12

Chapter 12

Reporting Student Achievement in Mathematics and Science

Eugenio J. Gonzalez, Joseph Galia, Alka Arora, Ebru Erberber, and Dana Diaconu

12.1 Overview

The *TIMSS 2003 International Mathematics Report* (Mullis, Martin, Gonzalez, and Chrostowski, 2004) and the *TIMSS 2003 International Science Report* (Martin, Mullis, Gonzalez, and Chrostowski, 2004) summarize eighth- and fourth-grade students' mathematics and science achievement in each participating country. This chapter provides information about the international benchmarks established to help users of the achievement results understand the meaning of the achievement scales, and describes the scale anchoring procedure applied to describe student performance at these benchmarks. The chapter also describes the jackknifing technique employed by TIMSS to capture the sampling and imputation variances that follow from TIMSS' complex student sampling and booklet design, and describes how important statistics used to compare student achievement across the participating countries were calculated.

12.2 Describing International Benchmarks of Student Achievement on the TIMSS 2003 Mathematics and Science Scales[1]

It is important for users of TIMSS achievement results to understand what the scores on the TIMSS mathematics and science achievement scales mean. That is, what does it mean 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 Bench-

1 The description of the scale anchoring procedure was adapted from Kelly (1999), Gregory and Mullis (2000), and Gonzalez and Kennedy (2003).

mark (550), Intermediate International Benchmark (475), and Low International Benchmark (400).

In brief, 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 (the international benchmarks) can answer correctly. The items, so identified, are grouped by content area within benchmarks for review by mathematics and science experts. For TIMSS, the Science and Mathematics Item Replacement Committee (SMIRC) conducted the review. They examined the content of each item and described the kind of mathematics or science knowledge demonstrated by students answering the item correctly. The panelists then summarized the detailed list in a brief description of performance at each anchor point. This procedure resulted in a content referenced interpretation of the achievement results that can be considered in light of the TIMSS 2003 Mathematics and Science Frameworks.

12.2.1 Identifying the Benchmarks

Identifying the scale points to serve as benchmarks has been a challenge in the context of measuring trends. For the TIMSS 1995 and 1999 assessments, the scales were anchored using percentiles. That is, the analysis was conducted using the Top 10 percent (90^{th} percentile), the Top Quarter (75^{th} percentile), the Top Half (50^{th} percentile), and the Bottom Quarter (25^{th} percentile). However, with different participating countries in each TIMSS cycle and different achievement for countries participating in previous cycles, it was pointed out by the National Research Coordinators (NRCs) that the percentile points were changing with each cycle and that stability was required.

It was clear that TIMSS needed a set of points to serve as benchmarks, that would not change in the future, that would look sensible, and that were similar to points used in 1999. After much consideration of points used in other international (IALS and PISA) and national assessments (e.g., NAEP in the United States), it was decided to use specific scale points with equal intervals as the international benchmarks. At the TIMSS Project Management Meeting in March 2004, a set of four points on the mathematics and science achievement scales was identified to be used as the international benchmarks, namely 400, 475, 550, and 625. These points were selected to be as close as possible to the percentile points anchored in 1999 at the eighth grade (i.e., Top 10% was 616 for mathematics and science, Top Quarter was 555 for mathematics and 558 for science, Top Half was 479 for mathematics and 488 for science, and Bottom Quarter was 396 for mathematics and 410 for science). The newly defined benchmark scale points were used as the

basis for the scale anchoring descriptions. Exhibit 12.1 shows the scale scores representing each international benchmark for both grades in mathematics and science.

Exhibit 12.1 **TIMSS 2003 International Benchmarks for Eighth and Fourth Grade Mathematics and Science**

Scale Score	International Benchmark
625	Advanced International Benchmark
550	High International Benchmark
475	Intermediate International Benchmark
400	Low International Benchmark

12.2.2 Identifying the Anchor Items

After selecting the benchmark points to be described on the TIMSS 2003 mathematics and science achievement scales, the first step in the scale-anchoring procedure was to establish criteria for identifying those students scoring at the international benchmarks. Following the procedure used in previous IEA studies, a student scoring within plus and minus five scale score points of a benchmark was identified for the benchmark analysis. The score ranges around each international benchmark and the number of students scoring in each range for mathematics and science are shown in Exhibit 12.2 for the eighth grade and in Exhibit 12.3 for the fourth grade. The range of plus and minus five points around a benchmark is intended to provide an adequate sample in each group, yet be small enough so that performance at each benchmark anchor point is still distinguishable from the next. The data analysis for the scale anchoring was based on these students scoring at each benchmark range.

Exhibit 12.2 **Range around Each Anchor Point and Number of Observations within Ranges – Eighth Grade**

	Low Benchmark	Intermediate Benchmark	High Benchmark	Advanced Benchmark
Range of Scale Scores	**395 - 405**	**470 - 480**	**545 - 555**	**620 – 630**
Mathematics Students	6372	8294	6955	3320
Science Students	5633	8731	8373	3477

Exhibit 12.3 **Range around Each Anchor Point and Number of Observations within Ranges – Fourth Grade**

	Low Benchmark	Intermediate Benchmark	High Benchmark	Advanced Benchmark
Range of Scale Scores	395 - 405	470 - 480	545 - 555	620 – 630
Mathematics Students	2352	4173	5169	2481
Science Students	2408	4559	4892	2085

12.2.3 Anchoring Criteria

Having identified the number of students scoring at each benchmark anchor point, the next step was establishing criteria for determining whether particular items anchor at each of the anchor points. An important feature of the scale anchoring method is that it yields descriptions of the performance demonstrated by students reaching the benchmarks on the TIMSS mathematics and science achievement scales, and that these descriptions reflect demonstrably different accomplishments of students reaching each successively higher benchmark. The process entails the delineation of sets of items that students at each benchmark anchor point are very likely to answer correctly and that discriminate between performance at the various benchmarks. Criteria were applied to identify the items that are answered correctly by most of the students at the anchor point, but by fewer students at the next lower point.

In scale anchoring, the anchor items for each point are intended to be those that differentiate between adjacent anchor points, e.g., between the Advanced and the High international benchmarks. To meet this goal, the criteria for identifying the items must take into consideration performance at more than one anchor point. Therefore, in addition to a criterion for the percentage of students at a particular benchmark correctly answering an item, it was necessary to use a criterion for the percentage of students scoring at the next lower benchmark who correctly answer an item. For multiple choice items, the criterion of 65% was used for the anchor point, since students would be likely (about two-thirds of the time) to answer the item correctly. The criterion of less than 50% was used for the next lower point, because with this response probability, students were more likely to have answered the item incorrectly than correctly. Because there is no possibility of guessing, for constructed response items the criterion of 50% was used for the anchor point and no criterion was used for the lower points.

