APPENDIX A

Overview of TIMSS Benchmarking 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. The Third International Mathematics and Science Study (TIMSS), conducted in 1994-1995, was the largest and most complex IEA study, and included both mathematics and science at third and fourth grades, seventh and eighth grades, and the final year of secondary school. In 1999, TIMSS again assessed eighth-grade students in both mathematics and science to measure trends in student achievement since 1995. TIMSS 1999 was also known as TIMSS-Repeat, or TIMSS-R.[1]

To provide U.S. states and school districts with an opportunity to benchmark the performance of their students against that of students in the high-performing TIMSS countries, the International Study Center at Boston College, with the support of the National Center for Education Statistics and the National Science Foundation, established the TIMSS 1999 Benchmarking Study. Through this project, the TIMSS mathematics and science achievement tests and questionnaires were administered to representative samples of students in participating states and school districts in the spring of 1999, at the same time the tests and questionnaires were administered in the TIMSS countries. Participation in TIMSS Benchmarking was intended to help states and districts understand their comparative educational standing, assess the rigor and effectiveness of their own mathematics and science programs in an international context, and improve the teaching and learning of mathematics and science.

Participants in TIMSS Benchmarking

Thirteen states availed of the opportunity to participate in the Benchmarking Study. Eight public school districts and six consortia also participated, for a total of fourteen districts and consortia. They are listed in Exhibit 1 of the Introduction, together with the 38 countries that took part in TIMSS 1999.

[1] The TIMSS 1999 results for mathematics and science, respectively, are reported in Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., Gregory, K.D., Garden, R.A., O'Connor, K.M., Chrostowski, S.J., and Smith, T.A. (2000), *TIMSS 1999 International Mathematics Report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade*, Chestnut Hill, MA: Boston College, and in Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., Gregory, K.D., Smith, T.A., Chrostowski, S.J., Garden, R.A., and O'Connor, K.M. (2000), *TIMSS 1999 International Science Report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade*, Chestnut Hill, MA: Boston College.

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

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 tests were endorsed by the NRCs of the participating countries.[3]

[3] 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, K.M. O'Connor, and S.E. Stemler (eds.), *TIMSS 1999 Benchmarking Technical Report*, Chestnut Hill, MA: Boston College.

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.2 Distribution of Mathematics Items by Content Reporting Category and Performance Category

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.

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

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.[5]

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, including the United States.[6] The eighth grade was the target population for all of the Benchmarking participants.

The selection of valid and efficient samples was essential to the success of TIMSS and of the Benchmarking Study. For TIMSS internationally, NRCs, including Westat, the sampling and data collection coordinator for TIMSS in the United States, received training in how to select the school and student samples and in the use of the sampling software, and worked in close consultation with Statistics Canada, the TIMSS sampling consultants, on all phases of sampling. As well as conducting the sampling and data collection for the U.S. national TIMSS sample, Westat was also responsible for sampling and data collection in each of the Benchmarking states, districts, and consortia.

5 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, K.M. O'Connor, and S.E. Stemler (eds.), *TIMSS 1999 Benchmarking Technical Report*, Chestnut Hill, MA: Boston College.

6 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. Sampling for the Benchmarking project is described in Fowler, J., Rizzo, L., and Rust, K. (2001), "TIMSS Benchmarking Sampling Design and Implementation" in M.O. Martin, K.D. Gregory, K.M. O'Connor, and S.E. Stemler (eds.), *TIMSS 1999 Benchmarking Technical Report*, Chestnut Hill, MA: Boston College.

To document the quality of the school and student samples in each of the TIMSS countries, staff from Statistics Canada and the International Study Center worked with the TIMSS sampling referee (Keith Rust, Westat) to review sampling plans, sampling frames, and sampling implementation. Particular attention was paid to coverage of the target population and to participation by the sampled schools and students. The data from the few countries that did not fully meet all of the sampling guidelines are annotated in the TIMSS international reports, and are also annotated in this report. The TIMSS samples for the Benchmarking participants were also carefully reviewed in light of the TIMSS sampling guidelines, and the results annotated where appropriate. Since Westat was the sampling contractor for the Benchmarking project, the role of sampling referee for the Benchmarking review was filled by Pierre Foy, of Statistics Canada.

