8 Reporting Student Achievement in Mathematics and Science Literacy, Advanced Mathematics, and Physics

Eugenio J. Gonzalez
Boston College

8.1 STANDARDIZING THE TIMSS INTERNATIONAL SCALE SCORES

The item response theory (IRT) scaling procedures described in Chapter 7 yielded imputed proficiency scores ("plausible values") in a logit metric, with the majority of scores falling in the range from -3 to +3. These scores were transformed onto an international achievement scale with mean 500 and standard deviation 100 – a scale that was more suited to reporting international results. This scale avoids negative values for student scale scores and eliminates the need for decimal points in reporting student achievement.

Since a plausible value is an imputed score that includes a random component, it is customary when using this method to draw a number of plausible values for each respondent (usually five). Each analysis is then carried out five times, once with each plausible value, and the results are averaged to get the best overall result. The variability among the five results is a measure of the error due to imputation and, where it is large, may be combined with jackknife estimates of sampling error to give a more realistic indication of the total variability of a statistic. Since the TIMSS final year of secondary school population (Population 3) showed significant variability between results from the five plausible values, it was decided to incorporate this variation in the analytic procedures.

In order to ensure that the mean of the TIMSS international achievement scale was close to the average student achievement level across countries, it was necessary to estimate the mean and standard deviation of the logit scores for all participating students. To accomplish this, the logit scores for all students from all countries were combined into a standardization sample. Each country was equally weighted. The means and standard deviations derived from this procedure are shown in Tables 8.1 through 8.12. These tables show the average logit for each of the five plausible values.

Table 8.1 Standardization Parameters of International Mathematics Literacy Scores

Scale	Mean Logit	Standard Deviation
Mathematics Literacy Plausible Value #1	0.3490	1.1086
Mathematics Literacy Plausible Value #2	0.3503	1.1012
Mathematics Literacy Plausible Value #3	0.3495	1.1027
Mathematics Literacy Plausible Value #4	0.3507	1.1038
Mathematics Literacy Plausible Value #5	0.3489	1.1040

Table 8.2 Standardization Parameters of International Science Literacy Scores

Scale	Mean Logit	Standard Deviation
Science Literacy Plausible Value #1	0.3393	0.9421
Science Literacy Plausible Value #2	0.3439	0.9407
Science Literacy Plausible Value #3	0.3425	0.9423
Science Literacy Plausible Value #4	0.3417	0.9435
Science Literacy Plausible Value #5	0.3414	0.9405

Table 8.3 Standardization Parameters of International Advanced Mathematics Scores

Scale	Mean Logit	Standard Deviation
Advanced Mathematics Plausible Value #1	-0.1156	0.8664
Advanced Mathematics Plausible Value #2	-0.1195	0.8657
Advanced Mathematics Plausible Value #3	-0.1134	0.8674
Advanced Mathematics Plausible Value #4	-0.1163	0.8684
Advanced Mathematics Plausible Value #5	-0.1191	0.8699

Table 8.4 Standardization Parameters of International Numbers and Equations Scores

Scale	Mean Logit	Standard Deviation
Numbers and Equations Plausible Value #1	0.0450	1.0782
Numbers and Equations Plausible Value #2	0.0567	1.0787
Numbers and Equations Plausible Value #3	0.0490	1.0788
Numbers and Equations Plausible Value #4	0.0552	1.0751
Numbers and Equations Plausible Value #5	0.0559	1.0817

Table 8.5 Standardization Parameters of International Calculus Scores

Scale	Mean Logit	Standard Deviation
Calculus Plausible Value #1	-0.3704	1.1983
Calculus Plausible Value #2	-0.3608	1.2005
Calculus Plausible Value #3	-0.3644	1.1984
Calculus Plausible Value #4	-0.3604	1.2015
Calculus Plausible Value #5	-0.3590	1.2062

Table 8.6 Standardization Parameters of International Geometry Scores

Scale	Mean Logit	Standard Deviation
Geometry Plausible Value #1	-0.1862	0.9357
Geometry Plausible Value #2	-0.1790	0.9334
Geometry Plausible Value #3	-0.1837	0.9345
Geometry Plausible Value #4	-0.1781	0.9327
Geometry Plausible Value #5	-0.1789	0.9371