The criteria used to identify multiple-choice items that "anchored" are outlined below:

For the Low International Benchmark (400), a multiple-choice item anchored if

- At least 65% of students scoring in the range answered the item correctly
- Because the Low International Benchmark was the lowest one described, items were not identified in terms of performance at a lower point

For the Intermediate International Benchmark (475), a multiple-choice item anchored if

- At least 65% of students scoring in the range answered the item correctly and
- Less than 50% of students at the Low International Benchmark answered the item correctly

For the High International Benchmark (550), a multiple-choice item anchored if

- At least 65% of students scoring in the range answered the item correctly and
- Less than 50% of students at the Intermediate International Benchmark answered the item correctly

For the Advanced International Benchmark (625), a multiple-choice item anchored if

- At least 65% of students scoring in the range answered the item correctly and
- Less than 50% of students at the High International Benchmark answered the item correctly

To include all of the items in the anchoring process and provide information about content areas and cognitive processes that might not have had many items anchor exactly, items that met a slightly less stringent set of criteria were also identified. The criteria to identify multiple-choice items that "almost anchored" were the following:

For the Low International Benchmark (400), a multiple-choice item almost anchored if

- At least 60% of students scoring in the range answered the item correctly
- Because Low International Benchmark was the lowest point, items were not identified in terms of performance at a lower point

For the Intermediate International Benchmark (475), a multiple-choice item almost anchored if

- At least 60% of students scoring in the range answered the item correctly and

- Less than 50% of students at the Low International Benchmark answered the item correctly

For the High International Benchmark (550), a multiple-choice item almost anchored if

- At least 60% of students scoring in the range answered the item correctly and
- Less than 50% of students at the Intermediate International Benchmark answered the item correctly

For the Advanced International Benchmark (625), a multiple-choice item almost anchored if

- At least 60% of students scoring in the range answered the item correctly and
- Less than 50% of students at the High International Benchmark answered the item correctly

To be completely inclusive for all items, items that met only the criterion that at least 60% of the students answered correctly (regardless of the performance of students at the next lower point) were also identified. The three categories of items were mutually exclusive, and ensured that all of the items were available to inform the descriptions of student achievement at the anchor levels. A multiple-choice item was considered to be "too difficult" to anchor if less than 60% of students at the Advanced Benchmark answered the item correctly.

Different criteria were used to identify constructed-response items that "anchored." A constructed-response item anchored at one of the international benchmarks if at least 50% of students at that benchmark answer the item correctly. A constructed-response item was considered to be "too difficult" to anchor if less than 50% of students at the Advanced Benchmark answered the item correctly.

12.2.4 Computing the Item Percent Correct At Each Anchor Level

The percentage of students scoring in the range around each anchor point that answered the item correctly was computed. To compute these percentages, students in each country were weighted to contribute proportional to the size of the student population in a country. Most of the TIMSS 2003 items are scored dichotomously. For these items, the percent of students at each anchor point who answered each item correctly was computed. For constructed-response items, percentages were computed for the students receiving full credit, even if the item was scored for partial as well as full credit.

12.2.5 Identifying Anchor Items

For the TIMSS 2003 mathematics and science scales, the criteria described above were applied to identify the items that anchored, almost anchored, and met only the 60 to 65 percent criterion. Exhibit 12.4 and Exhibit 12.5 present the number of these items, at the eighth grade, anchoring at each anchor point on the mathematics and science scales, respectively. Exhibit 12.6 and Exhibit 12.7 present the numbers at the fourth grade. All together, at the eighth grade, four mathematics items met the anchoring criteria at the Low International Benchmark, 40 did so for the Intermediate International Benchmark, 75 for the High International Benchmark, and 63 for the Advanced International Benchmark. Twelve items were too difficult for the Advanced International Benchmark. In science, 10 items met one of the criteria for anchoring at the Low International Benchmark, 23 for the Intermediate International Benchmark, 61 for the High International Benchmark, and 68 for the Advanced International Benchmark. Twenty-seven items were too difficult to anchor at the Advanced International Benchmark at the eighth grade.

At the fourth grade level, 17 mathematics items met the anchoring criteria at the Low International Benchmark, 43 did so for the Intermediate International Benchmark, 56 for the High International Benchmark, and 33 for the Advanced International Benchmark. Ten items were too difficult for the Advanced International Benchmark. In science, 32 items met one of the criteria for anchoring at the Low International Benchmark, 37 for the Intermediate International Benchmark, 28 for the High International Benchmark, and 37 for the Advanced International Benchmark. Sixteen items were too difficult to anchor at the Advanced International Benchmark at the fourth grade.

Including items meeting the less stringent anchoring criteria substantially increased the number of items that could be used to characterize performance at each benchmark, beyond what would have been available if only the items that met the 65 percent criteria were included. Even though these items did not meet the 65 percent anchoring criteria, they were still items that students scoring at the benchmarks had a high degree of probability of answering correctly.

Exhibit 12.4 **Number of Items Anchoring at Each Anchor Level Eighth Grade Mathematics**

	Anchored	Almost Anchored	Met 60-65% Criterion	Total
Low (400)	3*	1	-	4
Intermediate (475)	25	5*	10	40
High (550)	46	10	19*	75
Advanced (625)	41	5	17	63
Too Difficult to Anchor	12	-	-	12
Total	**127**	**21**	**46**	**194**

* These numbers where obtained based on the anchor points where the calculator-sensitive items anchor if considered without calculator (see Appendix A of the International Mathematics Report for more details on calculator use in TIMSS 2003 assessment))

Exhibit 12.5 **Number of Items Anchoring at Each Anchor Level Eighth Grade Science**

	Anchored	Almost Anchored	Met 60-65% Criterion	Total
Low (400)	6	4	-	10
Intermediate (475)	10	4	9	23
High (550)	35	5	21	61
Advanced (625)	40	5	23	68
Too Difficult to Anchor	27	-	-	27
Total	**118**	**18**	**53**	**189**

Exhibit 12.6 **Number of Items Anchoring at Each Anchor Level Fourth Grade Mathematics**

	Anchored	Almost Anchored	Met 60-65% Criterion	Total
Low (400)	15	2	-	17
Intermediate (475)	21	11	11	43
High (550)	36	7	13	56
Advanced (625)	23	1	9	33
Too Difficult to Anchor	10	-	-	10
Total	**105**	**21**	**33**	**159**[2]

2 Following the item review, two items were deleted out of 161 items in the Mathematics Grade 4 test, resulting in 159 items (see chapter 10 for more details on item review process).

Exhibit 12.7 **Number of Items Anchoring at Each Anchor Level Fourth Grade Science**

	Anchored	Almost Anchored	Met 60-65% Criterion	Total
Low (400)	26	6	-	32
Intermediate (475)	20	5	12	37
High (550)	18	2	8	28
Advanced (625)	25	3	9	37
Too Difficult to Anchor	16	-	-	16
Total	**105**	**16**	**29**	**150**[3]

12.2.6 Expert Review of Anchor Items by Content Area

Having identified the items that anchored at each of the international benchmarks, the next step was to have the items reviewed by the TIMSS 2003 Science and Mathematics Item Review Committee (SMIRC) to develop descriptions of student performance. In preparation for the review by the SMIRC, the mathematics and science items, respectively, were organized in binders grouped by benchmark anchor point and within anchor point, the items were sorted by content area and then by the anchoring criteria they met – items that anchored, followed by items that almost anchored, followed by items that met only the 60 to 65% criteria. The following information was included for each item: content area, main topic, cognitive domain, answer key, percent correct at each anchor point, and overall international percent correct. For open-ended items, the scoring guides were included.

The TIMSS & PIRLS International Study Center convened the SMIRC for a four-day meeting. The assignment consisted of three tasks: (1) work through each item in each binder and arrive at a short description of the knowledge, understanding, and/or skills demonstrated by students answering the item correctly; (2) based on the items that anchored, almost anchored, and met only the 60-65% criterion, draft a description of the level of comprehension demonstrated by students at each of the four benchmark anchor points; and (3) select example items to support and illustrate the anchor point descriptions. Following the meeting, these drafts were edited and revised as necessary for use in the TIMSS 2003 International Reports.