Although all countries and Benchmarking participants were expected to draw samples representative of the entire internationally desired population (all students in the upper of the two adjacent grades with the greatest proportion of 13-year-olds), the few countries where this was not possible were permitted to define a national desired population that excluded part of the internationally desired population. Exhibit A.3 shows any differences in coverage between the international and national desired populations. Almost all TIMSS countries achieved 100 percent coverage (36 out of 38), with Lithuania and Latvia the exceptions. Consequently, the results for Lithuania are annotated, and because coverage fell below 65 percent for Latvia, the Latvian results are labeled "Latvia (LSS)," for Latvian-Speaking Schools. Additionally, 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 to reflect this as well. Exhibit A.3 also shows that the sampling plans for the Benchmarking participants all incorporated 100 percent coverage of the desired population. Four of the 13 states (Idaho, Indiana, Michigan, and Pennsylvania) as well as the Southwest Pennsylvania Math and Science Collaborative included private schools as well as public schools.

In operationalizing their desired eighth-grade population, countries and Benchmarking participants could define a population to be sampled 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.3 also shows that the degree of such exclusions was small. Among countries, only Israel reached the 10 percent limit, and among Benchmarking participants, only Guilford County and Montgomery County did so. All three are annotated as such in the achievement chapters of 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.

States participating in the Benchmarking study were required to sample at least 50 schools and approximately 2,000 eighth-grade students. School districts and consortia were required to sample at least 25 schools and at least 1,000 students. Where there were fewer than 25 schools in a district or consortium, all schools were to be included, and the within-school sample increased to yield the total of 1,000 students.

Exhibits A.4 and A.5 present achieved sample sizes for schools and students, respectively, for the TIMSS countries and for the Benchmarking participants. Where a district or consortium was part of a state that also participated, the state sample was augmented by the district or consortium sample, properly weighted in accordance with its size. Schools in a state that were sampled as part of the U.S. national TIMSS sample were also used to augment the state sample. For example, the Illinois sample consists of 90 schools, 41 from the state Benchmarking sample (including five schools from the national TIMSS sample), 27 from the Chicago Public Schools, 17 from the First in the World Consortium, and five from the Naperville School District.

Exhibit A.6 shows the participation rates for schools, students, and overall, both with and without the use of replacement schools, for TIMSS countries and Benchmarking participants. All of the countries met the guideline for sampling participation – 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, and are annotated accordingly in the achievement chapters.

With the exception of Pennsylvania and Texas, all the Benchmarking participants met the sampling guidelines, although Indiana did so only after including replacement schools. Indiana is annotated to reflect this in the achievement chapters, and Pennsylvania and Texas are italicized in all exhibits in this report.

	International Desired Population		National Desired Population		
	Coverage	Notes on Coverage	School-Level Exclusions	Within-Sample Exclusions	Overall Exclusions
United States	100%		0%	4%	4%
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%

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

	International Desired Population		National Desired Population		
	Coverage	Notes on Coverage	School-Level Exclusions	Within-Sample Exclusions	Overall Exclusions
States					
Connecticut	100%		0%	5%	5%
Idaho	100%	Included private schools	0%	2%	2%
Illinois	100%		0%	4%	4%
Indiana	100%	Included private schools	0%	6%	6%
Maryland	100%		0%	6%	6%
Massachusetts	100%		0%	5%	5%
Michigan	100%	Included private schools	0%	2%	2%
Missouri	100%		0%	4%	4%
North Carolina	100%		0%	4%	4%
Oregon	100%		0%	5%	5%
Pennsylvania	100%	Included private schools	0%	6%	6%
South Carolina	100%		0%	2%	2%
Texas	100%		0%	4%	4%
Districts and Consortia					
Academy School Dist. #20, CO	100%		NA	2%	2%
Chicago Public Schools, IL	100%		NA	4%	4%
Delaware Science Coalition, DE	100%		NA	5%	5%
First in the World Consort., IL	100%		NA	2%	2%
Fremont/Lincoln/WestSide PS, NE	100%		NA	2%	2%
Guilford County, NC	100%		NA	10%	10%
Jersey City Public Schools, NJ	100%		NA	6%	6%
Miami-Dade County PS, FL	100%		NA	7%	7%
Michigan Invitational Group, MI	100%		NA	2%	2%
Montgomery County, MD	100%		NA	17%	17%
Naperville Sch. Dist. #203, IL	100%		NA	7%	7%
Project SMART Consortium, OH	100%		NA	2%	2%
Rochester City Sch. Dist., NY	100%		NA	1%	1%
SW Math/Sci. Collaborative, PA	100%	Included private schools	NA	4%	4%