Table 8.7 Standardization Parameters of International Physics Scores

Scale	Mean Logit	Standard Deviation
Physics Plausible Value #1	-0.5506	0.7215
Physics Plausible Value #2	-0.5457	0.7247
Physics Plausible Value #3	-0.5464	0.7240
Physics Plausible Value #4	-0.5505	0.7255
Physics Plausible Value #5	-0.5477	0.7249

Table 8.8 Standardization Parameters of International Mechanics Scores

Scale	Mean Logit	Standard Deviation
Mechanics Plausible Value #1	-0.7019	1.0645
Mechanics Plausible Value #2	-0.7052	1.0630
Mechanics Plausible Value #3	-0.7056	1.0599
Mechanics Plausible Value #4	-0.6994	1.0638
Mechanics Plausible Value #5	-0.7036	1.0636

Table 8.9 Standardization Parameters of International Electricity and Magnetism Scores

Scale	Mean Logit	Standard Deviation
Electricity and Magnetism Plausible Value #1	-0.6917	0.8441
Electricity and Magnetism Plausible Value #2	-0.6994	0.8490
Electricity and Magnetism Plausible Value #3	-0.6960	0.8472
Electricity and Magnetism Plausible Value #4	-0.6903	0.8482
Electricity and Magnetism Plausible Value #5	-0.6968	0.8455

Table 8.10 Standardization Parameters of International Heat Scores

Scale	Mean Logit	Standard Deviation
Heat Plausible Value #1	-0.3200	0.9414
Heat Plausible Value #2	-0.3243	0.9458
Heat Plausible Value #3	-0.3203	0.9432
Heat Plausible Value #4	-0.3183	0.9472
Heat Plausible Value #5	-0.3238	0.9405

Table 8.11 Standardization Parameters of International Wave Phenomena Scores

Scale	Mean Logit	Standard Deviation
Wave Phenomena Plausible Value #1	-0.3260	1.0758
Wave Phenomena Plausible Value #2	-0.3288	1.0774
Wave Phenomena Plausible Value #3	-0.3317	1.0711
Wave Phenomena Plausible Value #4	-0.3226	1.0753
Wave Phenomena Plausible Value #5	-0.3316	1.0750

Table 8.12 Standardization Parameters of International Particle, Quantum, Astrophysics and Relativity Scores

Scale	Mean Logit	Standard Deviation
Particle, Quantum, Astrophysics & Relativity Plausible Value #1	-0.6179	0.9492
Particle, Quantum, Astrophysics & Relativity Plausible Value #2	-0.6199	0.9469
Particle, Quantum, Astrophysics & Relativity Plausible Value #3	-0.6205	0.9439
Particle, Quantum, Astrophysics & Relativity Plausible Value #4	-0.6174	0.9466
Particle, Quantum, Astrophysics & Relativity Plausible Value #5	-0.6220	0.9406

Each country was weighted to contribute equally to the calculation of the international mean and standard deviation. The transformation applied to the plausible value logit scores was

$$S_{ijk} = 500 + 100 * \left(\frac{\theta_{ijk} - \bar{\theta}_j}{SD_{\theta_j}} \right)$$

where S_{ijk} is the standardized scale score with mean 500 and standard deviation 100 for student *i*, in plausible value *j*, in country *k*; θ_{ijk} is the logit score for the same student, $\bar{\theta}_j$ is the weighted average across all countries on plausible value *j*, and SD_{θ_j} is the standard deviation across all countries on plausible value *j*. Since five plausible values (logit scores) were drawn for each student, each of these was transformed so that the international mean of the result scores was 500, with standard deviation 100.

Because plausible values are actually random draws from the estimated distribution of student achievement and not actual student scores, student proficiency estimates were occasionally obtained that were unusually high or low. Where a transformed plausible value fell below 10, the value was recoded to 10, making 10 the lowest score on the transformed scale. This happened in very few cases across the countries. Where a transformed plausible value surpassed 990, the value was recoded to 990, making 990 the highest score on the transformed scale.

