Science Performance in the United States and Internationally

The TIMSS science assessment

Like the TIMSS mathematics assessment, the TIMSS science assessment is designed along two dimensions: the science topics or content that students are expected to learn and the cognitive skills students are expected to have developed. The content domains covered at grade four are *life science*, *physical science*, and *Earth science* (see table 10). At grade eight, the content domains are *biology*, *chemistry*, *physics*, and *Earth science*. The cognitive domains in each grade are *knowing*, *applying*, and *reasoning*. Example items from the TIMSS science assessment are included in appendix B (see items B8 through B14).

The proportion of items devoted to a domain, and therefore the contribution of the domain to the overall science scale score, differs somewhat across grades. For example, at grade four in 2007, 37 percent of the TIMSS science assessment focused on the *physical science* domain, while at grade eight, 46 percent of the assessment focused on the analogous *chemistry* and *physics* domains. The proportion of items devoted to each cognitive domain is similar across grades.

Also, within a content or cognitive domain, the makeup of items, in terms of difficulty and form of knowledge and skills addressed, differs across grade levels to reflect the nature, difficulty, and emphasis of the subject matter encountered in school at each grade. The *TIMSS 2007 Assessment Frameworks* (Mullis et al. 2005) provides a more detailed description of the content and cognitive domains assessed in TIMSS. The development and validation of the science cognitive domains is based on the same processes used in the development of the mathematics cognitive domains. Details of the development of the mathematics cognitive domains can be found in *IEA's TIMSS 2003 International Report on Achievement in the Mathematics Cognitive Domains: Findings From a Developmental Project* (Mullis, Martin, and Foy 2005).

TIMSS provides an overall science scale score as well as content and cognitive domain scores at each grade level. As with the mathematics scale, the TIMSS science scale is from 0 to 1,000, and the international mean score is set at 500, with an international standard deviation of 100. The scaling of data is conducted separately for each grade and each content domain. While the scales were created to each have a mean of 500 and a standard deviation of 100, the subject matter and the level of difficulty of items necessarily differ between the assessments at both grades. Therefore, direct comparisons between scores across grades should not be made. Comparability over time is established by linking the data from each assessment to the data from the assessment that preceded it. More information on how the TIMSS scale was created can be found in appendix A.

Average scores in 2007

The average science scores for both U.S. fourth- and eighth-graders were higher than the TIMSS scale average (table 11).

Table 10. Percentage of fourth- and eighth-grade TIMSS science assessment devoted to content and cognitive domains: 2007

Grade four		Grade eight	
Content domains	Percent of assessment	Content domains	Percent of assessment
Life science	43	Biology	36
Physical science	37	Chemistry	20
Earth science	21	Physics	26
		Earth science	19
Cognitive domains	**Percent of assessment**	**Cognitive domains**	**Percent of assessment**
Knowing	44	Knowing	39
Applying	36	Applying	40
Reasoning	20	Reasoning	21

NOTE: The content and cognitive domains are the foundation of the Trends in International Mathematics and Science Study (TIMSS) assessment. The content domains define the specific science subject matter covered by the assessment, and the cognitive domains define the sets of behaviors expected of students as they engage with the science content. Each science content domain has several topic areas. Each topic area is presented as a list of objectives covered in a majority of participating countries, at either grade four or grade eight. However, the cognitive domains of science are defined by the same three sets of expected behaviors—knowing, applying, and reasoning. Detail may not sum to totals because of rounding.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007.

Table 11. Average science scores of fourth- and eighth-grade students, by country: 2007

Grade four		Grade eight	
Country	Average score	Country	Average score
TIMSS scale average	500	TIMSS scale average	500
Singapore	587	Singapore	567
Chinese Taipei	557	Chinese Taipei	561
Hong Kong SAR[1]	554	Japan	554
Japan	548	Korea, Rep. of	553
Russian Federation	546	England[3]	542
Latvia[2]	542	Hungary	539
England	542	Czech Republic	539
United States[3,4]	**539**	Slovenia	538
Hungary	536	Hong Kong SAR[1,3]	530
Italy	535	Russian Federation	530
Kazakhstan[2]	533	**United States[3,4]**	**520**
Germany	528	Lithuania[2]	519
Australia	527	Australia	515
Slovak Republic	526	Sweden	511
Austria	526	Scotland[3]	496
Sweden	525	Italy	495
Netherlands[5]	523	Armenia	488
Slovenia	518	Norway	487
Denmark[3]	517	Ukraine	485
Czech Republic	515	Jordan	482
Lithuania[2]	514	Malaysia	471
New Zealand	504	Thailand	471
Scotland[3]	500	Serbia[2,4]	470
Armenia	484	Bulgaria[7]	470
Norway	477	Israel[7]	468
Ukraine	474	Bahrain	467
Iran, Islamic Rep. of	436	Bosnia and Herzegovina	466
Georgia[2]	418	Romania	462
Colombia	400	Iran, Islamic Rep. of	459
El Salvador	390	Malta	457
Algeria	354	Turkey	454
Kuwait[6]	348	Syrian Arab Republic	452
Tunisia	318	Cyprus	452
Morocco	297	Tunisia	445
Qatar	294	Indonesia	427
Yemen	197	Oman	423
		Georgia[2]	421
		Kuwait[6]	418
		Colombia	417
		Lebanon	414
		Egypt	408
		Algeria	408
		Palestinian Nat'l Auth.	404
		Saudi Arabia	403
		El Salvador	387
		Botswana	355
		Qatar	319
		Ghana	303

In 2007, the average score of U.S. fourth-graders was 539 and the average score of U.S. eighth-graders was 520, compared to the TIMSS scale average of 500 at each grade level.

At grade four, the average U.S. science score was higher than those in 25 of the 35 other countries, lower than the average scores in 4 countries (all of them in Asia), and not measurably different from the average scores of students in the remaining 6 countries.

At grade eight, the average U.S. science score was higher than those in 35 of the 47 other countries, lower than in 9 countries (all located in Asia or Europe), and not measurably different from the average scores in the other 3 countries.

■ Average score is higher than the U.S. average score ($p < .05$)
□ Average score is not measurably different from the U.S. average score ($p < .05$)
▒ Average score is lower than the U.S. average score ($p < .05$)

[1]Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.
[2]National Target Population does not include all of the International Target Population defined by the Trends in International Mathematics and Science Study (TIMSS) (see appendix A).
[3]Met guidelines for sample participation rates only after substitute schools were included (see appendix A).
[4]National Defined Population covers 90 percent to 95 percent of National Target Population (see appendix A).
[5]Nearly satisfied guidelines for sample participation rates only after substitute schools were included (see appendix A).
[6]Kuwait tested the same cohort of students as other countries, but later in 2007, at the beginning of the next school year.
[7]National Defined Population covers less than 90 percent of National Target Population (but at least 77 percent, see appendix A).
NOTE: Countries are ordered by 2007 average score. The tests for significance take into account the standard error for the reported difference. Thus, a small difference between the United States and one country may be significant while a large difference between the United States and another country may not be significant. The standard errors of the estimates are shown in tables E-20 and E-21 available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009001.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007.

Trends in scores since 1995

At grade four, 16 countries, including the United States, participated in both the first TIMSS in 1995 and the most recent TIMSS in 2007 and therefore can be compared over a 12-year period. Comparing 2007 with 1995, 7 of the 16 countries showed improvement in average science scores, 5 countries showed declines, and 4 countries, including the United States, had no measurable difference in average scores (table 12). In 2007, the U.S. fourth-grade average science score was 539, compared with 542 in 1995.

Table 12. Trends in average science scores of fourth- and eighth-grade students, by country: 1995 to 2007

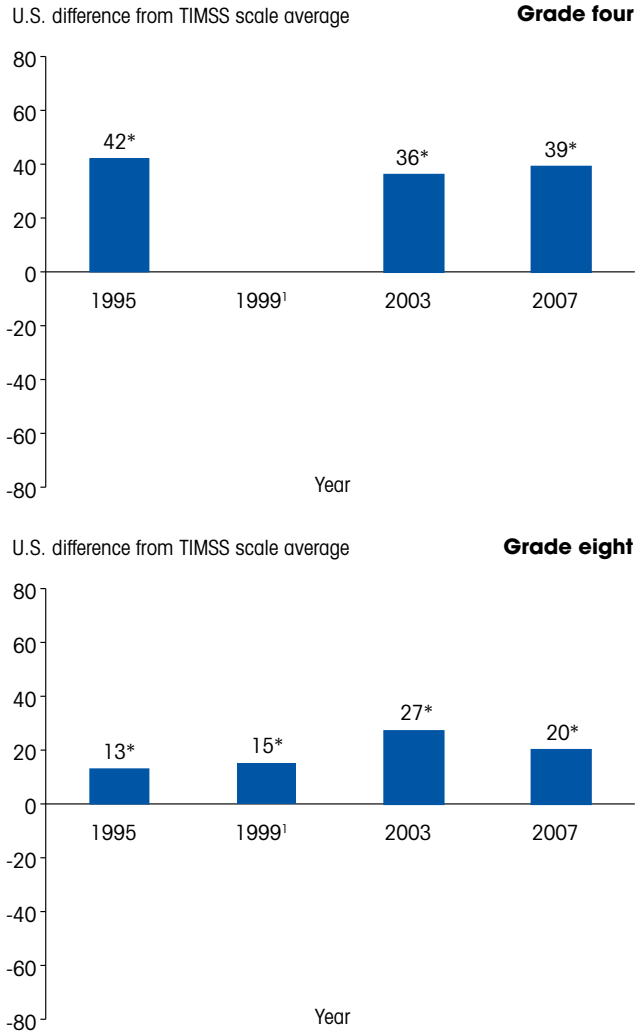
Grade four				Grade eight			
	Average score		Difference[1]		Average score		Difference[1]
Country	1995	2007	2007–1995	Country	1995	2007	2007–1995
Singapore	523	587	63*	Lithuania[2]	464	519	55*
Latvia[2]	486	542	56*	Colombia	365	417	52*
Iran, Islamic Rep. of	380	436	55*	Slovenia	514	538	24*
Slovenia	464	518	54*	Hong Kong SAR[3,4]	510	530	20*
Hong Kong SAR[3]	508	554	46*	England[4]	533	542	8
Hungary	508	536	28*	**United States[4,5]**	**513**	**520**	**7**
England	528	542	14*	Korea, Rep. of	546	553	7*
Australia	521	527	6	Russian Federation	523	530	7
New Zealand	505	504	-1	Hungary	537	539	2
United States[4,5]	**542**	**539**	**-3**	Australia	514	515	1
Japan	553	548	-5*	Cyprus	452	452	#
Netherlands[6]	530	523	-7	Japan	554	554	-1
Austria	538	526	-12*	Iran, Islamic Rep. of	463	459	-4
Scotland	514	500	-14*	Scotland[4]	501	496	-5
Czech Republic	532	515	-17*	Romania	471	462	-9
Norway	504	477	-27*	Singapore	580	567	-13
				Czech Republic	555	539	-16*
				Norway	514	487	-28*
				Sweden	553	511	-42*

■ Country difference in average scores between 1995 and 2007 is greater than analogous U.S. difference ($p < .05$)
□ Country difference in average scores between 1995 and 2007 is not measurably different from analogous U.S. difference ($p < .05$)
▒ Country difference in average scores between 1995 and 2007 is less than analogous U.S. difference ($p < .05$)
Rounds to zero.
*$p < .05$. Within-country difference between 1995 and 2007 average scores is significant.
[1]Difference calculated by subtracting 1995 from 2007 estimate using unrounded numbers.
[2]In 2007, National Target Population did not include all of the International Target Population defined by the Trends in International Mathematics and Science Study (TIMSS) (see appendix A).
[3]Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.
[4]In 2007, met guidelines for sample participation rates only after substitute schools were included (see appendix A).
[5]In 2007, National Defined Population covered 90 percent to 95 percent of National Target Population (see appendix A).
[6]In 2007, nearly satisfied guidelines for sample participation rates only after substitute schools were included (see appendix A).
NOTE: Bulgaria collected data in 1995 and 2007, but due to a structural change in its education system, comparable science data from 1995 are not available. Countries are ordered by the difference between 1995 and 2007 overall average scores. All countries met international sampling and other guidelines in 2007, except as noted. Data are not shown for some countries, because comparable data from previous cycles are not available. The tests for significance take into account the standard error for the reported difference. Thus, a small difference between the United States and one country may be significant while a large difference between the United States and another country may not be significant. Detail may not sum to totals because of rounding. The standard errors of the estimates are shown in tables E-20 and E-21 available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009001.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 1995 and 2007.

At grade eight, 19 countries, including the United States, participated in TIMSS in both 1995 and 2007. Five countries had higher average science scores in 2007 than in 1995, 3 countries showed declines in their average scores, and 11 countries, including the United States, had no measurable difference between average scores in 1995 and 2007. The U.S. eighth-grade average science score was 520, compared with 513 in 1995.

Figure 15 shows the difference between the average U.S. science scores and the TIMSS scale average at grades four and eight for each of the TIMSS administrations. The average size of difference in science scores between the U.S. fourth-graders and the TIMSS scale average shows no significant change across the data collection years, from 36 to 42 scale score points above the TIMSS scale average. Similarly, at grade eight, there has been no measurable change in the size of the difference, on average, across the data collection years.

Figure 15. Difference between average science scores of U.S. fourth- and eighth-grade students and the TIMSS scale average: 1995, 1999, 2003, and 2007



*p < .05. Difference between U.S. average and Trends in International Mathematics and Science Study (TIMSS) scale average is statistically significant.
[1]No fourth-grade assessment was conducted in 1999.
NOTE: The United States met guidelines for sample participation rates only after substitute schools were included. The National Defined Population covers 90 percent to 95 percent of National Target Population (see appendix A). Difference calculated by subtracting the TIMSS scale average (500) from the U.S. average science score. The standard errors of the estimates are shown in table E-40 available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009001.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 1995, 1999, 2003 and 2007.

Content and cognitive domain scores in 2007

As in mathematics, TIMSS also provides scores for science content and cognitive domains (see table 13 for a description of the science cognitive domains). U.S. fourth-graders scored higher than the TIMSS scale average across the science content domains in 2007 (table 14). U.S. fourth-graders' average scores in *life science*, *physical science*, and *Earth science* were between 33 and 40 scale score points above the TIMSS scale average of 500 in each content domain.

U.S. fourth-graders outperformed their peers in 25 countries in the *life science* domain, 24 countries in the *physical science* domain, and 21 countries in the *Earth science* domain. They were outperformed by their peers in 3 countries in the *life science* and *Earth science* domains, and 7 countries in the *physical science* domain.

U.S. fourth-graders' average scores in the cognitive domains of *knowing*, *applying*, and *reasoning* were, on average, between 33 and 41 scale score points higher than the TIMSS scale average of 500. U.S. fourth-graders outperformed students in 22 to 26 countries across the three cognitive domains. U.S. fourth-graders were outperformed by their peers in 1 country in the *knowing* domain, and 5 countries in the *applying* and *reasoning* domains.

At the eighth-grade level, U.S. students scored higher than the TIMSS scale average in three of the four science content domains and the three cognitive domains in 2007 (table 15). U.S. eighth-graders' average score in *biology*, *chemistry*, and *Earth science* was, on average, 10 to 30 scale score points above the TIMSS scale score average of 500. On the other hand, U.S. eighth-graders' average score in the *physics* domain was not measurably different from the TIMSS scale score average.

U.S. eighth-graders outperformed students in 36 countries in the *biology* and *Earth science* domains, 35 countries in the *chemistry* domain, and 32 countries in the *physics* domain. They were outperformed by their peers in 5 countries in the *biology* and *Earth science* domains, 9 countries in the *chemistry* domain, and 10 countries in the *physics* domain.

Table 13. Description of TIMSS science cognitive domains: 2007

Cognitive Domain	Description
Knowing	*Knowing* addresses the facts, information, concepts, tools, and procedures that students need to know to function scientifically. The key skills of this cognitive domain include making or identifying accurate statements about science facts, relationships, processes, and concepts; identifying the characteristics or properties of specific organisms, materials, and processes; providing or identifying definitions of scientific terms; recognizing and using scientific vocabulary, symbols, abbreviations, units, and scales in relevant contexts; describing organisms, physical materials, and science processes that demonstrate knowledge of properties, structure, function, and relationships; supporting or clarifying statements of facts or concepts with appropriate examples; identifying or providing specific examples to illustrate knowledge of general concepts; and demonstrating knowledge of the use of scientific apparatus, tools, equipment, procedures, measurement devices, and scales.
Applying	*Applying* focuses on students' ability to apply knowledge and conceptual understanding to solve problems or answer questions. The key skills of this cognitive domain include identifying or describing similarities and differences between groups of organisms, materials, or processes; distinguishing, classifying, or ordering individual objects, materials, organisms, and processes based on given characteristics and properties; using a diagram or model to demonstrate understanding of a science concept, structure, relationship, process, or biological or physical system or cycle; relating knowledge of an underlying biological or physical concept to an observed or inferred property, behavior, or use of objects, organisms, or materials; interpreting relevant textual, tabular, or graphical information in light of a science concept or principle; identifying or using a science relationship, equation, or formula to find a quantitative or qualitative solution involving the direct application or demonstration of a concept; providing or identifying an explanation for an observation or natural phenomena, demonstrating understanding of the underlying science concept, principle, law, or theory.
Reasoning	*Reasoning* goes beyond the cognitive processes involved in solving routine problems to include more complex tasks. The key skills of this cognitive domain include analyzing problems to determine the relevant relationships, concepts, and problem-solving steps; developing and explaining problem-solving strategies; providing solutions to problems that require consideration of a number of different factors or related concepts; making associations or connections between concepts in different areas of science; demonstrating understanding of unified concepts and themes across the domains of science; integrating mathematical concepts or procedures in the solutions to science problems; combining knowledge of science concepts with information from experience or observation to formulate questions that can be answered by investigation; formulating hypotheses as testable assumptions using knowledge from observation or analysis of scientific information and conceptual understanding; making predictions about the effects of changes in biological or physical conditions in light of evidence and scientific understanding; designing or planning investigations appropriate for answering scientific questions or testing hypotheses; detecting patterns in data; describing or summarizing data trends; interpolating or extrapolating from data or given information; making valid inferences based on evidence; drawing appropriate conclusions; demonstrating understanding of cause and effect; making general conclusions that go beyond the experimental or given conditions; applying conclusions to new situations; determining general formulas for expressing physical relationships; evaluating the impact of science and technology on biological and physical systems; evaluating alternative explanations and problem-solving strategies; evaluating the validity of conclusions through examination of the available evidence; and constructing arguments to support the reasonableness of solutions to problems.

NOTE: The descriptions of the cognitive domains are the same for grades four and eight.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007.

Table 14. Average science content and cognitive domain scores of fourth-grade students, by country: 2007

	Content domain			Cognitive domain		
Country	Life science	Physical science	Earth science	Knowing	Applying	Reasoning
TIMSS scale average	500	500	500	500	500	500
Singapore	582	585	554	587	579	568
Chinese Taipei	541	559	553	536	556	571
Hong Kong SAR[1]	532	558	560	546	549	561
Japan	530	564	529	528	542	567
Russian Federation	539	547	536	542	546	542
Latvia[2]	535	544	536	540	535	551
England	532	543	538	543	536	537
United States[3,4]	**540**	**534**	**533**	**541**	**533**	**535**
Hungary	548	529	517	540	531	529
Italy	549	521	526	530	539	526
Kazakhstan[2]	528	528	534	534	536	519
Germany	529	524	524	527	526	525
Australia	528	522	534	529	523	530
Slovak Republic	532	513	530	527	527	513
Austria	526	514	532	529	526	513
Sweden	531	508	535	526	521	527
Netherlands[5]	536	503	524	518	525	525
Slovenia	511	530	517	511	525	527
Denmark[3]	527	502	522	516	515	525
Czech Republic	520	511	518	520	516	510
Lithuania[2]	516	514	511	511	515	524
New Zealand	506	498	515	511	500	505
Scotland[3]	504	499	508	511	494	501
Armenia	489	492	479	486	487	484
Norway	487	469	497	485	478	480
Ukraine	482	475	474	476	477	478
Iran, Islamic Rep. of	442	454	433	437	451	436
Georgia[2]	427	414	432	434	424	388
Colombia	408	411	401	409	404	409
El Salvador	410	392	393	410	393	376
Algeria	351	377	365	350	379	357
Kuwait[6]	353	345	363	360	338	331
Tunisia	323	340	325	316	329	349
Morocco	292	324	293	291	311	318
Qatar	291	303	305	304	283	293
Yemen	—	—	—	—	—	—

■ Average score is higher than the U.S. average score ($p < .05$)
□ Average score is not measurably different from the U.S. average score ($p < .05$)
■ Average score is lower than the U.S. average score ($p < .05$)
— Not available. Average achievement could not be accurately estimated.
[1]Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.
[2]National Target Population does not include all of the International Target Population defined by the Trends in International Mathematics and Science Study (TIMSS) (see appendix A).
[3]Met guidelines for sample participation rates only after substitute schools were included (see appendix A).
[4]National Defined Population covers 90 percent to 95 percent of National Target Population (see appendix A).
[5]Nearly satisfied guidelines for sample participation rates only after substitute schools were included (see appendix A).
[6]Kuwait tested the same cohort of students as other countries, but later in 2007, at the beginning of the next school year.
NOTE: Countries are ordered by 2007 overall science average scale score. The tests for significance take into account the standard error for the reported difference. Thus, a small difference between averages for the United States and one country may be significant while a large difference between averages for the United States and another country may not be significant. The standard errors of the estimates are shown in table E-22 available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009001.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007.

Table 15. Average science content and cognitive domain scores of eighth-grade students, by country: 2007

	Content domain				Cognitive domain		
Country	Biology	Chemistry	Physics	Earth Science	Knowing	Applying	Reasoning
TIMSS scale average	500	500	500	500	500	500	500
Singapore	564	560	575	541	554	567	564
Chinese Taipei	549	573	554	545	565	560	541
Japan	553	551	558	533	534	555	560
Korea, Rep. of	548	536	571	538	543	547	558
England[1]	541	534	545	529	530	538	547
Hungary	534	536	541	531	524	549	530
Czech Republic	531	535	537	534	533	539	534
Slovenia	530	539	524	542	533	533	538
Hong Kong SAR[1,2]	527	517	528	532	532	522	533
Russian Federation	525	535	519	525	534	527	520
United States[1,3]	**530**	**510**	**503**	**525**	**512**	**516**	**529**
Lithuania[4]	527	507	505	515	513	512	527
Australia	518	505	508	519	501	510	530
Sweden	515	499	506	510	505	509	517
Scotland[1]	495	497	494	498	480	495	511
Italy	502	481	489	503	494	498	493
Armenia	490	478	503	475	493	502	459
Norway	487	483	475	502	486	486	491
Ukraine	477	490	492	482	477	488	488
Jordan	478	491	479	484	491	485	471
Malaysia	469	479	484	463	458	473	487
Thailand	478	462	458	488	473	472	473
Serbia[3,4]	474	467	467	466	485	469	455
Bulgaria[5]	467	472	466	480	489	471	448
Israel[5]	472	467	472	462	456	472	481
Bahrain	473	468	466	465	469	468	469
Bosnia and Herzegovina	464	468	463	469	486	463	452
Romania	459	463	458	471	451	470	460
Iran, Islamic Rep. of	449	463	470	476	468	454	462
Malta	453	461	470	456	436	462	473
Turkey	462	435	445	466	462	450	462
Syrian Arab Republic	459	450	447	448	474	445	440
Cyprus	447	452	458	457	438	456	460
Tunisia	452	458	432	447	441	445	458
Indonesia	428	421	432	442	426	425	438
Oman	414	416	443	439	428	423	428
Georgia[4]	423	418	416	425	440	422	394
Kuwait[6]	419	418	438	410	430	417	411
Colombia	434	420	407	407	418	417	428
Lebanon	405	447	431	389	403	422	420
Egypt	406	413	413	426	434	404	395
Algeria	411	414	397	413	409	410	414
Palestinian Nat'l Auth.	402	413	414	408	407	412	396
Saudi Arabia	407	390	408	423	417	403	395
El Salvador	398	377	380	400	394	388	384
Botswana	359	371	351	361	361	358	362
Qatar	318	322	347	312	325	322	—
Ghana	304	342	276	294	316	291	—

Average score is higher than the U.S. average score ($p < .05$)
Average score is not measurably different from the U.S. average score ($p < .05$)
Average score is lower than the U.S. average score ($p < .05$)
— Not available. Average achievement could not be accurately estimated.
[1]Met guidelines for sample participation rates only after substitute schools were included (see appendix A).
[2]Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.
[3]National Defined Population covers 90 percent to 95 percent of National Target Population (see appendix A).
[4]National Target Population does not include all of the International Target Population defined by the Trends in International Mathematics and Science Study (TIMSS) (see appendix A).
[5]National Defined Population covers less than 90 percent of National Target Population (but at least 77 percent, see appendix A).
[6]Kuwait tested the same cohort of students as other countries, but later in 2007, at the beginning of the next school year.
NOTE: Countries are ordered by 2007 overall science average scale score. The tests for significance take into account the standard error for the reported difference. Thus, a small difference between averages for the United States and one country may be significant while a large difference between averages for the United States and another country may not be significant. The standard errors of the estimates are shown in table E-23 available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009001.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007.

In the three cognitive domains, the average U.S. score at eighth grade was higher than the TIMSS scale average. In 2007, U.S. eighth-graders' average scores in the *knowing*, *applying*, and *reasoning* domains were between 12 and 29 scale score points higher than the TIMSS scale average of 500. U.S. eighth-graders outperformed students in 33 to 35 countries across the three cognitive domains. U.S. eighth-graders were outperformed by their peers in 6 to 10 countries across the three cognitive domains.

Performance on the TIMSS international benchmarks

The TIMSS international benchmarks distinguish four levels of student achievement: advanced, high, intermediate, and low, and provide a way to understand how students' proficiency in science varies along the TIMSS scale (table 16). The descriptions of the benchmarks differ between the two grade levels, as the science skills and knowledge needed to respond to the assessment items reflect the nature, difficulty, and emphasis at each grade.

Table 16. Description of TIMSS international science benchmarks, by grade: 2007

Benchmark (score cutpoint)	Grade four
Advanced (625)	*Students can apply knowledge and understanding of scientific processes and relationships in beginning scientific inquiry.* Students communicate their understanding of characteristics and life processes of organisms as well as of factors relating to human health. They demonstrate understanding of relationships among various physical properties of common materials and have some practical knowledge of electricity. Students demonstrate some understanding of the solar system and Earth's physical features and processes. They show a developing ability to interpret the results of investigations and draw conclusions as well as a beginning ability to evaluate and support an argument.
High (550)	*Students can apply knowledge and understanding to explain everyday phenomena.* Students demonstrate some understanding of plant and animal structure, life processes, and the environment and some knowledge of properties of matter and physical phenomena. They show some knowledge of the solar system, and of Earth's structure, processes, and resources. Students demonstrate beginning scientific inquiry knowledge and skills, and provide brief descriptive responses combining knowledge of science concepts with information from everyday experience of physical and life processes.
Intermediate (475)	*Students can apply basic knowledge and understanding to practical situations in the sciences.* Students recognize some basic information related to characteristics of living things and their interaction with the environment, and show some understanding of human biology and health. They also show some understanding of familiar physical phenomena. Students know some basic facts about the solar system and have a developing understanding of Earth's resources. They demonstrate some ability to interpret information in pictorial diagrams and apply factual knowledge to practical situations.
Low (400)	*Students have some elementary knowledge of life science and physical science.* Students can demonstrate knowledge of some simple facts related to human health and the behavioral and physical characteristics of animals. They recognize some properties of matter, and demonstrate a beginning understanding of forces. Students interpret labeled pictures and simple diagrams, complete simple tables, and provide short written responses to questions requiring factual information.
	Grade eight
Advanced (625)	*Students can demonstrate a grasp of some complex and abstract concepts in biology, chemistry, physics, and Earth science.* They have an understanding of the complexity of living organisms and how they relate to their environment. They show understanding of the properties of magnets, sound, and light, as well as demonstrating understanding the structure of matter and physical and chemical properties and changes. Students apply knowledge of the solar system and of Earth's features and processes, and apply understanding of major environmental issues. They understand some fundamentals of scientific investigation and can apply basic physical principles to solve some quantitative problems. They can provide written explanations to communicate scientific knowledge.
High (550)	*Students can demonstrate conceptual understanding of some science cycles, systems, and principles.* They have some understanding of biological concepts including cell processes, human biology and health, and the interrelationship of plants and animals in ecosystems. They apply knowledge to situations related to light and sound, demonstrate elementary knowledge of heat and forces, and show some evidence of understanding the structure of matter, and chemical and physical properties and changes. They demonstrate some understanding of the solar system, Earth's processes and resources, and some basic understanding of major environmental issues. Students demonstrate some scientific inquiry skills. They combine information to draw conclusions, interpret tabular and graphical information, and provide short explanations conveying scientific knowledge.
Intermediate (475)	*Students can recognize and communicate basic scientific knowledge across a range of topics.* They demonstrate some understanding of characteristics of animals, food webs, and the effect of population changes in ecosystems. They are acquainted with some aspects of sound and force and have elementary knowledge of chemical change. They demonstrate elementary knowledge of the solar system, Earth's processes, and resources and the environment. Students extract information from tables and interpret pictorial diagrams. They can apply knowledge to practical situations and communicate their knowledge through brief descriptive responses.
Low (400)	*Students can recognize some basic facts from the life and physical sciences.* They have some knowledge of the human body, and demonstrate some familiarity with everyday physical phenomena. Students can interpret pictorial diagrams and apply knowledge of simple physical concepts to practical situations.

