in both mathematics and science to measure trends in student achievement since 1995. Also, 1999 represented four years since the first TIMSS, and the population of students originally assessed as fourth-graders had advanced to the eighth grade. Thus, TIMSS 1999 also provided information about whether the relative performance of these students had changed in the intervening years.

TIMSS 2003, the third data collection in the TIMSS cycle of studies, was administered at the eighth and fourth grades. For countries that participated in previous assessments, TIMSS 2003 provides three-cycle trends at the eighth grade (1995, 1999, 2003) and data over two points in time at the fourth grade (1995 and 2003). In countries new to the study, the 2003 results can help policy makers and practitioners assess their comparative standing and gauge the rigor and effectiveness of their mathematics and science programs. TIMSS 2007 will again assess mathematics and science achievement at fourth and eighth grades, providing previously participating countries an opportunity to extend their trend lines and new countries an opportunity to join a valuable and exciting endeavor.

Participants in TIMSS

Exhibit A.1 lists all the countries that have participated in TIMSS in 1995, 1999, or 2003 at fourth or eighth grade. In all, 67 countries have participated in TIMSS at one time or another. Of the 49 countries that participated in TIMSS 2003, 48 participated at the eighth grade and 26 at the fourth grade. Yemen participated at the fourth but not the eighth grade. The exhibit shows that at the eighth grade 23 countries also participated in TIMSS 1995 and TIMSS 1999. For these participants, trend data across three points in time are available. Eleven countries participated in TIMSS 2003 and TIMSS 1999 only, while three countries participated in TIMSS 2003 and TIMSS 1995. These countries have trend data for two points in time. Of the 12 new countries participating in the study, 11 participated at eighth grade and 2 at the fourth grade.

Of the 26 countries participating in TIMSS 2003 at the fourth grade, 16 also participated in 1995, providing data at two points in time.

.1 Countries Participating in TIMSS 2003, 1999, and 1995

4&8

Countries	Grade 8			Grade 4	
	2003	1	1 5	2003	1 5
1 (	●	●			
(	●			●	
(\	●	●	●	●	●
(			●		●
(	●				
\ ( )	●	●	●	●	
\ ( )			●		
	●				
\ (	●	●	●		
		●	●		●
\	●	●			
	●	●		●	
\			●		
(	●	●	●	●	●
\		●	●		●
			●		
	●				
\	●	●	●	●	●
	●				
		●			
(			●		
			●		
	●				
(			●		●
(\	●	●	●	●	●
	●	●	●	●	●
(	●	●	●	●	●
\			●		●
	●	●			
	●	●	●		●
			●		●
	●	●	●	●	●
	●				
	●	●	●	●	
	●	●			
	●	●			
	●	●		●	
	●	●		●	
	●	●	●	●	●
	●	●	●	●	●
	●		●	●	●
\ \	●				
\	●	●		●	
(\			●		●

1 2003

	2003	1	1 5	2003	1 5
	●	●	●		

Developing the TIMSS 2003 Science Assessment

The development of the TIMSS 2003 science assessment was a collaborative process spanning a two-and-a-half-year period and involving science educators and development specialists from all over the world.[2] Central to this effort was a major updating and revision of the existing TIMSS assessment frameworks to address changes during the last decade in curricula and the way science is taught. The resulting publication, entitled *TIMSS Assessment Frameworks and Specifications 2003*, serves as the basis of TIMSS 2003 and beyond.[3]

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

Content Domain

Grade 8

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

Grade 4

- Life Science
- Physical Science
- Earth Science

Cognitive Domain

- Factual Knowledge
- Conceptual Understanding
- Reasoning and Analysis

item-writing guidelines for multiple-choice and constructed-response items and provided specific training in writing science items in accordance with the *TIMSS Assessment Frameworks and Specifications 2003*. In the weeks that followed, more than 2,000 items and scoring guides were drafted and reviewed by the task force. The items were further reviewed by the Science and Mathematics Item Review Committee, a group of internationally prominent mathematics and science educators nominated by participating countries to advise on subject-matter issues in the assessment. Committee members also contributed enormously to the quality of the assessment by helping to develop tasks and items to assess problem solving and scientific inquiry.

