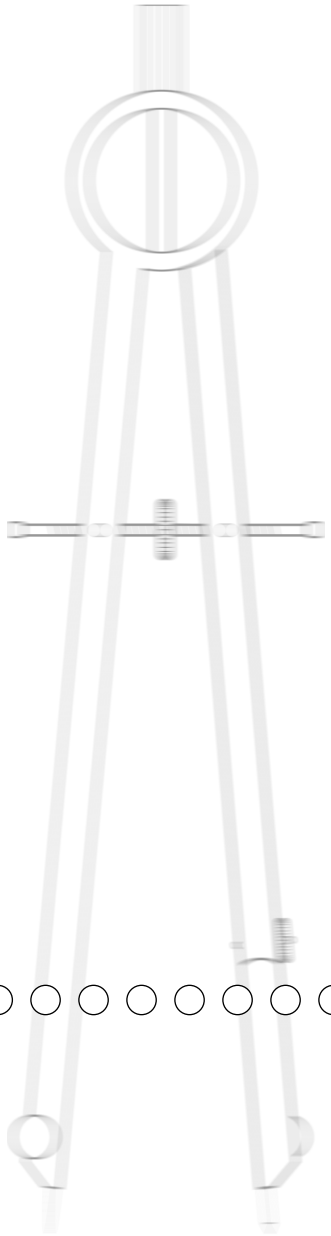

16 Reporting Student Achievement in Mathematics and Science

Eugenio J. Gonzalez
Kelvin D. Gregory

16.1 Overview

As described in earlier chapters, TIMSS 1999 makes extensive use of imputed student proficiency scores to report achievement in mathematics and science, both in the subjects overall and in content areas. This chapter describes the procedures followed in computing the major statistics used to summarize achievement in the international reports (Mullis et al., 2000; Martin et al., 2000), including country means based on plausible values, Bonferroni adjustments for multiple comparisons, reporting trends in achievement, estimating international benchmarks of achievement, and producing profiles of relative performance in subject matter content areas.

16.2 National and International Student Achievement

The item response theory (IRT) scaling procedure described in Chapter 14 yields five imputed scores or plausible values for each student. A national average for each plausible value was computed as the weighted mean

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

where

$\bar{X}_{pvl}$ 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 12

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

The country average is the mean of the five national plausible value means.

The international average for each plausible value was computed as the average of the plausible value for each country

$$\bar{X}_{\bullet pvl} = \frac{\sum_{k=1}^{N} \bar{X}_{pvl,k}}{N}$$

where

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

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

and N is the number of countries.

The international average was the average of the five international mean plausible values.

16.3 Achievement Differences Across Countries

An aim of the TIMSS 1999 international reports is to provide fair and accurate comparisons of student achievement across the participating countries. Most of the exhibits in the 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 multiple comparison charts presented in the TIMSS 1999 international reports allow the comparison of average performance of a country with that of other participating countries. The significance tests reported in these charts include a Bonferroni adjustment for multiple comparisons that holds to 5% the probability of erroneously declaring the mean of one country to be different from that of another country. The Bonferroni adjustment is necessary because that probability greatly increases as the number of simultaneous comparisons increases.

If repeated samples were taken from two populations with the same mean and variance, and in each one the hypothesis that the two means are significantly different at the $\alpha = .05$ level (i.e., with 95% confidence) was tested, then it would be expected in about 5% of the comparisons significant differences would be found between the sample means even though no difference exists in the populations. The probability of finding significant differences when none exist (the so-called type I error) is given by $\alpha = .05$. Con-

versely, the probability of not making such an error is 1 - α, which in the case of a single test is .95. However, comparing the means of three countries involves three tests (country A versus country B, country B versus country C, and country A versus country C). Since these are independent tests, the probability of not making a type I error in any of these tests is the product of the individual probabilities, which is $(1 - \alpha)(1 - \alpha)(1 - \alpha)$. With $\alpha = .05$, the overall probability of not making a type I error is only .873, which is considerably less than the probability for a single test. As the number of tests increases, the probability of not making a type I error decreases rapidly, and conversely, the probability of making such an error increases.