NOTE: Score cutpoints for the international benchmarks are determined through scale anchoring. Scale anchoring involves selecting benchmarks (scale points) on the achievement scales to be described in terms of student performance, and then identifying items that students scoring at the anchor points can answer correctly. The score cutpoints are set at equal intervals along the achievement scales. The score cutpoints were selected to be as close as possible to the standard percentile cutpoints (i.e., 90th, 75th, 50th, and 25th percentiles). More information on the setting of the score cutpoints can be found in appendix A and Martin et al. (2008).
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007.

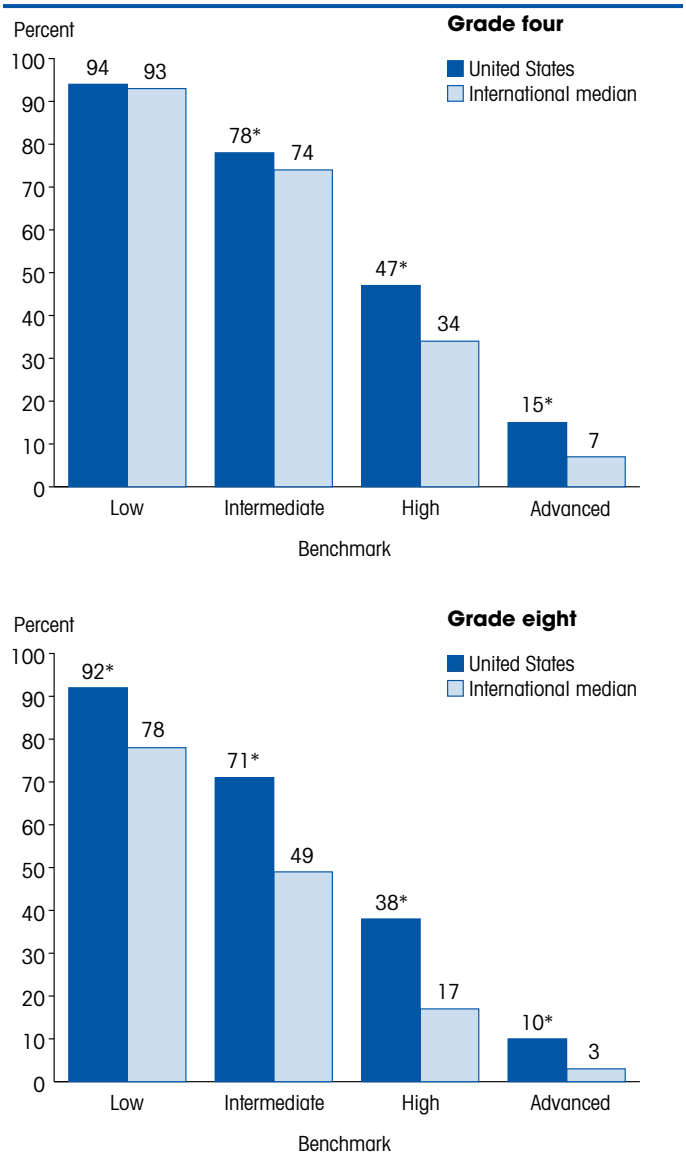
In 2007, there were higher percentages of U.S. fourth-graders performing at or above three of the four TIMSS international benchmarks than the international median percentage (figure 16).[14] For example, 15 percent of U.S. fourth-graders performed at or above the advanced benchmark (625) in science compared to the international median of 7 percent. These students demonstrated an ability to apply their knowledge and understanding of scientific processes and relationships in beginning scientific inquiry (see description in table 16). At the other end of the scale, 94 percent of U.S. fourth-graders performed at or above the low benchmark (400) which was not measurably different from the international median of 93 percent. These students showed at least some elementary knowledge of life science and physical science.

At the eighth grade, there were higher percentages of U.S. students performing at or above each of the four TIMSS international science benchmarks than the international median (figure 16). For example, 10 percent of U.S. eighth-graders performed at or above the advanced benchmark (625) compared to the international median of 3 percent. These students demonstrated a grasp of some complex and abstract concepts in biology, chemistry, physics, and Earth science (see description in table 14). At the other end of the scale, 92 percent of U.S. eighth-graders performed at or above the low benchmark (400) compared with the international median of 78 percent. These students recognized some basic facts from the life science and physical science.

At grade four, two countries had higher percentages of students performing at or above the advanced international science benchmark than the United States (figure 17). Fourth-graders in these two countries, Singapore and Chinese Taipei, were also found to outperform U.S. fourth-graders, on average, on the overall science scale (see table 11). At grade eight, six countries had higher percentages of students performing at or above the advanced science benchmark than the United States (figure 17). These six countries also had higher average overall eighth-grade science scores than the United States (see table 11).

In comparison with earlier data collections, a lower percentage of U.S. fourth-graders performed at or above the advanced benchmark in 2007 than in 1995 (15 v. 19 percent; data not shown). There were no measurable differences in the percentage of U.S. fourth-graders performing at or above the high, intermediate, or low international science benchmarks between 1995 and 2007 (high: 50 v. 47 percent; intermediate: 78 v. 78 percent; low: 92 v. 94 percent). At grade eight, there were fewer U.S. students performing at or above the advanced benchmark than in 1999 (10 v. 12 percent), but not between 1995 and 2007 (data not shown). On the other hand, there were more U.S. eighth-graders performing at or above the low science benchmark in 2007 than in 1995 (92 v. 87 percent). There was no measurable difference in the percentage of U.S. eighth-graders performing at or above the high or intermediate international benchmarks in 2007 than in 1995.

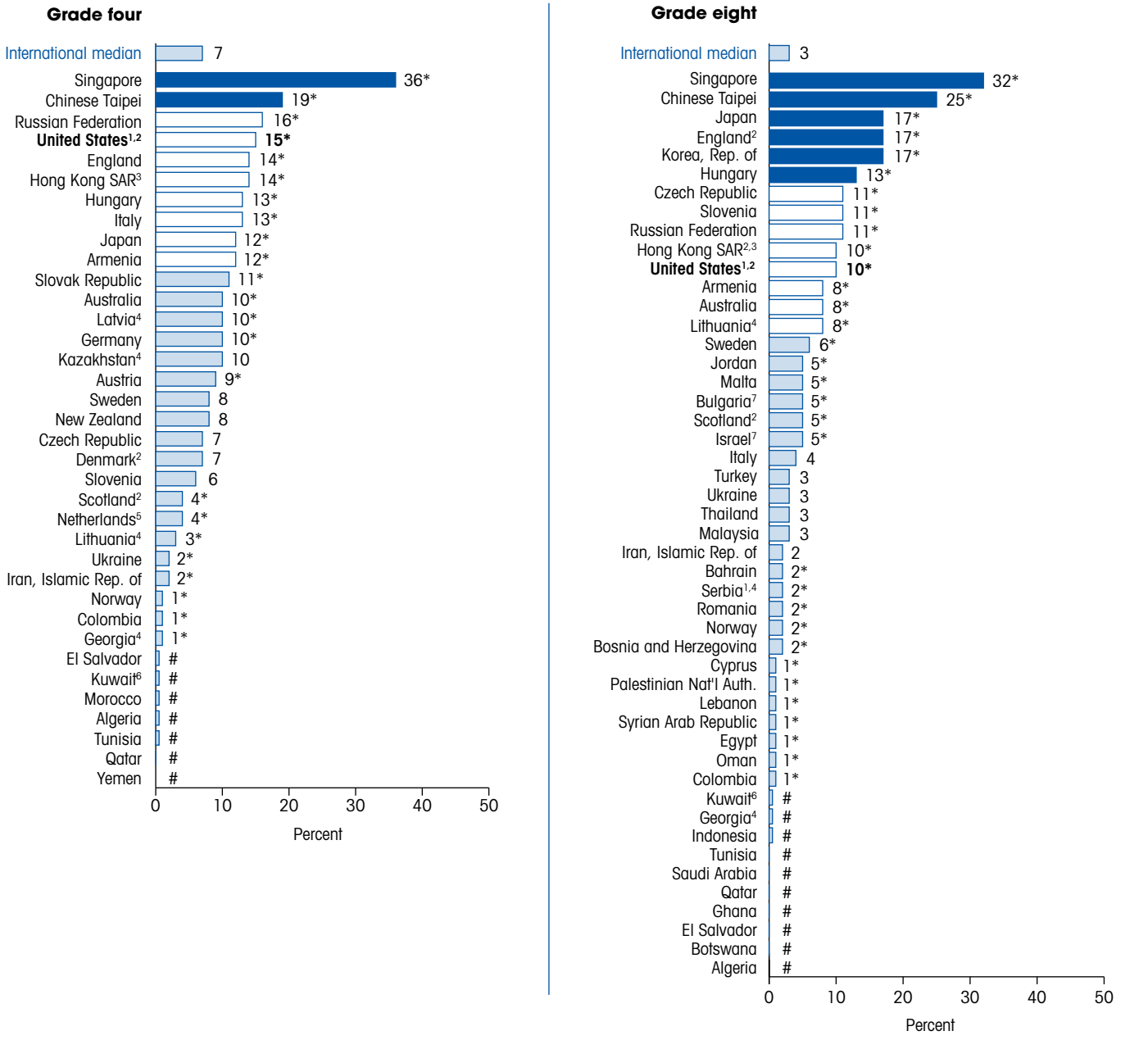
Figure 16. Percentage of U.S. fourth- and eighth-grade students who reached each TIMSS international science benchmark compared with the international median percentage: 2007



*p < .05. U.S. percentage is significantly different from the Trends in International Mathematics and Science (TIMSS) international median percentage.
NOTE: The United States met guidelines for sample participation rates only after substitute schools were included. The National Defined Population covered 90 percent to 95 percent of National Target Population (see appendix A). The TIMSS international median represents all participating TIMSS jurisdictions, including the United States. The international median represents the percentage at which half of the participating countries have that percentage of students at or above the median and half have that percentage of students below the median. The standard errors for the estimates are shown in table E-24 available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009001.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007.

[14]The international median at each benchmark represents the percentage at which half of the participating countries have that percentage of students at or above the median and half have that percentage of students below the median. For example, the low international benchmark median of 93 percent at grade four indicates that half of the countries have 93 percent or more of their students who met the low benchmark, and half have less than 93 percent of their students who met the low benchmark.

Figure 17. Percentage of fourth- and eighth-grade students who reached the TIMSS advanced international benchmark in science, by country: 2007



Percentage is higher than U.S. percentage ($p < .05$)
Percentage is not measurably different from U.S. percentage ($p < .05$)
Percentage is lower than U.S. percentage ($p < .05$)
Rounds to zero.
*$p < .05$. Percentage is significantly different from the international median percentage.
[1]National Defined Population covers 90 percent to 95 percent of National Target Population (see appendix A).
[2]Met guidelines for sample participation rates only after substitute schools were included (see appendix A).
[3]Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.
[4]National Target Population does not include all of the International Target Population defined by the Trends in International Mathematics and Science Study (TIMSS) (see appendix A).
[5]Nearly satisfied guidelines for sample participation rates only after substitute schools were included (see appendix A).
[6]Kuwait tested the same cohort of students as other countries, but later in 2007, at the beginning of the next school year.
[7]National Defined Population covers less than 90 percent of National Target Population (but at least 77 percent, see appendix A).
NOTE: The Trends in International Mathematics and Science Study (TIMSS) international median represents all participating TIMSS jurisdictions, including the United States. The international median represents the percentage at which half of the participating countries have that percentage of students at or above the median and half have that percentage of students below the median. The tests for significance take into account the standard error for the reported difference. Thus, a small difference between the United States and one country may be significant while a large difference between the United States and another country may not be significant. The standard errors for the estimates are shown in table E-42 available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009001.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007.

Performance within the United States

As with mathematics, the TIMSS science data were analyzed to investigate the performance of students grouped in four ways: the highest and lowest performing students; males and females; racial and ethnic groups; and public schools serving students with different low-income concentrations.

Scores of lower and higher performing students

To examine the science performance of each participating country's higher and lower performing students, cutpoint scores were calculated for students performing at or above the 90th percentile (the top 10 percent of students) and those performing at or below the 10th percentile (the bottom 10 percent of students). The 10th and 90th percentiles cutpoint scores were calculated for each country, rather than across all countries combined.

In 2007, the highest-performing U.S. fourth-graders (those performing at or above the 90th percentile) scored 643 or higher in science (table 17). This was higher than the 90th percentile score for fourth-graders in 27 countries and lower than 2 of the 35 other countries. Of the 4 countries that outperformed the United States, on average, in science at grade four (see table 11), 2 had higher 90th percentile cutpoint scores than the United States: Singapore and Chinese Taipei. Scores at the 90th percentile ranged between 379 (Yemen) and 701 (Singapore). The difference in scores between the highest-performing students in Singapore and the United States was 58 score points.

The lowest-performing U.S. fourth-graders in science (those performing at or below the 10th percentile) scored 427 or less in 2007 (table 17). The 10th percentile score for U.S. fourth-graders was higher than the 10th percentile score in 17 countries and lower than that in 7 countries: Singapore, Chinese Taipei, the Russian Federation, Hong Kong SAR, Japan, Latvia, and the Netherlands. The range in scores at the 10th percentile was between 20 (Yemen) and 466 (Hong Kong SAR). The difference in scores between the lowest-performing students in Hong Kong SAR and the United States was 39 score points.

Table 17. Science scores of fourth- and eighth-grade students defining 10th and 90th percentiles, by country: 2007

Grade four			Grade eight		
Country	**90th percentile**	**10th percentile**	**Country**	**90th percentile**	**10th percentile**
International average	586	359	International average	573	352
Singapore	701	464	Singapore	694	421
Chinese Taipei	653	457	Chinese Taipei	665	439
Russian Federation	646	443	England[1]	649	427
United States[1,2]	**643**	**427**	Japan	648	454
England	641	438	Korea, Rep. of	646	452
Armenia	640	336	Hungary	635	437
Hungary	637	425	Czech Republic	630	447
Hong Kong SAR[3]	637	466	Slovenia	628	442
Italy	636	429	Russian Federation	627	427
Japan	633	459	Hong Kong SAR[3]	625	419
Slovak Republic	627	416	**United States[1,2]**	**623**	**410**
Australia	626	423	Australia	617	410
Latvia[4]	625	454	Lithuania[4]	616	414
Kazakhstan[4]	623	433	Armenia	612	366
Germany	623	427	Sweden	608	405
Austria	620	423	Jordan	601	349
Sweden	617	429	Scotland[1]	597	388
New Zealand	614	382	Bulgaria[7]	595	330
Denmark[1]	610	417	Malta	595	298
Slovenia	610	416	Israel[7]	591	329
Czech Republic	610	416	Italy	590	393
Netherlands[5]	598	445	Ukraine	588	374
Lithuania[4]	595	428	Malaysia	581	357
Scotland[1]	593	400	Norway	578	389
Ukraine	576	364	Thailand	578	363
Norway	570	374	Turkey	577	336
Iran, Islamic Rep. of	558	304	Bahrain	575	351
Georgia[4]	524	306	Romania	572	345
Colombia	522	271	Serbia[2,4]	571	359
El Salvador	507	267	Iran, Islamic Rep. of	566	355
Kuwait[6]	505	182	Bosnia and Herzegovina	565	359
Tunisia	497	119	Cyprus	556	339
Algeria	483	220	Syrian Arab Republic	546	355
Morocco	465	139	Palestinian Nat'l Auth.	543	255
Qatar	464	121	Oman	541	293
Yemen	379	20	Lebanon	539	284
			Egypt	537	275
			Kuwait[6]	530	298
			Georgia[4]	527	309
			Tunisia	524	367
			Indonesia	520	330
			Colombia	514	319
			Saudi Arabia	503	300
			Algeria	488	327
			Qatar	480	146
			Botswana	478	220
			El Salvador	477	298
			Ghana	445	163

■ Percentile cutpoint score is higher than U.S. cutpoint score ($p < .05$)
□ Percentile cutpoint score is not measurably different from U.S. cutpoint score ($p < .05$)
▒ Percentile cutpoint score is lower than U.S. cutpoint score ($p < .05$)

[1]Met guidelines for sample participation rates only after substitute schools were included (see appendix A).
[2]National Defined Population covers 90 percent to 95 percent of National Target Population (see appendix A).
[3]Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.
[4]National Target Population does not include all of the International Target Population defined by the Trends in International Mathematics and Science Study (TIMSS) (see appendix A).
[5]Nearly satisfied guidelines for sample participation rates only after substitute schools were included (see appendix A).
[6]Kuwait tested the same cohort of students as other countries, but later in 2007, at the beginning of the next school year.
[7]National Defined Population covers less than 90 percent of National Target Population (but at least 77 percent, see appendix A).

NOTE: Countries are ordered based on the 90th percentile cutpoint for science scores. Cutpoints are calculated based on distribution of student scores within each country. The international average is the average of the cutpoint scores for all reported countries. The standard errors of the estimates are shown in tables E-25 and E-26 available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009001.

SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007.

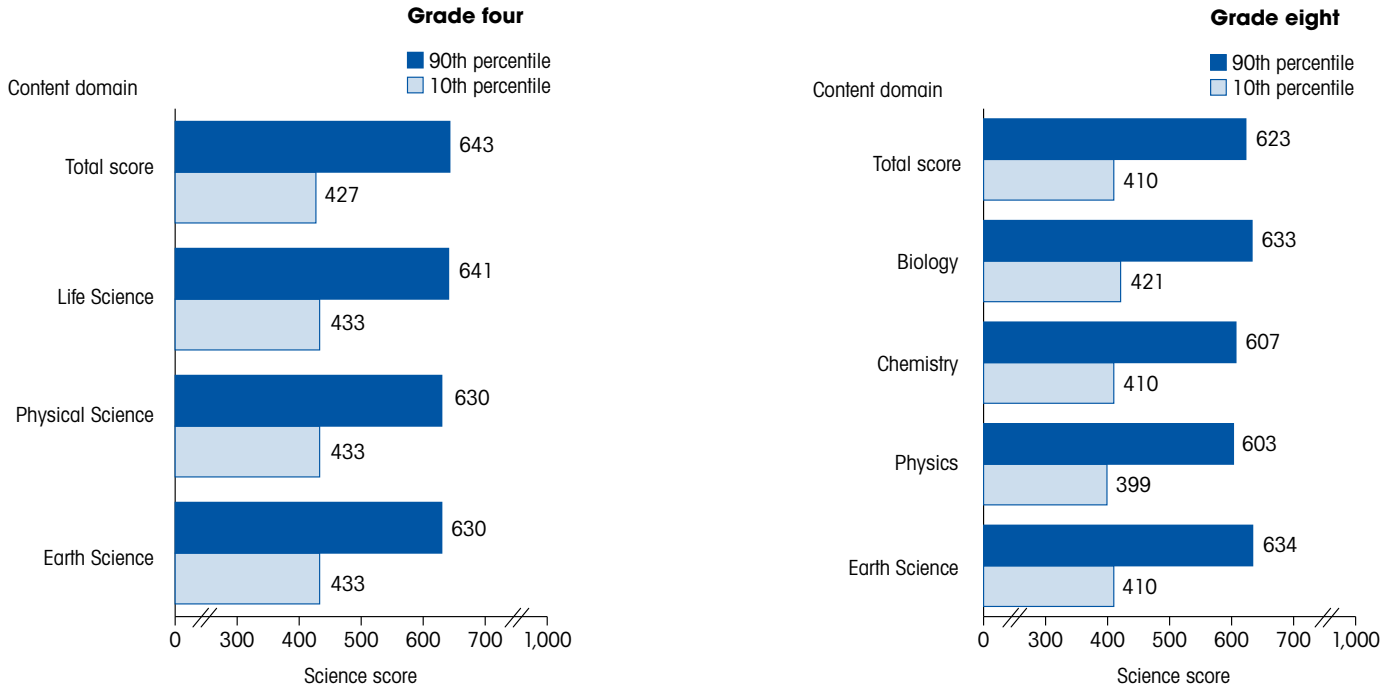
On the three science content domains at grade four in 2007, the highest-performing U.S. students (90th percentile or higher) scored 641 or higher on the *life science* domain and 630 or higher on both the *physical science* or *Earth science* domains (figure 18). The lowest-performing U.S. students (10th percentile or lower) scored 433 or lower on the *life science, physical science*, and *Earth science* domains.

At grade eight, the highest-performing U.S. students (90th percentile or higher) in science scored 623 or higher in 2007 (table 17). This was higher than the 90th percentile score in 34 countries and lower than in 6 countries: Singapore, Chinese Taipei, England, Japan, Korea, and Hungary. The range in 90th percentile scores was between 445 (Ghana) and 694 (Singapore). The difference in scores between the highest-performing students in Singapore and the United States was 71 score points.

At the other end of the scale, the lowest-performing U.S. eighth-graders (10th percentile or lower) scored 410 or lower in science in 2007 (table 17). The 10th percentile score for U.S. eighth-graders was higher than the 10th percentile score in 34 countries and lower than in 8 countries: Chinese Taipei, England, Japan, Korea, Hungary, the Czech Republic, Slovenia, and the Russian Federation. The range in 10th percentile scores was between 163 (Ghana) and 454 (Japan). The difference in scores between the lowest-performing students in Japan and the United States was 44 score points.

On the four science content domains at grade eight, the highest-performing U.S. eighth-graders (90th percentile or higher) scored 633 or higher on the *biology* domain, 607 or higher on the *chemistry* domain, 603 or higher on the *physics* domain, and 634 or higher on the *Earth science* domain (figure 18). The lowest-performing U.S. students (10th percentile or lower) scored 421 or lower on the *biology* domain, 410 or lower on the *chemistry* and *Earth science* domains, and 399 or lower on the *physics* domain in 2007.

Figure 18. Cutpoints at the 10th and 90th percentile for science content domain scores of U.S. fourth- and eighth-grade students: 2007



NOTE: The United States met guidelines for sample participation rates only after substitute schools were included. The National Defined Population covered 90 percent to 95 percent of National Target Population (see appendix A). The standard errors of the estimates are shown in table E-27 available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009001.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007.

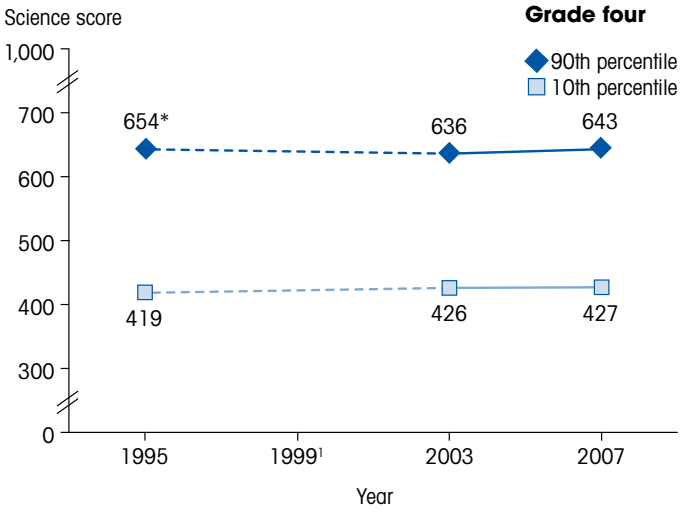
A comparison of 1995 and 2007 shows a decline in the 90th percentile cutpoint score for U.S. fourth graders in science, the point marking the top 10 percent of students (figure 19). In 2007, the 90th percentile score was 643, 11 score points lower than the analogous score of 654 in 1995. A comparison of the 10th percentile science scores for U.S. fourth-graders in 1995 and 2007 and 2003 and 2007 shows no measurable difference.

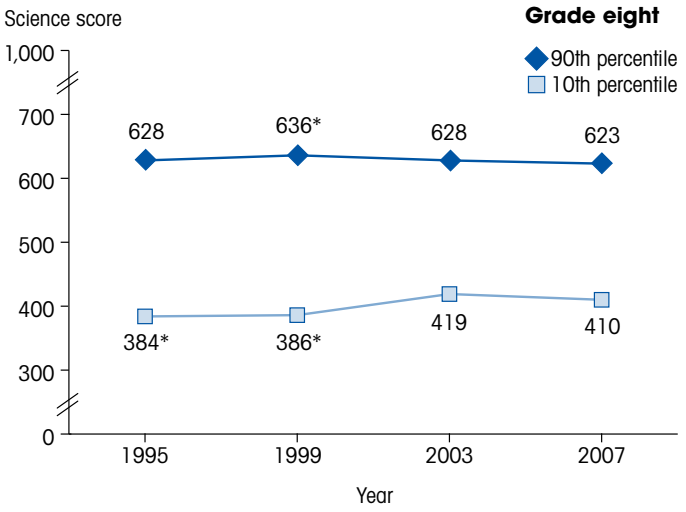
At grade eight, the data suggest a different picture. The 90th percentile cutpoint score in science showed no measurable differences in comparisons of 2007 to 1995 or 2003, but showed a decrease when the 2007 score was compared to the 1999 score (636 v. 623). The score identifying the lowest-performing U.S. eighth-graders in science was higher in 2007 than in 1995 (410 v. 384) and in 1999 (410 v. 386).

Average scores of male and female students

In 2007, U.S. fourth-grade males and females showed no measurable difference in their average science performance (figure 20). Fourteen of the 35 other countries participating at grade four showed a significant difference in average science scores of males and females: 8 countries in favor of males and 6 in favor of females. The largest differences were 64 score points in Kuwait (in favor of females) and 15 score points in Colombia (in favor of males).

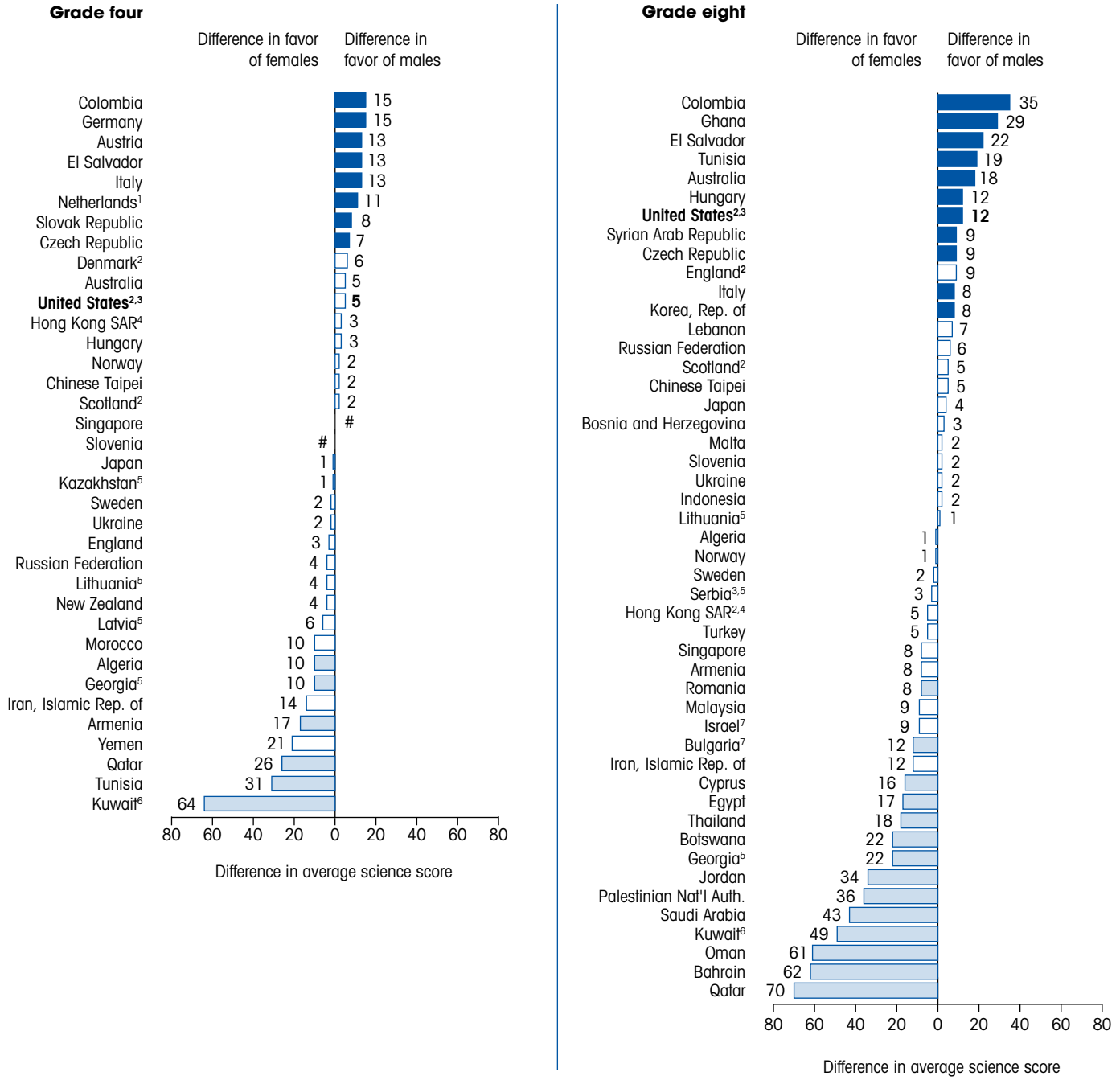
Figure 19. Trends in 10th and 90th percentile science scores of U.S. fourth- and eighth-grade students: 1995, 1999, 2003, and 2007





*p < .05. Percentile cutpoint score is significantly different from 2007 percentile cutpoint score.
[1]No fourth-grade assessment was conducted in 1999.
NOTE: In 2007, the United States met guidelines for sample participation rates only after substitute schools were included. The National Defined Population covered 90 percent to 95 percent of National Target Population (see appendix A). Cutpoints are calculated based on distribution of U.S. student scores. The standard errors of the estimates are shown in table E-28 available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009001.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 1995, 1999, 2003, and 2007.