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

	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
United States	250	246	202	19	221
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

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

	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
States					
Connecticut	54	54	52	0	52
Idaho	54	54	47	0	47
Illinois	90	90	85	0	85
Indiana	61	61	39	13	52
Maryland	79	77	73	0	73
Massachusetts	59	58	57	0	57
Michigan	66	62	55	2	57
Missouri	57	55	43	8	51
North Carolina	71	68	67	0	67
Oregon	51	51	45	0	45
Pennsylvania	116	113	80	0	80
South Carolina	53	53	49	0	49
Texas	71	70	51	1	52
Districts and Consortia					
Academy School Dist. #20, CO	4	4	4	0	4
Chicago Public Schools, IL	27	27	26	0	26
Delaware Science Coalition, DE	25	25	25	0	25
First in the World Consort., IL	17	17	15	0	15
Fremont/Lincoln/WestSide PS, NE	12	12	12	0	12
Guilford County, NC	17	17	17	0	17
Jersey City Public Schools, NJ	25	25	24	0	24
Miami-Dade County PS, FL	25	25	25	0	25
Michigan Invitational Group, MI	21	21	21	0	21
Montgomery County, MD	25	25	25	0	25
Naperville Sch. Dist. #203, IL	5	5	5	0	5
Project SMART Consortium, OH	24	24	24	0	24
Rochester City Sch. Dist., NY	7	7	7	0	7
SW Math/Sci. Collaborative, PA	50	49	39	0	39

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

	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
United States	94%	9981	115	142	9724	652	9072
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

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

	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
States							
Connecticut	94%	2190	6	43	2141	124	2023
Idaho	95%	1968	17	27	1924	94	1847
Illinois	96%	5144	30	136	4978	227	4781
Indiana	95%	2175	9	27	2139	102	2046
Maryland	94%	3877	21	339	3517	221	3317
Massachusetts	95%	2538	18	54	2466	131	2353
Michigan	96%	2811	7	44	2760	143	2623
Missouri	94%	2147	27	40	2080	128	1979
North Carolina	94%	3502	34	191	3277	214	3097
Oregon	93%	2044	24	29	1991	126	1889
Pennsylvania	95%	3463	18	60	3385	167	3236
South Carolina	94%	2177	18	36	2123	130	2011
Texas	93%	2189	18	44	2127	149	1996
Districts and Consortia							
Academy School Dist. #20, CO	94%	1329	0	15	1314	81	1233
Chicago Public Schools, IL	94%	1227	13	21	1193	74	1132
Delaware Science Coalition, DE	92%	1389	16	18	1355	103	1268
First in the World Consort., IL	96%	782	1	2	779	30	750
Fremont/Lincoln/WestSide PS, NE	95%	1178	20	25	1133	60	1093
Guilford County, NC	92%	1215	17	121	1077	76	1018
Jersey City Public Schools, NJ	94%	1116	5	47	1064	65	1004
Miami-Dade County PS, FL	91%	1356	23	10	1323	117	1229
Michigan Invitational Group, MI	91%	994	0	11	983	80	903
Montgomery County, MD	94%	1481	13	254	1214	72	1155
Naperville Sch. Dist. #203, IL	96%	1343	9	84	1250	47	1212
Project SMART Consortium, OH	94%	1188	11	18	1159	74	1096
Rochester City Sch. Dist., NY	84%	1165	8	9	1148	190	966
SW Math/Sci. Collaborative, PA	95%	1638	14	21	1603	79	1538