Several methods can be used to correct for the increased probability of a type I error while making many simultaneous comparisons. Dunn (1961) developed a procedure that is appropriate for testing a set of a priori hypotheses while controlling the probability that the type I error will occur. In this procedure, the value α is adjusted to compensate for the increase in the probability of making the error (the Dunn-Bonferroni procedure for multiple a priori comparisons; Winer, Brown, and Michels, 1991).

The TIMSS 1999 international reports contain multiple comparison exhibits that show the statistical significance of the differences between all possible combinations of the 38 participating countries. There were (38*37)/2 = 703 possible differences. In the Bonferroni procedure the significance level (α) of a statistical test is adjusted by the number of comparisons that are planned and then looking up the appropriate quantile from the normal distribution. In deciding on the appropriate adjustment of the significance level for TIMSS, it was necessary to decide how the multiple comparison exhibits would most likely be used. A very conservative approach would be to adjust the significance level to compensate for all of the 703 possible comparisons among the 38 countries concerned. However, this risks an error of a different kind, that of concluding that a difference in sample means is not significant when in fact there is a difference in the population means.

Since most users are likely to be interested in comparing a single country with all other countries, rather than in making all possible between-country comparisons at once, the more realistic approach of using the number of countries (minus one) to adjust the significance level was adopted. This meant that the number

of simultaneous comparisons to be adjusted for was 37 instead of 703. The critical value for a 95% significance test adjusted for 37 simultaneous comparisons is 3.2049, from the appropriate quantiles from the normal (Gaussian) distribution.

Mean proficiencies were considered significantly different if the absolute difference between them, divided by the standard error of the difference, was greater than the critical value. For differences between countries, which can be considered as independent samples, the standard error of the difference in means was 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. Exhibit 16.1 shows the means and standard errors for mathematics and science used in the calculation of statistical significance. By applying the Bonferroni adjustment, it was possible to state that, for any given row or column of the multiple comparison chart, the differences n countries are statistically significant at the 95% level of confidence.

Exhibit 16.1 Means and Standard Errors for Multiple Comparisons Exhibits

Country	Math		Science	
	Mean	S.E.	Mean	SE
Australia	525.080	4.840	540.258	4.395
Belgium (Flemish)	557.958	3.291	534.858	3.074
Bulgaria	510.591	5.850	518.011	5.355
Canada	530.753	2.460	533.082	2.063
Chile	392.494	4.364	420.372	3.720
Chinese Taipei	585.117	4.033	569.076	4.425
Cyprus	476.382	1.792	460.238	2.350
Czech Republic	519.874	4.176	539.417	4.171
England	496.330	4.150	538.468	4.750
Finland	520.452	2.743	535.207	3.471
Hong Kong, SAR	582.056	4.280	529.547	3.655
Hungary	531.601	3.674	552.381	3.693
Indonesia	403.070	4.896	435.472	4.507
Iran, Islamic Rep.	422.148	3.397	448.003	3.765
Israel	466.336	3.932	468.062	4.936
Italy	479.479	3.829	493.281	3.881
Japan	578.604	1.654	549.653	2.227
Jordan	427.664	3.592	450.343	3.832
Korea, Rep. of	587.152	1.969	548.642	2.583
Latvia (LSS)	505.059	3.435	502.693	4.837
Lithuania	481.567	4.281	488.152	4.105
Macedonia, Rep. of	446.604	4.224	458.095	5.240
Malaysia	519.256	4.354	492.431	4.409
Moldova	469.231	3.883	459.137	4.029
Morocco	336.597	2.573	322.816	4.319
Netherlands	539.875	7.147	544.749	6.870
New Zealand	490.967	5.178	509.634	4.905
Philippines	344.905	5.979	345.229	7.502
Romania	472.440	5.787	471.865	5.823
Russian Federation	526.023	5.935	529.220	6.395
Singapore	604.393	6.259	567.894	8.034
Slovak Republic	533.953	3.959	535.009	3.290
Slovenia	530.113	2.777	533.255	3.218
South Africa	274.503	6.815	242.640	7.850
Thailand	467.377	5.088	482.314	3.983
Tunisia	447.925	2.430	429.512	3.436
Turkey	428.606	4.343	432.951	4.268
United States	501.633	3.971	514.915	4.553