Figure 20. Difference in average science scores of fourth- and eighth-grade students, by sex and country: 2007



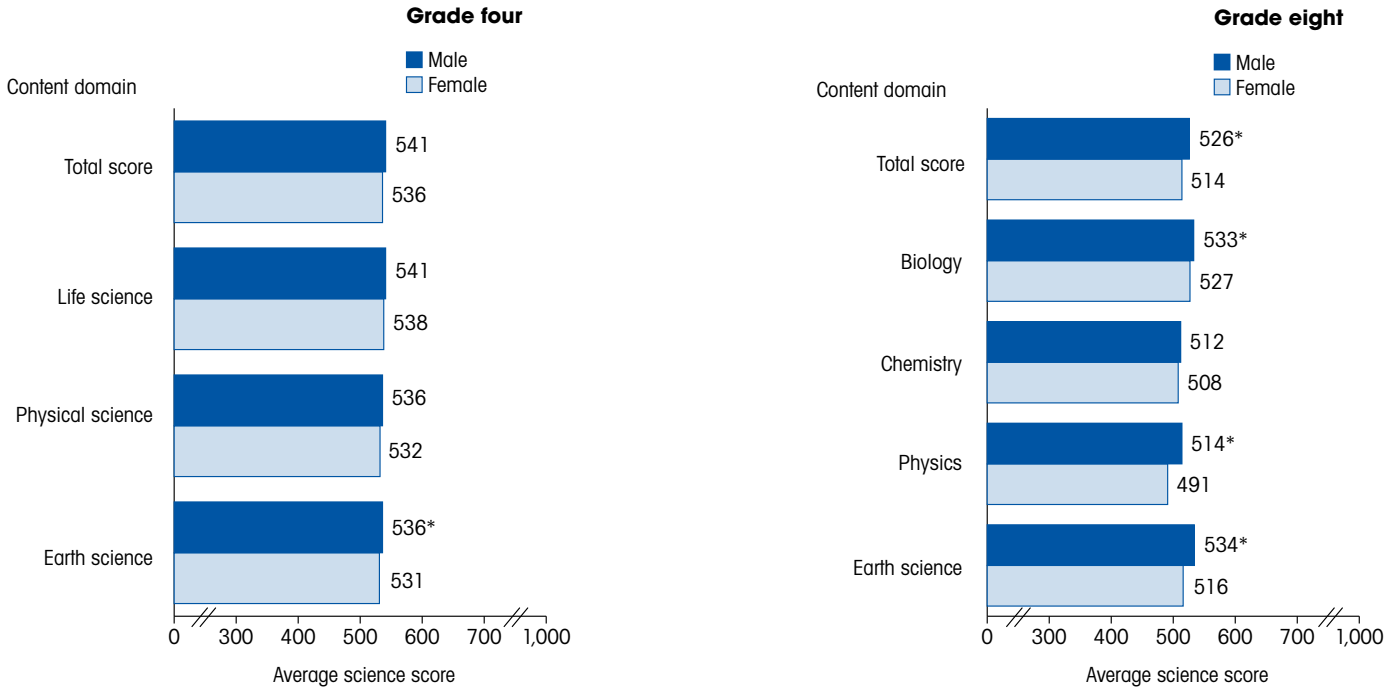
Male-female difference in average science scores favors males and is statistically significant ($p < .05$)
Male-female difference in average science scores is not measurably different ($p < .05$)
Male-female difference in average science scores favors females and is statistically significant ($p < .05$)
Rounds to zero.
[1]Nearly satisfied guidelines for sample participation rates only after substitute schools were included (see appendix A).
[2]Met guidelines for sample participation rates only after substitute schools were included (see appendix A).
[3]National Defined Population covers 90 percent to 95 percent of National Target Population (see appendix A).
[4]Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.
[5]National Target Population does not include all of the International Target Population defined by the Trends in International Mathematics and Science Study (TIMSS) (see appendix A).
[6]Kuwait tested the same cohort of students as other countries, but later in 2007, at the beginning of the next school year (see appendix A).
[7]National Defined Population covers less than 90 percent of National Target Population (but at least 77 percent, see appendix A).
NOTE: The standard errors of the estimates are shown in tables E-29 and E-30 available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009001.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007.

Although there was no measurable sex difference on the total average science score, U.S. males outperformed U.S. females in one content area: *Earth science* (536 v. 531; figure 21). There was no measurable difference detected in the average scores of U.S. fourth-grade males and females in either the *life science* or *physical science* domains.

Unlike their fourth-grade counterparts, U.S. eighth-grade males outperformed their female classmates in science in 2007 (figure 20). Among the 47 other countries participating in TIMSS, 24 showed a difference in the average science scores of males and females: 10 countries in favor of males and 14 in favor of females. The largest differences were 70 score points in Qatar (in favor of females) and 35 score points in Colombia and Germany (in favor of males).

Like the overall science scale at grade eight, U.S. males scored higher, on average, than their female classmates in three of the four science content domains: *biology* (533 v. 527), *physics* (514 v. 491), and *Earth science* (534 v. 516; figure 21). There was no measurable difference detected in the average science scores of U.S. eighth-grade males and females in the *chemistry* domain.

Figure 21. Average science scores of U.S. fourth- and eighth-grade students, by content domain and sex: 2007

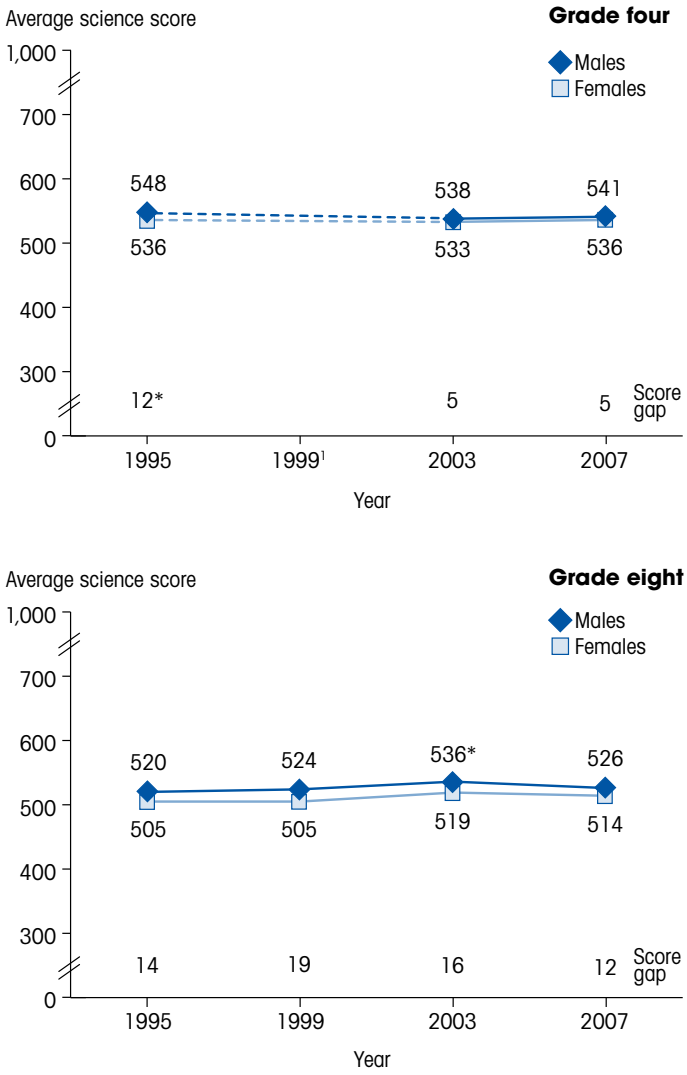


*$p < .05$. Difference between average science scores for males and females is statistically significant and favors males.
NOTE: The United States met guidelines for sample participation rates only after substitute schools were included. The National Defined Population covered 90 percent to 95 percent of National Target Population (see appendix A). The standard errors of the estimates are shown in table E-31 available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009001.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007.

There was no measurable change in the average scores of either U.S. males or females at grade four when 2007 scores were compared to those from 1995 and 2003 (figure 22). However, the advantage for males decreased, from 12 scale score points in 1995 to 5 scale score points in 2003 and 2007.

At grade eight, there was also no measurable change in the average science scores of U.S. males and females or the gap between them when 2007 scores were compared to 1995 (figure 22). However, the average science score for males was lower in 2007 than it was in 2003 (526 v. 536).

Figure 22. Trends in sex differences in average science scores of U.S. fourth- and eighth-grade students: 1995, 1999, 2003, and 2007



*p < .05. Significantly different from 2007.
[1]No fourth-grade assessment was conducted in 1999.
NOTE: In 2007, the United States met guidelines for sample participation rates only after substitute schools were included. The National Defined Population covered 90 percent to 95 percent of National Target Population (see appendix A). Detail may not sum to totals due to rounding. The standard errors of the estimates are shown in table E-32 available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009001.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 1995, 1999, 2003, and 2007.

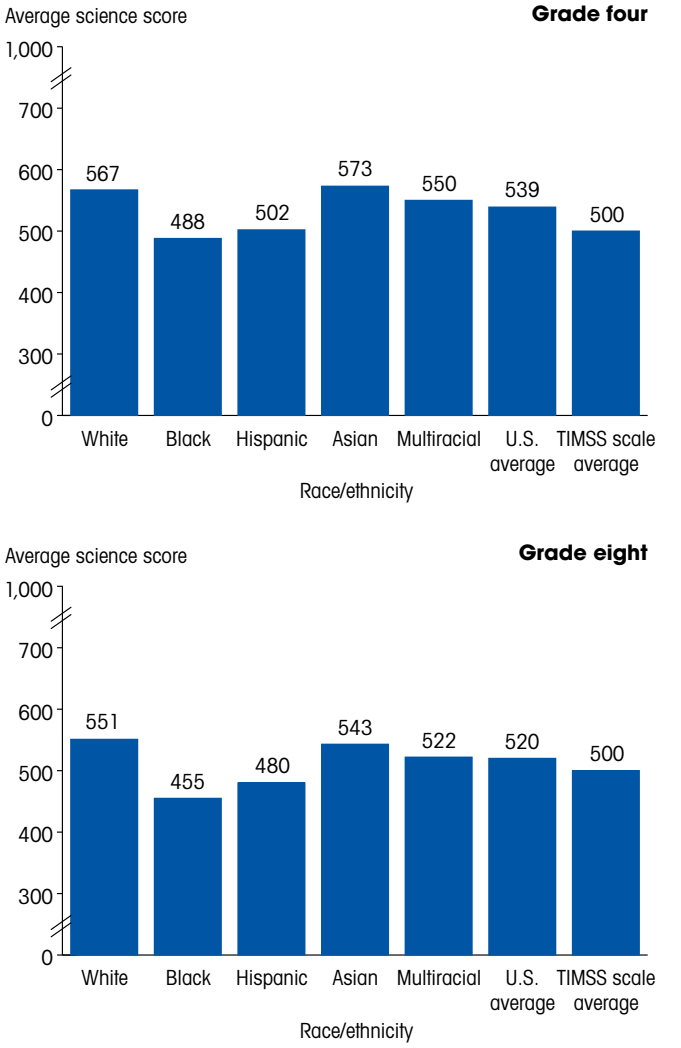
Average scores of students of different races and ethnicities

In 2007, in comparison to the TIMSS scale average, U.S. White, Asian, and multiracial fourth-graders scored higher in science, on average, while U.S. Black fourth-graders scored lower (figure 23). U.S. Hispanic fourth-graders' average score showed no measurable difference from the TIMSS scale average. In comparison to the U.S. national average, U.S. White and Asian fourth-graders scored higher in science, on average, while U.S. Black and Hispanic fourth-graders scored lower. U.S. multiracial fourth-graders' average score showed no measurable difference from the U.S. national average.

At grade eight, U.S. White, Asian, and multiracial students scored higher, on average, than the TIMSS scale average in science and U.S. Black and Hispanic eighth-graders scored lower, on average (figure 23). In comparison to the U.S. national average, U.S. White and Asian eighth-graders scored higher in science, on average, while U.S. Black and Hispanic eighth-graders scored lower. U.S. multiracial eighth-graders' average score showed no measurable difference from the U.S. national average.

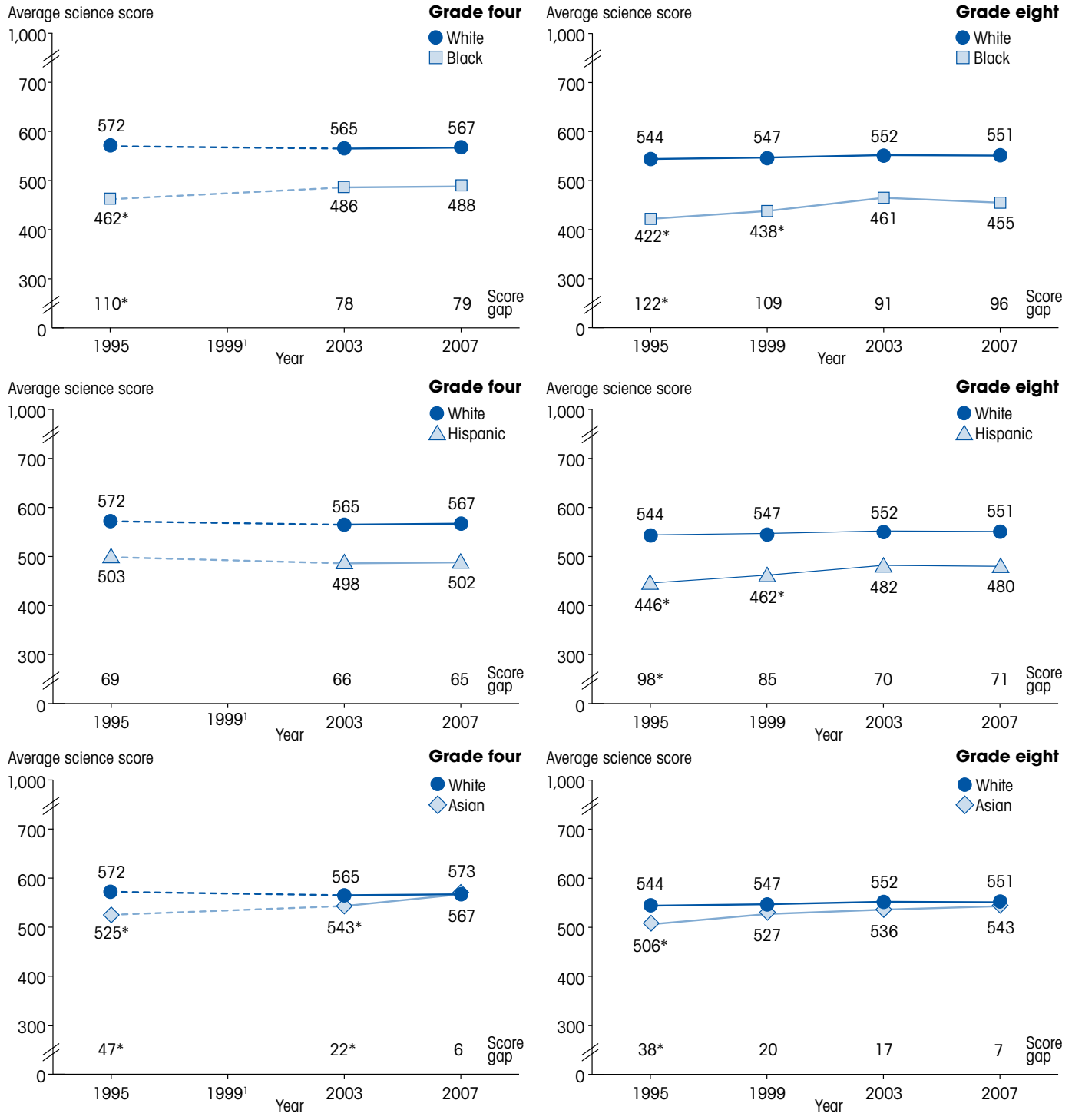
Examination of performance over time shows that U.S. Black and Asian fourth-graders, and U.S. Black, Hispanic, and Asian eighth-graders had an overall pattern of improvement in science, on average (figure 24). There was no measurable change in the average science scores of White and Hispanic fourth-graders, and White eighth-graders when 2007 scores were compared to those from the earlier assessments. Moreover, though significant differences remain in the average scores of White students compared with most of their classmates, the score gap between White students and their counterparts decreased from 1995, at both grades. The exception is the score gap in science between White and Hispanic fourth-graders, which showed no measurable change over the data collection years.

Figure 23. Average science scores of U.S. fourth- and eighth-grade students, by race/ethnicity: 2007



NOTE: Reporting standards were not met for American Indian/Alaska Native and Native Hawaiian/Other Pacific Islander. Black includes African American. Racial categories exclude Hispanic origin. Students who identified themselves as being of Hispanic origin were classified as Hispanic, regardless of their race. Although data for some race/ethnicities are not shown separately because the reporting standards were not met, they are included in the U.S. totals shown throughout the report. The United States met guidelines for sample participation rates only after substitute schools were included. The National Defined Population covered 90 percent to 95 percent of the National Target Population (see appendix A). See appendix A in this report for more information. The standard errors of the estimates are shown in table E-33 available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009001.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007.

Figure 24. Trends in differences in average science scores of U.S. fourth- and eighth-grade students, by selected race/ethnicity: 1995, 1999, 2003, and 2007



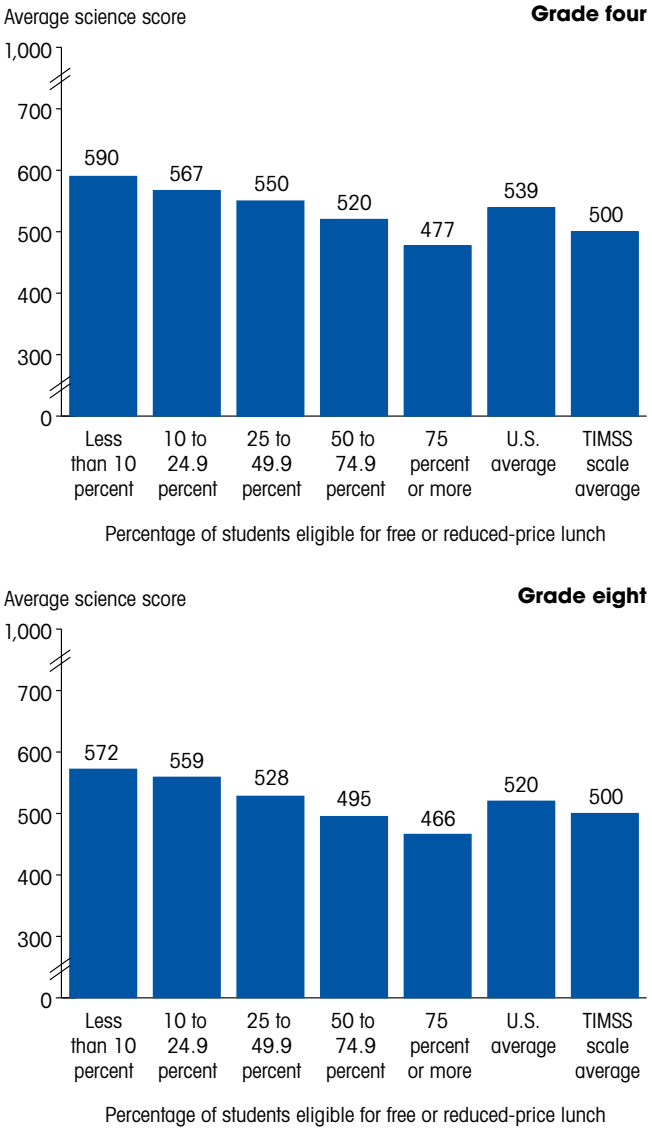
*p < .05. Significantly different from 2007.
[1]No fourth-grade assessment was conducted in 1999.
NOTE: Only the four numerically largest racial categories are shown. Multiracial data were not collected in 1995 and 1999. Reporting standards were not met for American Indian/Alaska Native and Native Hawaiian/Other Pacific Islander. Black includes African American. Racial categories exclude Hispanic origin. Students who identified themselves as being of Hispanic origin were classified as Hispanic, regardless of their race. Although data for some race/ethnicities are not shown separately because the reporting standards were not met, they are included in the U.S. totals shown throughout the report. In 2007, the United States met guidelines for sample participation rates only after substitute schools were included. The National Defined Population covered 90 percent to 95 percent of National Target Population (see appendix A). The tests for significance take into account the standard error for the reported difference. Thus, a small difference between averages for one student group may be significant while a large difference for another student group may not be significant. See appendix A in this report for more information. The standard errors of the estimates are shown in table E-34 available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009001.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 1995, 1999, 2003, and 2007.

Average scores of students attending public schools of various poverty levels

The U.S. results are also arrayed by the concentration of low-income enrollment in the public schools, as measured by eligibility for free or reduced-price lunch, and shown in relation to the TIMSS scale average and the U.S. national average. In comparison to the TIMSS scale average, the average science score of U.S. fourth graders in the highest poverty public schools (at least 75 percent of students eligible for free or reduced-price lunch) in 2007 was lower; the average scores of fourth-graders in each of the other categories of school poverty was higher than the TIMSS scale average (figure 25). In comparison to the U.S. national average score, fourth-graders in schools with 50 percent or more students eligible for free or reduced-price lunch scored lower in science, on average, while those in schools with lower proportions of poor students scored higher, on average, than the U.S. national average.

In comparison to the TIMSS scale average, U.S. eighth-graders attending public schools with fewer than 50 percent of students eligible for the free or reduced-price lunch program scored higher in science, on average (figure 25). On the other hand, U.S. eighth-graders in public schools with 75 percent or more of students eligible scored lower in science, on average, than the TIMSS scale average. In comparison to the U.S. national average, U.S. eighth-graders in public schools with fewer than 25 percent of students eligible scored higher in science, on average, while students in public schools with at least 50 percent eligible scored lower, on average.

Figure 25. Average science scores of U.S. fourth- and eighth-grade students, by percentage of students in public school eligible for free or reduced-price lunch: 2007



NOTE: Analyses are limited to public schools only, based on school reports of the percentage of students in public school eligible for the federal free or reduced-price lunch program. The United States met guidelines for sample participation rates only after substitute schools were included. The National Defined Population covered 90 percent to 95 percent of the National Target Population (see appendix A). The standard errors of the estimates are shown in table E-35 available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009001.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007.

Comparisons of the 2007 average science scores to those for the earlier years within each school poverty level revealed no measurable change in the average science scores at either grade four or eight, with one exception (figure 26).[15] At grade eight, students in public schools with the highest poverty levels (75 percent or more) had a higher average science score in 2007 than in 1999 (466 v. 440).

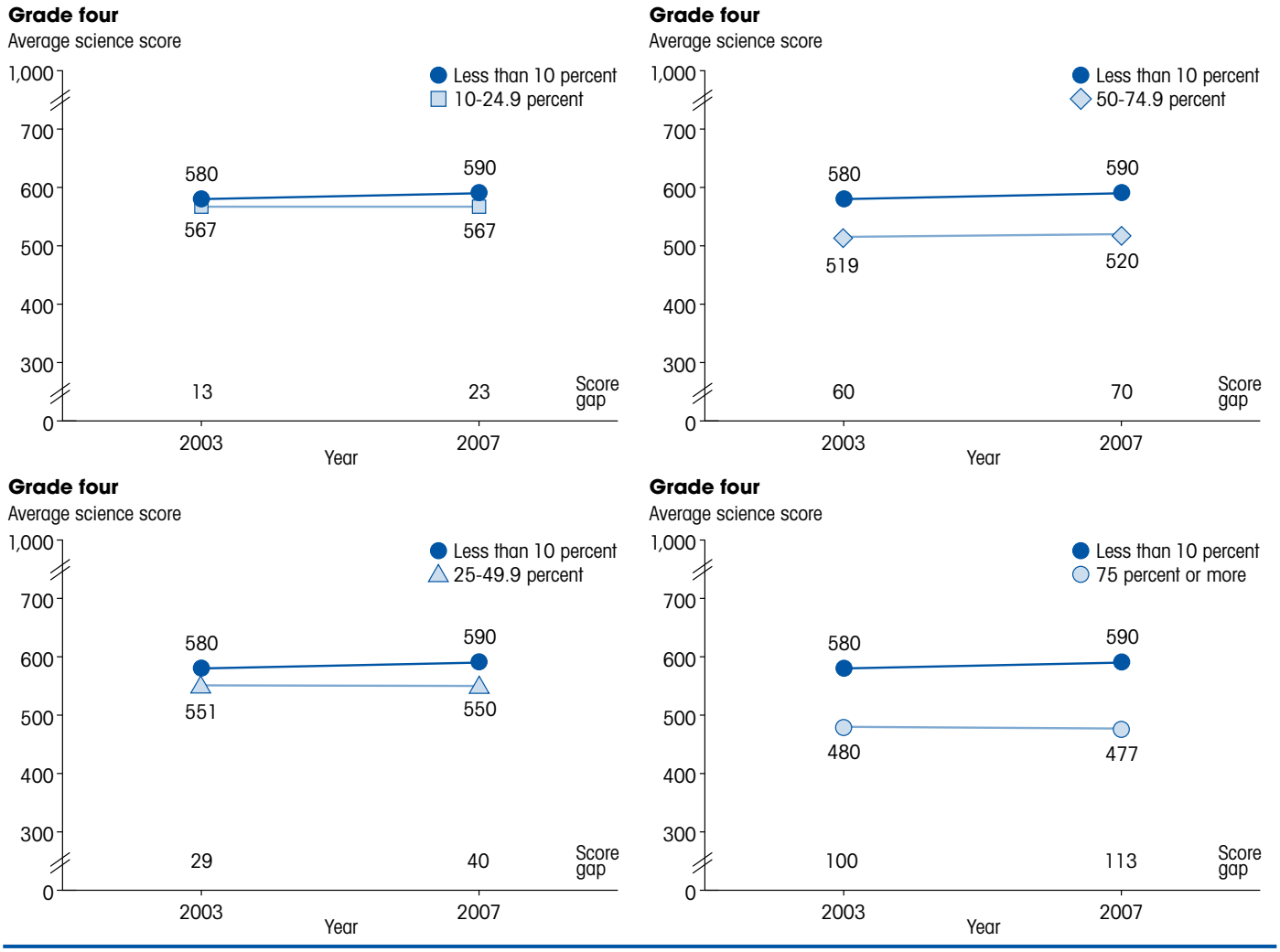
In addition, the size of the difference in average scores, or the score gap, between U.S. fourth- and eighth-graders in public schools with the lowest poverty level (less than 10 percent) and their peers attending public schools with higher poverty levels showed no measurable change (figure 26).

Effect size of the difference in average scores

As noted in the mathematics section of this report, statistically significant results do not necessarily indicate those findings that are important or large enough to consider as informing policy or practice. Small differences may be statistically significant, but may not have much practical import.

As discussed earlier, the highest scoring countries outpaced the United States on a number of measures. The difference at grade four between the U.S. average science score (539) and the Singapore average score (587) was 48 score points (see table 11). The gap between the United States and Singapore is also apparent in the percentage of students scoring at the advanced level: 15 percent of U.S. fourth-graders met the advanced international benchmark compared with 36 percent

Figure 26. Trends in differences in average science scores of U.S. fourth- and eighth-grade students, by school poverty level: 1999, 2003, and 2007



See notes at end of table.

