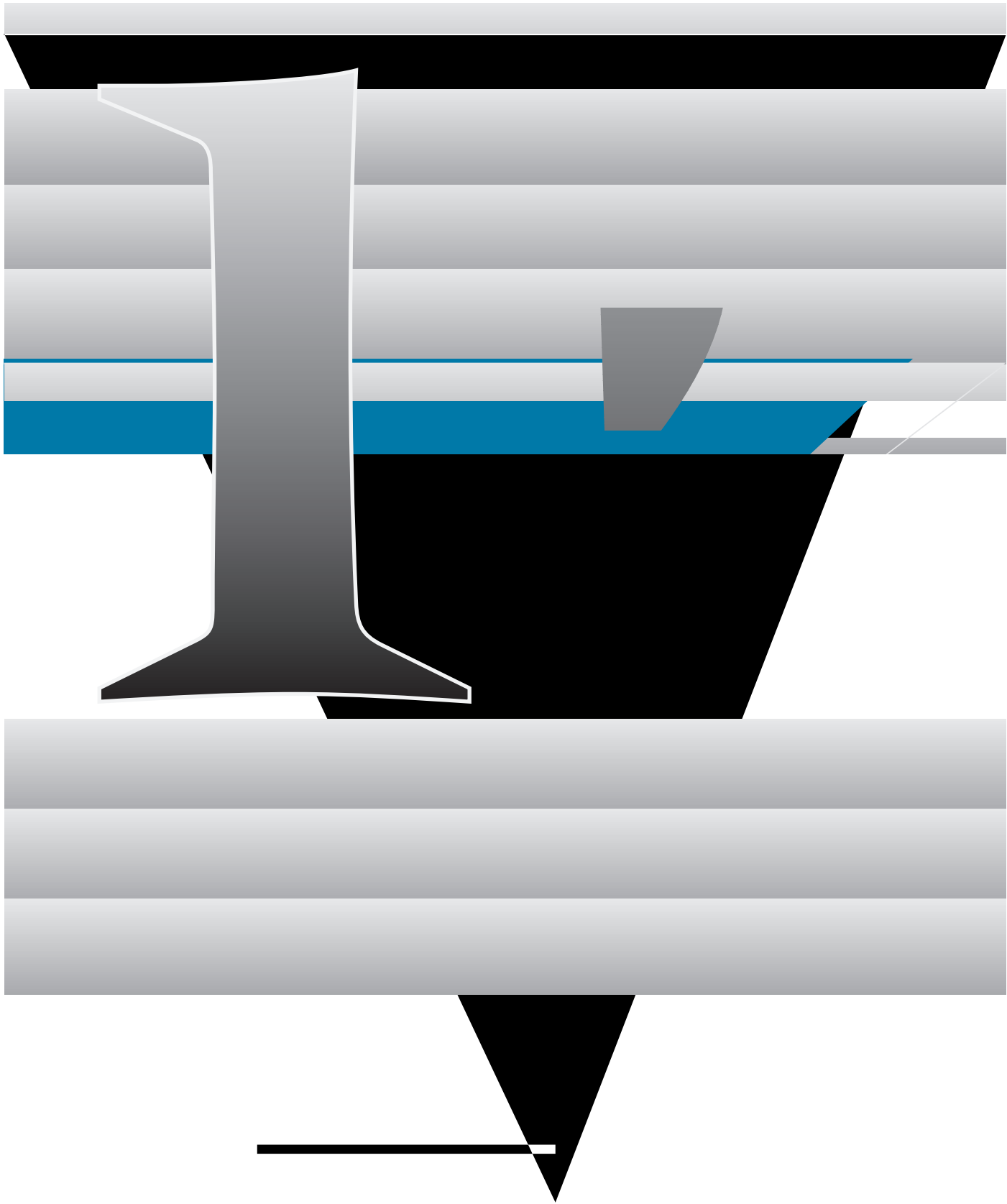

Chapter 12

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 stb d ! % ˜ t sa v h- ` dh t

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.

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

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

Exhibit 12.7

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

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 estim " il ɔ µp a ! W

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 math-

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 12.16 **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				

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

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

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

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

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}

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

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.