16.4 Comparing Achievement with the International Mean

Many of the data exhibits in the TIMSS 1999 international reports show countries' mean achievement compared with the international mean. Since this results in 38 simultaneous comparisons, the critical value was adjusted to 3.2125 using the Dunn-Bonferroni procedure.

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* was

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

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*

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

The imputation component of the standard error 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:

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

16.5 Trends in Achievement

TIMSS 1999 was designed to enable comparisons between a country's achievement on the 1995 and 1999 assessments. A total of 26 countries participated at the eighth grade in both assessments. Although all countries had acceptable sampling participation in 1999, three countries – Israel, South Africa, and Thailand – failed to meet sampling guidelines in 1995, and were omitted from the calculation of trends.

When assessing whether eighth-grade achievement had significantly changed from 1995 to 1999, TIMSS applied a Bonferroni correction for 23 simultaneous comparisons.

Of the 23 countries with eighth-grade data in both 1995 and 1999, 17 also had fourth-grade data from 1995.To show how countries' relative performance changed from fourth to eighth grade, TIMSS calculated the significance of the difference between each country's mean and the mean across all 17 countries, adjusting for 17 simultaneous comparisons.

The means and standard errors of the 1995 fourth- and eighth-grade students and the 1999 eighth-grade students for countries included in the trend exhibits from the international reports are shown in Exhibit 16.2 and 16.3 for mathematics and science, respectively.

Exhibit 16.2 Means and Standard Errors for Mathematics Trend Exhibits

Country	4th Grade 1995		8th Grade 1995		8th Grade 1999	
	Mean	S.E.	Mean	S.E.	Mean	S.E.
Australia	517.190	2.991	518.872	3.803	525.080	4.840
Belgium (Flemish)	--	--	549.679	5.867	557.958	3.291
Bulgaria	--	--	526.780	5.798	510.591	5.850
Canada	505.693	3.385	520.544	2.174	530.753	2.460
Cyprus	474.930	3.221	467.533	2.237	476.382	1.792
Czech Republic	540.503	3.065	545.551	4.521	519.874	4.176
England	483.980	3.345	497.669	2.975	496.330	4.150
Hong Kong, SAR	556.993	3.986	568.886	6.136	582.056	4.280
Hungary	521.326	3.607	526.626	3.182	531.601	3.674
Iran, Islamic Rep.	386.969	4.992	418.450	3.871	422.148	3.397
Israel	--	--	513.315	6.224	481.609	4.706
Italy	510.028	4.681	491.015	3.370	485.411	4.825
Japan	567.219	1.855	581.069	1.575	578.604	1.654
Korea, Rep. of	580.904	1.802	580.720	1.962	587.152	1.969
Latvia (LSS)	498.939	4.557	488.281	3.578	505.059	3.435
Lithuania	--	--	471.839	4.101	481.567	4.281
Netherlands	549.233	2.959	528.843	6.147	539.875	7.147
New Zealand	469.180	4.367	500.944	4.722	490.967	5.178
Romania	--	--	473.729	4.571	472.440	5.787
Russian Federation	--	--	523.618	5.331	526.023	5.935
Singapore	590.187	4.536	608.593	3.978	604.393	6.259
Slovak Republic	--	--	533.991	3.076	533.953	3.959
Slovenia	525.162	3.174	530.953	2.756	530.113	2.777
South Africa	--	--	277.705	9.212	274.503	6.815
Thailand	--	--	516.216	6.050	467.377	5.088
United States	517.847	2.950	492.318	4.746	501.633	3.971
International Avg.	517.428	0.892	519.413	0.861	521.303	0.922