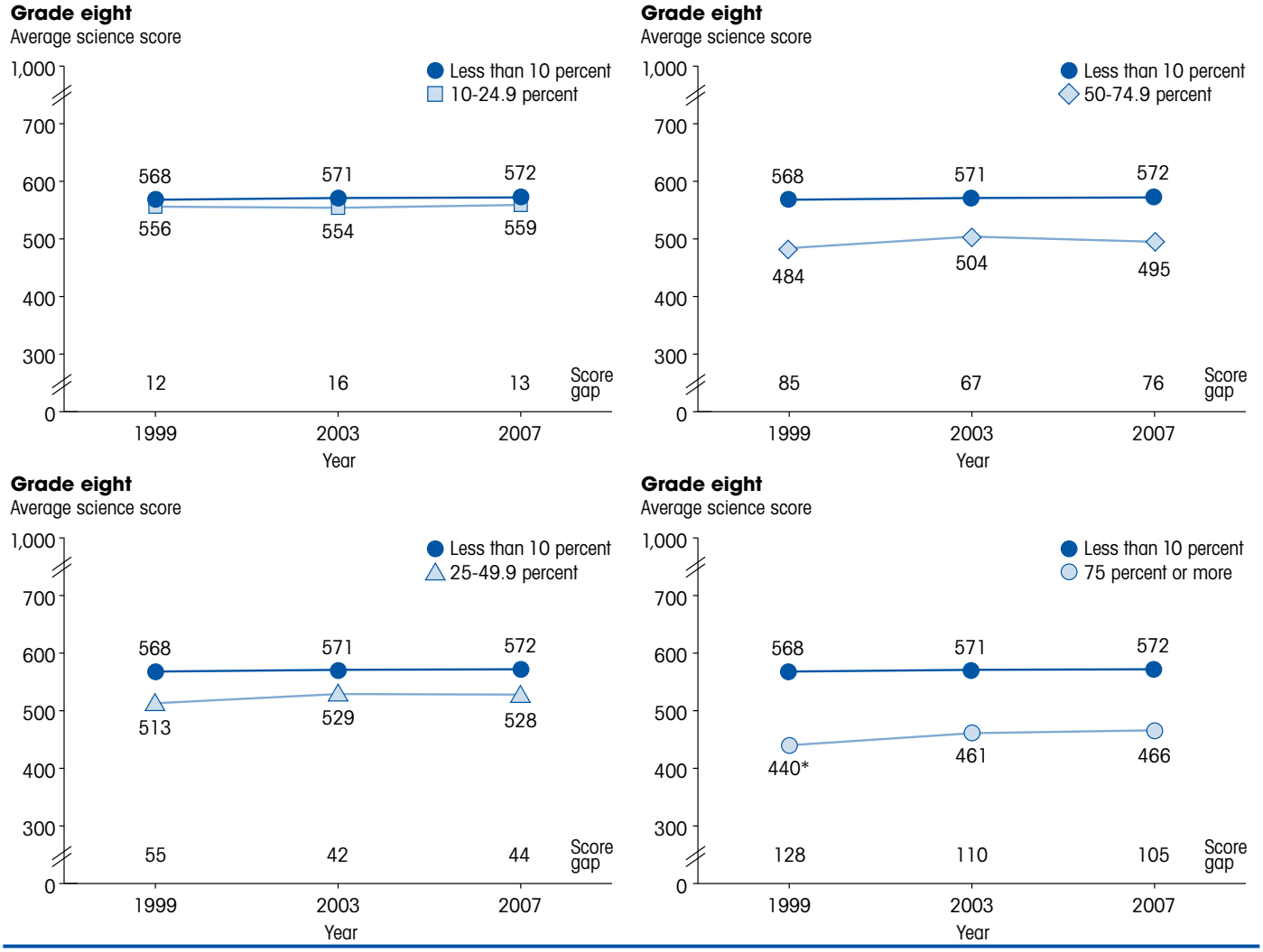
[15]Information on the percentage of students eligible for the federal free or reduced-price lunch program was not collected in 1995 for either grade. Thus, comparisons over time on the poverty measure are limited to a 8-year period.

in Singapore (see figure 17). Are differences within the United States between groups of students (e.g., by race/ethnicity or poverty concentration in schools) bigger or smaller than these international differences? Effect sizes help make these comparisons. Figure 27 shows the effect size of the difference in science only for those groups with statistically significant score differences. Appendix A includes a discussion of how effect sizes were calculated.

As shown in figure 27, and as observed in mathematics, the effect sizes between groups vary considerably. For example, in grade four science, the effect size of the difference between U.S. White and Black students is 2.2 times and between U.S. White and Hispanic students is 1.6 times the effect size between the United States and Singapore, the country with the highest estimated score. The largest observed effect size, between U.S. fourth-graders in schools with the lowest and highest poverty levels, is 3 times the effect size between the United States and Singapore.

At grade eight, the effect size of the difference in science scores between U.S. White and Black students is 2.6 times and between U.S. White and Hispanic students is 2 times the effect size between the United States and Singapore, the country with the highest estimated score. The largest observed effect size, between U.S. eighth-graders in schools with the lowest and highest poverty levels, is 2.8 times the effect size between the United States and Singapore.

Figure 26. Trends in differences in average science scores of U.S. fourth- and eighth-grade students, by school poverty level: 1999, 2003, and 2007—Continued

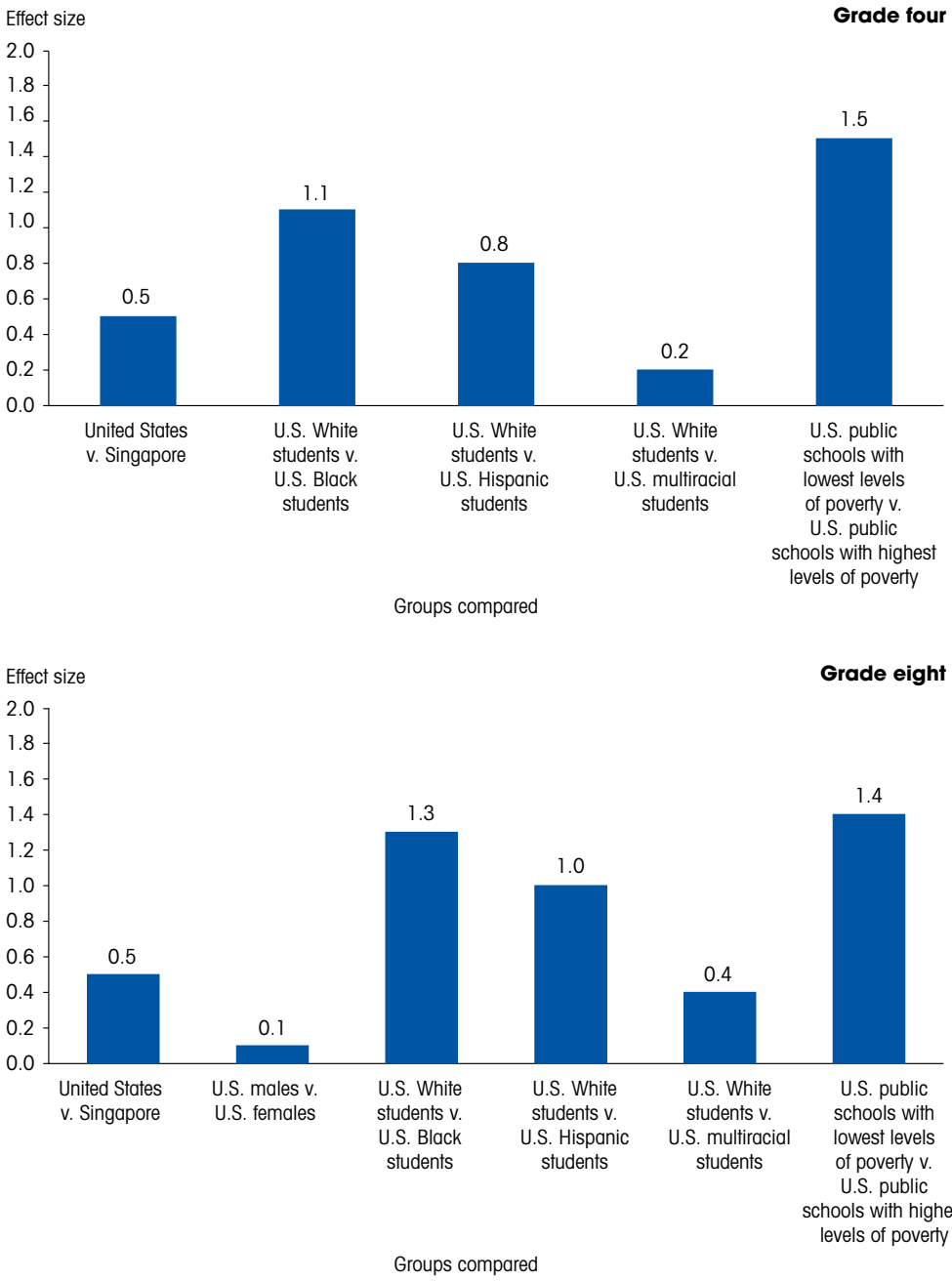


*p < .05. Significantly different from 2007.

NOTE: Information on the percentage of students in school eligible for free or reduced-price lunch was not collected in 1995. No fourth-grade assessment was conducted in 1999. Analyses are limited to public schools only, based on school reports of the percentage of students in school eligible for the federal free or reduced-price lunch program. In 2007, the United States met guidelines for sample participation rates only after substitute schools were included. The National Defined Population covered 90 percent to 95 percent of National Target Population (see appendix A). The standard errors of the estimates are shown in table E-36 available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009001.

SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 1999, 2003, and 2007.

Figure 27. Effect size of difference in average science achievement of fourth- and eighth-grade students, by country, sex, race/ethnicity, and school poverty level: 2007



NOTE: Effect size is shown only for statistically significant differences between group means. Effect size is calculated by dividing the raw difference between group means by the pooled standard deviation (see appendix A). Black includes African American. Racial categories exclude Hispanic origin. Students who identified themselves as being of Hispanic origin were classified as Hispanic, regardless of their race. High-poverty schools are those in which 75 percent or more of students are eligible for the federal free or reduced-price lunch program. Low-poverty schools are those in which less than 10 percent of students are eligible. The United States met guidelines for sample participation rates only after substitute schools were included. The National Defined Population covered 90 percent to 95 percent of the National Target Population. See table E-37 (available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009001) for standard deviations of the U.S. and other countries' student populations. See table E-38 (available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009001) for standard deviations of U.S. student subpopulations.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007.

Page intentionally left blank

References

Beaton, A.E., and González, E. (1995). *The NAEP Primer*. Chestnut Hill, MA: Boston College.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd ed*. Hillsdale, NJ: Erlbaum.

Ferraro, D., and Van de Kerckhove, W. (2006). *Trends in International Mathematics and Science Study (TIMSS) 2003: Nonresponse Bias Analysis* (NCES 2007-044). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Foy, P., Joncas, M., and Zuhlke, O. (2005).*TIMSS 2007 School Sampling Manual*. Unpublished Manuscript, Chestnut Hill, MA: Boston College.

IEA Data Processing Center. (2006). *TIMSS 2007 Data Entry Manager Manual*. Hamburg, Germany: Author.

Martin, M.O., Mullis, I.V.S., and Foy, P. (2008). *TIMSS 2007 International Science Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Chestnut Hill, MA: Boston College.

Matheson, N., Salganik, L., Phelps, R., Perie, M., Alsalam, N., and Smith, T. (1996). *Education Indicators: An International Perspective* (NCES 96-003). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Mullis, I.V.S., Martin, M.O., and Foy, P. (2005). *IEA's TIMSS 2003 International Report on Achievement in the Mathematics Cognitive Domains: Findings From a Developmental Project*. Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Ruddock, G.J., O'Sullivan, C.Y., Arora, A., and Erberber, E. (2005). *TIMSS 2007 Assessment Frameworks*. Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., and Foy, P. (2008). *TIMSS 2007 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Chestnut Hill, MA: Boston College.

National Center for Education Statistics. (2002). *NCES Statistical Standards* (NCES 2003-601). Institute of Education Sciences, U.S. Department of Education. Washington, DC: Author.

Olson, J.F., Martin, M.O., and Mullis, I.V.S. (2008). *TIMSS 2007 Technical Report*. Chestnut Hill, MA: Boston College.

Rosnow, R.L., and Rosenthal, R. (1996). Computing Contrasts, Effect Sizes, and Counternulls on Other People's Published Data: General Procedures for Research Consumers. *Psychological Methods, 1*:331-340.

United Nations Educational, Scientific and Cultural Organization (UNESCO). (1999). *Classifying Educational Programmes Manual for ISCED-97 Implementation in OECD Countries* (1999 Edition). Paris: Author. Retrieved April 9, 2008 from http://www.oecd.org/dataoecd/7/2/1962350.pdf.

Westat. (2007). *WesVar 5.0 User's Guide*. Rockville, MD: Author.

Page intentionally left blank

Appendix A: Technical Notes

Introduction

The Trends in International Mathematics and Science Study (TIMSS) is a cross-national comparative study of the performance and schooling contexts of fourth- and eighth-grade students in mathematics and science. In this fourth cycle of TIMSS, mathematics and science assessments and associated questionnaires were administered in 43 jurisdictions at the fourth-grade level and 56 jurisdictions at the eighth-grade level during 2007. TIMSS is coordinated by the International Association for the Evaluation of Educational Achievement (IEA), with national sponsors in each participating jurisdiction. In the United States, TIMSS is sponsored by the National Center for Education Statistics (NCES), in the Institute of Education Sciences at the U.S. Department of Education. This appendix provides an overview of the technical aspects of TIMSS 2007, including the sampling, data collection, test development and administration, weighting and variance estimation, scaling, and statistical testing procedures used to collect and analyze the data. More detailed information can be found in the *TIMSS 2007 Technical Report* (Olson, Martin, and Mullis 2008).

International requirements for sampling, data collection, and response rates

In order to ensure comparability of the data across countries, the IEA provided detailed international guidelines on the various aspects of data collection described here, and implemented quality control procedures. Participating countries were obliged to follow these guidelines.

Target populations

In order to identify comparable populations of students to be sampled, the IEA defined the target populations as follows (Olson, Martin, and Mullis 2008):

Fourth-grade student population. The international desired target population is all students enrolled in the grade that represents 4 years of schooling, counting from the first year of the International Standard Classification of Education (ISCED)[1] Level 1, providing that the mean age at the time of testing is at least 9.5 years. For most countries, the target grade should be the fourth grade, or its national equivalent. All students enrolled in the target grade, regardless of their age, belong to the international desired target population.

Eighth-grade student population. The international desired target population is all students enrolled in the grade that represents 8 years of schooling, counting from the first year of ISCED Level 1, providing that the mean age at the time of testing is at least 13.5 years. For most countries, the target grade should be the eighth grade, or its national equivalent. All students enrolled in the target grade, regardless of their age, belong to the international desired target population.

Teacher population. The mathematics and science teachers linked to the selected students. Note that these teachers are not a representative sample of teachers within the country. Rather, they are the mathematics and science teachers who teach a representative sample of students in two grades within the country (grades four and eight in the United States).

School population. All eligible schools[2] containing either of the following: one or more fourth-grade classrooms; or one or more eighth-grade classrooms.

Sampling

The sample design employed by the TIMSS 2007 assessment is generally referred to as a three-stage stratified cluster sample. The sampling units at each stage were defined as follows.

First-stage sampling units. The first-stage sampling units consisted of individual schools selected with probability proportionate to size (PPS), size being the estimated number of students enrolled in the target grade. Prior to sampling, schools in the sampling frame could be assigned to a predetermined number of explicit or implicit strata. Schools were to be sampled using a PPS systematic sampling method. Substitution schools—schools selected to replace those that were originally sampled but refused to participate—were to be identified simultaneously.

Second-stage sampling units. The second-stage sampling units were classrooms within sampled schools. Countries were required to randomly select a minimum of one eligible classroom per target grade per school from a list of eligible classrooms prepared for each target grade. However, countries also had the option of selecting more than one eligible classroom per target grade per school and were encouraged to do so.

[1]The ISCED was developed by the United Nations Educational, Scientific, and Cultural Organization (UNESCO) to facilitate the comparability of educational levels across countries. ISCED Level 1 begins with the first year of formal, academic learning (UNESCO 1999). In the United States, ISCED Level 1 begins at grade one.
[2]Some sampled schools may be considered ineligible for reasons noted in the section below titled "School exclusions."

Third-stage sampling units. The third-stage sampling units were students within sampled classrooms. Generally, all students in a sampled classroom were to be selected for the assessment though it was possible to sample a subgroup of students within a classroom, but only after consultation with Statistics Canada, the organization serving as the sampling referee.

Sample size for the main survey

TIMSS guidelines call for a minimum of 150 schools to be sampled per grade, with a minimum of 4,000 students assessed per grade. The basic sample design of one classroom per target grade per school was designed to yield a otal sample of approximately 4,500 students per population. Countries with small class sizes or less than 30 students per school, were directed to consider sampling more schools, more classrooms per school, or both, to meet the minimum target of 4,000 tested students.

In 2007, countries that had participated in TIMSS 2003 were required to increase the size of their student samples to provide data for a bridge study. This study was designed to evaluate the effect of a small change in the assessment design between 2003 and 2007. Countries that participated in TIMSS 2003 were asked to include four additional booklets from 2003 in with the 14 booklets for TIMSS 2007 at each grade. As a result, student sample sizes needed to be increased to ensure that the number of students taking each booklet was sufficient for the purposes of scaling. The 2003-07 Bridge Study is described below in the section on "Scaling".

Exclusions

The following discussion draws on the *TIMSS 2007 School Sampling Manual* (Foy, Joncas, and Zuhlke 2005). All schools and students excluded from the national defined target population are referred to as the excluded population. Exclusions could occur at the school level, with entire schools being excluded, or within schools, with specific students or entire classrooms excluded. TIMSS 2007 did not provide accommodations for students with disabilities or students who were unable to read or speak the language of the test. The IEA requirement with regard to exclusions is that they should not exceed more than 5 percent of the national desired target population (Foy, Joncas, and Zuhlke 2005).

School exclusions. Countries could exclude schools that

- are geographically inaccessible;
- are of extremely small size;
- offer a curriculum, or school structure, radically different from the mainstream educational system; or
- provide instruction only to students in the excluded categories defined under "within-school exclusions," such as schools for the blind.

Within-school exclusions. Countries were asked to adapt the following international within-school exclusion rules to define excluded students:

- Students with intellectual disabilities—Students who, in the professional opinion of the school principal or other qualified staff members, are considered to have intellectual disabilities or who have been tested psychologically as such. This includes students who are emotionally or mentally unable to follow even the general instructions of the test. Students were not to be excluded solely because of poor academic performance or normal disciplinary problems.
- Students with functional disabilities—Students who are permanently physically disabled in such a way that they cannot perform in the TIMSS testing situation. Students with functional disabilities who are able to respond were to be included in the testing.
- Non-native-language speakers—Students who are unable to read or speak the language(s) of the test and would be unable to overcome the language barrier of the test. Typically, a student who had received less than 1 year of instruction in the language(s) of the test was to be excluded.

Defined participation rates

In order to minimize the potential for response biases, the IEA developed participation or response rate standards that apply to all countries and govern whether or not a nation's data are included in the TIMSS 2007 international dataset and the way in which national statistics are presented in the international reports. These standards were set using composites of response rates at the school, classroom, and student and teacher levels and response rates were calculated with and without the inclusion of substitute schools that were selected to replace schools refusing to participate.

The response rate standards determine how a jurisdiction's data will be reported in the international reports. These standards take the following two forms, distinguished primarily by whether or not meeting the school response rate of 85 percent requires the counting of substitute schools.

Category 1: Met requirements. Countries that meet all of the following conditions are considered to have fulfilled the IEA requirements: (a) a minimum school participation rate of 85 percent, based on original sampled schools only; *and* (b) a minimum classroom participation rate of 95 percent, from both original and substitute schools; *and* (c) a minimum student participation rate of 85 percent, from both original and substitute schools.

Category 2: Met requirements after substitutes. In the case of countries not meeting the category 1 requirements, provided that at least 50 percent of schools in the original sample participate, a country's data are considered acceptable if the following requirements are met: a minimum combined school, classroom and student participation rate of 75 percent, based on the product of the participation rates described above. That is, the product of (a), (b) and (c), as defined in the Category 1 standard, must be greater than or equal to 75 percent.

Countries satisfying the Category 1 standard are included in the international tabular presentations without annotation. Those only able to satisfy the Category 2 standard are included as well but are annotated to indicate their response rate status. The data from countries failing to meet either standard are presented separately in the international tabular presentations.

Sampling, data collection, and response rates in the United States and other countries

The U.S. TIMSS sample design

In the United States and most other countries, the target populations of students corresponded to the fourth and eighth grades. In sampling these populations TIMSS used a three-stage stratified cluster sampling design.[3] While the U.S. sampling frame was not explicitly stratified it was implicitly stratified (that is, sorted for sampling) by four categorical stratification variables: type of school (public or private), region of the country (Northeast, Central, West, Southeast);[4] community type (eight levels);[5] and minority status (above or below 15 percent of the student population).

The first stage made use of a systematic PPS technique to select schools for the original sample. Using a sampling frame based on the 2006 National Assessment of Educational Progress (NAEP) school sampling frame,[6] schools were selected with a probability proportionate to the school's estimated enrollment of fourth- or eighth-grade students. Data for public schools were taken from the Common Core of Data (CCD), and data for private schools were taken from the Private School Universe Survey (PSS). In addition, for each original school selected, the two neighboring schools in the sampling frame were designated as substitute schools. The first school following the original sample school was the first substitute and the first school preceding it was the second substitute. If an original school refused to participate, the first substitute was contacted. If that school also refused to participate, the second substitute was contacted. There were several constraints on the assignment of substitutes. One sampled school was not allowed to substitute for another, and a given school could not be assigned to substitute for more than one sampled school. Furthermore, substitutes were required to be in the same implicit stratum as the sampled school.

The second stage consisted of selecting intact mathematics classes within each participating school. Schools provided lists of fourth- or eighth-grade classrooms. Within schools, classrooms with fewer than 15 students were collapsed into pseudo-classrooms, so that each classroom on the school's classroom sampling frame had at least 20 students.[7] An equal probability sample of two classrooms (pseudo-classrooms) was identified from the classroom frame for the school. In schools where there was only one classroom, this classroom was selected with certainty. At the fourth-grade level, 30 pseudo-classrooms were created prior to classroom sampling with 20 of these being selected in the final fourth-grade classroom sample. At the eighth-grade level, 253 pseudo-classrooms were created, of which 58 were included in the final classroom sample.

All students in sampled classrooms (pseudo-classrooms) were selected for assessment. In this way, the overall sample design for the United States was intended to approximate a self-weighting sample of students as much as possible, with each fourth- or eighth-grade student having an equal probability of selection.

[3]The primary purpose of stratification is to improve the precision of the survey estimates. If explicit stratification of the population is used, the units of interest (schools, for example) are sorted into mutually exclusive subgroups–strata. Units in the same stratum are as homogeneous as possible, and units in different strata are as heterogeneous as possible, with respect to the characteristics of interest to the survey. Separate samples are then selected from each stratum. In the case of implicit stratification, the units of interest are simply sorted with respect to one or more variables known to have a high correlation with the variable of interest. In this way, implicit stratification guarantees that the sample of units selected will be spread across the categories of the stratification variables.
[4]The Northeast region consists of Connecticut, Delaware, the District of Columbia, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, and Vermont. The Central region consists of Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, Wisconsin, and South Dakota. The West region consists of Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oklahoma, Oregon, Texas, Utah, Washington, and Wyoming. The Southeast region consists of Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Virginia, and West Virginia.
[5]Eight community types are distinguished: large city of 250,000+; midsize city of < 250,000; urban fringe of large city; urban fringe of mid-size city; large town of 25,000+; small town of 2,500-25,000; rural outside metropolitan statistical area (MSA); rural inside MSA.
[6]In order to maximize response rates from both districts and schools it was necessary to begin the recruitment of both prior to the end of the 2005-06 school year. Since the 2007 NAEP sampling frame was not available until March 2006, it was necessary to base the TIMSS samples on the 2006 NAEP sampling frame.
[7]Since classrooms are sampled with equal probability within schools, small classrooms would have the same probability of selection as large classrooms. Selecting classrooms under these conditions would likely mean that student sample size would be reduced, and some instability in the sampling weights created. To avoid these problems, pseudo-classes are created for the purposes of classroom sampling. Following sampling, the pseudo-class combinations are dissolved and the small classes involved retain their own identity. In this way, data on students, teachers, and classroom practices are linked in small classes in the same way as with larger classes.

U.S.TIMSS fourth-grade sample

School sample. The fourth-grade school sample consisted of 300 schools. Ten ineligible schools were identified on the basis that they served special student populations, or had closed or altered their grade makeup since the sampling frame was developed. This left 290 schools eligible to participate, and 202 agreed to do so. The school response rate before substitution then was 70 percent unweighted. The analogous weighted school response rate was also 70 percent (see table A-1) and is given by the following formula:

$$\text{weighted school response rate before replacement} = \frac{\sum_{i \in Y} W_i E_i}{\sum_{i \in (Y \cup N)} W_i E_i},$$

where Y denotes the set of responding original-sample schools; N denotes the set of eligible non-responding original sample schools; W_i denotes the base weight for school i; $W_i = 1/P_i$, where P_i denotes the school selection probability for school i; and E_i denotes the enrollment size of age-eligible students, as indicated on the sampling frame.

In addition to the 202 participating schools from the original sample, 55 substitute schools participated for a total of 257 participating schools at the fourth grade in the United States (see table A-2). This gives a weighted (and unweighted) school participation rate after substitution of 89 percent (see table A-1).[8]

Classroom sample. Schools agreeing to participate were asked to list their fourth-grade mathematics classes as the basis for sampling at the classroom level, resulting in the identification of a total of 1,108 mathematics classrooms. At this time, schools were given the opportunity to identify special classes–classes in which all or most of the students had intellectual or functional disabilities or were non-native-language speakers. While these classes were regarded as eligible, the students as a group were treated as "excluded" since, in the opinion of the school, their disabilities or language capabilities would render meaningless their performance on the assessment. Some 876 fourth-grade students in a total of 99 classrooms in 63 schools were excluded in this way. Schools identified 32 classrooms containing 222 students with intellectual disabilities (25 percent), 41 classrooms containing 221 students with functional disabilities (25 percent) and 26 classrooms containing 433 non-native-language speakers (50 percent). The remaining 1,009 classrooms served as the pool from which the classroom sample was drawn.

Classrooms with fewer than 15 students were collapsed into pseudo-classrooms prior to sampling so that each eligible classroom in a school had at least 20 students. Two classrooms (pseudo-classrooms) were selected per school where possible. In schools with only one classroom, this classroom was selected with certainty. Some 521 classrooms were selected as a result of this process. All selected classrooms participated in TIMSS yielding a classroom response rate of 100 percent (Olson, Martin, and Mullis 2008, exhibit A.6).

Student sample. Schools were asked to list the students in each of these 521 classrooms, along with the teachers who taught mathematics and science to these students. A total of 11,454 students were listed as a result. Subsequently, 2,454 of these students were allocated to the bridge study since they completed a TIMSS 2003 assessment booklet rather than the TIMSS 2007 assessment (see the description of the 2003-07 bridge study in the section on Scaling below). Eliminating these students from further consideration leaves 9,000 fourth-grade students as the pool of students selected to take part in TIMSS 2007 proper. These students are identified by IEA as "sampled students in participating schools" (Olson, Martin, and Mullis 2008, exhibit A.5).

This pool of students is reduced by within-school exclusions and withdrawals. At the time schools listed the students in the sampled classrooms, they had the opportunity to identify particular students who were not suited to take the test because of physical or intellectual disabilities (i.e., students with disabilities who had been mainstreamed) or because they were non-English-language speakers. Schools identified a total of 543 students they wished to have excluded from the assessment; 323 students with intellectual disabilities (59 percent), 92 students with functional disabilities (17 percent), and 128 students who were non-English-language speakers (24 percent). And, by the time of the assessment a further 140 of the listed students had withdrawn from the school or classroom. In total then, the pool of 9,000 sampled students was reduced by 683 students (543 excluded and 140 withdrawn) to yield 8,317 "eligible" students. The number of eligible students is used as the base for calculating student response rates (Olson, Martin, and Mullis 2008, exhibit A.5).

The number of eligible students was further reduced on assessment day by 421 student absences, leaving 7,896 "assessed students" identified as having completed a TIMSS 2007 assessment booklet (see Table A-2). IEA defines the student response rate as the number of students assessed as a percentage of the number of eligible students which, in this case yields a weighted (and unweighted) student response rate of 95 percent (see table A-1).

[8]Substitute schools are matched pairs and do not have an independent probability of selection. NCES standards (Standard 1-3-8) indicate that, in these circumstances, response rates should be calculated without including substitute schools (National Center for Education Statistics 2002). TIMSS response rates denoted as "before replacement" conform to this standard. TIMSS response rates denoted as "after replacement" are not consistent with NCES standards since, in the calculation of these rates, substitute schools are treated as the equivalent of sampled schools.