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

	School Participation		Student Participation	Overall Participation	
	Before Replacement	After Replacement		Before Replacement	After Replacement
United States	83%	90%	94%	78%	85%
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%

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

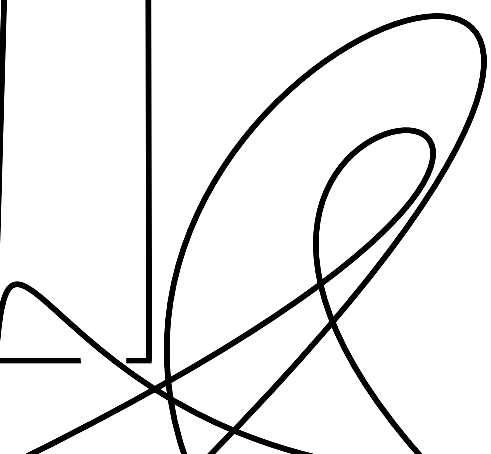
	School Participation		Student Participation	Overall Participation	
	Before Replacement	After Replacement		Before Replacement	After Replacement
States					
Connecticut	96%	96%	94%	90%	90%
Idaho	88%	88%	95%	83%	83%
Illinois	95%	95%	96%	91%	91%
Indiana	61%	83%	95%	58%	79%
Maryland	94%	94%	94%	88%	88%
Massachusetts	98%	98%	95%	93%	93%
Michigan	89%	92%	96%	85%	88%
Missouri	79%	94%	94%	75%	88%
North Carolina	98%	98%	94%	92%	92%
Oregon	89%	89%	93%	83%	83%
Pennsylvania	66%	66%	95%	63%	63%
South Carolina	92%	92%	94%	86%	86%
Texas	73%	74%	93%	67%	69%
Districts and Consortia					
Academy School Dist. #20, CO	100%	100%	94%	94%	94%
Chicago Public Schools, IL	95%	95%	94%	90%	90%
Delaware Science Coalition, DE	100%	100%	92%	92%	92%
First in the World Consort., IL	93%	93%	96%	90%	90%
Fremont/Lincoln/WestSide PS, NE	100%	100%	95%	95%	95%
Guilford County, NC	100%	100%	92%	92%	92%
Jersey City Public Schools, NJ	97%	97%	94%	91%	91%
Miami-Dade County PS, FL	100%	100%	91%	91%	91%
Michigan Invitational Group, MI	100%	100%	91%	91%	91%
Montgomery County, MD	100%	100%	94%	94%	94%
Naperville Sch. Dist. #203, IL	100%	100%	96%	96%	96%
Project SMART Consortium, OH	100%	100%	94%	94%	94%
Rochester City Sch. Dist., NY	100%	100%	84%	84%	84%
SW Math/Sci. Collaborative, PA	78%	78%	95%	75%	75%

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. As the data collection contractor for the U.S. national TIMSS, Westat was fully acquainted with the TIMSS procedures, and applied them in each of the Benchmarking jurisdictions in the same way as in the national data collection.

Each country was responsible for conducting quality control procedures and describing this effort in the NRC's report documenting procedures used in the study. In addition, the International Study Center considered it essential to monitor compliance with standardized procedures through an international program of quality control site visits. 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



As a parallel quality control effort for the Benchmarking project, the International Study Center recruited and trained a team of 18 quality control observers, and sent them to observe the data collection activities of the Westat test administrators in a sample of about 10 percent of the schools in the study (98 schools in all).[8] In line with the experience internationally, the observers reported that the data collection was conducted successfully according to the prescribed procedures, and that no serious problems were encountered.

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.

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. In the United States, the scoring was conducted by National Computer Systems (NCS) under contract to Westat. To ensure that student responses from the Benchmarking participants were scored in the same way as those from the U.S. national sample, NCS had both sets of data scored at the same time and by the same scoring staff.