Exhibit 16.3 Means and Standard Errors for Science Trend Exhibits

Country	4th Grade 1995		8th Grade 1995		8th Grade 1999	
	Mean	S.E.	Mean	S.E.	Mean	S.E.
Australia	541.322	3.630	526.502	4.028	540.258	4.395
Belgium (Flemish)	--	--	532.897	6.391	534.858	3.074
Bulgaria	--	--	545.245	5.203	518.011	5.355
Canada	525.343	3.053	513.988	2.638	533.082	2.063
Cyprus	450.029	3.202	452.012	2.091	460.238	2.350
Czech Republic	531.713	3.038	554.955	4.547	539.417	4.171
England	527.670	3.094	533.348	3.570	538.468	4.750
Hong Kong, SAR	507.824	3.330	509.730	5.785	529.547	3.655
Hungary	507.744	3.405	536.754	3.106	552.381	3.693
Iran, Islamic Rep.	380.184	4.553	462.872	3.628	448.003	3.765
Israel	--	--	508.957	6.349	484.303	5.652
Italy	523.826	4.601	497.248	3.551	497.900	4.752
Japan	553.183	1.765	554.475	1.754	549.653	2.227
Korea, Rep. of	575.571	2.119	545.778	2.045	548.642	2.583
Latvia (LSS)	486.383	4.905	476.156	3.332	502.693	4.837
Lithuania	---	--	463.564	4.049	488.152	4.105
Netherlands	530.332	3.173	541.418	6.029	544.749	6.870
New Zealand	505.117	5.299	510.862	4.858	509.634	4.905
Romania	--	--	470.926	5.134	471.865	5.823
Russian Federation	--	--	522.581	4.486	529.220	6.395
Singapore	523.400	4.803	580.352	5.483	567.894	8.034
Slovak Republic	--	--	531.913	3.309	535.009	3.290
Slovenia	521.966	4.030	540.980	2.794	533.255	3.218
South Africa	--	--	262.941	11.092	242.640	7.850
Thailand	--	--	510.045	4.704	482.314	3.983
United States	541.863	3.258	512.587	5.560	514.915	4.553
International Avg.	513.734	0.888	518.137	0.889	521.211	0.897

Because of differences from 1995 to 1999 in the sampling of student populations, the results for Israel and Italy in exhibits of trend data differ from those containing just 1999 data. In TIMSS 1995, Israel tested only Hebrew-speaking students, while in 1999 the target population included both Hebrew and Arab speaking students. To provide meaningful trend analysis, TIMSS compared 1995 and 1999 using the Hebrew-speaking part of the population only. In Italy, the 1995 assessment sampled students from most but not all provinces, whereas in 1999 all provinces were included. The TIMSS 1999 trend data for Italy represents those provinces that participated in TIMSS 1995 only.

16.6 International Benchmarks of Achievement

In order to provide more information about student achievement, TIMSS identified four points on each of the mathematics and science scales for use as international benchmarks. The Top 10% benchmark was defined as the 90th percentile on the TIMSS scale, computed across all students in all participating countries, with countries weighted in proportion to the size of

their eighth-grade population. This point on each scale (mathematics and science) is the point above which the top 10% of students in the 1999 TIMSS assessment scored. The upper quarter benchmark is the 75th percentile on the scale, above which the top 25% of students scored. The median benchmark is the 50th percentile, above which the top half of students scored. Finally, the lower quarter benchmark is the 25th percentile, the point reached by the top 75% of students.

The percentage of students in each country meeting or exceeding the marker levels were reported. In computations of the international benchmarks of achievement, each country was weighted to contribute as many students as there were students in the target population. In other words, each country's contribution to setting the international benchmarks was proportional to the estimated population enrolled in the eighth grade. Exhibit 16.4 shows the contribution of each country to the estimation of the international benchmarks.