Table A-1. Coverage of target populations and participation rates, by grade and country: 2007

	Grade four						
Country	Years of formal schooling	Percentage of international desired population coverage	National desired population overall exclusion rate	Weighted school participation rate before substitution	Weighted school participation rate after substitution	Weighted student participation rate	Combined weighted school and student participation rate[1]
Algeria	4	100	2.1	99	99	97	97
Armenia	4	100	3.4	93	100	96	96
Australia	4	100	4.0	99	100	95	95
Austria	4	100	5.0	98	99	98	97
Chinese Taipei	4	100	2.8	100	100	100	100
Colombia	4	100	2.1	93	99	98	97
Czech Republic	4	100	4.9	89	98	94	92
Denmark	4	100	4.1	71	91	94	85
El Salvador	4	100	2.3	99	100	98	98
England	5	100	2.1	83	90	93	84
Georgia	4	85	4.8	92	100	98	98
Germany	4	100	1.3	96	100	97	96
Hong Kong SAR	4	100	5.4	81	84	96	81
Hungary	4	100	4.4	93	99	97	96
Iran, Islamic Rep. of	4	100	3.0	100	100	99	99
Italy	4	100	5.3	91	100	97	97
Japan	4	100	1.1	97	99	97	95
Kazakhstan	4	94	5.3	99	100	100	100
Kuwait	4	100	0.0	100	100	85	85
Latvia	4	72	4.6	93	97	95	92
Lithuania	5	93	5.4	99	100	94	94
Morocco	4	100	1.4	81	81	96	77
Netherlands	4	100	4.8	48	95	97	91
New Zealand	4.5-5.5	100	5.4	97	100	96	96
Norway	4	100	5.1	88	97	95	92
Qatar	4	100	1.8	100	100	97	97
Russian Federation	4	100	3.6	100	100	98	98
Scotland	5	100	4.5	77	94	94	88
Singapore	4	100	1.5	100	100	96	96
Slovak Republic	4	100	3.3	98	100	97	97
Slovenia	4	100	2.1	92	99	95	93
Sweden	4	100	3.1	98	100	97	97
Tunisia	4	100	2.9	100	100	99	99
Ukraine	4	100	0.6	96	96	97	93
United States	4	100	9.2	70	89	95	84
Yemen	4	100	2.0	99	100	98	98

(See notes at end of table)

Table A-1. Coverage of target populations and participation rates, by grade and country: 2007 —Continued

	Grade eight						
Country	Years of formal schooling	Percentage of international desired population coverage	National desired population overall exclusion rate	Weighted school participation rate before substitution	Weighted school participation rate after substitution	Weighted student participation rate	Combined weighted school and student participation rate[1]
Algeria	8	100	0.1	99	99	96	95
Armenia	8	100	3.3	94	100	96	96
Australia	8	100	1.9	100	100	93	93
Bahrain	8	100	1.5	100	100	97	97
Bosnia and Herzegovina	8 or 9	100	1.5	100	100	98	98
Botswana	8	100	0.1	100	100	99	99
Bulgaria	8	100	20.3	94	98	96	94
Chinese Taipei	8	100	3.3	100	100	99	99
Colombia	8	100	1.6	96	100	98	98
Cyprus	8	100	2.5	100	100	96	96
Czech Republic	8	100	4.6	92	100	95	95
Egypt	8	100	0.5	99	100	98	98
El Salvador	8	100	2.8	99	100	98	98
England	9	100	2.3	78	86	88	75
Georgia	8	85	3.9	97	100	97	97
Ghana	8	100	0.9	100	100	98	98
Hong Kong SAR	8	100	3.8	73	79	96	75
Hungary	8	100	3.9	92	99	97	96
Indonesia	8	100	3.4	100	100	97	97
Iran, Islamic Rep. of	8	100	0.5	100	100	98	98
Israel	8	100	22.8	94	97	94	91
Italy	8	100	5.0	93	100	96	96
Japan	8	100	3.5	96	97	93	91
Jordan	8	100	2.0	100	100	96	96
Korea, Rep. of	8	100	1.6	100	100	99	99
Kuwait	8	100	0.3	97	97	87	84
Lebanon	8	100	1.4	81	92	93	85
Lithuania	8	92	4.2	98	99	91	90
Malaysia	8	100	3.3	100	100	98	98
Malta	9	100	2.9	100	100	95	94
Norway	8	100	2.6	88	93	93	86
Oman	8	100	1.2	100	100	99	99
Palestinian Nat'l Auth.	8	100	1.0	100	100	98	98
Qatar	9	100	0.8	100	100	97	97
Romania	8	100	1.8	99	99	97	97
Russian Federation	7 or 8	100	2.3	100	100	97	97
Saudi Arabia	8	100	0.5	99	99	95	94
Scotland	9	100	1.7	74	86	90	77
Serbia	8	80	6.8	100	100	98	98
Singapore	8	100	1.8	100	100	95	95
Slovenia	7 or 8	100	1.9	92	99	93	92
Sweden	8	100	3.6	100	100	94	94
Syrian Arab Republic	8	100	0.6	100	100	96	96
Thailand	8	100	3.4	90	100	99	99
Tunisia	8	100	0.0	100	100	98	98
Turkey	8	100	2.6	100	100	98	98
Ukraine	8	100	0.2	98	98	97	95
United States	8	100	7.9	68	83	93	77

[1]The combined weighted school and student participation rate is derived by multiplying the unrounded weighted school and student participation rates.
NOTE: Only countries that completed the necessary steps for their data to appear in the reports from the International Study Center are listed. In addition to the countries listed above, seven separate jurisdictions participated in TIMSS 2007: the provinces of British Columbia, Ontario, and Quebec in Canada; the Basque region of Spain; Dubai, UAE; and the states of Massachusetts and Minnesota. Information on these seven jurisdictions can be found in the international TIMSS 2007 reports (Mullis, Martin, and Foy 2008; Martin, Mullis, and Foy 2008). Countries could participate at either grade level. Countries were required to sample students enrolled in the grade that represents 4 years of schooling, counting from the first year of the International Standard Classification of Education (ISCED) Level 1, providing that the mean age at the time of testing is at least 9.5 years, or students enrolled in the grade that represents eight years of schooling, counting from the first year of ISCED Level 1. In the United States and most countries, this corresponds to grade four and grade eight, respectively. In Bulgaria, the science assessment was administered to a diminished number of schools and students. The weighted school participation rate before substitution shown above refers to the mathematics assessment. This number should be reduced to 93 percent in describing the science assessment.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007.

Note that the 876 students excluded because whole classes were excluded do not figure in the calculation of student response rates. They do, however, figure in the calculation of the coverage of the International Target Population. Together, these 876 students excluded prior to classroom sampling, plus the 543 within-class exclusions resulted in an overall student exclusion rate of 9.2 percent (see table A-1 and Olson, Martin, and Mullis 2008, exhibit A.3). The reported coverage of the International Target Population then is 90.8 percent (see Olson, Martin, and Mullis 2008, Exhibit A.3). IEA standards define this degree of coverage as acceptable though falling outside the desired range of 95 percent or better.

Combined participation rates. The combined school, classroom, and student weighted response rate standard of 75 percent used by TIMSS in situations in which it is necessary to recruit substitute schools was met in this instance. Both the weighted and unweighted product of the separate response rates (84 percent) exceeded this 75 percent standard (see table A-1). The application of international guidelines means, however, that U.S. statistics describing fourth-grade students are annotated in international reports to indicate that coverage of the defined student population was less than the IEA standard of 95 percent and that participation rates were met only after substitute schools were included.

Tables A-1 and A-2 are extracts from the international report Exhibits noted above and are designed to summarize information on school and student responses rates and coverage of the fourth- and eighth-grade target populations in each nation.

Table A-2. Total number of schools and students, by grade and country: 2007

	Grade four						
Country	Schools in original sample	Eligible schools in original sample	Schools in original sample that participated	Substitute schools	Total schools that participated	Sampled students in participating schools	Students assessed
Algeria	150	150	149	0	149	4,366	4,223
Armenia	150	148	143	5	148	4,253	4,079
Australia	230	229	226	3	229	4,511	4,108
Austria	199	197	194	2	196	5,158	4,859
Chinese Taipei	150	150	150	0	150	4,260	4,131
Colombia	150	143	132	10	142	5,320	4,801
Czech Republic	150	147	132	12	144	4,583	4,235
Denmark	150	150	105	32	137	3,907	3,519
El Salvador	150	148	146	2	148	4,467	4,166
England	160	159	131	12	143	4,784	4,316
Georgia	152	144	131	13	144	4,384	4,108
Germany	250	247	239	7	246	5,464	5,200
Hong Kong SAR	150	150	122	4	126	3,965	3,791
Hungary	150	145	135	9	144	4,221	4,048
Iran, Islamic Rep. of	240	224	224	0	224	3,939	3,833
Italy	170	170	155	15	170	4,912	4,470
Japan	150	150	145	3	148	4,677	4,487
Kazakhstan	150	141	140	1	141	4,063	3,990
Kuwait	150	150	149	0	149	4,468	3,803
Latvia	150	150	140	6	146	4,188	3,908
Lithuania	163	156	154	2	156	4,345	3,980
Morocco	226	224	184	0	184	4,282	3,894
Netherlands	150	148	72	69	141	3,608	3,349
New Zealand	220	220	213	7	220	5,347	4,940
Norway	150	150	131	14	145	4,462	4,108
Qatar	114	114	114	0	114	7,411	7,019
Russian Federation	206	206	206	0	206	4,659	4,464
Scotland	150	148	114	25	139	4,320	3,929
Singapore	177	177	177	0	177	5,235	5,041
Slovak Republic	184	184	181	3	184	5,269	4,963
Slovenia	150	150	138	10	148	4,664	4,351
Sweden	160	155	151	4	155	4,965	4,676
Tunisia	150	150	150	0	150	4,242	4,134
Ukraine	150	150	144	0	144	4,459	4,292
United States	300	290	202	55	257	9,000	7,896
Yemen	150	144	143	1	144	6,128	5,811

See notes at end of table.

Table A-2. Total number of schools and students, by grade and country: 2007—Continued

Country	Grade eight						
	Schools in original sample	Eligible schools in original sample	Schools in original sample that participated	Substitute schools	Total schools that participated	Sampled students in participating schools	Students assessed
Algeria	150	150	149	0	149	5,793	5,447
Armenia	150	148	143	5	148	4,898	4,689
Australia	230	228	228	0	228	4,549	4,069
Bahrain	74	74	74	0	74	4,434	4,230
Bosnia and Herzegovina	150	150	150	0	150	4,373	4,220
Botswana	150	150	150	0	150	4,310	4,208
Bulgaria	170	166	158	5	163	4,312	4,019
Chinese Taipei	150	150	150	0	150	4,164	4,046
Colombia	150	148	142	6	148	5,343	4,873
Cyprus	67	67	67	0	67	4,755	4,399
Czech Republic	150	147	135	12	147	5,182	4,845
Egypt	237	233	231	2	233	6,906	6,582
El Salvador	150	145	143	2	145	4,329	4,063
England	160	160	126	11	137	4,768	4,025
Georgia	152	135	131	4	135	4,533	4,178
Ghana	163	163	163	0	163	5,678	5,294
Hong Kong SAR	152	152	112	8	120	3,657	3,470
Hungary	150	145	133	11	144	4,321	4,111
Indonesia	150	149	149	0	149	4,419	4,203
Iran, Islamic Rep. of	220	208	208	0	208	4,140	3,981
Israel	150	150	140	6	146	3,708	3,294
Italy	170	170	159	11	170	4,873	4,408
Japan	150	150	144	2	146	4,656	4,312
Jordan	200	200	200	0	200	5,733	5,251
Korea, Rep. of	150	150	150	0	150	4,358	4,240
Kuwait	163	163	158	0	158	4,721	4,091
Lebanon	150	148	120	16	136	4,062	3,786
Lithuania	150	144	141	1	142	4,537	3,991
Malaysia	150	150	150	0	150	4,589	4,466
Malta	60	59	59	0	59	5,053	4,670
Norway	150	150	133	6	139	5,085	4,627
Oman	150	146	146	0	146	4,894	4,752
Palestinian Nat'l Auth.	155	148	147	1	148	4,572	4,378
Qatar	67	67	66	0	66	7,558	7,184
Romania	150	150	149	0	149	4,447	4,198
Russian Federation	210	210	210	0	210	4,706	4,472
Saudi Arabia	167	166	165	0	165	4,515	4,243
Scotland	150	150	109	20	129	4,700	4,070
Serbia	150	147	147	0	147	4,246	4,045
Singapore	164	164	164	0	164	4,828	4,599
Slovenia	150	150	138	10	148	4,414	4,043
Sweden	160	159	158	1	159	5,712	5,215
Syrian Arab Republic	150	150	150	0	150	5,025	4,650
Thailand	150	150	134	16	150	5,579	5,412
Tunisia	150	150	150	0	150	4,258	4,080
Turkey	150	146	146	0	146	4,682	4,498
Ukraine	150	150	146	0	146	4,598	4,424
United States	300	287	197	42	239	8,447	7,377

NOTE: Only countries that completed the necessary steps for their data to appear in the reports from the International Study Center are listed. In addition to the countries listed above, seven separate jurisdictions participated in TIMSS 2007: the provinces of British Columbia, Ontario, and Quebec in Canada; the Basque region of Spain; Dubai, UAE; and the states of Massachusetts and Minnesota. Information on these seven jurisdictions can be found in the international TIMSS 2007 reports (Mullis, Martin, and Foy 2008; Martin, Mullis, and Foy 2008). Countries could participate at either grade level. Countries were required to sample students enrolled in the grade that represents 4 years of schooling, counting from the first year of the International Standard Classification of Education (ISCED) Level 1, providing that the mean age at the time of testing is at least 9.5 years, or students enrolled in the grade that represents eight years of schooling, counting from the first year of ISCED Level 1. In the United States and most countries, this corresponds to grade four and grade eight, respectively. In Bulgaria, the science assessment was administered to a diminished number of schools and students. The numbers shown in the table refer to the mathematics assessment. These should be reduced accordingly to describe the science assessment, as follows:eligible schools=142; participating schools in original sample=134; total participating schools=134; sampled students in participating schools=3,426; students assessed=3,079.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007.

U.S.TIMSS eighth-grade sample

School sample. The eighth-grade school sample consisted of 300 schools. Thirteen ineligible schools were identified on the basis that they served special student populations, or had closed or altered their grade makeup since the sampling frame was developed. This left 287 schools eligible to participate and 197 agreed to do so. The unweighted school response rate before substitution then was 69 percent. The analogous weighted school response rate was 68 percent (see table A-1).

In addition to the 197 participating schools from the original sample, 42 substitute schools participated for a total of 239 participating schools at the eighth grade in the United States (see table A-2). This gives a weighted (and unweighted) school participation rate after substitution of 83 percent (see table A-1).[9]

Classroom sample. Schools agreeing to participate were asked to list their eighth-grade mathematics classes as the basis for sampling at the classroom level, resulting in the identification of a total of 3,125 mathematics classrooms. At this time, schools were given the opportunity to identify special classes–classes in which all or most of the students had intellectual or functional disabilities or were non-English-language speakers. While these classes were regarded as eligible, the students as a group were treated as "excluded" since, in the opinion of the school, their disabilities or language capabilities would render meaningless their performance on the assessment. Some 2,834 eighth-grade students in a total of 308 classrooms in 133 schools were excluded in this way. Schools identified 106 classrooms containing 788 students with intellectual disabilities (28 percent), 136 classrooms containing 989 students with functional disabilities (35 percent) and 66 classrooms containing 1,057 non-native-language speakers (37 percent). The remaining 2,775 classrooms served as the pool from which the sample was drawn.

Classrooms with fewer than 15 students were collapsed into pseudo-classrooms prior to sampling so that each eligible classroom in a school had at least 20 students. Two classrooms (pseudo-classrooms) were selected per school where possible. In schools where there was only one classroom, this classroom was selected with certainty. Some 539 classrooms were selected as a result of this process. All selected classrooms participated in TIMSS yielding a classroom response rate of 100 percent (Olson, Martin, and Mullis 2008, exhibit A6).

Subsequently, schools were asked to list the students in each sampled classroom, along with the teachers who taught mathematics and science to these students. At this time, schools were given the opportunity to identify particular students in these classrooms who were not suited to take the test because of physical or intellectual disabilities (i.e., students with disabilities who had been mainstreamed) or because they were non-native- language speakers.

Student sample. Schools were asked to list the students in each of these 539 sampled classrooms, along with the teachers who taught mathematics and science to these students. A total of 10,793 students were listed as being in the selected classrooms. Subsequently, 2,346 of these students were allocated to the bridge study since they completed a TIMSS 2003 assessment booklet rather than the TIMSS 2007 assessment (see the description of the 2003-07 bridge study in the section on Scaling below). Eliminating these students from further consideration leaves 8,447 eighth-grade students as the pool of students selected to take part in TIMSS 2007 proper. These students are identified by IEA as "sampled students in participating schools" (Olson, Martin, and Mullis 2008, exhibit A5).

This pool of students is reduced by within-school exclusions and withdrawals. At the time schools listed the students in sampled classrooms, they had the opportunity to identify particular students who were not suited to take the test because of physical or intellectual disabilities (i.e., students with disabilities who had been mainstreamed) or because they were non-native-language speakers. Schools identified a total of 272 students they wished to have excluded from the assessment; 154 students with intellectual disabilities (57 percent), 48 students with functional disabilities (18 percent) and 70 students who were non-English-language speakers (26 percent). And, by the time of the assessment a further 202 of the listed students had withdrawn from the school or classroom. In total then, the pool of 8,447 sampled students was reduced by 474 students (272 excluded and 202 withdrawn) to yield 7,973 "eligible" students. The number of eligible students is used as the base for calculating student response rates (Olson, Martin, and Mullis 2008, exhibit A5). .

The number of eligible students was further reduced on assessment day by 596 student absences, leaving 7,377 "assessed students" identified as having completed a TIMSS 2007 assessment booklet (see table A-2). The IEA defines the student response rate as the number of students assessed as a percentage of the number of eligible students which, in this case yields a weighted (and unweighted) student response rate of 93 percent (see table A-1).

Note that the 2,834 students excluded because whole classes were excluded do not figure in the calculation of student response rates. They do, however, figure in the calculation of the coverage of the International Target Population. Together, these 2,834 students excluded prior to classroom sampling, plus the 272 within-class exclusions resulted in an overall student exclusion rate of 7.9 percent (see table A-1 and Olson,

[9]Substitute schools are matched pairs and do not have an independent probability of selection. NCES standards (Standard 1-3-8) indicate that, in these circumstances, response rates should be calculated without including substitute schools (National Center for Education Statistics 2002). TIMSS response rates denoted as "before replacement" conform to this standard. TIMSS response rates denoted as "after replacement" are not consistent with NCES standards since, in the calculation of these rates, substitute schools are treated as the equivalent of sampled schools.

Martin, and Mullis 2008, exhibit A.3). The reported coverage of the International Target Population then is 92.1 percent (see Olson, Martin, and Mullis 2008, exhibit A.3). IEA standards define this degree of coverage as acceptable though falling outside the desired range of 95 percent or better.

Combined participation rates. The combined school, classroom and student weighted response rate standard of 75 percent used by TIMSS in situations where substitute schools were necessary was met in this instance. Both the weighted and unweighted product of the separate response rates (77 percent) exceeded this 75 percent standard (see table A-1). The application of international guidelines means, however, that U.S. statistics describing eighth-grade students are annotated in international reports to indicate that coverage of the defined student population was less than the IEA standard of 95 percent and that participation rates were met only after substitute schools were included. Table A-2 summarizes information on the coverage of the eighth-grade target populations in each nation.

Nonresponse bias in the U.S. TIMSS samples

NCES standards require a nonresponse bias analysis if the school-level response rate falls below 85 percent of the sampled schools (standard 2-2-2; National Center for Education Statistics 2002), as they did for both fourth- and eighth-grade samples. As a consequence a nonresponse bias analysis was initiated and took a form similar to that adopted for TIMSS 2003 (Ferraro and Van de Kerckhove 2006). A full report of this study will be included in a technical report to be released with the U.S. national TIMSS dataset.

Three methods were chosen to perform this analysis. The first method focused exclusively on the sampled schools and ignored substitute schools. The schools were weighted by their school base weights, excluding any nonresponse adjustment factor. The second method focused on *sampled schools plus substitute schools*, treating as nonrespondents those schools from which a final response was not received. Again, schools were weighted by their base weights, with the base weight for each substitute school set to the base weight of the original school that it replaced. The third method repeated the analyses from the second method using nonresponse adjusted weights.[10]

In order to compare TIMSS respondents and nonrespondents, it was necessary to match the sample of schools back to the sample frame to identify as many characteristics as possible that might provide information about the presence of nonresponse bias.[11] The characteristics available for analysis in the sampling frame were taken from the CCD for public schools, and from the PSS for private schools. For categorical variables, the distribution of the characteristics for respondents was compared with the distribution for all schools. The hypothesis of independence between a given school characteristic and the response status (whether or not the school participated) was tested using a Rao-Scott modified chi-square statistic. For continuous variables, summary means were calculated and the difference between means was tested using a t test. Note that this procedure took account of the fact that the two samples in question were not independent samples, but in fact the responding sample was a subsample of the full sample. This effect was accounted for in calculating the standard error of the difference. Note also that in those cases where both samples were weighted using just the base weights, the test is exactly equivalent to testing that the mean of the respondents was equal to the mean of the nonrespondents.

In addition, multivariate logistic regression models were set up to identify whether any of the school characteristics were significant in predicting response status when the effects of all potential influences were considered simultaneously.

Public and private schools were modeled together using the following variables:[12] community type (central city, urban fringe/large town, rural/small town); control of school (public or private); NAEP region (Northeast, Southeast, Central, West); poverty level (percentage of students in school eligible for free or reduced-price lunch);[13] number of students enrolled in fourth or eighth grade; total number of students; and, percentage minority students.[14]

[10]A detailed treatment of the meaning and calculation of sampling weights, including the nonresponse adjustment factors, is provided in the *TIMSS 2007 Technical Report* (Olson, Martin, and Mullis 2008).
[11]Comparing characteristics for respondents and nonrespondents is not always a good measure of nonresponse bias if the characteristics are either unrelated or weakly related to more substantive items in the survey. Nevertheless, this is often the only approach available.
[12]NAEP region and community type were dummy coded for the purposes of these analyses. In the case of NAEP region, "West" was used as the omitted group. For community type, "urban fringe/large town" was chosen as the omitted group.
[13]The measure of school poverty is based on the proportion of students in a school eligible for the Free or Reduced-Price Lunch (FRPL) program, a federally assisted meal program that provides nutritionally balanced, low-cost or free lunches to eligible children each school day. For the purposes of the nonresponse bias analyses, schools were classified as "low poverty" if less than 50 percent of the students were eligible for FRPL, and "high poverty" if 50 percent or more of students were eligible. Since the nonresponse bias analyses involve both participating and nonparticipating schools, they are based, out of necessity, on data from the sampling frame. TIMSS data are not available for nonparticipating schools. The school frame data are derived from the CCD and PPS. The CCD data provide information on the percentage of students in each school who are eligible for free- or reduced-price lunch, but are limited to public schools. The PPS data do not provide the same information for private schools. In the interest of retaining all of the schools and students in these analyses, private schools were assumed to be low-poverty schools–that is, they were assumed to be schools in which less than 50 percent of students were eligible for FRPL. Separate analyses of the TIMSS data for participating private schools suggest the reasonableness of this assumption. Of the 21 grade four private schools, only one reports having 50 percent or more of students eligible for FRPL. Among the 21 grade eight private schools, only two report having 50 percent or more of students eligible for FRPL.
[14]Two forms of this school attribute were used in the analyses. In the bivariate analyses the percentage of each race/ethnic group was related separately to participation status. In the logistic regression analyses a single measure was used to characterize each school, namely, "percentage of minority students."

Results for the original sample of schools. In the analyses for the original sample of schools, all substituted schools were treated as nonresponding schools. The results of these analyses follow.

- **Fourth grade.** In the investigation into nonresponse bias at the school level for TIMSS fourth-grade schools, comparisons between schools in the eligible sample and participating schools showed that there was no relationship between response status and the majority of school characteristics available for analysis. In separate variable-by-variable bivariate analyses, three variables were found to be related to participation: community type, region, and racial/ethnic composition. Central city schools were underrepresented among participating schools by almost 4 percent and rural small-town schools were overrepresented by the same amount. Similarly, schools in the Central region were overrepresented by close to 5 percent, and schools in the West underrepresented by about 3.5 percent in the original sample of participating schools. And, in regard to racial/ethnic composition, both the percentage of White, non-Hispanic and the percentage of American Indian or Alaska Native students were higher in participating schools than in the eligible sample. Although each of these findings indicates some potential for nonresponse bias, when all of these factors were considered simultaneously in a regression analysis, the results indicated that the only independent source of bias lay with the fact that, relative to schools in the West, schools in the Central region were somewhat overrepresented among the participating schools.
- **Eighth grade.** The bivariate analyses for eighth-grade schools showed no relationship between participation and any of the school characteristics examined. However, the multivariate regression analysis showed that, relative to urban fringe/large town schools, central city schools were overrepresented among the participating schools. And, relative to schools in the West region, schools in the Central region were similarly overrepresented.

Results for the final sample of schools. In the analyses for the final sample of schools, all substitute schools were included with the original schools as responding schools, leaving nonresponding schools as those for which no assessment data were available. The results of these analyses follow and are somewhat more complicated than the analyses for the original sample of schools.

- **Fourth grade.** The bivariate results for the final sample of fourth-grade schools indicated that two of the three variables were still found to be related to participation: community type, and racial/ethnic composition. As in the earlier analysis, central city schools were underrepresented among participating schools (by some 2.5 percent) and rural small-town schools were overrepresented (by some 2 percent). Similarly, both the percentage of White, non-Hispanic and the percentage of American Indian or Alaska Native students were higher in participating schools than in the eligible sample. In each instance the differences were substantially reduced over those seen in connection with the original sample. These same differences could not be demonstrated in the multivariate regression analysis which failed to show any variables as significant predictors of participation.

  For the final sample of schools with school nonresponse adjustments applied to the weights,[15] the results were identical. These results suggest that there is some potential for nonresponse bias in the fourth-grade original sample based on the characteristics studied. It also suggests that the use of substitute schools reduced the potential for bias. The school nonresponse adjustment had no effect on the characteristics of the weighted responding sample of schools.
- **Eighth grade.** The bivariate results for the final sample indicated that two variables were related to participation: community type, and the percentage of American Indian or Alaska Native students. Central city schools were overrepresented among participating schools by some 4 percent, and schools in urban fringe/large town were underrepresented by nearly 4 percent, And, in regard to racial/ethnic composition, the percentage of American Indian or Alaska Native students in participating schools was higher than in all eligible schools. The multivariate regression analysis indicated that, relative to urban fringe/large town schools, central city schools were overrepresented among the participating schools, and that the percentage of minority students in participating schools was lower than in all eligible schools.

  With school nonresponse adjustments applied to the weights,[16] the results were identical. These results suggest that there is some potential for nonresponse bias in the original sample based on the characteristics studied. It also suggests that, while there is no evidence that the use of substitute schools reduced the potential for bias, it has not added to it substantially. The school nonresponse adjustment had no effect on the characteristics of the weighted responding sample of schools.

[15]The international weighting procedures created a nonresponse adjustment class for each explicit stratum; see the *TIMSS 2007 Technical Report* (Olson, Martin, and Mullis 2008) for details. In the case of the U.S. fourth-grade sample, there was no explicit stratification and thus a single adjustment class. The procedures could not be varied for individual countries to account for any specific needs. Therefore, the U.S. nonresponse bias analyses could have no influence on the weighting procedures and were undertaken after the weighting process was complete.

[16]The international weighting procedures created a nonresponse adjustment class for each explicit stratum. For the eighth grade, there was no explicit stratification and thus a single adjustment class. Again, the procedures were not varied for individual countries to account for any specific needs. As with the fourth grade, the nonresponse bias analyses for the eighth grade could have no influence on the weighting procedures

Test development

TIMSS is a cooperative effort involving representatives from every country participating in the study. For TIMSS 2007, the test development effort began with a revision of the frameworks that are used to guide the construction of the assessment (Mullis et al. 2005). The frameworks were updated to reflect changes in the curriculum and instruction of participating countries. Extensive input from experts in mathematics and science education, assessment, and curriculum, and representatives from national educational centers around the world contributed to the final shape of the frameworks. Maintaining the ability to measure change over time was an important factor in revising the frameworks.

As part of the TIMSS dissemination strategy, approximately one half of the 2003 assessment items were released for public use. To replace assessment items that had been released, countries submitted items for review by subject-matter specialists, and additional items were written by the IEA Science and Mathematics Review Committee in consultation with item-writing specialists in various countries to ensure that the content, as explicated in the frameworks, was covered adequately. Items were reviewed by an international Science and Mathematics Item Review Committee and field-tested in most of the participating countries. Results from the field test were used to evaluate item difficulty, how well items discriminated between high- and low-performing students, the effectiveness of distracters in multiple-choice items, scoring suitability and reliability for constructed-response items, and evidence of bias toward or against individual countries or in favor of boys or girls. As a result of this review, 196 new fourth-grade items were selected for inclusion in the international assessment. In total, 353 mathematics and science items were included in the fourth-grade TIMSS assessment booklets. At the eighth grade, the review of the item statistics from the field test led to the inclusion 240 new eighth-grade items in the assessment. In total, 429 mathematics and science items were included in the eighth-grade TIMSS assessment booklets. More detail on the distribution of new and trend items is included in table A-3.