Participating countries field-tested the items with representative samples of students, and all of the potential new items were again reviewed by the Science and Mathematics Item Review Committee. The NRCs had several opportunities to review the items and scoring criteria. The resulting TIMSS 2003 science tests contained 189 items at the eighth grade and 152 items at the fourth grade.

Exhibit A.3 presents the number and percentage of items, the number of multiple-choice and constructed-response items, and the number of score points in each of the science content domains for eighth and fourth grades. Comparable information is presented for the three cognitive domains. About two-fifths of the items at each grade level were in constructed-response format, requiring students to generate and write their own answers. Some constructed-response questions asked for short answers while others required extended responses with students showing their work or providing explanations for their answers. The remaining questions used a multiple-choice format. In scoring the items, correct answers to most questions were worth one point. However, responses to some constructed-response questions (particularly those requiring extended responses) were evaluated for partial credit, with a fully correct answer being awarded two points (see later section on scoring). The total number of score points available for analysis thus somewhat exceeds the number of items (211 and 168

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

To ensure reliable measurement of trends over time, the TIMSS 2003 assessment included items that had been used in the 1995 and 1999 assessments as well as items developed for the first time in

Distribution of Science Items by Content Domain and Cognitive Domain

Content Domain	Percentage of Items	Total Number of Items	Number of Multiple-Choice Items	Number of Constructed-Response Items[1]	Number of Score Points[2]
Life Science	2	54	2	25	65
Chemistry	16	31	20	11	34
Physics	24	46	28	18	4
Earth Science	16	31	22		33
Environmental Science	14	27	10	17	30
Total	100	18	10	80	211

Cognitive Domain	Percentage of Items	Total Number of Items	Number of Multiple-Choice Items	Number of Constructed-Response Items[1]	Number of Score Points[2]
Factual Knowledge	30	57	50	7	5
Conceptual Understanding	3	73	42	31	80
Reasoning and Analysis	31	5	17	42	72
Total	100	18	10	80	211

1

2

Life Science

Distribution of Score Points in TIMSS 2003 from Each Assessment Year by Science Content Domain

Grade 8

Content Domain	From 1995	From 1999	New in 2003	Total
Life Science	6	12	47	65
Chemistry	4	11	1	34
Physics	5	17	27	4
Earth Science	6	6	21	33
Environmental Science	3	6	21	30
Total	24	52	135	211

Grade 4

Content Domain	From 1995	From 1999	New in 2003	Total
Life Science	12	\ /	60	72
Physical Science		\ /	50	5
Earth Science	12	\ /	25	37
Total	33	\ /	135	168

TIMSS 2003 Assessment Design

Not all of the students in the TIMSS assessment responded to all of the science items. To ensure broad subject-matter coverage without overburdening individual students, TIMSS 2003, as in the 1995 and 1999 assessments, used a matrix-sampling technique that assigns each assessment item to one of a set of item blocks, and then assembles student test booklets by combining the item blocks according to a balanced design. Each student takes one booklet containing both mathematics and science items. Thus, the same students participated in both the mathematics and science testing.

Exhibit A.5 summarizes the TIMSS 2003 assessment design, presenting both the matrix-sampling item blocks for mathematics and science and the item block-to-booklet assignment plan. According to the design, the 313 mathematics and science items at fourth grade and 383 items at eighth grade are divided among 28 item blocks at each grade, 14 mathematics blocks labeled M01 through M14, and 14 science blocks labeled S01 through S14. Each block contains either mathematics items only or science items only. This general block design is the same for both grades, although the planned assessment time per block is 12 minutes for fourth grade and 15 minutes for eighth grade. At the eighth grade, six blocks in each subject (blocks 01 – 06) contain secure items from 1995 and 1999 to measure trends and eight blocks (07 – 14) contain new items developed for TIMSS 2003. Since fourth grade was not included in the 1999 assessment, trend items from 1995 only were available, and these were placed in the first three blocks. The remaining 11 blocks contain items new in 2003.