Exhibits 12.8 and 12.9 present, for each scale, the number of items per content area that met one of the anchoring criteria discussed above, at each International Benchmark, and the number of items that were too difficult for the Advanced International Benchmark, at the eighth grade level. Exhibits 12.10 and 12.11 present the same information for the fourth grade. The descriptions for each item developed by SMIRC and the summaries are

3 Following the item review, two items were deleted out of 152 items in the Science Grade 4 test, resulting in 150 items (see chapter 10 for more details on item review process).

presented in the TIMSS 2003 International Reports.

Exhibit 12.8 **Number of Items Anchoring at Each Anchor Level, by Content Area Eighth Grade Mathematics**

	Low (400)	Intermediate (475)	High (550)	Advanced (625)	Too Difficult to Anchor	Total
Number	2*	11*	22*	20	2	57
Algebra	0	11	16	16	4	47
Measurement	1	4	14	10	2	31
Geometry	0	8	12	10	1	31
Data	1	6	11	7	3	28
Total	**4**	**40**	**75**	**63**	**12**	**194**

* These numbers where obtained based on the anchor points where the calculator-sensitive items anchor if considered without calculator (see Appendix A of the International Mathematics Report for more details on calculator use in TIMSS 2003 assessment)

Exhibit 12.9 **Number of Items Anchoring at Each Anchor Level, by Content Area Eighth Grade Science**

	Low (400)	Intermediate (475)	High (550)	Advanced (625)	Too Difficult to Anchor	Total
Life Science	4	4	19	19	8	54
Chemistry	1	1	8	16	5	31
Physics	3	7	17	13	6	46
Earth Science	1	7	9	10	4	31
Environmental Science	1	4	8	10	4	27
Total	**10**	**23**	**62**	**68**	**27**	**189**

Exhibit 12.10 **Number of Items Anchoring at Each Anchor Level, by Content Area Fourth Grade Mathematics**

	Low (400)	Intermediate (475)	High (550)	Advanced (625)	Too Difficult to Anchor	Total
Number	7	18	22	12	4	63
Patterns and Relationships	1	6	8	4	4	23
Measurement	2	5	11	13	1	32
Geometry	5	6	10	2	1	24
Data	2	8	5	2	0	17
Total	**17**	**43**	**56**	**33**	**10**	**159**[4]

4 Following the item review, two items were deleted out of 161 items in the Mathematics Grade 4 test, resulting in 159 items (see chapter 10 for more details on item review process).

Exhibit 12.11 **Number of Items Anchoring at Each Anchor Level, by Content Area Fourth Grade Science**

	Low (400)	Intermediate (475)	High (550)	Advanced (625)	Too Difficult to Anchor	Total
Life Science	17	14	11	15	7	64
Physical Science	9	13	12	13	6	53
Earth Science	6	10	5	9	3	33
Total	**32**	**37**	**28**	**37**	**16**	**150**[5]

12.3 Capturing the Uncertainty in the TIMSS Student Achievement Measures

To obtain estimates of students' proficiency in mathematics and science that were both accurate and cost-effective, TIMSS 2003 made extensive use of probability sampling techniques to sample students from national eighth- and fourth-grade student populations, and applied matrix sampling methods to target individual students with a subset of the entire set of assessment materials. Statistics computed from these student samples were used to estimate population parameters. This approach made an efficient use of resources, in particular keeping student response burden to a minimum, but at a cost of some variance or uncertainty in the statistics. To quantify this uncertainty, each statistic in the TIMSS 2003 international reports (Mullis et al., 2004; Martin et al., 2004) is accompanied by an estimate of its standard error. These standard errors incorporate components reflecting the uncertainty due to generalizing from student samples to the entire eighth- or fourth-grade student population (sampling variance), and to inferring students' performance on the entire assessment from their performance on the subset of items that they took (imputation variance).

12.3.1 Estimating Sampling Variance

The TIMSS 2003 sampling design applied a stratified multistage cluster-sampling technique to the problem of selecting efficient and accurate samples of students while working with schools and classes. This design capitalized on the structure of the student population (i.e., students grouped in classes within schools) to derive student samples that permitted efficient and economical data collection. Unfortunately, however, such a complex sampling design complicates the task of computing standard errors to quantify sampling variability.

When, as in TIMSS, the sampling design involves multistage cluster sampling, there are several options for estimating sampling errors that avoid the assumption of simple random sampling (Wolter, 1985). The jackknife

5 Following the item review, two items were deleted out of 152 items in the Science Grade 4 test, resulting in 150 items (see chapter 10 for more details on item review process).

repeated replication technique (JRR) was chosen by TIMSS because it is computationally straightforward and provides approximately unbiased estimates of the sampling errors of means, totals, and percentages.

The variation on the JRR technique used in TIMSS 2003 is described in Johnson and Rust (1992). It assumes that the primary sampling units (PSUs) can be paired in a manner consistent with the sample design, with each pair regarded as members of a pseudo-stratum for variance estimation purposes. When used in this way, the JRR technique appropriately accounts for the combined effect of the between- and within-PSU contributions to the sampling variance. The general use of JRR entails systematically assigning pairs of schools to sampling zones, and randomly selecting one of these schools to have its contribution doubled and the other to have its contribution zeroed, so as to construct a number of "pseudo-replicates" of the original sample. The statistic of interest is computed once for all of the original sample, and once again for each pseudo-replicate sample. The variation between the estimates for each of the replicate samples and the original sample estimate is the jackknife estimate of the sampling error of the statistic.

12.3.1.1 Constructing Sampling Zones for Sampling Variance Estimation

To apply the JRR technique used in TIMSS 2003, the sampled schools had to be paired and assigned to a series of groups known as sampling zones. This was done at Statistics Canada by working through the list of sampled schools in the order in which they were selected and assigning the first and second schools to the first sampling zone, the third and fourth schools to the second zone, and so on. In total 75 zones were used, allowing for 150 schools per country. When more than 75 zones were constructed, they were collapsed to keep the total number to 75.

Sampling zones were constructed within design domains, or explicit strata. Where there was an odd number of schools in an explicit stratum, either by design or because of school nonresponse, the students in the remaining school were randomly divided to make up two "quasi" schools for the purposes of calculating the jackknife standard error. Each zone then consisted of a pair of schools or "quasi" schools. Exhibit 12.12 shows the range of sampling zones used in each country.

Exhibit 12.12 **Number of Sampling Zones Used in Each Country**

Country	TIMSS 2003 Sampling Zones	TIMSS 1999 Sampling Zones	TIMSS 1995 Sampling Zones
Armenia	75	-	-
Australia	75	-	74
Bahrain	75	-	-
Belgium (Flemish)	75	74	71
Botswana	73	-	-
Bulgaria	75	75	58
Chile	75	75	-
Chinese Taipei	75	75	-
Cyprus	75	61	55
Egypt	75	-	-
England	44	64	64
Estonia	75	-	-
Ghana	75	-	-
Hong Kong, SAR	63	69	43
Hungary	75	74	75
Indonesia	75	75	-
Iran, Islamic Rep. of	75	75	75
Israel	74	70	-
Italy	75	75	-
Japan	74	71	75
Jordan	70	74	-
Korea, Rep. of	75	75	75
Latvia	70	73	64
Lebanon	75	-	-
Lithuania	72	75	73
Macedonia, Rep. of	75	75	-
Malaysia	75	75	-
Moldova, Rep. of	75	75	-
Morocco	67	75	-
Netherlands	65	63	48
New Zealand	75	75	75
Norway	69	-	74
Palestinian Nat'l Auth.	73	-	-
Philippines	69	75	-
Romania	74	74	72
Russian Federation	69	56	41
Saudi Arabia	75	-	-