Exhibit 16.4 Estimated Enrollment at the Eighth Grade Within Country

Country	Sample Size	Estimated Enrollment
Australia	4032	260130
Belgium (Flemish)	5259	65539
Bulgaria	3272	88389
Canada	8770	371062
Chile	5907	208910
Chinese Taipei	5772	310429
Cyprus	3116	9786
Czech Republic	3453	119462
England	2960	552231
Finland	2920	59665
Hong Kong, SAR	5179	79097
Hungary	3183	111298
Indonesia	5848	1956221
Iran, Islamic Rep.	5301	1655741
Israel	4195	81486
Italy	3328	548711
Japan	4745	1416819
Jordan	5052	89171
Korea, Rep. of	6114	609483
Latvia (LSS)	2873	18122
Lithuania	2361	40452
Macedonia, Rep. of	4023	30280
Malaysia	5577	397762
Moldova	3711	59956
Morocco	5402	347675
Netherlands	2962	198144
New Zealand	3613	51553
Philippines	6601	1078093
Romania	3425	2596
Russian Federation	4332	2057413
Singapore	4966	41346
Slovak Republic	3497	72521
Slovenia	3109	23514
South Africa	8146	844706
Thailand	5732	727087
Tunisia	5051	139639
Turkey	7841	618058
United States	9072	3336295

If all countries had the same distribution of student achievement, approximately 10% of students within each country would be above the 90th percentile in the international distribution, regardless of the country's population size. That this is not the case, and that countries vary considerably, is evident from the fact that, 46% of students in Singapore reached the top 10% benchmark, compared to less than 1% in Tunisia, the Philippines, South Africa, and Morocco.

Because of the imputation technology used to derive the proficiency scores, the international benchmarks had to be computed once for each of the five plausible values, and the results averaged to arrive at the final figure. The standard errors presented in the exhibits are computed taking into account the sampling design as well as the variance due to imputation. The international benchmarks are presented in Exhibit 16.5 and 16.6 for mathematics and science, respectively.

Exhibit 16.5 International Benchmarks of Mathematics Achievement for the Eighth Grade

Proficiency Score	25th Percentile	50th Percentile	75th Percentile	90th Percentile
Plausible Value 1	396.86	479.20	554.49	615.15
Plausible Value 2	395.76	478.79	554.74	615.37
Plausible Value 3	395.62	478.56	554.83	616.23
Plausible Value 4	394.57	478.09	554.03	615.02
Plausible Value 5	396.30	479.10	554.56	615.76
Mean Plausible Value	395.82	478.75	554.53	615.51

Exhibit 16.6 International Benchmarks of Science Achievement for the Eighth Grade

Proficiency Score	25th Percentile	50th Percentile	75th Percentile	90th Percentile
Plausible Value 1	409.03	487.76	558.66	617.01
Plausible Value 2	409.87	487.61	557.60	615.88
Plausible Value 3	410.38	488.04	557.27	616.12
Plausible Value 4	410.05	487.54	557.47	615.82
Plausible Value 5	410.87	487.59	557.79	615.88
Mean Plausible Value	410.04	487.71	557.76	616.14

16.7 Gender Differences within Countries

TIMSS reported gender differences in overall student achievement in mathematics and science overall, as well as in 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. The significance test was adjusted for multiple comparisons based on the number of countries presented.

Because in most countries males and females attend the same schools, the samples of males and females cannot be treated as independent for the purpose of statistical tests. Accordingly, TIMSS used a jackknife procedure applicable to correlated samples for estimating the standard error of the male-female difference. This involves computing the differences between boys and girls once for each of the 75 replicate samples, and five more times, once for each plausible value, as described in Chapter 12.