In the TIMSS 2003 design, the 28 blocks of items are distributed across 12 student booklets, as shown in Exhibit A.5. Each booklet consists of six blocks of items. To enable linking between booklets, each block appears in two, three, or four different booklets. The assessment time for individual students is 72 minutes at fourth grade (six 12-minute blocks) and 90 minutes at eighth grade (six 15-minute blocks), which is comparable to that in the 1995 and 1999 assessments. The

Booklet Design for TIMSS 2003

Student Booklet	Part I			Part II		
1	01	0	0	07	0	07
	0	0	0	0	0	0
	0	0	0	0	1	11
	0	0	0	10	1	1
	0	0	0	11	0	1
	0	01	01	1	10	1
7	01	0	0	07	0	07
	0	0	0	0	0	0
	0	0	0	0	1	11
10	0	0	0	10	1	1
11	0	0	0	11	0	1
1	0	01	01	1	10	1

Background Questionnaires

As in previous assessments, TIMSS in 2003 administered a broad array of questionnaires to collect data on the educational context for student achievement. For TIMSS 2003, a concerted effort was made to streamline and upgrade the questionnaires. This work began with articulating the information to be collected in the TIMSS 2003 framework and continued with extensive field testing.[4]

Across the two grades and two subjects, TIMSS 2003 involved 11 questionnaires. *National Research Coordinators* completed four questionnaires. With the assistance of their curriculum experts, they provided detailed information on the organization, emphasis, and content coverage of the mathematics and science curriculum at fourth and eighth grades. The *fourth- and eighth-grade students* who were tested answered questions pertaining to their attitudes towards mathematics and science, their academic self-concept, classroom activities, home background, and out-of-school activities. The *mathematics and science teachers* of sampled students responded to questions about teaching emphasis on the topics in the curriculum frameworks, instructional practices, professional training and education, and their views on mathematics and science. Separate questionnaires for mathematics and science teachers were administered at the eighth grade, while to reflect the fact that most younger students are taught all subjects by the same teacher, a single questionnaire was used at the fourth grade. The principals or heads of schools at the fourth and eighth grades responded to questions about school staffing and resources, school safety, mathematics and science course offerings, and teacher support.

4 *TIMSS 2003 Technical Report*, (2004),

Population Defi

.7 School Sample Sizes

Countries	Number of Schools in Original Sample	Number of Eligible Schools in Original Sample	Number of Schools in Original Sample That Participated	Number of Replacement Schools That Participated	Total Number of Schools That Participated
	150	150	148	0	148
	230	227	178	26	204
	150	150	133	16	14
	150	150	150	0	150
	150	150	150	0	150
	150	150	7	44	123
	150	150	116	16	132
	160	15	156	1	157
	176	171	171	0	171
	172	171	165	6	171
	150	150	150	0	150
	150	14	137	3	140
	160	160	147	6	153
	153	151	147	4	151
	227	225	1 7	0	1 7
	150	14	77	53	130
	228	228	1 4	26	220
	150	150	134	5	13
	160	160	122	13	135
	206	205	204	1	205
	150	150	4	31	125
	182	182	182	0	182
	177	177	16	5	174
	150	150	150	0	150
	310	300	212	36	248
	150	150	150	0	150
Benchmarking Participants					
	56	56	56	0	56
	200	1 6	17	10	18
	1 8	1 4	1 2	1	1 3