Exhibit 12.12 **Number of Sampling Zones Used in Each Country** *(...Continued)*

Country	TIMSS 2003 Sampling Zones	TIMSS 1999 Sampling Zones	TIMSS 1995 Sampling Zones
Scotland	65	-	64
Serbia	75	-	-
Singapore	75	73	69
Slovak Republic	75	73	73
Slovenia	75	-	61
South Africa	75	75	-
Sweden	75	-	60
Tunisia	75	75	-
United States	75	53	55

12.3.1.2 Computing Sampling Variance Using the JRR Method

The JRR algorithm used in TIMSS 2003 assumes that there are H sampling zones within each country, each containing two sampled schools selected independently. To compute a statistic t from the sample for a country, the formula for the JRR variance estimate of the statistic t is then given by the following equation:

$$Var_{jrr}(t) = \sum_{h=1}^{H} [t(J_h) - t(S)]^2$$

where H is the number of pairs in the sample for the country. The term $t(S)$ corresponds to the statistic for the whole sample (computed with any specific weights that may have been used to compensate for the unequal probability of selection of the different elements in the sample or any other post-stratification weight). The element $t(J_h)$ denotes the same statistic using the h^{th} jackknife replicate. This is computed using all cases except those in the h^{th} zone of the sample; for those in the h^{th} zone, all cases associated with one of the randomly selected units of the pair are removed, and the elements associated with the other unit in the zone are included twice. In practice, this is accomplished by recoding to zero the weights for the cases of the element of the pair to be excluded from the replication, and multiplying by two the weights of the remaining element within the h^{th} pair.

The computation of the JRR variance estimate for any statistic in TIMSS 2003 required the computation of the statistic up to 76 times for any given country: once to obtain the statistic for the full sample, and up to 75 times to obtain the statistics for each of the jackknife replicates (J_h). The number of times a statistic needed to be computed for a given country depended on the number of implicit strata or sampling zones defined for that country.

Doubling and zeroing the weights of the selected units within the sampling zones was accomplished by creating replicate weights that were then used in the calculations. In this approach, a set of temporary replicate weights are created for each pseudo-replicate sample. Each replicate weight is equal to k times the overall sampling weight, where k can take values of 0, 1, or 2 depending on whether the case is to be removed from the computation, left as it is, or have its weight doubled. The value of k for an individual student record for a given replicate depends on the assignment of the record to the specific PSU and zone.

Within each zone the members of the pair of schools are assigned an indicator (u_i), coded randomly to 1 or 0 so that one of them has a value of 1 on the variable u_i, and the other a value of 0. This indicator determines whether the weights for the elements in the school in this zone are to be doubled or zeroed. The replicate weight $W_h^{g,i,j}$ for the elements in a school assigned to zone h is computed as the product of k_h times their overall sampling weight, where k_h can take values of 0, 1, or 2 depending on whether the school is to be omitted, be included with its usual weight, or have its weight doubled for the computation of the statistic of interest. In TIMSS 2003, the replicate weights were not permanent variables, but were created temporarily by the sampling variance estimation program as a useful computing device.

To create replicate weights, each sampled student was first assigned a vector of 75 weights, $W_h^{g,i,j}$, where h takes values from 1 to 75. The value of $W_0^{g,i,j}$ is the overall sampling weight, which is simply the product of the final school weight, classroom weight, and student weight, as described in Chapter 9.

The replicate weights for a single case were then computed as

$$W_h^{g,i,j} = W_0^{g,i,j} \cdot k_{hi}$$

where the variable k_h for an individual i takes the value $k_{hi} = 2*u_i$ if the record belongs to zone h, and $k_{hi} = 1$ otherwise.

In the TIMSS 2003 analysis, 75 replicate weights were computed for each country regardless of the number of actual zones within the country. If a country had fewer than 75 zones, then the replicate weights W_h, where h was greater than the number of zones within the country, were each the same as the overall sampling weight. Although this involved some redundant computation, having 75 replicate weights for each country had no effect on the size of the error variance computed using the jackknife formula, but it facilitated the computation of standard errors for a number of countries at a time.

Standard errors presented in the international reports were computed using SAS programs developed at the TIMSS & PIRLS International Study Center. As a quality control check, results were verified using the WesVarPC software (Westat, 1997).

12.3.2 Estimating Imputation Variance

The TIMSS 2003 item pool was far too extensive to be administered in its entirety to any one student, and so a matrix-sampling test design was developed whereby each student was given a single test booklet containing only a part of the entire assessment.[6] The results for all of the booklets were then aggregated using item response theory to provide results for the entire assessment. Since each student responded to just a subset of the assessment items, multiple imputation (the generation of "plausible values") was used to derive reliable estimates of student performance on the assessment as a whole. Since every student proficiency estimate incorporates some uncertainty, TIMSS followed the customary procedure of generating five estimates for each student and using the variability among them as a measure of this imputation uncertainty, or error. In the TIMSS 2003 international report the imputation error for each variable has been combined with the sampling error for that variable to provide a standard error incorporating both.

The general procedure for estimating the imputation variance using plausible values is the following (Mislevy, R.J., Beaton, A.E., Kaplan, B., and Sheenan, K.M., 1992). First compute the statistic t, for each set of M plausible values. The statistics t_m, where m = 1, 2, …, 5, can be anything estimable from the data, such as a mean, the difference between means, percentiles, and so forth.

Once the statistics are computed, the imputation variance is then computed as:

$$Var_{imp} = (1 + 1/M)\, Var(t_1, \ldots, t_M)$$

where M is the number of plausible values used in the calculation, and is the variance of the M estimates computed using each plausible value.

12.3.3 Combining Sampling and Imputation Variance

The standard errors of the mathematics and science proficiency statistics reported by TIMSS include both sampling and imputation variance components. The standard errors were computed using the following formula:[7]

$$Var(t_{pv}) = Var_{jrr}(t_1) + Var_{imp}$$

6 Details of the TIMSS test design may be found in Chapter 2.

7 Under ideal circumstances and with unlimited computing resources, the imputation variance for the plausible values and the JRR sampling variance for each of the plausible values would be computed. This would be equivalent to computing the same statistic up to 380 times (once overall for each of the five plausible values using the overall sampling weights, and then 75 times more for each plausible value using the complete set of replicate weights). An acceptable shortcut, however, is to compute the JRR variance component using one plausible value, and then the imputation variance using the five plausible values. Using this approach, a statistic needs to be computed only 80 times.

where $Var_{jrr}(t_1)$ is the sampling variance for the first plausible value and Var_{imp} is the imputation variance. The User Guide for the TIMSS 2003 International Database contains programs in SAS and SPSS that compute each of these variance components for the TIMSS 2003 data.

Exhibits 12.13 through 12.16 show basic summary statistics for mathematics and science achievement in the TIMSS 2003 assessment for the eighth and fourth grades. Each exhibit presents the student sample size, the mean and standard deviation, averaged across the five plausible values, the jackknife standard error for the mean, and the overall standard errors for the mean including imputation error. Appendix E contains tables showing the same summary statistics for the mathematics and science content areas for the eighth and fourth grades.

12.4 Calculating National and International Statistics for Student Achievement

As described in earlier chapters, TIMSS 2003 made extensive use of imputed proficiency scores to report student achievement, both in the major content domains (number, algebra, measurement, geometry, and data for mathematics and life science, chemistry, physics, earth science, and environmental science for science) and mathematics and science as overall subjects. This section describes the procedures followed in computing the principal statistics used to summarize achievement in the International Reports (Mullis, et al., 2004; Martin et al., 2004), including means based on plausible values, gender differences, performance in content domains, and performance on example items.

