coefficients for all countries.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

IRT Scaling and Data Analysis

Two general analysis approaches were used for this report – item response theory scaling methods and average percent correct technology. The overall science results were summarized using an item response theory (IRT) scaling method (Rasch model). This scaling method produces a science score by averaging the responses of each student to the items which they took in a way that takes into account the difficulty 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. Analyses of the response patterns of students from participating countries indicated that, although the items in the test address a wide range of science content, the performance of the students across the items was sufficiently consistent that it could be usefully summarized in a single science score.

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. The scale was standardized using students from both the grades tested. When all participating countries and grades are treated equally, the TIMSS scale average is 500 and the standard deviation is 100. Since the countries varied in size, each country was reweighted to contribute equally to the mean and standard deviation of the scale. The average of the scale scores was constructed to be the average of the 41 means of participants that were available at the eighth grade and the 39 means at the seventh grade. The average and standard deviation of the scale scores are arbitrary and do not affect scale interpretations.

The analytic approach underlying the results in Chapters 2 and 3 of this report involved calculating the percentage of correct answers for each item for each participating country (as well as the percentages of different types of incorrect responses). The percents correct were averaged to summarize science performance overall and in each of the content areas for each country as a whole and by gender. For items with more than one part, each part was analyzed separately in calculating the average percents correct. Also, for items with more than one point awarded for full credit, the average percents correct reflect an average of the points received by students in each country. This was achieved by including the percent of students receiving one score point as well as the percentage receiving two score points and three score points in the calculations. Thus, the average percents correct are based on the number of score points rather than the number of items, per se. An exception to this is the international average percents correct reported for example items, where the values reflect the percent of students receiving full credit.

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 would have answered every question, it is important to have measures of the degree of uncertainty of the estimates. The jackknife procedure was used to estimate the standard error associated with each statistic presented in this report. The 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% confidence interval for the corresponding population result.