Countries	Within-School Student Participation (Weighted Percentage)	Number of Sampled Students in Participating Schools	Number of Students Withdrawn from Class/School	Number of Students Excluded	Number of Eligible Students	Number of Students Absent	Number of Students Assessed
	1%	6275	57	0	6218	544	5674
	4%	4675	6	3	4567	246	4321
	8%	4866	17	20	482	117	4712
	%	47 3	11	88	46 4	33	4661
	7%	4536	27	60	444	121	4328
	3%	3 17	45	0	3872	287	3585
	5%	4 01	23	4	4874	266	4608
	4%	3603	11	67	3525	206	331
	8%	4587	83	80	4424	72	4352
	7%	4641	23	185	4433	151	4282
	7%	46 0	16	16	4658	123	6. 602

151Scot8Japan94%34 3917 60 (7%)-6 62 13.407 163.70.

97%6 3600

4567

	Before Replacement	After Replacement			Before Replacement	After Replacement
	99%	99%	99%	90%	89%	89%
	81%	90%	100%	93%	75%	83%
	100%	100%	100%	98%	98%	98%
	82%	99%	98%	97%	77%	94%
	98%	98%	100%	98%	96%	96%
	97%	97%	99%	96%	92%	92%
	98%	100%	100%	99%	97%	99%
	100%	100%	100%	99%	99%	99%
	100%	100%	100%	96%	96%	96%
	99%	100%	100%	97%	97%	97%
	40%	54%	99%	86%	34%	46%
	99%	99%	100%	96%	95%	95%
	100%	100%	100%	93%	93%	93%
	74%	83%	99%	97%	72%	80%
	98%	99%	100%	95%	94%	94%
	98%	100%	100%	99%	97%	99%
	100%	100%	100%	98%	98%	98%

Participation Rates (Weighted)

Countries	School Participation		Class Participation	Student Participation	Overall Participation	
	Before Replacement	After Replacement			Before Replacement	After Replacement
	%	%	100%	1%	0%	0%
	78%	0%	100%	4%	73%	85%
( )	8 %	%	100%	8%	87%	7%
	100%	100%	100%	%	%	%
	100%	100%	100%	7%	7%	7%
	54%	82%	100%	3%	50%	76%
	77%	88%	%	5%	73%	83%
	8%	%	100%	4%	2%	3%
	100%	100%	100%	8%	8%	8%
	7%	100%	100%	7%	3%	7%
	100%	100%	100%	7%	7%	7%
	1%	4%	100%	4%	85%	88%
	2%	6%	%	2%	84%	87%
	7%	100%	100%	7%	4%	7%
	87%	87%	100%	3%	81%	81%
	52%	87%	100%	6%	50%	84%
	87%	8%	100%	5%	82%	3%
	8 %	3%	100%	5%	85%	88%
	78%	85%	100%	5%	75%	81%
	%	100%	100%	7%	6%	7%
	64%	83%	100%	2%	5 %	77%
	100%	100%	100%	8%	8%	8%
	5%	%	100%	2%	87%	1%
	100%	100%	100%	%	%	%
	70%	82%	%	5%	66%	78%
	100%	100%	100%	3%	3%	3%
Benchmarking Participants						
	100%	100%	100%	8%	8%	8%
	8 %	4%	100%	6%	85%	0%
	%	100%	100%	1%	0%	1%

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

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

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

TIMSS & PIRLS INTERNATIONAL STUDY CENTER,

of the schedule and shortages of resources, were able to conduct the data collection efficiently and professionally. Similarly, the TIMSS tests

constructed-response items in the science test for the TIMSS participants. The exhibit shows agreement for both the correctness score (the first digit) and for the two-digit diagnostic score. A high percentage of exact agreement was observed, with an overall average of 97 percent for correctness score and 92 percent for diagnostic score at the eighth grade and 96 and 92 percent, respectively at the fourth grade. The 0h

A A: A

.11 TIMSS 2003 Trend Scoring Reliability (1999–2003) for the Constructed-Response