For each of the TIMSS 2003 mathematics and science scales, the item response theory (IRT) scaling procedure described in Chapter 11 yields five imputed scores or plausible values for each student. The difference between the five values reflects the degree of uncertainty in the imputation process. When the process yields consistent results, the differences between the five values are very small. To obtain the best estimate for each of the TIMSS statistics, each one was computed five times, using each of the five plausible values in turn, and the results averaged to derive the reported value. The standard errors that accompany each reported statistic include two components as described in the previous section: one quantifying sampling variation and the other quantifying imputation variation.

Exhibit 12.13 **Summary Statistics and Standard Errors for Proficiency in Mathematics - Eighth Grade**

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Armenia	5726	478.127	83.522	2.952	2.997
Australia	4791	504.703	81.538	4.613	4.638
Bahrain	4199	401.196	76.317	1.571	1.727
Belgium (Flemish)	4970	536.710	73.494	2.696	2.772
Botswana	5150	366.345	71.554	2.189	2.581
Bulgaria	4117	476.169	84.077	4.222	4.315
Chile	6377	386.880	83.233	3.060	3.269
Chinese Taipei	5379	585.252	99.969	4.507	4.607
Cyprus	4002	459.366	81.377	1.474	1.653
Egypt	7095	406.168	92.754	3.423	3.505
England	2830	498.464	77.231	4.653	4.674
Estonia	4040	530.915	69.334	2.931	2.997
Ghana	5100	275.704	90.996	4.339	4.657
Hong Kong, SAR	4972	586.051	71.924	3.245	3.324
Hungary	3302	529.275	79.506	3.212	3.221
Indonesia	5762	410.702	88.789	4.796	4.844
Iran, Islamic Rep. of	4942	411.447	74.303	2.316	2.351
Israel	4318	495.648	84.682	3.360	3.422
Italy	4278	483.599	76.675	3.145	3.192
Japan	4856	569.921	79.874	1.985	2.074
Jordan	4489	424.352	89.007	4.068	4.086
Korea, Rep. of	5309	589.092	83.855	1.853	2.191
Latvia	3630	508.327	73.094	3.131	3.174
Lebanon	3814	433.045	66.747	3.040	3.091
Lithuania	4964	501.615	78.291	2.442	2.458
Macedonia, Rep. of	3893	434.983	88.380	3.500	3.542
Malaysia	5314	508.336	74.263	4.035	4.079
Moldova, Rep. of	4033	459.895	80.563	4.006	4.050
Morocco	2943	386.539	68.126	2.134	2.483
Netherlands	3065	536.273	69.391	3.788	3.820
New Zealand	3801	494.040	78.318	5.264	5.275
Norway	4133	461.470	70.859	2.427	2.499
Palestinian Nat'l Auth.	5357	390.486	91.839	3.037	3.104
Philippines	6917	377.690	87.339	5.164	5.208
Romania	4104	475.282	90.230	4.786	4.822
Russian Federation	4667	508.041	76.619	3.532	3.709
Saudi Arabia	4295	331.682	78.324	4.466	4.574
Scotland	3516	497.654	74.820	3.585	3.711

Exhibit 12.13 **Summary Statistics and Standard Errors for Proficiency in Mathematics - Eighth Grade** *(...Continued)*

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Serbia	4296	476.637	88.850	2.477	2.595
Singapore	6018	605.450	80.090	3.508	3.583
Slovak Republic	4215	507.740	82.382	3.250	3.308
Slovenia	3578	492.956	71.101	2.089	2.193
South Africa	8952	263.614	107.151	5.330	5.490
Sweden	4256	499.058	71.182	2.550	2.622
Tunisia	4931	410.329	60.340	2.121	2.186
United States	8912	504.366	79.993	3.270	3.309

Exhibit 12.14 **Summary Statistics and Standard Errors for Proficiency in Mathematics - Fourth Grade**

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Armenia	5674	455.925	86.681	3.473	3.489
Australia	4321	498.663	80.862	3.821	3.882
Belgium (Flemish)	4712	550.601	58.948	1.773	1.783
Chinese Taipei	4661	563.949	63.029	1.696	1.752
Cyprus	4328	509.810	85.391	2.399	2.424
England	3585	531.182	87.407	3.701	3.736
Hong Kong, SAR	4608	574.782	63.389	3.080	3.161
Hungary	3319	528.502	77.251	3.045	3.130
Iran, Islamic Rep. of	4352	389.052	85.697	4.012	4.153
Italy	4282	502.762	82.050	3.662	3.679
Japan	4535	564.556	73.749	1.515	1.598
Latvia	3687	535.855	72.517	2.789	2.835
Lithuania	4422	534.017	73.806	2.797	2.804
Moldova, Rep. of	3981	504.149	87.334	4.818	4.879
Morocco	4264	346.807	90.250	4.940	5.081
Netherlands	2937	540.373	54.625	2.013	2.109
New Zealand	4308	493.464	84.230	2.139	2.151
Norway	4342	451.342	80.240	2.260	2.298
Philippines	4572	358.195	109.709	7.861	7.911
Russian Federation	3963	531.682	78.249	4.734	4.746
Scotland	3936	490.321	77.541	3.166	3.252
Singapore	6668	594.427	84.222	5.558	5.597
Slovenia	3126	478.795	77.946	2.575	2.619
Tunisia	4334	339.300	99.591	4.567	4.730
United States	9829	518.284	76.272	2.429	2.436

Exhibit 12.15 **Summary Statistics and Standard Errors for Science Proficiency - Eighth Grade**

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Armenia	5726	461.267	81.041	3.413	3.465
Australia	4791	527.014	75.307	3.763	3.800
Bahrain	4199	438.255	74.470	1.625	1.793
Belgium (Flemish)	4970	515.506	66.954	2.457	2.487
Botswana	5150	364.569	86.472	2.771	2.840
Bulgaria	4117	478.843	92.987	5.072	5.151
Chile	6377	412.851	84.096	2.827	2.890
Chinese Taipei	5379	571.092	79.064	3.381	3.457
Cyprus	4002	441.474	79.496	1.589	2.049
Egypt	7095	421.117	103.720	3.825	3.898
England	2830	543.896	76.832	4.070	4.140
Estonia	4040	552.258	65.049	2.382	2.456
Ghana	5100	255.324	120.145	5.726	5.882
Hong Kong, SAR	4972	556.089	65.545	2.965	3.039
Hungary	3302	542.761	75.903	2.800	2.837
Indonesia	5762	420.221	78.769	3.981	4.055
Iran, Islamic Rep. of	4942	453.428	72.593	2.176	2.329
Israel	4318	488.200	84.965	3.028	3.082
Italy	4278	490.891	78.125	2.996	3.062
Japan	4856	552.178	71.011	1.691	1.739
Jordan	4489	474.845	89.396	3.755	3.848
Korea, Rep. of	5309	558.399	69.575	1.581	1.641
Latvia	3630	512.363	67.343	2.532	2.551
Lebanon	3814	393.399	92.556	4.271	4.315
Lithuania	4964	519.380	69.632	2.126	2.143
Macedonia, Rep. of	3893	449.373	91.641	3.575	3.596
Malaysia	5314	510.452	65.855	3.643	3.651
Moldova, Rep. of	4033	472.423	73.553	3.258	3.365
Morocco	2943	396.474	69.138	2.141	2.501
Netherlands	3065	535.765	61.278	3.046	3.077
New Zealand	3801	519.730	73.716	5.010	5.044
Norway	4133	493.863	69.755	2.107	2.170
Palestinian Nat'l Auth.	5357	435.387	92.463	3.215	3.240