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.7 shows the average and range of the within-country percent of exact agreement between scorers on the

[8] Quality control measures for the Benchmarking project are described in O'Connor, K. and Stemler, S. (2001), "Quality Control in the TIMSS Benchmarking Data Collection" in M.O. Martin, K.D. Gregory, K.M. O'Connor, and S.E. Stemler (eds.), *TIMSS 1999 Benchmarking Technical Report*, Chestnut Hill, MA: Boston College.

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. In the United States, the average percent exact agreement was 99 percent for the correctness score and 96 percent for the diagnostic score. Since the Benchmarking data were combined with the U.S. national TIMSS sample for scoring purposes, this high level of scoring reliability applies to the Benchmarking data also.

	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
United States	99	96	100	97	89	100
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
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

Exhibit A.8 displays the mathematics test reliability coefficient for each country and Benchmarking participant. This coefficient is the median KR-20 reliability across the eight test booklets. Among countries, 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. Reliability coefficients among Benchmarking participants were generally close to the international median, ranging from 0.88 to 0.91 across states, and from 0.84 to 0.91 across districts and consortia.

Exhibit A.8 Cronbach's Alpha Reliability Coefficient – TIMSS 1999 Mathematics Test

8th Grade Mathematics

Countries	Reliability Coefficient[1]
United States	0.90
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
International Median	0.89

States	Reliability Coefficient[1]
Connecticut	0.90
Idaho	0.89
Illinois	0.89
Indiana	0.89
Maryland	0.91
Massachusetts	0.90
Michigan	0.90
Missouri	0.88
North Carolina	0.90
Oregon	0.90
Pennsylvania	0.90
South Carolina	0.91
Texas	0.91

Districts and Consortia	
Academy School Dist. #20, CO	0.89
Chicago Public Schools, IL	0.84
Delaware Science Coalition, DE	0.89
First in the World Consort., IL	0.91
Fremont/Lincoln/WestSide PS, NE	0.90
Guilford County, NC	0.90
Jersey City Public Schools, NJ	0.89
Miami-Dade County PS, FL	0.86
Michigan Invitational Group, MI	0.87
Montgomery County, MD	0.90
Naperville Sch. Dist. #203, IL	0.88
Project SMART Consortium, OH	0.89
Rochester City Sch. Dist., NY	0.86
SW Math/Sci. Collaborative, PA	0.89

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

[1] For each country and jurisdiction, the reliability coefficient is the median KR-20 reliability across the eight test booklets.

Data Processing

To ensure the availability of comparable, high-quality data for analysis, TIMSS took rigorous quality control steps to create the international database.[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. In the United States, the creation of the data files for both the Benchmarking participants and the U.S. national TIMSS effort was the

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. Students from the Benchmarking samples were assigned scores on the overall mathematics scale as well as in each of the five mathematics content areas using the same item parameters and estimation procedures as for TIMSS internationally.

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 performance based on samples of students, rather than the values that could be calculated if every student in every country or Benchmarking jurisdiction 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 popu-

[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, K.M. O'Connor, and S.E. Stemler (eds.), *TIMSS 1999 Benchmarking Technical Report,* Chestnut Hill, MA: Boston College.

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

[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, K.M. O'Connor, and S.E. Stemler (eds.), *TIMSS 1999 Benchmarking Technical Report*, Chestnut Hill, MA: Boston College.

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]

Mathematics Curriculum Questionnaire

In an effort to collect information about the content of the intended curriculum in mathematics, TIMSS asked National Research Coordinators and Coordinators from the Benchmarking jurisdictions to complete a questionnaire about the structure, organization, and content coverage of their curricula. Coordinators 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 Coordinators found it necessary to expand on or qualify the topic description to describe their situation accurately. The country-specific adaptations to the mathematics curriculum questionnaire are presented in Exhibit A.9. No adaptations to the list of topics were necessary for the U.S. national version, nor were any adaptations made by any Benchmarking participants.

[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, K.M. O'Connor, and S.E. Stemler (eds.), *TIMSS 1999 Benchmarking 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.

	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 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 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 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 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.