Table A-3. Number of new and trend mathematics and science items in the TIMSS grade four and grade eight assessments, by type: 2007

	Grade four					
	All items		New items		Trend items	
	Number	Percent	Number	Percent	Number	Percent
All items						
Total	353	100	196	100	157	100
Multiple choice	189	54	108	55	81	52
Constructed response	164	46	88	45	76	48
Mathematics items						
Total	179	100	98	100	81	100
Multiple choice	96	54	55	56	41	51
Constructed response	83	46	43	44	40	49
Science items						
Total	174	100	98	100	76	100
Multiple choice	93	53	53	54	40	53
Constructed response	81	47	45	46	36	47

	Grade eight					
	All items		New items		Trend items	
	Number	Percent	Number	Percent	Number	Percent
All items						
Total	429	100	240	100	189	100
Multiple choice	224	52	117	49	107	57
Constructed response	205	48	123	51	82	43
Mathematics items						
Total	215	100	120	100	95	100
Multiple choice	117	54	61	51	56	59
Constructed response	98	46	59	49	39	41
Science items						
Total	214	100	120	100	94	100
Multiple choice	107	50	56	47	51	54
Constructed response	107	50	64	53	43	46

SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS) 2007.

Design of instruments

TIMSS 2007 included booklets containing assessment items as well as self-administered background questionnaires for principals, teachers, and students.

Assessment booklets

The assessment booklets were constructed such that not all of the students responded to all of the items. This is consistent with other large-scale assessments, such as the NAEP. To keep the testing burden to a minimum, and to ensure broad subject-matter coverage, TIMSS used a rotated block design that included both mathematics and science items. That is, students encountered both mathematics and science items during the assessment.

The 2007 fourth-grade assessment consisted of 14 booklets, each requiring approximately 72 minutes of response time. To ensure that TIMSS 2007 maintains the trend, and to provide for a correction through equating, if necessary, four additional "bridge" booklets were required but only for countries that participated in TIMSS 2003.[17] These bridge study booklets were identical to booklets used in 2003. Performance on the bridge booklets did not contribute to the overall score for TIMSS 2007 but the data were used in the trend scaling that placed the 2007 results on the same scale as previous TIMSS assessments and so allowed for comparisons across the years.

For the United States and other countries participating in the 2003 assessment, this meant a total of 18 booklets. The 18 booklets were rotated among students, with each participating student completing 1 booklet only. The mathematics and science items were each assembled separately into 14 blocks, or clusters, of items. Each block contained either mathematics items or science items only. The secure, or trend, items used in prior assessments were included in 3 blocks, with the other 11 blocks containing new items. Each of the 14 TIMSS 2007 booklets contained 4 blocks in total. The 4 additional bridge study booklets from TIMSS 2003 contained 6 blocks of items each.

The 2007 eighth-grade assessment followed the same pattern and consisted of 18 booklets, each requiring approximately 90 minutes of response time. The 18 booklets were rotated among students, with each participating student completing 1 booklet only. The mathematics and science items were assembled into 14 blocks, or clusters, of items. Each block contained either mathematics items or science items only. The secure, or trend, items used in prior assessments were included in 3 blocks, with the other 11 blocks containing new items. Each of the 14 TIMSS 2007 booklets contained 4 blocks in total. The 4 additional bridge study booklets from TIMSS 2003 contained 6 blocks of items each. Performance on the bridge booklets did not contribute to the overall score for TIMSS 2007 but the data were used in the trend scaling that placed the 2007 results on the same scale as previous TIMSS assessments and so allowed for comparisons across the years.

As part of the design process, it was necessary to ensure that the booklets showed a distribution across the mathematics and science content domains as specified in the frameworks. The number of mathematics and science items in the fourth- and eighth-grade TIMSS 2007 assessments is shown in table A-4.

Table A-4. Number of mathematics and science items in the TIMSS grade four and grade eight assessments, by type and content domain: 2007

Grade four			
		Response type	
Content domain	Total	Multiple choice	Constructed response
Total	353	189	164
Mathematics	179	96	83
Number	78	50	28
Geometric shapes and measures	44	32	12
Data display	97	14	83
Science	174	93	81
Life science	74	42	32
Physical science	64	35	29
Earth science	36	16	20

Grade eight			
		Response type	
Content domain	Total	Multiple choice	Constructed response
Total	429	224	205
Mathematics	215	117	98
Number	63	35	28
Algebra	64	34	30
Geometry	47	31	16
Data and chance	41	17	24
Science	214	107	107
Biology	76	36	40
Chemistry	42	21	21
Physics	55	31	24
Earth science	41	19	22

SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007.

[17] A detailed description of the bridge study and the use of the data obtained through the bridge booklets in scaling the 2007 assessment can be found in the *TIMSS 2007 Technical Report* (Olson, Martin, and Mullis 2008).

Background questionnaires

As in prior administrations of TIMSS, TIMSS 2007 included self-administered questionnaires for principals, teachers, and students. To create the questionnaires for 2007, the 2003 versions were reviewed extensively by the national research coordinators from the participating countries as well as a Questionnaire Item Review Committee (QIRC). Based on this review, the QIRC deleted or revised some questions, and added several new ones. Like the assessment items, all questionnaire items were field tested, and the results reviewed carefully. As a result, some of the questionnaire items needed to be revised prior to their inclusion in the final questionnaires. The questionnaires requested information to help provide a context for the performance scores, focusing on such topics as students' attitudes and beliefs about learning, their habits and homework, and their lives both in and outside of school; teachers' attitudes and beliefs about teaching and learning, teaching assignments, class size and organization, instructional practices, and participation in professional development activities; and principals' viewpoints on policy and budget responsibilities, curriculum and instruction issues and student behavior, as well as descriptions of the organization of schools and courses. Detailed results from the student, teacher, and school surveys are not discussed in this report but are available in the two international reports: the *TIMSS 2007 International Mathematics Report* (Mullis, Martin, and Foy 2008) and *TIMSS 2007 International Science Report* (Martin, Mullis, and Foy 2008).

Calculator usage

Calculators were not permitted during the TIMSS fourth-grade assessment. However, the TIMSS policy on calculator use at the eighth grade was to give students the best opportunity to operate in settings that mirrored their classroom experiences. Calculators were permitted but not required for the eighth-grade assessment materials. In the United States, students assigned one of the 14 TIMSS 2007 booklets were allowed, but not required, to use calculators. However, students assigned one of the trend booklets from the 2003 assessment were required to follow the 2003 rules in this respect. These students could use a calculator only for the second half of the booklet.

Translation

Source versions of all instruments (assessment booklets, questionnaires, and manuals) were prepared in English and translated into the primary language or languages of instruction in each country. In addition, it was sometimes necessary to adapt the instrument for cultural purposes, even in countries that use English as the primary language of instruction. All adaptations were reviewed and approved by the International Study Center to ensure they did not change the substance or intent of the question or answer choices. For example, proper names were sometimes changed to names that would be more familiar to students (e.g., Marja-leena to Maria).

Each country prepared translations of the instruments according to translation guidelines established by the International Study Center. Adaptations to the instruments were documented by each country and submitted for review. The goal of the translation guidelines was to produce translated instruments of the highest quality that would provide comparable data across countries.

Translated instruments were verified by an independent, professional translation agency prior to final approval and printing of the instruments. Countries were required to submit copies of the final printed instruments to the International Study Center. Further details on the translation process can be found in the *TIMSS 2007 Technical Report* (Olson, Martin, and Mullis 2008).

Recruitment, test administration, and quality assurance

TIMSS 2007 emphasized the use of standardized procedures in all countries. Each country collected its own data, based on comprehensive manuals and trainings provided by the international project team to explain the survey's implementation, including precise instructions for the work of school coordinators and scripts for test administrators to use in testing sessions.

Recruitment of schools and students

With the exception of private schools, the recruitment of schools required several steps. Beginning with the sampled schools, the first step entailed obtaining permission from the school district to approach the sampled school(s) in that district. If a district refused permission, then the district of the first substitute school was approached and the procedure was repeated. With permission from the district, the school(s) was contacted in a second step. If a sampled school refused to participate, the district of the first substitute was approached and the permission procedure repeated. During most of the recruitment period sampled schools and substitute schools were being recruited concurrently. Each participating school was asked to nominate a School Coordinator as the main point of contact for the study. The school coordinator worked with project staff to arrange logistics and liaise with staff, students and parents as necessary.

On the advice of the school, parental permission for students to participate was sought with one of three approaches to parents: a simple notification; a notification with a refusal form; and a notification with a consent form for parents to sign. In each approach, parents were informed that their students could opt out of participating.

Gifts to schools, School Coordinators, and students. Schools, School Coordinators, and students were provided with small gifts as a sign of appreciation for their willingness to participate. Schools were provided with an all-in-one printer/photocopier/scanner/fax, School Coordinators received a TIMSS satchel, and students were given a clock-compass carabiner.

Test administration

Test administration in the United States was carried out by professional staff trained according to the international guidelines. School personnel were asked only to assist with listings of students, identifying space for testing in the school, and specifying any parental consent procedures needed for sampled students.

Quality assurance

The International Study Center monitored compliance with the standardized procedures. National research coordinators 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 2-day training sessions about TIMSS, the responsibilities of the national centers in conducting the study, and their own roles and responsibilities. Some 30 schools in the U.S. samples were visited by the monitors—15 of the 257 schools in the fourth-grade sample, and 15 of the 239 schools in the eighth-grade sample. These schools were scattered geographically across the nation. In addition, each country conducted its own separate quality control procedures.

Scoring and scoring reliability

The TIMSS assessment items included both multiple-choice and constructed-response items. A scoring rubric (guide) was created for every item included in the TIMSS assessments. The rubrics were carefully written and reviewed by national research coordinators and other experts as part of the field test of items, and revised accordingly.

The national research coordinator in each country was responsible for the scoring and coding of data in that country, following established guidelines. The national research coordinator and, sometimes, additional staff attended scoring training sessions held by the International Study Center. The training sessions focused on the scoring rubrics and coding system employed in TIMSS. Participants in these training sessions were provided extensive practice in scoring example items over several days. Information on within-country agreement among coders was collected and documented by the International Study Center. Information on scoring and coding reliability was also used to calculate cross-country agreement among coders. Information on scoring reliability for constructed-response scoring in TIMSS 2007 is provided in table A-5.

Data entry and cleaning

The national research coordinator from each country oversaw data entry. The data collected for TIMSS 2007 were entered into data files with a common international format, as specified in the *Data Entry Manager Manual* (IEA Data Processing Center 2006), which accompanied data entry software (WinDEM) available to all participating countries. The software facilitated the checking and correction of data by providing various data consistency checks. The data were then sent to the IEA Data Processing Center (DPC) in Hamburg, Germany, for cleaning. The DPC checked that the international data structure was followed; checked the identification system within and between files; corrected single case problems manually; and applied standard cleaning procedures to questionnaire files. Results of the data cleaning process were documented by the DPC. This documentation was shared with the national research coordinator with specific questions to be addressed. The national research coordinator then provided the DPC with revisions to coding or solutions for anomalies. The DPC subsequently compiled background univariate statistics and preliminary test scores based on classical and Rasch item analyses. Detailed information on the entire data entry and cleaning process can be found in the *TIMSS 2007 Technical Report* (Olson, Martin, and Mullis 2008).

Weighting, scaling, and plausible values

Before the data were analyzed, responses from the groups of students assessed were assigned sampling weights to ensure that their representation in TIMSS 2007 results matched their actual percentage of the school population in the grade assessed. With these sampling weights in place, the analyses of TIMSS 2007 data proceeded in two phases: scaling and estimation. During the scaling phase, item response theory (IRT) procedures were used to estimate the measurement characteristics of each assessment question. During the estimation phase, the results of the scaling were used to produce estimates of student achievement. Subsequent analyses related these achievement results to the background variables collected by TIMSS 2007.

Weighting

Responses from the groups of students were assigned sampling weights to adjust for over- or under-representation during the sampling of a particular group. The use of sampling weights is necessary for the computation of sound, nationally representative estimates. The weight assigned to a student's responses is the inverse of the probability that the student

Table A-5. Within-country constructed-response scoring reliability for TIMSS grade four and grade eight mathematics and science items, by exact percent score agreement and country: 2007

	Grade four					
	Mathematics			Science		
	Average	Range		Average	Range	
Country	across items	Min	Max	across items	Min	Max
TIMSS average	98	88	100	96	81	100
Algeria	92	58	99	88	69	98
Armenia	99	94	100	98	93	100
Australia	100	98	100	99	95	100
Austria	99	95	100	98	90	100
Chinese Taipei	98	84	100	97	74	100
Colombia	99	93	100	98	50	100
Czech Republic	98	90	100	94	78	100
Denmark	97	83	100	91	72	100
El Salvador	99	96	100	99	78	100
England	99	91	100	98	88	100
Georgia	97	88	100	92	68	100
Germany	97	75	100	93	73	100
Hong Kong SAR	100	98	100	99	98	100
Hungary	100	97	100	99	96	100
Iran, Islamic Rep. of	99	96	100	97	83	100
Italy	99	94	100	98	85	100
Japan	99	94	100	97	88	100
Kazakhstan	99	96	100	99	97	100
Kuwait	100	98	100	99	94	100
Latvia	95	41	100	85	42	100
Lithuania	98	88	100	95	80	100
Morocco	95	33	100	93	75	100
Netherlands	97	86	100	92	71	100
New Zealand	99	95	100	97	90	100
Norway	99	92	100	97	88	100
Qatar	99	91	100	99	94	100
Russian Federation	100	98	100	100	99	100
Scotland	99	91	100	97	87	100
Singapore	99	93	100	96	90	100
Slovak Republic	99	92	100	99	97	100
Slovenia	100	99	100	99	93	100
Sweden	98	89	100	93	65	100
Tunisia	98	86	100	92	77	100
Ukraine	100	98	100	100	98	100
United States	98	83	100	94	68	100
Yemen	98	83	100	96	85	100

See notes at end of table.

Table A-5. Within-country constructed-response scoring reliability for TIMSS grade four and grade eight mathematics and science items, by exact percent score agreement and country: 2007—Continued

	Grade eight					
	Mathematics			Science		
		Range			Range	
Country	Average across items	Min	Max	Average across items	Min	Max
TIMSS average	98	89	100	96	82	100
Algeria	95	60	100	94	75	100
Armenia	99	94	100	98	89	100
Australia	99	93	100	97	88	100
Bahrain	100	97	100	94	78	100
Bosnia and Herzegovina	98	90	100	95	74	100
Botswana	98	84	100	95	79	100
Bulgaria	96	70	100	91	69	100
Chinese Taipei	98	47	100	94	66	100
Colombia	99	92	100	98	88	100
Czech Republic	98	86	100	93	75	100
Egypt	99	94	100	97	88	100
El Salvador	100	98	100	100	98	100
England	99	94	100	97	88	100
Georgia	97	76	100	92	67	100
Ghana	100	98	100	99	96	100
Hong Kong SAR	99	95	100	99	96	100
Hungary	98	84	100	95	86	100
Indonesia	98	90	100	97	81	100
Iran, Islamic Rep. of	99	93	100	97	86	100
Israel	96	82	100	92	74	100
Italy	99	85	100	96	63	100
Japan	97	84	100	91	54	100
Jordan	100	97	100	99	93	100
Korea, Rep. of	99	96	100	99	95	100
Kuwait	99	96	100	99	88	100
Lebanon	100	97	100	100	97	100
Lithuania	98	94	100	97	90	100
Malaysia	99	96	100	99	96	100
Malta	97	81	100	93	81	100
Norway	99	94	100	97	88	100
Oman	99	95	100	99	95	100
Palestinian Nat'l Auth.	98	89	100	94	82	100
Qatar	99	91	100	99	95	100
Romania	99	96	100	99	89	100
Russian Federation	100	98	100	99	93	100
Saudi Arabia	100	97	100	99	90	100
Scotland	99	95	100	97	84	100
Serbia	99	94	100	97	74	100
Singapore	98	93	100	96	90	100
Slovenia	100	98	100	100	95	100
Sweden	98	86	100	92	70	100
Syrian Arab Republic	99	95	100	99	92	100
Thailand	98	89	100	90	73	100
Tunisia	97	87	100	91	61	100
Turkey	100	95	100	97	81	100
Ukraine	98	80	100	92	68	100
United States	97	86	100	93	73	100

NOTE: The reliability of constructed-reponse scoring was determined by having two scorers independently score a random sample of some 200 student responses to each item. Table A-5 displays the average and range of the within-country exact percent of inter-rater agreement across all items. To gather and document within-country agreement among scorers, systematic subsamples of at least 100 students' responses to each constructed-response item were coded independently by two readers. The agreement score indicates the degree of agreement.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007.

would be selected for the sample. When responses are weighted, none are discarded, and each contributes to the results for the total number of students represented by the individual student assessed. Weighting also adjusts for various situations (such as school and student nonresponse) because data cannot be assumed to be randomly missing. The internationally defined weighting specifications for TIMSS require that each assessed student's sampling weight should be the product of (1) the inverse of the school's probability of selection, (2) an adjustment for school-level nonresponse, (3) the inverse of the classroom's probability of selection, and (4) an adjustment for student-level nonresponse.[18] All TIMSS 1995, 1999, 2003, and 2007 analyses are conducted using sampling weights. A detailed description of this process is provided in the TIMSS Technical 2007 Report (Olson, Martin, and Mullis 2008).

Scaling

In TIMSS, scale scores were estimated for each student using an item response theory (IRT) model. With IRT the difficulty of each item is deduced using information about how likely it is for students to get some items correct versus other items. Once the difficulty of each item is determined, the ability of each student can be estimated even when different students have been administered different items. At this point in the estimation process achievement scores are expressed in a standardized logit scale which ranges from -4 to +4. In order to make the scores more meaningful and to facilitate their interpretation, the scores are transformed to a new scale with a mean of 500 and a standard deviation of 100.

The procedures TIMSS used for the analyses were developed to produce accurate results for groups of students while limiting the testing burden on individual students. Furthermore, these procedures provided data that could be readily used in secondary analyses. IRT scaling provides estimates of item parameters (e.g., difficulty, discrimination) that define the relationship between the item and the underlying variable measured by the test. Parameters of the IRT model are estimated for each test question, with an overall scale being established as well as scales for each content area and cognitive domain specified in the assessment framework. For example, the TIMSS 2007 eighth-grade assessment had four scales describing four mathematics content areas and four science content areas, as well as three cognitive domains in each of mathematics and science.

In order to allow for the calculation of trends in achievement, comparisons of scores were necessary across the four TIMSS assessments conducted in 1995, 1999, 2003 and 2007. IRT estimation procedures were used to place scores from the multiple administrations on the same scale (the scale of the 1995 administration). This is made possible by the inclusion of common test items in successive administrations. This allows comparison of item parameters (such as the relative difficulty of items compared with each other and how well individual items predict overall scores) across administrations. This comparison of item parameters is used to drop items whose item parameters change dramatically across administrations and to equate scales across years. It is important to note that the item parameters do not depend directly on the average ability level of the students tested, though they may depend on the range of abilities among students tested (for example, to determine which of two difficult items is more difficult, it is important to test students of sufficient ability to get at least one of the items correct). Therefore, even if the average ability levels of students in countries participating in TIMSS over time changes, the scales still can be equated across administrations.

In TIMSS, scales are equated across administrations by linking the data from each administration to the data from the administration that preceded it, as follows. Data for students in adjacent assessments are pooled together and scaled using IRT to determine the difficulty and discrimination of each item. This puts the scores from adjacent assessments on the same scale. The achievement scores estimated from the new item parameters are then put on the original 1995 TIMSS metric by a linear transformation.

For example, in order to allow an examination of trends in eighth-grade achievement between 1995 and 1999, the TIMSS 1999 eighth-grade data were placed on the 1995 TIMSS scale by first scaling the 1995 and 1999 data for countries that participated in both years together to determine the item parameters. Ability estimates for all students (those assessed in 1995 and those assessed in 1999) based on the new item parameters were then estimated. In order to put these jointly calibrated 1995 and 1999 scores on the 1995 metric, a linear transformation is applied. This transformation is designed to give the jointly calibrated 1995 scores the same mean and standard deviation as the original 1995 scores that were reported in the 1995 assessment cycle. Once this linear transformation is established it is applied to the 1999 assessment scores for all countries participating in 1999. This puts the 1999 scores on the 1995 (longitudinal) metric while preserving any growth that has occurred between assessments.

Following this same procedure, TIMSS 2003 scores were jointly calibrated with the 1999 scores to place them on the same (1995) metric and, finally, TIMSS 2007 scores were jointly calibrated with the 2003 scores to place these on the same (1995) metric. By linking scores for each adjacent pair of assessments, all four sets of scores are placed on the same

[18]These adjustments are for overall response rates and did not include any of the characteristics associated with differential nonresponse as identified in the nonresponse bias analyses reported above.

longitudinal scale. As a result, even if the makeup of the countries participating in TIMSS changes over time, achievement comparisons within and between countries are legitimate at a single point in time and across time.

Information obtained from the bridge study described below was incorporated into this scaling to ensure strict comparability of scores across the four assessments. Details are provided in the *TIMSS 2007 Technical Report* (Olson, Martin, and Mullis 2008).

The 2003-07 Bridge Study. As the name suggests, TIMSS places a great deal of emphasis on the measurement of trends in achievement within and between countries. TIMSS provides for the measurement of these trends across the four TIMSS assessment years (1995, 1999, 2003, and 2007) by placing the scores from each assessment on the same scale. However, the TIMSS assessment design changed a little in 2007, and it was considered prudent to devise a procedure to measure the effect of this change, if any, on the comparability of the 2007 assessment scores with those from previous years. Given an effect, the intent was to incorporate a correction into the scaling procedures which establish the comparability of the 2007 achievement scores with those from 1995, 1999, and 2003.

In order to evaluate the effect of the change in assessment design in TIMSS 2007, a bridge study was incorporated into the main survey to allow a comparison of the 2007 assessment with the 2003 assessment. Countries that participated in TIMSS 2003 were asked to include four additional booklets from 2003 in with the 14 booklets for TIMSS 2007 at each grade. As a result, sample sizes needed to be increased to ensure that the number of students taking each booklet was sufficient for the purposes of scaling.

The findings from the bridge study indicated a small effect from the change in the assessment design. To accommodate this, a correction was introduced into the scaling procedures which placed the 2007 assessment scores on the same scale as the scores from the 1995, 1999 and 2003 assessments. A detailed description of the bridge study is provided in the *TIMSS 2007 Technical Report* (Olson, Martin, and Mullis 2008).

Plausible values

To keep student burden to a minimum, TIMSS administered a limited number of assessment items to each student—too few to produce accurate content-related scale scores for each student. To accommodate this situation, during the scaling process plausible values were estimated to characterize students participating in the assessment. Plausible values are imputed values and not test scores for individuals in the usual sense. In fact, they are biased estimates of the proficiencies of individual students. Plausible values do, however, provide unbiased estimates of population characteristics.

Plausible values represent what the true performance of an individual might have been, had it been observed. They are estimated as random draws (usually five) from an empirically derived distribution of score values based on the student's observed responses to assessment items and on background variables. Each random draw from the distribution is considered a representative value from the distribution of potential scale scores for all students in the sample who have similar characteristics and identical patterns of item responses. Differences between the plausible values quantify the degree of precision (the width of the spread) in the underlying distribution of possible scale scores that could have caused the observed performances.

An accessible treatment of the derivation and use of plausible values can be found in Beaton and González (1995). A more technical treatment can be found in the *TIMSS 2007 Technical Report* (Olson, Martin, and Mullis 2008).

International benchmarks

International benchmarks for achievement were developed in an attempt to provide a concrete interpretation of what the scores on the TIMSS mathematics and science achievement scales mean (for example, what it means to have a scale score of 513 or 426). To describe student performance at various points along the TIMSS mathematics and science achievement scales, TIMSS used scale anchoring to summarize and describe student achievement at four points on the mathematics and science scales—Advanced International Benchmark (625), High International Benchmark (550), Intermediate International Benchmark (475), and Low International Benchmark (400). Scale anchoring involves selecting benchmarks (scale points) on the TIMSS achievement scales to be described in terms of student performance and then identifying items that students scoring at the anchor points can answer correctly. Subsequently, these items are grouped by content area within benchmarks and reviewed by mathematics and science experts. These experts focus on the content of each item and describe the kind of mathematics or science knowledge demonstrated by students answering the item correctly. The experts then provide a summary description of performance at each anchor point leading to a content-referenced interpretation of the achievement results. Detailed information on the creation of the benchmarks is provided in the international TIMSS reports (Mullis, Martin, and Foy 2008; Martin, Mullis, and Foy 2008).

Data limitations

As with any study, there are limitations to TIMSS 2007 that researchers should take into consideration. Estimates produced using data from TIMSS 2007 are subject to two types of error—nonsampling and sampling errors.

Nonsampling errors can be due to errors made in collecting and processing data. Sampling errors can occur because the data were collected from a sample rather than a complete census of the population.

Nonsampling errors

Nonsampling error is a term used to describe variations in the estimates that may be caused by population coverage limitations, nonresponse bias, and measurement error, as well as data collection, processing, and reporting procedures. The sources of nonsampling errors are typically problems like unit and item nonresponse, the difference in respondents' interpretations of the meaning of the survey questions, response differences related to the particular time the survey was conducted, and mistakes in data preparation.

Missing data. Five kinds of missing data were identified by separate missing data codes: omitted, uninterpretable, not administered, not applicable, and not reached. An item was considered omitted if the respondent was expected to answer the item but no response was given (e.g., no box was checked in the item which asked "Are you a girl or a boy?"). Items with invalid responses (e.g., multiple responses to a question calling for a single response) were coded as *uninterpretable*. The *not administered* code was used to identify items not administered to the student, teacher or principal (e.g., those items excluded from the student's test booklet because of the BIB-spiraling of the items). An item was coded as *not applicable* when it is not logical that the respondent answer the question (e.g., when the opportunity to make the response is dependent on a filter question). Finally, items that are *not reached* were identified by a string of consecutive items without responses continuing through to the end of the assessment or questionnaire.

Missing background data on other than key variables[19] are not included in the analyses for this report and are not imputed. Item response rates for variables discussed in this report exceeded the NCES standard of 85 percent and so can be reported without notation. Of the three key variables identified in the TIMSS 2007 data for the United State—sex, race/ethnicity and the percentage of students eligible for free- or reduced-price lunch (FRPL)—as table A-6 indicates, sex has no missing responses and race/ethnicity missing responses are minimal at some 2 percent. The FRPL variable, however, has some 17 percent missing responses among the public schools in the sample and these were imputed by substituting values taken from the CCD for the schools in question. Note, however, that the CCD provides this information only for public schools. The comparable database for private schools (PPS) does not include data on participation in the FRPL program. While most private schools are ineligible for this Federal program, a few indicated that some of their students were taking part—6 of the 18 fourth-grade schools and 3 of the 14 eighth-grade schools. The reported values for these schools are included along with the zero values for schools who reported that they had no students taking part. Missing value codes then are assigned only to the 3 fourth-grade and 7 eighth-grade private schools who did not respond to the question.

Table A-6. Weighted response rates for unimputed variables for TIMSS grade four and grade eight: 2007

			Grade four		Grade eight	
Variable	Variable ID	Source of information	U.S. response rate	Range of response rates in other countries	U.S. response rate	Range of response rates in other countries
Sex	ITSEX	Classroom tracking form	100	99.5 - 100[1]	100	100
Race/ethnicity	STRACE	Student questionnaire	98	†	98	†
Free or reduced-price lunch	FRLUNCH	School questionnaire	83	†	83	†

†Not applicable.
[1]All countries other than Morocco achieved 100 percent response on this variable.
NOTE: FRLUNCH variable available for public schools only.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007.

[19]Key variables include survey-specific items for which aggregate estimates are commonly published by NCES. They include, but are not restricted to, variables most commonly used in table row stubs. Key variables also include important analytic composites and other policy-relevant variables that are essential elements of the data collection. For example, the National Assessment of Educational Progress (NAEP) consistently uses gender, race-ethnicity, urbanicity, region, and school type (public/private) as key reporting variables.

Sampling errors

Sampling errors arise when a sample of the population, rather than the whole population, is used to estimate some statistic. Different samples from the same population would likely produce somewhat different estimates of the statistic in question. This fact means that there is a degree of uncertainty associated with statistics estimated from a sample. This uncertainty is referred to as sampling variance and is usually expressed as the standard error of a statistic estimated from sample data. The approach used for calculating standard errors in TIMSS was Jackknife Repeated Replication (JRR). Standard errors can be used as a measure for the precision expected from a particular sample. Standard errors for all of the reported estimates are included in appendix C.

Confidence intervals provide a way to make inferences about population statistics in a manner that reflects the sampling error associated with the statistic. Assuming a normal distribution, the population value of this statistic can be inferred to lie within the confidence interval in 95 out of 100 replications of the measurement on different samples drawn from the same population.

That is, there is a 95 percent chance that the population value of the statistic lies within the range of 1.96 times the standard error above or below the estimated score. For example, the average mathematics score for the U.S. eighth-grade students was 508 in 2007, and this statistic had a standard error of 2.8. Therefore, it can be stated with 95 percent confidence that the actual average of U.S. eighth-grade students in 2007 was between 503 and 514 (1.96 x 2.8 = 5.5; confidence interval = 508 +/- 5.5).