Exhibit 12.15 **Summary Statistics and Standard Errors for Science Proficiency - Eighth Grade** *(...Continued)*

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Philippines	6917	377.373	102.264	5.659	5.803
Romania	4104	469.604	91.090	4.865	4.936
Russian Federation	4667	513.621	75.184	3.561	3.679
Saudi Arabia	4295	397.741	72.491	3.618	3.985
Scotland	3516	511.546	75.689	3.319	3.351
Serbia	4296	467.686	83.688	2.412	2.467
Singapore	6018	577.849	91.817	4.249	4.262
Slovak Republic	4215	516.785	75.587	3.159	3.215
Slovenia	3578	520.498	66.696	1.725	1.786
South Africa	8952	243.664	131.640	6.357	6.683
Sweden	4256	524.258	73.901	2.587	2.688
Tunisia	4931	403.547	60.483	1.914	2.082
United States	8912	527.298	80.681	3.095	3.143

Exhibit 12.16 **Summary Statistics and Standard Errors for Science Proficiency - Fourth Grade**

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Armenia	5674	436.528	95.954	4.219	4.299
Australia	4321	520.691	82.093	4.137	4.206
Belgium (Flemish)	4712	518.342	54.858	1.542	1.769
Chinese Taipei	4661	551.355	68.622	1.589	1.727
Cyprus	4328	480.485	74.171	2.214	2.379
England	3585	540.240	83.167	3.383	3.608
Hong Kong, SAR	4608	542.483	59.804	2.907	3.059
Hungary	3319	529.727	79.351	2.887	2.979
Iran, Islamic Rep. of	4352	413.923	96.600	4.070	4.104
Italy	4282	515.640	84.861	3.749	3.766
Japan	4535	543.469	73.117	1.343	1.509
Latvia	3687	531.521	68.794	2.464	2.489
Lithuania	4422	512.106	66.362	2.171	2.551
Moldova, Rep. of	3981	496.420	84.966	4.576	4.599
Morocco	4264	304.392	124.834	6.582	6.705
Netherlands	2937	525.125	53.351	1.816	2.001
New Zealand	4308	519.671	85.050	2.375	2.460
Norway	4342	466.346	83.994	2.154	2.619
Philippines	4572	331.620	145.326	9.293	9.433
Russian Federation	3963	526.187	82.019	5.115	5.167
Scotland	3936	501.975	77.719	2.808	2.887
Singapore	6668	565.148	86.786	5.517	5.548
Slovenia	3126	490.365	77.195	2.462	2.530
Tunisia	4334	313.989	125.686	5.583	5.655
United States	9829	535.631	81.247	2.408	2.526

National averages were computed as the average of the weighted means for each of the five plausible values. The weighted mean for each plausible value was computed as follows:

$$\overline{X}_{pvl} = \frac{\sum_{j=1}^{N} W^{i,j} . pv_{lj}}{\sum_{j=1}^{N} W^{i,j}}$$

where

$\overline{X}_{pvl}$ is the country mean for plausible value l

pv_{lj} is the l-th plausible value for the j-th student

$W^{i,j}$ is the weight associated with the j-th student in class i, described in Chapter 9

N is the number of students in the country's sample.

These five weighted means were then averaged to obtain the national average for each country. To provide a reference point for comparison purposes, TIMSS presented the international average of many of the national statistics (means and percentages). International averages were calculated by first computing the national average for each plausible value for each country and then averaging across countries. These five estimates of the international average were then themselves averaged to derive the international average presented in the TIMSS reports, as shown below:

$$\overline{X}_{\bullet pvl} = \frac{\sum_{k=1}^{K} \overline{X}_{pvl,k}}{K}$$

where

$\overline{X}_{\bullet pvl}$ is the international mean for plausible value l

$\overline{X}_{pvl,k}$ is the k-th country mean for plausible value l

and K is the number of countries.

12.4.1 Comparing Achievement Differences Across Countries

A basic aim of the TIMSS 2003 International Reports is to provide fair and accurate comparisons of student achievement across the participating countries. Most of the exhibits in the TIMSS reports summarize student achievement by means of a statistic such as a mean or percentage, and each statistic is accompanied by its standard error, which is a measure of the uncertainty due to student sampling and the imputation process. In comparisons of performance across countries, standard errors can be used to assess the statistical significance of the difference between the summary statistics.

The exhibits presented in the TIMSS 2003 international reports allow comparisons of average performance of a country with that of other participating countries. If repeated samples were taken from two populations with the same mean and variance and in each one the hypothesis that the means from the two samples are significantly different at the $\alpha = .05$ level (i.e. with 95% confidence) was tested, then in about five percent of the comparisons it

would be expected to find significant differences between the sample means even though no difference exists in the population. In such a test of the difference between two means, the probability of finding significant differences in the samples when none exist in the populations (the so-called type I error) is given by $\alpha = .05$. Conversely, the probability of not making such an error is $1 - \alpha$, which in the case of a single test is .95.

Mean proficiencies are considered significantly different if the absolute difference between them, divided by the standard error of the difference, is greater than the critical value. For differences between countries, which can be considered as independent samples, the standard error of the difference between means is computed as the square root of the sum of the squared standard errors of each mean:

$$se_{diff} = \sqrt{se_1^2 + se_2^2}$$

where se_1 and se_2 are the standard errors of the means. Exhibits 12.17 and 12.18 show the means and standard errors used in the calculation of statistical significance for mathematics and science achievement in the eighth and fourth grades.

In contrast to the practice in previous TIMSS reports, the significance tests presented in the TIMSS 2003 International Reports have NOT been adjusted for multiple comparisons among countries. Although adjustments such as the Bonferroni procedure guard against misinterpreting the outcome of multiple simultaneous significance tests, and have been used in previous TIMSS studies, the results vary depending on the number of countries included in the adjustment, leading to apparently conflicting results from comparisons using different numbers of countries.

12.4.2 Comparing National Achievement Against the International Mean

Many of the data exhibits in the TIMSS 2003 international reports show countries' mean achievement compared with the international mean, together with a test of the statistical significance between the two. These significance tests were based on the standard errors of the national and international means.

Exhibit 12.17 **Means and Standard Errors for Country Comparisons of Mathematics and Science Achievement in the Eighth Grade**

	Mathematics		Science	
Country	**Mean**	**S.E.**	**Mean**	**S.E.**
Armenia	478.127	2.997	461.267	3.465
Australia	504.703	4.638	527.014	3.800
Bahrain	401.196	1.727	438.255	1.793
Basque Country, Spain	487.061	2.732	488.754	2.678
Belgium (Flemish)	536.710	2.772	515.506	2.487
Botswana	366.345	2.581	364.569	2.840
Bulgaria	476.169	4.315	478.843	5.151
Chile	386.880	3.269	412.851	2.890
Chinese Taipei	585.252	4.607	571.092	3.457
Cyprus	459.366	1.653	441.474	2.049
Egypt	406.168	3.505	421.117	3.898
England	498.464	4.674	543.896	4.140
Estonia	530.915	2.997	552.258	2.456
Ghana	275.704	4.657	255.324	5.882
Hong Kong, SAR	586.051	3.324	556.089	3.039
Hungary	529.275	3.221	542.761	2.837
Indiana State, US	508.257	5.215	530.609	4.769
Indonesia	410.702	4.844	420.221	4.055
Iran, Islamic Rep. of	411.447	2.351	453.428	2.329
Israel	495.648	3.422	488.200	3.082
Italy	483.599	3.192	490.891	3.062
Japan	569.921	2.074	552.178	1.739
Jordan	424.352	4.086	474.845	3.848
Korea, Rep. of	589.092	2.191	558.399	1.641
Latvia	508.327	3.174	512.363	2.551
Lebanon	433.045	3.091	393.399	4.315
Lithuania	501.615	2.458	519.380	2.143
Macedonia, Rep. of	434.983	3.542	449.373	3.596
Malaysia	508.336	4.079	510.452	3.651
Moldova, Rep. of	459.895	4.050	472.423	3.365
Morocco	386.539	2.483	396.474	2.501
Netherlands	536.273	3.820	535.765	3.077
New Zealand	494.040	5.275	519.730	5.044
Norway	461.470	2.499	493.863	2.170
Ontario Province, Can.	520.932	3.105	532.920	2.656
Palestinian Nat'l Auth.	390.486	3.104	435.387	3.240