16.8 Relative Performance by Content Areas

In addition to performance in mathematics and science overall, it was of interest to see how countries performed on the content areas relative to performance on the subject overall. Five content areas in mathematics and six in science were used in this analysis. Relative performance on the content areas was examined separately for the two subjects. TIMSS 1999 computed the average across content area scores for each country, and then displayed country performance in each content area as the difference between that 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 column vector to form a matrix called R_{ks}, where a row contains the average proficiency score for country k on scale s for a specific subject. This R_{ks} matrix had also a "zeroth" row and column. The elements in r_{k0} contains the average of the elements on the kth row of the R_{ks} matrix. These are the country averages across the content areas. The elements in r_{0s} contains 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_{0j} 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 over-

all performance. A positive value for an element i_{ks} indicates a better than expected performance for country k on the content areas. This procedure was applied to each of the 5 plausible values and the results were averaged.

To construct confidence intervals the standard error for each content area in each country first had to be estimated. These were then combined with a Bonferroni adjustment, based on the number of content areas. The imputation portion of the error was obtained from combining the results from the five calculations, one with each separate plausible value.

To compute the sampling portion of the standard error, the vector of average proficiency was computed for each of the country replicates for each content area on the test. For each country and each content area 75 replicates were created.[1] Each replicate was randomly reassigned to one of 75 sampling zones or replicates (h). These column vectors were then joined to form a new set of matrices each called R^h_{ks}, where a row contains the average proficiency for country k on content area s for a specific subject, for the hth international set of replicates. Each of these R^h_{ks} matrices has also a "zeroth" row and column. The elements in r^h_{k0} contain the average of the elements on the kth row of the R^h_{ks} matrix. These are the country averages across the content areas. The elements in r^h_{0s} contain the average of the elements of the sth column of the R^h_{ks} matrix. These are the content area averages across all countries. The element r^h_{00} contains the overall average for the elements in vector r^h_{0j} or r^h_{k0}. Based on this information the set of matrices $R^h_{ks'}$ were constructed, in which the elements were computed as

$$i^h_{ks} = r^h_{ks} + r^h_{00} - r^h_{0s} - r^h_{k0}$$

The JRR standard error is then given by the formula

$$jse_{r_{ks}} = \sqrt{\sum_h (i_{ks} - i^h_{ks})^2}$$

○○○

1. 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.

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 to and subtracting from 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 according the Dunn-Bonferroni procedure. The critical value for mathematics, with five content scales, was 2.5758, and for science with six content scales, was 2.6383.

16.9 Percent Correct for Individual Items

To portray student achievement as fully as possible, the TIMSS 1999 international reports present many examples of the items used in the TIMSS 1999 tests, together with the percentage of students in each country responding correctly to the item. This percentage was based on the total number of students tested on the items. Omitted and not-reached items were treated as incorrect. For multiple-choice items the percentage was the weighted percentage of students that answered the item correctly. For free-response items with more than one score level, it was the weighted percentage of students that achieved the highest score possible.

When the% correct for example items was computed, student responses were classified in the following way. For multiple-choice items, a response to item j was classified as correct (C_j) when the correct option was selected, incorrect (W_j) when the incorrect option or no option was selected, invalid (I_j) when two or more options were selected, 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 free-response items, student responses to item j were classified as correct (C_j) when the maximum number of points was obtained, 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 response was not legible or interpretable or was 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% 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.

16.10 The Test-Curriculum Matching Analysis

TIMSS 1999 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 1999 Subject Matter Item Replacement 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 (NRCs) of the participating countries. Although every effort was made in TIMSS 1999 to ensure the widest possible subject matter coverage, no test can measure all that is taught or learned in every participating country. The question therefore arises how well the items on the tests match the curricula of the participating countries. To address this issue, TIMSS 1999 asked each country to indicate which items on the tests, if any, were inappropriate to its curriculum. For each country, TIMSS 1999 then took the list of remaining items and computed the average percentage correct on those 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 that of the students in the other participating countries. However, 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 1999 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 tends to improve the results for all other countries also, 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 Martin, M.O. and Kelly, D.L. (Eds.), *TIMSS technical report, volume II: Implementation and Analysis.* Chestnut Hill, MA: Boston College.

Dunn, O.J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association, 56,* 52-64.

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

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

Winer, B.J., Brown, D.R., & Michels, K.M. (1991). *Statistical principles in experimental design.* New York: McGraw Hill.