Description of background variables

The international versions of the TIMSS 2007 student, teacher, and school questionnaires are available at http://timss.bc.edu. The U.S. versions of these questionnaires are available at http://nces.ed.gov/timss.

Race/ethnicity

Students' race/ethnicity was obtained through student responses to a two-part question. Students were asked first whether they were Hispanic or Latino, and then whether they were members of the following racial groups: American Indian or Alaska Native; Asian; Black or African American; Native Hawaiian or other Pacific Islander; or White. Multiple responses to the race classification question were allowed. Results are shown separately for Blacks, Hispanics, Whites, Asians and Mixed-Race as distinct groups. The small numbers of students indicating that they were American Indian or Alaska Native or Native Hawaiian or other Pacific Islander were combined into a group labeled "Other." This category is treated as a residual category and is not reported separately in the analyses.

Poverty level in public schools (percentage of students eligible for free or reduced-price lunch)

The poverty level in public schools was obtained from principals' responses to the school questionnaire. The question asked the principal to report, as of approximately the first of October 2006, the percentage of students at the school eligible to receive free or reduced-price lunch through the National School Lunch Program. The answers were grouped into five categories: less than 10 percent; 10 to 24.9 percent; 25 to 49.9 percent; 50 to 74.9 percent; and 75 percent or more. Analysis was limited to public schools only. Missing data on this variable were replaced with measures taken from the CCD. The effect of this replacement on the confidentiality of the data was examined as part of the confidentiality analyses described in the following section.

Confidentiality and disclosure limitations

In accord with NCES standard 4-2-6 (National Center for Education Statistics 2002), confidentiality analyses for the United States were implemented to provide reasonable assurance that public-use data files issued by the IEA would not allow identification of individual U.S. schools or students when compared against publicly available data collections. Disclosure limitations included the identification and masking of potential disclosure risks for TIMSS schools and adding an additional measure of uncertainty of school, teacher, and student identification through random swapping of a small number of data elements within the student, teacher, and school files.

Statistical procedures

Tests of significance

Comparisons made in the text of this report were tested for statistical significance. For example, in the commonly made comparison of country averages against the average of the United States, tests of statistical significance were used to establish whether or not the observed differences from the U.S. average were statistically significant. The estimation of the standard errors that are required in order to undertake the tests of significance is complicated by the complex sample and assessment designs, both of which generate error variance. Together they mandate a set of statistically complex procedures in order to estimate the correct standard errors. As a consequence, the estimated standard errors contain a sampling variance component estimated by the jackknife repeated replication (JRR) procedure; and, where the assessments are concerned, an additional imputation variance component arising from the assessment design. Details on the procedures used can be found in the *WesVar 5.0 User's Guide* (Westat 2007).

In almost all instances, the tests for significance used were standard *t* tests.[20] These fell into two categories according to the nature of the comparison being made: comparisons of independent and nonindependent samples. Before describing the t tests used, some background on the two types of comparisons is provided below.

The variance of a difference is equal to the sum of the variances of the two initial variables minus two times the covariance between the two initial variables. A sampling distribution has the same characteristics as any distribution, except that units consist of sample estimates and not observations. Therefore,

$$\sigma^2_{(\hat{\mu}_X - \hat{\mu}_Y)} = \sigma^2_{(\hat{\mu}_X)} + \sigma^2_{(\hat{\mu}_Y)} - 2\mathrm{cov}(\hat{\mu}_X, \hat{\mu}_Y)$$

The sampling variance of a difference is equal to the sum of the two initial sampling variances minus two times the covariance between the two sampling distributions on the estimates.

If one wants to determine whether girls' performance differs from boys' performance, for example, then, as for all statistical analyses, a null hypothesis has to be tested. In this particular example, it consists of computing the difference between the boys' performance mean and the girls' performance mean (or the inverse). The null hypothesis is

$$H_0 : \hat{\mu}_{(boys)} - \hat{\mu}_{(girls)} = 0$$

To test this null hypothesis, the standard error on this difference is computed and then compared to the observed difference. The respective standard errors on the mean estimate for boys and girls ($\sigma_{(\hat{\mu}_{boys})}$, $\sigma_{(\hat{\mu}_{girls})}$) can be easily computed.

The expected value of the covariance will be equal to 0 if the two sampled groups are independent. If the two groups are not independent, as is the case with girls and boys attending the same schools within a country, or comparing a country mean with the international mean that includes that particular country, the expected value of the covariance might differ from 0.

In TIMSS, country samples are independent. Therefore, for any comparison between two countries, the expected value of the covariance will be equal to 0, and thus the standard error on the estimate is

$$\sigma_{(\hat{\theta}_i - \hat{\theta}_j)} = \sqrt{\sigma^2_{(\hat{\theta}_i)} + \sigma^2_{(\hat{\theta}_j)}}$$

with θ being any statistic.

Within a particular country, any subsamples will be considered as independent only if the categorical variable used to define the subsamples was used as an explicit stratification variable.

If sampled groups are not independent, the estimation of the covariance between, for instance, $\hat{\mu}_{(boys)}$ and $\hat{\mu}_{(girls)}$ would require the selection of several samples and then the analysis of the variation of $\hat{\mu}_{(boys)}$ in conjunction with $\hat{\mu}_{(girls)}$. Such a procedure is, of course, unrealistic. Therefore, as for any computation of a standard error in TIMSS, replication methods using the supplied replicate weights are used to estimate the standard error on a difference. Use of the replicate weights implicitly incorporates the covariance between the two estimates into the estimate of the standard error on the difference.

Thus, in simple comparisons of independent averages, such as the U.S. average with other country averages, the following formula was used to compute the *t* statistic:

$$t = \frac{(est_1 - est_2)}{\sqrt{(se_1)^2 + (se_2)^2}}$$

Est_1 and est_2 are the estimates being compared (e.g., average of country A and the U.S. average), and se_1 and se_2 are the corresponding standard errors of these averages.

The second type of comparison used in this report occurred when comparing differences of nonsubset, nonindependent groups (e.g., when comparing the average scores of males versus females within the United States). In such comparisons, the following formula was used to compute the *t* statistic:

$$t = \frac{(est_{grp1} - est_{grp2})}{se(est_{grp1} - est_{grp2})}$$

Est_{grp1} and est_{grp2} are the nonindependent group estimates being compared. $Se(est_{grp1} - est_{grp2})$ is the standard error of the difference calculated using a JRR procedure, which accounts for any covariance between the estimates for the two nonindependent groups.

[20] Adjustments for multiple comparisons were not applied in any of the t-tests undertaken.

Effect size

Tests of statistical significance are, in part, influenced by sample sizes. To provide the reader with an increased understanding of the importance of the significant difference between student populations in the United States, effect sizes are included in the report. Effect sizes use standard deviations, rather than standard errors and, therefore, are not influenced by the size of the student population samples. Following Cohen (1988) and Rosnow and Rosenthal (1996), effect size is calculated by finding the difference between the means of two groups and dividing that result by the pooled standard deviation of the two groups:

$$d = \frac{est_{grp1} - est_{grp2}}{sd_{pooled}}$$

Est_{grp1} and est_{grp2} are the student group estimates being compared. Sd_{pooled} is the pooled standard deviation of the groups being compared. The formula for the pooled standard deviation is as follows (Rosnow and Rosenthal 1996):

$$sd_{pooled} = \sqrt{\frac{sd_1^2 + sd_2^2}{2}}$$

where sd_1 and sd_2 are the standard deviations of the groups being compared.

For example, to calculate the effect size between the 2007 fourth-grade U.S. average and Hong Kong SAR average in mathematics, the difference in the estimated averages (607-529 = 78) is divided by the pooled standard deviation. The pooled standard deviation is calculated by finding the square root of the sum of the squared standard deviations for the United States (sd = 75) and Hong Kong SAR (sd = 67) divided by 2. Using this formula, the pooled standard deviation is 71. Dividing the difference in average scores (78) by the pooled standard deviation (71) produces an effect size of 1.1.

Table A-7 shows the differences in average scores, standard deviations, pooled standard deviations, and effect sizes for the comparisons reported in figures 14 and 27. The standard deviations for all countries and U.S. student subpopulations discussed in this report are provided in tables E-18 and E-19 (mathematics) and E-37 and E-38 (science).

Table A-7. Difference between average scores, standard deviations, and pooled standard deviations used to calculate effects sizes of mathematics and sciences scores of fourth- and eighth-grade students, by country, sex, race/ethnicity, and school poverty level: 2007

Subject/grade and groups compared	Difference in average scores	Standard deviation of group 1	Standard deviation of group 2	Pooled standard deviation	Effect size
Mathematics grade four					
United States v. Hong Kong SAR	78	75	67	71	1.1
U.S. males v. U.S. females	6	77	74	76	0.1
U.S. White students v. U.S. Black students	67	68	70	69	1.0
U.S. White students v. U.S. Hispanic students	46	68	70	69	0.7
U.S. White students v. U.S. Asian students	33	68	74	71	0.5
U.S. White students v. U.S. multiracial students	15	68	84	76	0.2
U.S. public schools with lowest levels of poverty v. U.S. schools with highest levels of poverty	103	64	72	68	1.5
Mathematics grade eight					
United States v. Chinese Taipei	90	77	106	93	1.0
U.S. White students v. U.S. Black students	76	69	70	70	1.1
U.S. White students v. U.S. Hispanic students	58	69	73	71	0.8
U.S. White students v. U.S. Asian students	16	69	68	69	0.2
U.S. White students v. U.S. multiracial students	27	69	73	71	0.4
U.S. public schools with lowest levels of poverty v. U.S. schools with highest levels of poverty	92	65	74	70	1.3
Science grade four					
United States v. Singapore	48	84	93	89	0.5
U.S. White students v. U.S. Black students	79	73	76	75	1.1
U.S. White students v. U.S. Hispanic students	65	73	81	77	0.8
U.S. White students v. U.S. multiracial students	17	73	85	79	0.2
U.S. public schools with lowest levels of poverty v. U.S. schools with highest levels of poverty	113	67	81	74	1.5
Science grade eight					
United States v. Singapore	47	82	104	94	0.5
U.S. males v. U.S. females	12	85	79	82	0.1
U.S. White students v. U.S. Black students	96	70	73	72	1.3
U.S. White students v. U.S. Hispanic students	71	70	77	74	1.0
U.S. White students v. U.S. multiracial students	29	70	77	74	0.4
U.S. public schools with lowest levels of poverty v. U.S. schools with highest levels of poverty	105	68	79	74	1.4

NOTE: Difference calculated by subtracting average score of group 1 from average score of group 2. Standard deviations and pooled standard deviations are shown only for statistically significant differences between group means. The pooled standard deviation is calculated by finding the square root of the sum of the squared standard deviations for the groups being compared divided by 2, following Rosnow and Rosenthal (1996). Black includes African American. Racial categories exclude Hispanic origin. Students who identified themselves as being of Hispanic origin were classified as Hispanic, regardless of their race. High-poverty schools are those in which 75 percent or more of students are eligible for the federal free or reduced-lunch program. Low-poverty schools are those in which less than 10 percent of students are eligible. The United States met guidelines for sample participation rates only after substitutes schools were included. The National Defined Population covered 90 to 95 percent of the National Target
Population. See tables E-18 and E-19 (available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009001) for standard deviations of the U.S. and other countries' student populations in mathematics. See tables E-37 and E-38 (available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009001) for the analogous standard deviations in science.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007.

Appendix B: Example Items

Exhibit B1. Example fourth-grade mathematics item: 2007

Content Domain	Number
Cognitive Domain	Applying

M031301

Al wanted to find how much his cat weighed. He weighed himself and noted that the scale read 57 kg. He then stepped on the scale holding his cat and found that it read 62 kg.

What was the weight of the cat in kilograms?

Answer: 5 kilograms

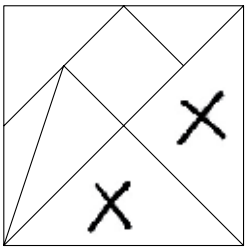
Country	Percent full credit
International average	60
Chinese Taipei	95
Singapore	87
Russian Federation	86
Hong Kong SAR[1]	86
Kazakhstan[2]	85
Netherlands[3]	85
Japan	83
Lithuania[2]	81
Austria	80
Germany	80
Latvia[2]	80
Czech Republic	76
Denmark[4]	75
Hungary	73
Slovenia	69
Italy	68
Ukraine	68
Norway	67
Sweden	66
Armenia	65
Scotland[4]	64
England	63
Australia	61
Slovak Republic	60
United States[4,5]	**60**
Georgia[2]	59
New Zealand	53
Iran, Islamic Rep. of	43
Tunisia	28
Algeria	23
El Salvador	21
Morocco	19
Colombia	18
Kuwait[6]	12
Qatar	9
Yemen	5

[1]Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.
[2]National Target Population does not include all of the International Target Population defined by TIMSS (see appendix A).
[3]Nearly satisfied guidelines for sample participation rates only after replacement schools were included (see appendix A).
[4]Met guidelines for sample participation rates only after replacement schools were included (see appendix A).
[5]National Defined Population covers 90 percent to 95 percent of National Target Population (see appendix A).
[6]Kuwait tested the same cohort of students as other countries, but late in 2007, at the beginning of the next school year.
NOTE: Countries are sorted by 2007 average percent correct. The answer shown illustrates the type of student response that was given full credit.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA). Trends in International Mathematics and Science Study (TIMSS), 2007.

Exhibit B2. Example fourth-grade mathematics item: 2007

Content Domain	Geometric Shapes and Measures
Cognitive Domain	Knowing

The square is cut into 7 pieces. Put an X on each of the 2 triangles that are the same size and shape.



M031271

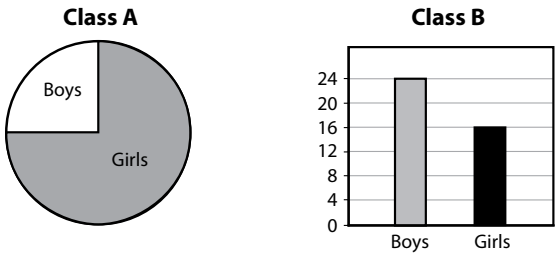
Country	Percent full credit
International average	72
Hong Kong SAR[1]	91
Slovenia	91
Lithuania[2]	89
Denmark[3]	88
Scotland[3]	88
England	88
Singapore	88
Japan	87
Italy	87
Sweden	86
Australia	85
United States[3,4]	**85**
Slovak Republic	84
Norway	84
Czech Republic	83
Austria	82
Chinese Taipei	81
Hungary	81
Latvia[2]	81
Russian Federation	81
New Zealand	81
Netherlands[5]	79
Kazakhstan[2]	77
Germany	76
Armenia	74
Ukraine	67
Colombia	59
Georgia[2]	59
Iran, Islamic Rep. of	58
El Salvador	50
Algeria	44
Kuwait[6]	40
Morocco	39
Tunisia	38
Qatar	32
Yemen	13

[1]Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.
[2]National Target Population does not include all of the International Target Population defined by TIMSS (see appendix A).
[3]Met guidelines for sample participation rates only after replacement schools were included (see appendix A).
[4]National Defined Population covers 90 percent to 95 percent of National Target Population (see appendix A).
[5]Nearly satisfied guidelines for sample participation rates only after replacement schools were included (see appendix A).
[6]Kuwait tested the same cohort of students as other countries, but late in 2007, at the beginning of the next school year.
NOTE: Countries are sorted by 2007 average percent correct. The answer shown illustrates the type of student response that was given full credit.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA). Trends in International Mathematics and Science Study (TIMSS), 2007.

Exhibit B3. Example fourth-grade mathematics item: 2007

Content Domain	Data Display
Cognitive Domain	Reasoning

Class A and B each have 40 students.



There are more girls in Class A than in Class B. How many more?

- ● 14
- Ⓑ 16
- Ⓒ 24
- Ⓓ 30

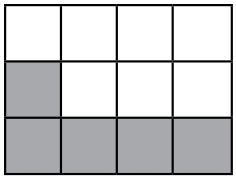
M041336

Country	Percent full credit
International average	32
Singapore	63
Hong Kong SAR[1]	63
Kazakhstan[2]	51
Chinese Taipei	47
Lithuania[2]	46
Netherlands[3]	44
Russian Federation	42
Japan	41
England	40
Slovak Republic	39
United States[4,5]	**38**
Hungary	37
Sweden	37
Latvia[2]	37
Australia	36
Slovenia	35
Germany	35
Denmark[4]	34
Scotland[4]	34
Austria	34
Armenia	33
Ukraine	32
New Zealand	32
Norway	31
Czech Republic	31
Georgia[2]	26
Italy	26
Algeria	21
Morocco	15
Iran, Islamic Rep. of	15
Tunisia	14
Qatar	13
Kuwait[6]	12
Yemen	9
El Salvador	9
Colombia	9

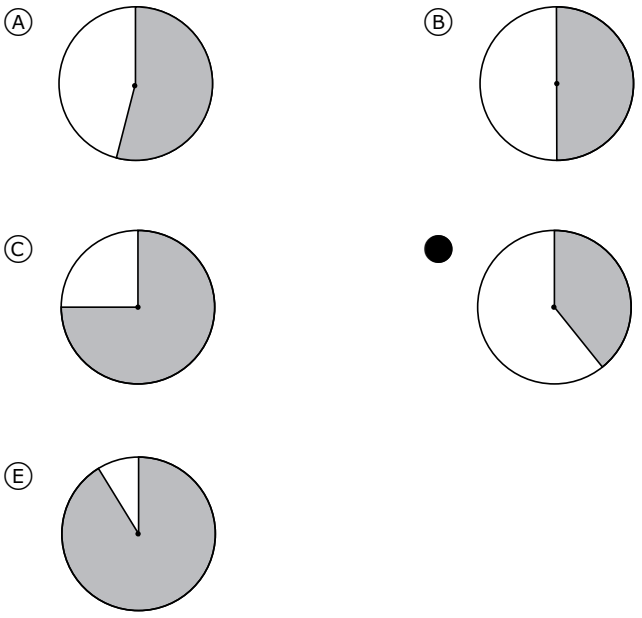
[1]Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.
[2]National Target Population does not include all of the International Target Population defined by TIMSS (see appendix A).
[3]Nearly satisfied guidelines for sample participation rates only after replacement schools were included (see appendix A).
[4]Met guidelines for sample participation rates only after replacement schools were included (see appendix A).
[5]National Defined Population covers 90 percent to 95 percent of National Target Population (see appendix A).
[6]Kuwait tested the same cohort of students as other countries, but late in 2007, at the beginning of the next school year.
NOTE: Countries are sorted by 2007 average percent correct. The answer shown illustrates the type of student response that was given full credit.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA). Trends in International Mathematics and Science Study (TIMSS), 2007.

Exhibit B4. Example eighth-grade mathematics item: 2007

Content Domain	Number
Cognitive Domain	Knowing



Which circle has approximately the same fraction of its area shaded as the rectangle above?



M022043

Country	Percent full credit
International average	63
Korea, Rep. of	89
Japan	85
Hong Kong SAR[1,2]	82
Chinese Taipei	81
United States[2,3]	**81**
Singapore	81
Sweden	77
England[2]	77
Hungary	77
Australia	75
Czech Republic	74
Lithuania[4]	74
Malaysia	74
Scotland[2]	74
Norway	73
Russian Federation	73
Slovenia	72
Malta	72
Italy	70
Cyprus	70
Thailand	68
Israel[5]	66
Turkey	64
Ukraine	63
Romania	62
Bahrain	61
Tunisia	61
Serbia[3,4]	60
Bulgaria	59
Kuwait[6]	56
Iran, Islamic Rep. of	55
Lebanon	55
Colombia	54
Algeria	54
Bosnia and Herzegovina	53
Indonesia	52
Syrian Arab Republic	51
Georgia[4]	51
Jordan	48
El Salvador	47
Oman	46
Armenia	46
Qatar	44
Egypt	44
Saudi Arabia	41
Botswana	41
Palestinian Nat'l Auth.	41
Ghana	34

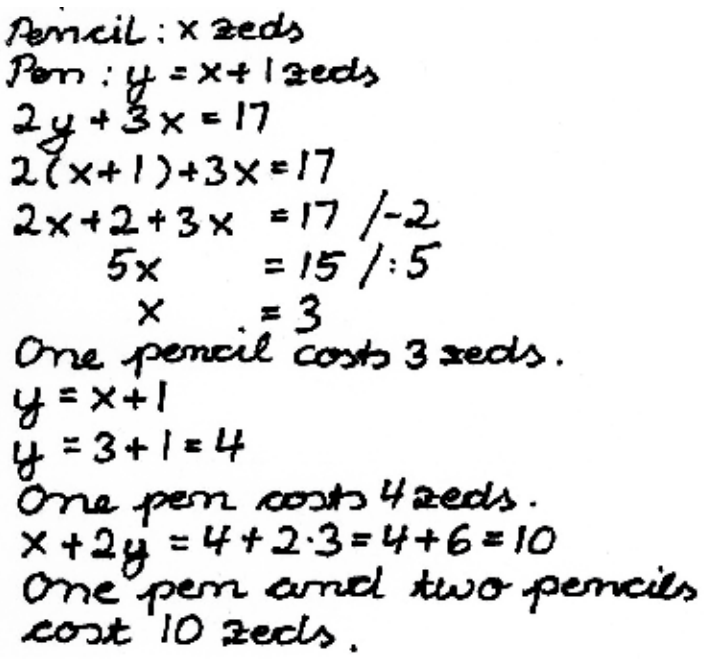
[1]Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.
[2]Met guidelines for sample participation rates only after replacement schools were included (see appendix A).
[3]National Defined Population covers 90 percent to 95 percent of National Target Population (see appendix A).
[4]National Target Population does not include all of the International Target Population defined by TIMSS (see appendix A).
[5]National Defined Population covers less than 90 percent of National Target Population (but at least 77 percent, see appendix A).
[6]Kuwait tested the same cohort of students as other countries, but later in 2007, at the beginning of the next school year.
NOTE: Countries are sorted by 2007 average percent correct. The answer shown illustrates the type of student response that was given full credit.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA). Trends in International Mathematics and Science Study (TIMSS), 2007.

Exhibit B5. Example eighth-grade mathematics item: 2007

Content Domain	Algebra
Cognitive Domain	Reasoning

Joe knows that a pen costs 1 zed more than a pencil.
His friend bought 2 pens and 3 pencils for 17 zeds.
How many zeds will Joe need to buy 1 pen and 2 pencils?

Show your work.



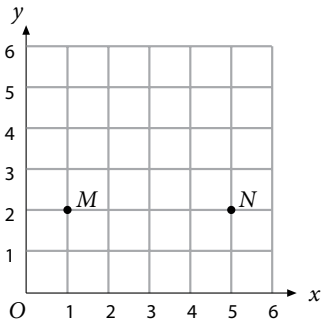
M042263

Country	Percent full credit
International average	18
Chinese Taipei	68
Korea, Rep. of	68
Singapore	59
Hong Kong SAR[1,2]	53
Japan	42
United States[2,3]	**37**
Australia	36
England[2]	34
Sweden	34
Slovenia	30
Scotland[2]	29
Czech Republic	25
Hungary	24
Israel[4]	24
Malta	21
Armenia	21
Italy	19
Russian Federation	19
Norway	18
Turkey	18
Bulgaria	17
Lithuania[5]	15
Serbia[3,5]	15
Romania	14
Malaysia	14
Thailand	13
Cyprus	11
Ukraine	11
Colombia	9
Georgia[5]	8
Indonesia	8
Bosnia and Herzegovina	8
Tunisia	6
Lebanon	5
Jordan	5
Oman	4
Bahrain	4
Iran, Islamic Rep. of	3
Saudi Arabia	3
Syrian Arab Republic	3
El Salvador	2
Algeria	2
Egypt	2
Kuwait[6]	2
Botswana	2
Qatar	2
Ghana	1
Palestinian Nat'l Auth.	1

[1]Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.
[2]Met guidelines for sample participation rates only after replacement schools were included (see appendix A).
[3]National Defined Population covers 90 percent to 95 percent of National Target Population (see appendix A).
[4]National Defined Population covers less than 90 percent of National Target Population (but at least 77 percent, see appendix A).
[5]National Target Population does not include all of the International Target Population defined by TIMSS (see appendix A).
[6]Kuwait tested the same cohort of students as other countries, but later in 2007, at the beginning of the next school year.
NOTE: Countries are sorted by 2007 average percent correct. The answer shown illustrates the type of student response that was given full credit.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA).
Trends in International Mathematics and Science Study (TIMSS), 2007.

Exhibit B6. Example eighth-grade mathematics item: 2007

Content Domain	Geometry
Cognitive Domain	Applying



Two points M and N are shown in the figure above. John is looking for a point P such that MNP is an isosceles triangle. Which of these points could be point P?

● (3,5)
Ⓑ (3,2)
Ⓒ (1,5)
Ⓓ (5,1)

M032294

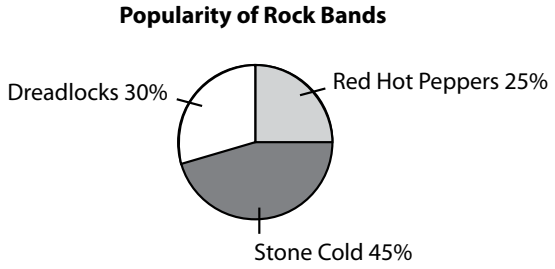
Country	Percent full credit
International average	57
Chinese Taipei	86
Korea, Rep. of	82
Japan	81
Hong Kong SAR[1,2]	80
Slovenia	80
Lithuania[3]	78
Singapore	77
Russian Federation	77
Hungary	74
Malaysia	73
Scotland[2]	68
Ukraine	68
Serbia[3,4]	67
Malta	65
Lebanon	65
Israel[5]	64
England[2]	63
Czech Republic	63
Kuwait[6]	63
Romania	62
Italy	61
Bahrain	59
Indonesia	59
Oman	59
Bulgaria	58
Syrian Arab Republic	58
Egypt	58
Norway	56
Bosnia and Herzegovina	55
Thailand	55
Jordan	54
Armenia	53
Australia	51
Cyprus	51
Algeria	50
Iran, Islamic Rep. of	49
Sweden	48
Saudi Arabia	46
United States[2,4]	**45**
Georgia[3]	41
Palestinian Nat'l Auth.	41
Turkey	38
Qatar	38
El Salvador	33
Colombia	30
Botswana	30
Tunisia	26
Ghana	26

[1]Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.
[2]Met guidelines for sample participation rates only after replacement schools were included (see appendix A).
[3]National Target Population does not include all of the International Target Population defined by TIMSS (see appendix A).
[4]National Defined Population covers 90 percent to 95 percent of National Target Population (see appendix A).
[5]National Defined Population covers less than 90 percent of National Target Population (but at least 77 percent, see appendix A).
[6]Kuwait tested the same cohort of students as other countries, but later in 2007, at the beginning of the next school year.
NOTE: Countries are sorted by 2007 average percent correct. The answer shown illustrates the type of student response that was given full credit.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA). Trends in International Mathematics and Science Study (TIMSS), 2007.

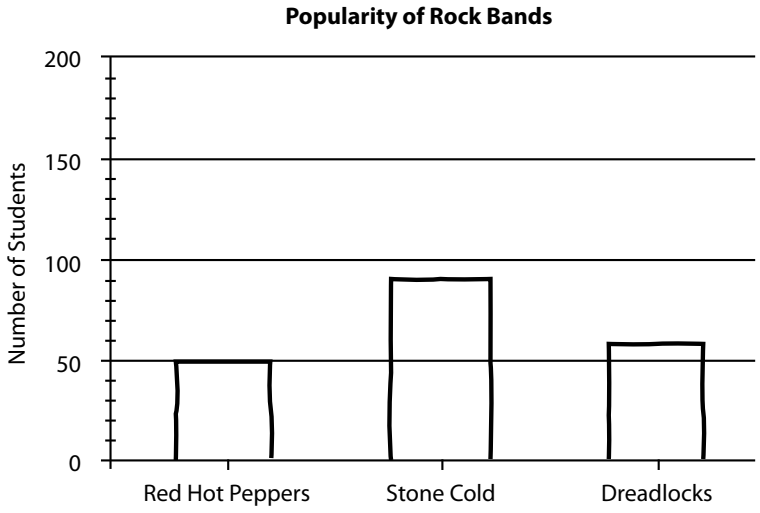
Exhibit B7. Example eighth-grade mathematics item: 2007

Content Domain	Data and Chance
Cognitive Domain	Applying

The results of a survey of 200 students are shown in the pie chart.



Make a bar chart showing the number of students in each category in the pie chart.