Exhibit 12.17 **Means and Standard Errors for Country Comparisons of Mathematics and Science Achievement in the Eighth Grade** *(...Continued)*

	Mathematics		Science	
Country	**Mean**	**S.E.**	**Mean**	**S.E.**
Philippines	377.690	5.208	377.373	5.803
Quebec Province, Can.	543.075	3.031	531.013	3.044
Romania	475.282	4.822	469.604	4.936
Russian Federation	508.041	3.709	513.621	3.679
Saudi Arabia	331.682	4.574	397.741	3.985
Scotland	497.654	3.711	511.546	3.351
Serbia	476.637	2.595	467.686	2.467
Singapore	605.450	3.583	577.849	4.262
Slovak Republic	507.740	3.308	516.785	3.215
Slovenia	492.956	2.193	520.498	1.786
South Africa	263.614	5.490	243.664	6.683
Sweden	499.058	2.622	524.258	2.688
Tunisia	410.329	2.186	403.547	2.082
United States	504.366	3.309	527.298	3.143

Exhibit 12.18 **Means and Standard Errors for Country Comparisons of Mathematics and Science Achievement in the Fourth Grade**

	Mathematics		Science	
Country	**Mean**	**S.E.**	**Mean**	**S.E.**
Armenia	455.925	3.489	436.528	4.299
Australia	498.663	3.882	520.691	4.206
Belgium (Flemish)	550.601	1.783	518.342	1.769
Chinese Taipei	563.949	1.752	551.355	1.727
Cyprus	509.810	2.424	480.485	2.379
England	531.182	3.736	540.240	3.608
Hong Kong, SAR	574.782	3.161	542.483	3.059
Hungary	528.502	3.130	529.727	2.979
Indiana State, US	532.874	2.806	553.287	3.710
Iran, Islamic Rep. of	389.052	4.153	413.923	4.104
Italy	502.762	3.679	515.640	3.766
Japan	564.556	1.598	543.469	1.509
Latvia	535.855	2.835	531.521	2.489
Lithuania	534.017	2.804	512.106	2.551
Moldova, Rep. of	504.149	4.879	496.420	4.599
Morocco	346.807	5.081	304.392	6.705
Netherlands	540.373	2.109	525.125	2.001
New Zealand	493.464	2.151	519.671	2.460
Norway	451.342	2.298	466.346	2.619
Ontario Province, Can.	511.184	3.830	540.205	3.746
Philippines	358.195	7.911	331.620	9.433
Quebec Province, Can.	505.848	2.409	500.392	2.484
Russian Federation	531.682	4.746	526.187	5.167
Scotland	490.321	3.252	501.975	2.887
Singapore	594.427	5.597	565.148	5.548
Slovenia	478.795	2.619	490.365	2.530
Tunisia	339.300	4.730	313.989	5.655
United States	518.284	2.436	535.631	2.526

When comparing each country's mean with the international average, TIMSS took into account the fact that the country contributed to the international standard error. To correct for this contribution, TIMSS adjusted the

standard error of the difference. The sampling component of the standard error of the difference for country *j* is

$$se_{s\_dif\_j} = \frac{\sqrt{((N-1)^2 - 1)se_j^2 + \sum_{k=1}^{N} se_k^2}}{N}$$

where

$se_{s\_dif\_j}$ is the standard error of the difference due to sampling when country *j* is compared to the international mean,

N is the number of countries,

se_k^2 is the sampling standard error for country *k*, and

se_j^2 is the sampling standard error for country *j*.

The imputation component of the standard error for country *j* was computed by taking the square root of the imputation variance calculated as follows

$$se_{i\_dif\_j} = \sqrt{\frac{6}{5} Var(d_1,...,d_l,...,d_5)},$$

where d_l is the difference between the international mean and the country mean for plausible value *l*.

Finally, the standard error of the difference was calculated as

$$se_{dif\_j} = \sqrt{se_{i\_dif\_j}^2 + se_{s\_dif\_j}^2}.$$

12.4.3 Reporting Gender Differences Within Countries

TIMSS reported gender differences in overall student achievement in mathematics and science overall, as well as in mathematics and science content areas. Gender differences were presented in an exhibit showing mean achievement for males and females and the differences between them, with an accompanying graph indicating whether the difference was statistically significant.

Because in most countries males and females attend the same schools, the samples of males and females cannot be treated as independent samples for the purpose of statistical tests. Accordingly, TIMSS used a jackknife procedure applicable to correlated samples for estimating the standard errors of the male-female differences. This involved computing the average difference between boys and girls in each country once for every one of the 75 replicate samples, and five more times, once for each plausible value, as described above.

12.4.4 Examining Profiles of Relative Performance by Content Areas

In addition to performance on mathematics and science overall, it was of interest to see how countries performed in the content areas or domains within each subject relative to their performance on the subject overall. There were five content areas in mathematics and five content areas for science that were used in this analysis.[8] The relative performance of the countries in the content areas was examined separately for each subject. TIMSS 2003 computed the average across content area scores for each country, and then displayed country performance in each content area as the difference between the content area average and the overall average. Confidence intervals were estimated for each difference.

In order to do this, TIMSS computed the vector of average proficiencies for each of the content areas on the test, and joined each of these column vectors to form a matrix R_{ks}, where a row contains the average proficiency score for country *k* on scale *s* for a specific subject. This R_{ks} matrix also had a "zeroth" row and column. The elements in r_{k0} contained the average of the elements on the k^{th} row of the R_{ks} matrix. These were the country averages across the content areas. The elements in r_{0s} contained the average of the elements of the *sth* column of the R_{ks} matrix. These are the content area averages across all countries. The element r_{00} contains the overall average for the elements in vector r_{0s} or r_{k0}. Based on this information the matrix I_{ks} was constructed in which the elements are computed as

$$i_{ks} = r_{ks} + r_{00} - r_{0s} - r_{k0}$$

Each of these elements can be considered as the interaction between the performance of country *k* on content area *s*. A value of zero for an element i_{ks} indicates a level of performance for country *k* on content area *s* that would be expected given its performance on other content areas and its performance relative to other countries on that content area. A negative value for an element i_{ks} indicates a performance for country *k* on content area *s* lower than would be expected on the basis of the country's overall performance. A positive value for an element i_{ks} indicates a performance for country *k* on content area *s* better than expected. This procedure was applied to each of the five plausible values and the results averaged.

To construct confidence intervals it was necessary first to estimate the standard error for each content area in each country. These were then combined with an adjustment for multiple comparisons, based on the number of content areas.[9] The imputation portion of the error was obtained from combining the results from the five calculations, one with each separate plausible value.