Countries	Correctness Score Agreement			Diagnostic Score Agreement		
	3	75	100	81	56	100
	2	7	100	83	68	100
	6	87	100	83	45	100
	1	80	100	77	47	100
	2	70	100	80	38	100
	0	70		7	50	
	8	74	100	80	58	100
	2	74	100	84	64	100
	0	63	100	75	41	7
	2	68	100	82	55	
	3	80	100	81	46	100
	4	86	100	88	73	100
	2	72	100	84	62	100
	6	0	100	87	76	
	3	77	100	85	56	100
	7	36	100	65	21	8
	86	66	100	74	40	100
		8	100	8	80	100
	2	80	100	74	35	100
	4	87		7	52	8
	0	44	100	76	32	100
	6	1	100	0	73	100
	3	80	100	7	55	
	7	3	100	88	61	100
	8	73	100	76	56	100
	4	71	100	0	72	100
	3	71	100	7	1	100
	4	83	100	84	70	100
	2	75	100	81	54	100
Benchmarking Participants						
	1	76	100	81	60	100
	1	76	100	81	60	100

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

Test Reliability

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

Data Processing

To ensure the availability of comparable, high-quality data for analysis, TIMSS took rigorous quality control steps to create the international database.[11] TIMSS prepared manuals and softwareTw(d -aourolie anal0.0rs.hNlTeN)

IRT Scaling and Data Analysis

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

The TIMSS science achievement scale was designed to provide a reliable measure of student achievement spanning 1995, 1999, and 2003. The metric of the scale was established originally with the 1995 assessment. When all countries participating in 1995 at the eighth grade are treated equally, the TIMSS scale average over those countries is 500 and the standard deviation is 100. The same applies for the fourth-grade assessment. Since the countries varied in size, each country was weighted to contribute equally to the mean and standard deviation of the scale. The average and standard deviation of the scale scores are arbitrary and do not affect scale interpretation. To preserve the metric of the original 1995 scale, the 1999 eighth-grade assessment was scaled using students from the countries that participated in both 1995 and 1999. Then students from the countries that tested in 1999 but not 1995 were assigned scores on the basis of the scale.

At the eighth grade, TIMSS developed the 2003 scale in the same way as in 1999, preserving the metric first with students from countries that participated in both 1999 and 2003,[13] and then assigning scores on the basis of the scale to students tested in 2003 but not the earlier assessment. At fourth grade, because there was no assessment in 1999, the 2003 and 1995 data were linked directly together using students from countries that participated in both assessments, and the

13 1 5 1 1 5 1 - 2003 .

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

To allow more accurate estimation of summary statistics for student subpopulations, the TIMSS scaling made use of plausible-value technology, whereby five separate estimates of each student's score were generated on each scale, based on the student's responses to the items in the student's booklet and the student's background characteristics. The five score estimates are known as "plausible values," and the variability between them encapsulates the uncertainty inherent in the score estimation process.

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

Estimating Sampling Error

Because the statistics presented in this report are estimates of national performance based on samples of students, rather than the values that could be calculated if every student in every country had answered every question, it is important to have measures of the degree of uncertainty of the estimates. The jackknife procedure was used to estimate the standard error associated with each statistic presented in this report.[14] The jackknife standard errors also include an error component due to variation among the five plausible values generated for each student. The use of confidence intervals, based on the standard errors, provides a way to make inferences about the population means and proportions in a manner that reflects the uncertainty associated with the sample estimates. An estimated sample statistic plus or minus two standard errors represents a 95 percent confidence interval for the corresponding population result.

14 , . (2004), *Report*, .

Assessing Statistical Significance

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

Setting International Benchmarks of Student Achievement

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

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

the scale. Scale anchoring is a way of describing students' performance at different points on a scale in terms of what they know and can do. It involves a statistical component, in which items that discriminate between successive points on the scale are identified, and a judgmental component in which subject-matter experts examine the items and generalize to students' knowledge and understandings.[15]

15 - . (2004), (.), *TIMSS 2003 Technical Report*, 1 5 . (2000), *TIMSS International Benchmarks: U.S. Performance and Standards in an International Context*,