M04220

Country	Percent full credit
International average	27
Korea, Rep. of	76
Singapore	75
Chinese Taipei	70
Japan	68
Hong Kong SAR[1,2]	66
Sweden	56
Lithuania[3]	51
Hungary	48
Czech Republic	45
England[2]	45
Slovenia	44
Norway	41
United States[2,4]	**40**
Malta	40
Australia	38
Scotland[2]	38
Russian Federation	35
Malaysia	35
Cyprus	33
Israel[5]	31
Romania	29
Serbia[3,4]	27
Italy	27
Thailand	26
Ukraine	24
Bulgaria	23
Jordan	22
Turkey	17
Lebanon	15
Georgia[3]	15
Indonesia	14
Bosnia and Herzegovina	13
Armenia	12
Iran, Islamic Rep. of	11
Colombia	10
Egypt	10
Bahrain	9
Tunisia	8
Palestinian Nat'l Auth.	8
Botswana	7
Syrian Arab Republic	7
Oman	6
El Salvador	4
Qatar	4
Saudi Arabia	3
Algeria	3
Kuwait[6]	3
Ghana	2

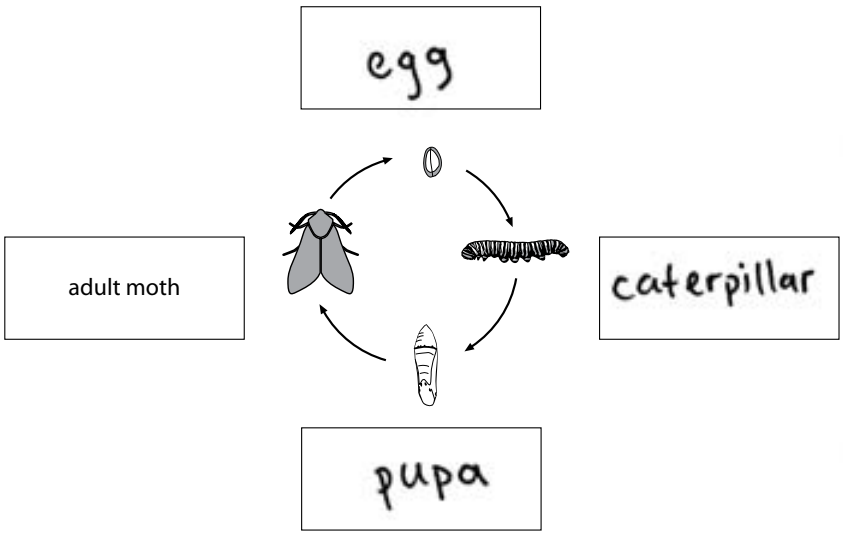
[1]Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.
[2]Met guidelines for sample participation rates only after replacement schools were included (see appendix A).
[3]National Target Population does not include all of the International Target Population defined by TIMSS (see appendix A).
[4]National Defined Population covers 90 percent to 95 percent of National Target Population (see appendix A).
[5]National Defined Population covers less than 90 percent of National Target Population (but at least 77 percent, see appendix A).
[6]Kuwait tested the same cohort of students as other countries, but later in 2007, at the beginning of the next school year.
NOTE: Countries are sorted by 2007 average percent correct. The answer shown illustrates the type of student response that was given full credit.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA).
Trends in International Mathematics and Science Study (TIMSS), 2007.

Exhibit B8. Example fourth-grade science item: 2007

Content Domain	Life Science
Cognitive Domain	Knowing

The diagram below shows the life cycle of a moth.

Write the name of each stage in the boxes provided.
One stage has been completed for you.



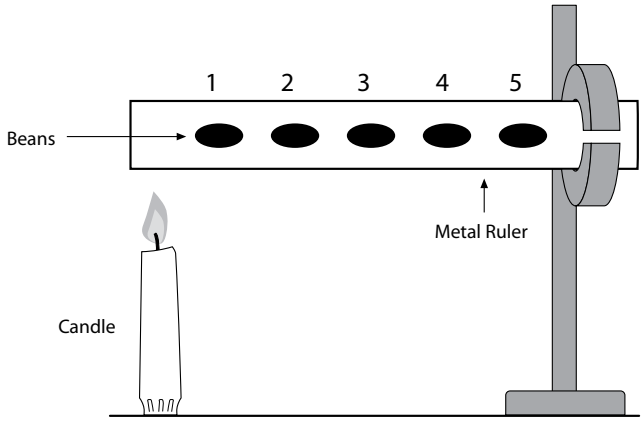
S041018

Country	Percent full credit
International average	33
Japan	93
Slovak Republic	66
Singapore	64
Chinese Taipei	61
Hungary	56
Australia	56
Sweden	53
New Zealand	52
United States[1,2]	**48**
Denmark[1]	45
Lithuania[3]	43
Czech Republic	40
Latvia[3]	39
Germany	38
Netherlands[4]	37
Austria	36
England	36
Scotland[1]	33
Kuwait[5]	32
Italy	32
Kazakhstan[3]	26
Slovenia	25
Iran, Islamic Rep. of	23
Russian Federation	23
Hong Kong SAR[6]	22
Armenia	21
Norway	20
Ukraine	18
Georgia[3]	16
Qatar	7
El Salvador	5
Colombia	4
Algeria	1
Tunisia	1
Yemen	#
Morocco	#

Rounds to zero.
[1]Met guidelines for sample participation rates only after replacement schools were included (see appendix A).
[2]National Defined Population covers 90 percent to 95 percent of National Target Population (see appendix A).
[3]National Target Population does not include all of the International Target Population defined by TIMSS (see appendix A).
[4]Nearly satisfied guidelines for sample participation rates only after replacement schools were included (see appendix A).
[5]Kuwait tested the same cohort of students as other countries, but late in 2007, at the beginning of the next school year.
[6]Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.
NOTE: Countries are sorted by 2007 average percent correct. The answer shown illustrates the type of student response that was given full credit.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA). Trends in International Mathematics and Science Study (TIMSS), 2007.

Exhibit B9. Example fourth-grade science item: 2007

Content Domain	Physical Science
Cognitive Domain	Reasoning



Beans are fixed on a metal ruler with butter as shown in the figure above. The ruler is heated at one end. In which order will the beans fall off?

- ● 1, 2, 3, 4, 5
- Ⓑ 5, 4, 3, 2, 1
- Ⓒ 1, 3, 5, 4, 2
- Ⓓ All at the same time

S031078

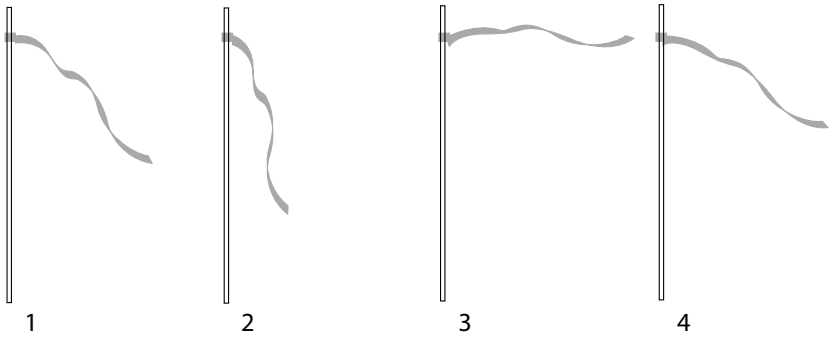
Country	Percent full credit
International average	57
Japan	92
Singapore	88
Hong Kong SAR[1]	75
Russian Federation	70
Slovenia	70
Czech Republic	69
Latvia[2]	69
Hungary	67
Kazakhstan[2]	67
England	67
United States[3,4]	**66**
Netherlands[5]	65
Chinese Taipei	65
Italy	65
Ukraine	65
Germany	64
Austria	63
Lithuania[2]	63
Slovak Republic	63
Denmark[3]	62
Australia	59
Scotland[3]	58
New Zealand	58
Armenia	56
Sweden	55
Norway	53
Georgia[2]	41
Qatar	40
Colombia	39
El Salvador	36
Algeria	35
Kuwait[6]	35
Tunisia	31
Morocco	24
Iran, Islamic Rep. of	24
Yemen	20

[1]Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.
[2]National Target Population does not include all of the International Target Population defined by TIMSS (see appendix A).
[3]Met guidelines for sample participation rates only after replacement schools were included (see appendix A).
[4]National Defined Population covers 90 percent to 95 percent of National Target Population (see appendix A).
[5]Nearly satisfied guidelines for sample participation rates only after replacement schools were included (see appendix A).
[6]Kuwait tested the same cohort of students as other countries, but late in 2007, at the beginning of the next school year.
NOTE: Countries are sorted by 2007 average percent correct. The answer shown illustrates the type of student response that was given full credit.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA). Trends in International Mathematics and Science Study (TIMSS), 2007.

Exhibit B10. Example fourth-grade science item: 2007

Content Domain	Earth Science
Cognitive Domain	Applying

A ribbon is tied to a pole to measure the wind strength as shown below.



Write the numbers 1, 2, 3, and 4 in the correct order that shows the wind strength from the **strongest** to **weakest**.

Answer : 3, 4, 1, 2

S031081

Country	Percent full credit
International average	58
Chinese Taipei	90
Singapore	88
Japan	88
Hong Kong SAR[1]	82
Australia	80
England	78
Scotland[2]	76
Latvia[3]	76
Russian Federation	75
United States[2,4]	**75**
Netherlands[5]	75
Kazakhstan[3]	74
Sweden	72
Slovak Republic	72
New Zealand	70
Italy	70
Slovenia	68
Hungary	68
Denmark[2]	68
Lithuania[3]	67
Czech Republic	64
Austria	63
Germany	57
Norway	53
Ukraine	53
Georgia[3]	49
Armenia	44
Colombia	37
Tunisia	29
Iran, Islamic Rep. of	29
Kuwait[6]	24
El Salvador	23
Qatar	20
Algeria	16
Yemen	15
Morocco	12

[1]Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.
[2]Met guidelines for sample participation rates only after replacement schools were included (see appendix A).
[3]National Target Population does not include all of the International Target Population defined by TIMSS (see appendix A).
[4]National Defined Population covers 90 percent to 95 percent of National Target Population (see appendix A).
[5]Nearly satisfied guidelines for sample participation rates only after replacement schools were included (see appendix A).
[6]Kuwait tested the same cohort of students as other countries, but late in 2007, at the beginning of the next school year.
NOTE: Countries are sorted by 2007 average percent correct. The answer shown illustrates the type of student response that was given full credit.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA). Trends in International Mathematics and Science Study (TIMSS), 2007.

Exhibit B11. Example eighth-grade science item: 2007

Content Domain	Biology
Cognitive Domain	Knowing

S032385

Which characteristic is found ONLY in mammals?

Ⓐ eyes that detect color

● glands that make milk

Ⓒ skin that absorbs oxygen

Ⓓ bodies that are protected by scales

Country	Percent full credit
International average	63
Chinese Taipei	91
Hong Kong SAR[1,2]	86
Thailand	84
Turkey	82
Syrian Arab Republic	79
Hungary	78
Lithuania[3]	76
Slovenia	76
Japan	75
Czech Republic	74
Armenia	73
Cyprus	72
Jordan	72
Saudi Arabia	72
Kuwait[4]	70
Bulgaria[5]	70
Korea, Rep. of	70
Georgia[3]	69
Israel[5]	68
Serbia[3,6]	67
Bosnia and Herzegovina	67
Bahrain	66
Romania	66
Italy	65
Russian Federation	63
Iran, Islamic Rep. of	60
Singapore	60
Lebanon	60
Algeria	58
Australia	56
Palestinian Nat'l Auth.	55
Indonesia	55
Malaysia	55
Colombia	54
Ukraine	54
Botswana	53
United States[2,6]	**53**
El Salvador	53
Sweden	53
England[2]	53
Norway	51
Qatar	49
Oman	49
Tunisia	48
Malta	44
Scotland[2]	41
Egypt	40
Ghana	31

[1]Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.
[2]Met guidelines for sample participation rates only after replacement schools were included (see appendix A).
[3]National Target Population does not include all of the International Target Population defined by TIMSS (see appendix A).
[4]Kuwait tested the same cohort of students as other countries, but later in 2007, at the beginning of the next school year.
[5]National Defined Population covers less than 90 percent of National Target Population (but at least 77 percent, see appendix A).
[6]National Defined Population covers 90 percent to 95 percent of National Target Population (see appendix A).
NOTE: Countries are sorted by 2007 average percent correct. The answer shown illustrates the type of student response that was given full credit.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA). Trends in International Mathematics and Science Study (TIMSS), 2007.

Exhibit B12. Example eighth-grade science item: 2007

Content Domain	Chemistry
Cognitive Domain	Applying

The mass of substances A and B are measured on a balance, as shown in Figure 1. Substance B is put into the beaker and substance C is formed. The empty beaker is put back on the balance, as shown in Figure 2.

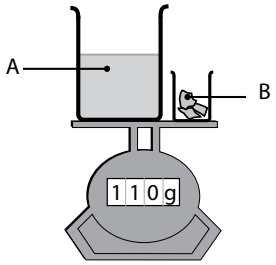


Figure 1

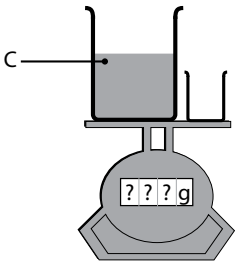


Figure 2

The scale in Figure 1 shows a mass of 110 grams.

What will it show in Figure 2?

(Check one box.)

- ☐ More than 110 grams
- ☑ 110 grams
- ☐ Less than 110 grams

Explain your answer.

The mass will be the same because the mass of reactants equals the mass of products.

S042106

Country	Percent full credit
International average	23
Japan	65
Korea, Rep. of	51
Chinese Taipei	51
Italy	46
Czech Republic	43
Slovenia	39
Hungary	39
Russian Federation	39
Sweden	38
Singapore	37
Lithuania[1]	37
Israel[2]	33
Hong Kong SAR[3,4]	30
Ukraine	29
England[4]	28
Armenia	28
Malta	27
Australia	25
Norway	25
Thailand	25
United States[4,5]	**24**
Cyprus	24
Scotland[4]	22
Tunisia	22
Romania	22
Serbia[1,5]	20
Jordan	19
Bulgaria[2]	19
Bahrain	18
Lebanon	18
Bosnia and Herzegovina	17
Colombia	16
Turkey	16
Malaysia	14
Iran, Islamic Rep. of	13
Syrian Arab Republic	13
Palestinian Nat'l Auth.	11
El Salvador	9
Oman	9
Egypt	8
Algeria	7
Kuwait[6]	7
Indonesia	6
Saudi Arabia	5
Georgia[1]	4
Qatar	3
Ghana	3
Botswana	1

[1]National Target Population does not include all of the International Target Population defined by TIMSS (see appendix A).
[2]National Defined Population covers less than 90 percent of National Target Population (but at least 77 percent, see appendix A).
[3]Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.
[4]Met guidelines for sample participation rates only after replacement schools were included (see appendix A).
[5]National Defined Population covers 90 percent to 95 percent of National Target Population (see appendix A).
[6]Kuwait tested the same cohort of students as other countries, but later in 2007, at the beginning of the next school year.
NOTE: Countries are sorted by 2007 average percent correct. The answer shown illustrates the type of student response that was given full credit.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA). Trends in International Mathematics and Science Study (TIMSS), 2007.

Exhibit B13. Example eighth-grade science item: 2007

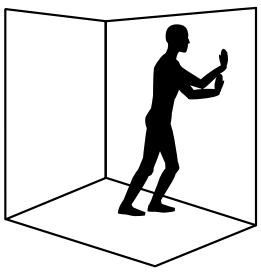
Content Domain	Physics
Cognitive Domain	Applying

Work is done when an object is moved in the direction of an applied force. A person performed different tasks as shown in the diagrams below. In which diagram is the person doing work?

Ⓐ

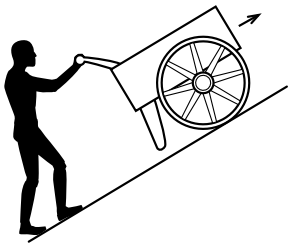
Holding a heavy object

Ⓑ



Pushing against a wall

● Pushing a cart up a ramp



Ⓓ Reading a book

S032392

Country	Percent full credit
International average	78
Singapore	96
United States[1,2]	**91**
Bulgaria[3]	91
Russian Federation	91
Korea, Rep. of	91
Hungary	90
Ukraine	90
Lithuania[4]	89
Slovenia	88
Turkey	88
Serbia[2,4]	87
Italy	87
Indonesia	86
Iran, Islamic Rep. of	86
Czech Republic	86
Australia	86
Lebanon	86
Malta	86
England[1]	85
Malaysia	84
Scotland[1]	83
Georgia[4]	82
Sweden	82
Japan	82
Chinese Taipei	81
Armenia	80
Romania	79
Syrian Arab Republic	79
Jordan	79
Bosnia and Herzegovina	78
Norway	76
Hong Kong SAR[1,5]	75
Thailand	74
Cyprus	72
Algeria	71
Israel[3]	71
Bahrain	70
Egypt	70
Colombia	70
El Salvador	68
Kuwait[6]	67
Palestinian Nat'l Auth.	65
Botswana	64
Ghana	63
Saudi Arabia	61
Oman	58
Qatar	55
Tunisia	49

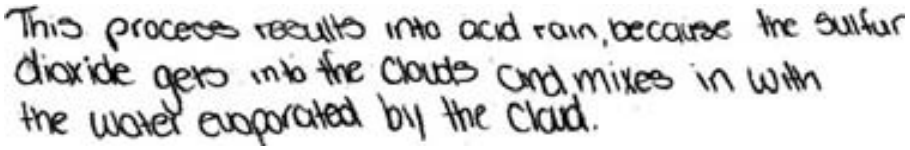
[1]Met guidelines for sample participation rates only after replacement schools were included (see appendix A).
[2]National Defined Population covers 90 percent to 95 percent of National Target Population (see appendix A).
[3]National Defined Population covers less than 90 percent of National Target Population (but at least 77 percent, see appendix A).
[4]National Target Population does not include all of the International Target Population defined by TIMSS (see appendix A).
[5]Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.
[6]Kuwait tested the same cohort of students as other countries, but later in 2007, at the beginning of the next school year.
NOTE: Countries are sorted by 2007 average percent correct. The answer shown illustrates the type of student response that was given full credit.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA). Trends in International Mathematics and Science Study (TIMSS), 2007.

Exhibit B14. Example eighth-grade science item: 2007

Content Domain	Earth science
Cognitive Domain	Reasoning

The Central City Power Station burns coal to generate power for the city. As the coal burns, sulfur that is present in the coal reacts with oxygen to form sulfur dioxide, which is given off as a gas.

How does this process result in acid rain?



S022244

Country	Percent full credit
International average	20
Korea, Rep. of	48
Singapore	47
Hong Kong SAR[1,2]	42
Lithuania[3]	42
Japan	39
Slovenia	38
England[2]	38
Chinese Taipei	35
Hungary	34
Australia	32
Jordan	30
Scotland[2]	28
Italy	27
Russian Federation	25
Czech Republic	25
Sweden	24
United States[2,4]	**23**
Bulgaria	23
Malta	22
Bosnia and Herzegovina	21
Norway	20
Armenia	20
Romania	19
Ukraine	18
Thailand	18
Bahrain	17
Israel[5]	17
Egypt	17
Serbia[3,4]	16
Malaysia	16
Iran, Islamic Rep. of	15
Syrian Arab Republic	13
Algeria	13
Georgia[3]	12
Indonesia	11
Palestinian Nat'l Auth.	11
Oman	11
Turkey	10
Lebanon	9
Saudi Arabia	8
Cyprus	7
Colombia	7
Kuwait[6]	5
Tunisia	5
El Salvador	4
Botswana	3
Ghana	3
Qatar	2

[1]Hong Kong is a Special Administrative Region (SAR) of the People's Republic of China.
[2]Met guidelines for sample participation rates only after replacement schools were included (see appendix A).
[3]National Target Population does not include all of the International Target Population defined by TIMSS (see appendix A).
[4]National Defined Population covers 90 percent to 95 percent of National Target Population (see appendix A).
[5]National Defined Population covers less than 90 percent of National Target Population (but at least 77 percent, see appendix A).
[6]Kuwait tested the same cohort of students as other countries, but later in 2007, at the beginning of the next school year.
NOTE: Countries are sorted by 2007 average percent correct. The answer shown illustrates the type of student response that was given full credit.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA). Trends in International Mathematics and Science Study (TIMSS), 2007.

Page intentionally left blank

Appendix C: TIMSS-NAEP Comparison

How Does the Content of TIMSS Compare with That of Other Assessments?

It is often asked how TIMSS compares with other assessments that measure similar subjects and populations, in particular, the National Assessment of Educational Progress (NAEP). The various assessments in which the United States participates, including NAEP, TIMSS, and the Program for International Student Assessment (PISA), vary in some obvious ways, such as the goals of the studies (and whether they are focused on national objectives or shared international objectives); the precise definitions of the populations they are measuring; the degree of precision required for estimates and resulting different sample sizes; their frameworks and specifications; and, for TIMSS and PISA, the different groups of countries that participate. However, there also are differences that are less obvious and that can only be found by comparing the content of the assessments through examination of the items.

In a recent comparison study, TIMSS 2007 mathematics and science items were classified to the NAEP assessment frameworks (2005/2007 for mathematics and 2005 for science) in terms of content topics and objectives, grade-level expectations, and cognitive dimensions in order to allow a direct comparison of the two assessments. In other studies (one past and one recent), PISA mathematics and science items also were placed on the NAEP frameworks, which allows content comparison of the TIMSS and PISA via the national frameworks. This section highlights some of the main findings; additional details on the comparison study will be included in a technical report to be released with the U.S. national TIMSS dataset at a later date.

Although the TIMSS and NAEP fourth- and eighth-grade *mathematics* frameworks are organized similarly and, broadly, cover the same range of content (e.g., number, measurement, geometry, algebra, and data), there are some differences in the relative emphases on the different topic areas between the assessments. For example, at the fourth grade, NAEP has a greater percentage of items that focus on measurement topics than does TIMSS (21 versus 14 percent, respectively), whereas TIMSS has a greater percentage of items focusing on geometry than NAEP (20 versus 16 percent, respectively). There are similar examples at the eighth-grade level among TIMSS, NAEP, and PISA, which focuses on an older group of students.

As with mathematics, the TIMSS and NAEP *science* frameworks cover the same range of major content areas, including Earth, physical (including chemistry), and life sciences. However, again, there are differences in the distribution of items even at the broad content level. These differences tend to be larger for science than for mathematics, with differences between the two assessments in the percentage of items in a given content area reaching 14 percent or more in Earth science and 8 percent or more in physical sciences at both grades. As an example, 37 percent of the TIMSS fourth-grade assessment is devoted to physical science compared to 29 percent of NAEP's fourth-grade assessment. This pattern continues at eighth grade. NAEP, on the other hand, has higher percentages of Earth science items than does TIMSS at both grades. PISA's focus (with 47 percent of items) tends to be on life science.

There is one other notable finding from the comparison study of science assessments. Twelve and 20 percent of fourth- and eighth-grade TIMSS items, respectively, could not be placed within the more detailed objectives of the NAEP framework, indicating that there are some differences at the item level between the two assessments, not just in distribution of items across content areas.

Page intentionally left blank

Appendix D: Online Resources and Publications

Online Resources

The NCES website (http://nces.ed.gov/timss) provides background information on the TIMSS surveys, copies of NCES publications that relate to TIMSS, information for educators about ways to use TIMSS in the classroom, and data files. The international TIMSS website (http://www.timss.org) includes extensive information on the study, including the international reports and databases.

NCES Publications

The following publications are intended to serve as examples of some of the numerous reports that have been produced in relation to the Trends in International Mathematics and Science Study (TIMSS) by NCES. All of the publications listed here are available at http://nces.ed.gov/timss.

TIMSS 2003 Achievement Report

Gonzales, P., Guzmán, J.C., Partelow, L., Pahlke, E., Jocelyn, L., Kastberg, D., and Williams, T. (2004). *Highlights From the Trends in International Mathematics and Science Study (TIMSS) 2003* (NCES 2005–005). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

TIMSS 1999 Achievement Reports

Gonzales, P., Calsyn, C., Jocelyn, L., Mak, K., Kastberg, D., Arafeh, S., Williams, T., and Tsen, W. (2000). *Pursuing Excellence: Comparisons of International Eighth-Grade Mathematics and Science Achievement From a U.S. Perspective, 1995 and 1999* (NCES 2001–028). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Gonzales, P., Calsyn, C., Jocelyn, L., Mak, D., Kastberg, D., Arafeh, S., Williams, T., and Tsen, W. (2000). *Highlights From TIMSS-R* (NCES 2001–027). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

TIMSS 1995 Achievement Reports

National Center for Education Statistics, U.S. Department of Education. (1997). *Pursuing Excellence: A Study of U.S. Fourth-Grade Mathematics and Science Achievement in International Context* (NCES 97–255). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Peak, L. (1996). *Pursuing Excellence: A Study of U.S. Eighth-Grade Mathematics and Science Teaching, Learning, Curriculum, and Achievement in International Context* (NCES 97–198). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Takahira, S., Gonzales, P., Frase, M., and Salganik, L.H. (1998). *Pursuing Excellence: A Study of U.S. Twelfth-Grade Mathematics and Science Achievement in International Context* (NCES 98–049). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

TIMSS Videotape Classroom Study Reports

Hiebert, J., Gallimore, R., Garnier, H., Givvin Bogard, K., Hollingsworth, H., Jacobs, J., Miu-Ying Chui, A., Wearne, D., Smith, M., Kersting, N., Manaster, A., Tseng, E., Etterbeek, W., Manaster, C., Gonzales, P., and Stigler, J. (2003). *Teaching Mathematics in Seven Countries: Results From the TIMSS 1999 Video Study* (NCES 2003–013 Revised). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

National Center for Education Statistics, U.S. Department of Education. (2000). *Highlights From the TIMSS Videotape Classroom Study* (NCES 2000–094). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

Roth, K.J., Druker, S.L., Garnier, H., Lemmens, M., Chen, C., Kawanaka, T., Rasmussen, D., Trubacova, S., Warvi, D., Okamoto, Y., Gonzales, P., Stigler, J., and Gallimore, R. (2006). *Teaching Science in Five Countries: Results From the TIMSS 1999 Video Study* (NCES 2006-011). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Stigler, J.W., Gonzales, P., Kawanaka, T., Knoll, S., and Serrano, A. (1999). *The TIMSS Videotape Classroom Study: Methods and Findings From an Exploratory Research Project on Eighth-Grade Mathematics Instruction in Germany, Japan, and the United States* (NCES 1999–074). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

IEA Publications

The following publications are intended to serve as examples of some of the numerous reports that have been produced in relation to TIMSS by the IEA. All of the publications listed here are available at http://timss.bc.edu.

TIMSS 2007 Achievement Reports

Martin, M.O., Mullis, I.V.S., and Foy, P. (2008). *TIMSS 2007 International Science Report: Findings From IEA's Trends in International Mathematics and Science Study at the Eighth and Fourth Grades*. Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., and Foy, P. (2008). *TIMSS 2007 International Mathematics Report: Findings From IEA's Trends in International Mathematics and Science Study at the Eighth and Fourth Grades*. Chestnut Hill, MA: Boston College.

TIMSS 2003 Achievement Reports

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

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

TIMSS 1999 Achievement Reports

Martin, M.O., Mullis, I.V.S., González, 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.

Mullis, I.V.S., Martin, M.O., González, 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.

TIMSS 1995 Achievement Reports

Beaton, A.E., Martin, M.O., Mullis, I.V.S., González, E.J., Smith, T.A., and Kelly, D.L. (1996). *Science Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College.

Beaton, A.E., Mullis, I.V.S., Martin, M.O., González, E.J., Kelly, D.L., and Smith, T.A. (1996). *Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College.

Martin, M.O., Mullis, I.V.S., Beaton, A.E., González, E.J., Smith, T.A., and Kelly, D.L. (1997). *Science Achievement in the Primary School Years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Beaton, A.E., González, E.J., Kelly, D.L., and Smith, T.A. (1997). *Mathematics Achievement in the Primary School Years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Beaton, A.E., González, E.J., Kelly, D.L., and Smith, T.A. (1998). *Mathematics and Science Achievement in the Final Year of Secondary School: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College.

TIMSS Technical Reports and Frameworks

Martin, M.O., and Kelly, D.L. (Eds.). (1996). *Third International Mathematics and Science Study Technical Report, Volume I: Design and Development*. Chestnut Hill, MA: Boston College.

Martin, M.O., and Kelly, D.L. (Eds.). (1998). *Third International Mathematics and Science Study Technical Report, Volume II: Implementation and Analysis, Primary and Middle School Years*. Chestnut Hill, MA: Boston College.

Martin, M.O., and Kelly, D.L. (Eds.). (1999). *Third International Mathematics and Science Study Technical Report, Volume III: Implementation and Analysis, Final Year of Secondary School*. Chestnut Hill, MA: Boston College

Martin, M.O., Gregory, K.D., and Stemler, S.E. (2000). *TIMSS 1999 Technical Report*. Chestnut Hill, MA: Boston College.

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

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

Mullis, I.V.S., Martin, M.O., Ruddock, G.J., O'Sullivan, C.Y., Arora, A., and Erberber, E. (2005). *TIMSS 2007 Assessment Frameworks*. Chestnut Hill, MA: Boston College.

Olson, J.F., Martin, M.O., and Mullis, I.V.S. (2008). *TIMSS 2007 Technical Report*. Chestnut Hill, MA: Boston College.

TIMSS Encyclopedia

Mullis, I.V.S., Martin, M.O., Olson, J.F., Berger, D.R., Milne, D., and Stanco, G.M. (Eds.). (2008). *TIMSS 2007 Encyclopedia: A Guide to Mathematics and Science Education Around the World*. Chestnut Hill, MA: Boston College.