8 Science at fourth grade had just three content areas.

9 Note that the adjustment was for multiple comparisons between content areas, and not across countries.

To compute the JRR portion of the standard error, the vector of average proficiency was computed for each of the country replicates for each of the content areas on the test. For each country and each content area 75 replicates were created.[10] Each replicate was randomly reassigned to one of 75 sampling zones or replicates. These column vectors were then joined to form a new set of matrices each called R_{ks}^{h} where a row contains the average proficiency for country *k* on content area *s* for a specific subject, for the h^{th} international set of replicates. Each of these R_{ks}^{h} matrices had also a "zeroth" row and column. The elements in r_{k0}^{h} contained the average of the elements on the *kth* row of the R_{ks}^{h} matrix. These are the country averages across the content areas. The elements in r_{0s}^{h} contained the average of the elements of the *sth* column of the R_{ks}^{h} matrix. These were the content area averages across all countries. The element r_{0}^{h} contains the overall average for the elements in vector r_{0s}^{h} or r_{k0}^{h}. Based on this information the set of matrices R_{ks}^{h} were constructed, in which the elements were computed as

$$i_{ks}^{h} = r_{ks}^{h} + r_{00}^{h} - r_{0s}^{h} - r_{k0}^{h}$$

The JRR standard error is then given by the formula

$$jse_{r_{ks}} = \sqrt{\sum_{h}\left(i_{ks} - i_{ks}^{h}\right)^{2}}$$

The overall standard error was computed by combining the JRR and imputation variances. A relative performance was considered significantly different from the expected if the 95% confidence interval built around it did not include zero. The confidence interval for each of the i_{ks} elements was computed by adding and subtracting to the i_{ks} element its corresponding standard error multiplied by the critical value for the number of comparisons.

The critical values were determined by adjusting the critical value for a two-tailed test, at the alpha 0.05 level of significance for multiple comparisons. The critical value for mathematics and science with five content scales was 2.5758. For the three content scales in fourth grade science, the critical value was 2.3939.

12.4.5 Reporting Student Performance on Individual Items

To portray student achievement as fully as possible, the TIMSS 2003 international reports present many examples of the items used in the TIMSS 2003 tests, together with the percentages of students in each country responding correctly to or earning full credit on the items. The base of these percentages was the total number of students that were administered the item. For multiple-choice items, the weighted percentage of students that answered the item correctly was reported. For constructed-response items with more than one

10 In countries where the were less than 75 jackknife zones, 75 replicates were also created by assigning the overall mean to the as many replicates as were necessary to have 75.

score level, it was the weighted percentage of students that achieved full credit on the item. Omitted and not-reached items were treated as incorrect.

When the percent correct for example items was computed, student responses were classified in the following way. For multiple-choice items, the responses to item j were classified as correct (C_j) when the correct option for an item was selected, incorrect (W_j) when the incorrect option or no option at all was selected, invalid (I_j) when two or more choices were made on the same question, not reached (R_j) when it was assumed that the student stopped working on the test before reaching the question, and not administered (A_j) when the question was not included in the student's booklet or had been mistranslated or misprinted. For constructed-response items, student responses to item j were classified as correct (C_j) when the maximum number of points was obtained on the question, incorrect (W_j) when the wrong answer or an answer not worth all the points in the question was given, invalid (N_j) when the student's response was not legible or interpretable, or simply left blank, not reached (R_j) when it was determined that the student stopped working on the test before reaching the question, and not administered (A_j) when the question was not included in the student's booklet or had been mistranslated or misprinted. The percent correct for an item (P_j) was computed as

$$P_j = \frac{c_j}{c_j + w_j + i_j + r_j + n_j}$$

where c_j, w_j, i_j, r_j and n_j are the weighted counts of the correct, wrong, invalid, not reached, and not interpretable responses to item j, respectively.

As described in Chapters 10 and 11, student responses to items in block positions 3 and 6 of the student booklets were found to have different properties to student responses than the same items located in other positions in the booklets. Although these student responses were included in the IRT scaling, albeit with different item parameters, they were not included in the calculation of percent correct on individual example items.

12.5 Examining the TIMSS 2003 Test in the Light of National Curricula

TIMSS 2003 developed international tests of mathematics and science that reflect, as far as possible, the various curricula of the participating countries. The subject matter coverage of these tests was reviewed by the TIMSS 2003 Science and Mathematics Item Review Committee, which consisted of mathematics and science educators and practitioners from around the world, and the tests were approved for use by the National Research Coordinators of the participating countries. Although every effort was made in TIMSS 2003

to ensure the widest possible subject matter coverage, no test can measure all that is taught or learned in every participating country. Given that no test can cover the curriculum in every country completely, the question arises as to how well the items on the tests match the curricula of each of the participating countries. To address this issue, TIMSS 2003 asked each country to indicate which items on the tests, if any, were inappropriate to its curriculum. For each country, in turn, TIMSS 2003 took the list of remaining items, and computed the average percentage correct on these items for that country and all other countries. This allowed each country to select only those items on the tests that they would like included, and to compare the performance of their students on those items with the performance of the students in each of the other participating countries on that set of items. In addition to comparing the performance of all countries on the set of items chosen by each country, the Test-Curriculum Matching Analysis (TCMA) also shows each country's performance on the items chosen by each of the other countries. In these analyses, each country was able to see not only the performance of all countries on the items appropriate for its curriculum, but also the performance of its students on items judged appropriate for the curriculum in other countries. The analytical method of the TCMA is described in Beaton and Gonzalez (1997).

The TCMA results show that the TIMSS 2003 tests provide a reasonable basis for comparing achievement across the participating countries. The analysis shows that omitting items considered by one country to be difficult for their students tends to improve the results for that country, but also tends to improve the results for all other countries as well, so that the overall pattern of relative performance is largely unaffected.

References

Beaton, A.E. & Gonzalez, E. J. (1997). TIMSS Test-Curriculum Matching Analysis. In M. O. Martin & D.L. Kelly (Eds.), *TIMSS technical report, volume II: Implementation and analysis.* Chestnut Hill, MA: Boston College.

Gregory, K. D. & Mullis, I. V. S. (2000). Describing international benchmarks of student achievement. In M. O. Martin, K. D. Gregory, & S. E. Stemler (Eds.), *TIMSS 1999 technical report*. Chestnut Hill, MA: Boston College.

Gonzalez, E. J. & Kennedy, A.M. (2003). Statistical analysis and reporting of the PIRLS data. In M.O. Martin, I.V.S. Mullis, & A.M. Kennedy (Eds.), *PIRLS 2001 technical report*. Chestnut Hill, MA: Boston College.

Johnson, E. G., and Rust, K.F. (1992). Population references and variance estimation for NAEP data. *Journal of Educational Statistics*, 17, 175-190.

Kelly, D. L. (1999). *Interpreting the Third International Mathematics and Science Study (TIMSS) achievement scales using scale anchoring*. Unpublished doctoral dissertation, Boston College.

Mislevy, R.J., Beaton, A.E., Kaplan, B., and Sheenan, K.M. (1992). Estimating Population Characteristics from Sparse Matrix Samples of Item Responses, *Journal of Educational Measurement*, 29, 133-161.

Martin, M.O., Mullis, I.V.S., Gonzalez, E. J., Chrostowski, S.J. (2004). *TIMSS 2003 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.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Chrostowski, S. J. (2004). *TIMSS 2003 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.

Westat, Inc. (1997). *A user's guide to WesVarPC*. Rockville, MD: Westat, Inc.

Wolter, K. M. (1985). *Introduction to variance estimation*. New York: Springer-Verlag.