8.2 STANDARDIZING THE INTERNATIONAL ITEM DIFFICULTIES

To help readers of the TIMSS international reports understand the international achievement scales, TIMSS produced item difficulty maps that showed the location on the scales of several items from the subject matter content areas covered by the mathematics and science tests. In order to locate the example items on the achievement scales, the item difficulty parameter for each item had to be transformed from its original logit metric to the metric of the international achievement scales (a mean of 500 and standard deviation of 100).

The procedure for deriving the international item difficulties is described in Chapter 7. The international item difficulties obtained from the scaling procedure represent the proficiency level of a person who has a 50 percent chance of responding to the item correctly. For the item difficulty maps it was preferred that the difficulty correspond to the proficiency level of a person showing greater mastery of the item. For this reason it was decided to calibrate these item difficulties in terms of the proficiency of a person with a 65 percent chance of responding correctly.

In order to derive the item difficulties for the item difficulty maps, the original item difficulties obtained from the scaling procedure were transformed in two ways. First they were moved along the logit scale from the point where a student with that proficiency would have a 50 percent chance of responding correctly to the point where the student would have a 65 percent chance of responding correctly. This was achieved by adding the natural log of the odds of a 65 percent response rate to the original log odds, since the logit metric allows this addition to take place in a straightforward manner. Second, the new logit item difficulty was transformed into the international achievement scale. This was done five times, once with the mean and standard deviation of each plausible value (shown in Tables 8.1 through 8.12). The average of this transformation was taken as the transformed international item difficulty:

$$d'_i = \left(\frac{1}{5}\right) \times \sum_{j=1}^{5} \left(500 + 100 \times \left(\frac{d_i + \ln(0.65/0.35) - \bar{\theta}_j}{SD_{\theta_j}}\right)\right)$$

where d'_i is the item difficulty for item i transformed onto the international standardized scale metric, d'_i is the item difficulty in the original logit metric, $\bar{\theta}_j$ is the mean logit score on each plausible value for the scale to which the item is assigned, and SD_{θ_j} is the standard deviation of the plausible values. For the purpose of transforming the item difficulties, only the difficulty of the items on the overall scale was used. That is, the difficulty for an item is presented as part of one of the four overall scales reported: mathematics literacy, science literacy, advanced mathematics, or physics.

8.3 MULTIPLE COMPARISONS OF ACHIEVEMENT

An essential purpose of the TIMSS international reports is to provide fair and accurate comparisons of student achievement across the participating countries. Most of the tables in the reports summarize student achievement by means of a statistic such as a mean or percentage, and each summary statistic is accompanied by its standard error, which is a measure of the variability in the statistic resulting from the sampling process. When comparing the performance of students from two 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 international report for Population 3 are designed to help the reader compare the average performance of a country with that of other participating countries of interest. The significance tests reported in these charts are based on a Bonferroni procedure for multiple comparisons that holds to 5 percent the probability of erroneously declaring the mean of one country to be different from that of another country.

If we were to take repeated samples from two populations with the same mean and test the hypothesis that the means from these two samples are significantly different at the $\alpha = .05$ level, i.e. with 95 percent confidence, then in about 5 percent of the comparisons we would expect to find significant differences between the sample means even though we know that there is no difference between the population means. In this example with one 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 a type I error is $1 - \alpha$, which in the case of a single test is .95. However, if we wish to compare the means of three countries, this 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, and conversely, the probability of making a type I 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. When using this procedure, the researcher adjusts the value α when making multiple simultaneous comparisons to compensate for the increase in the probability of making a type I error. This is known as the Dunn-Bonferroni procedure for multiple a priori comparisons (Winer, Brown, and Michels, 1991).

In this procedure the significance level of the test of the difference between means is adjusted by dividing the significance level (α) by the number of comparisons that are planned and then looking up the appropriate quantile from the normal distribution. In

deciding the number of comparisons, and hence the appropriate adjustment to the significance level for TIMSS, it was necessary to decide how the multiple comparison tables would most likely be used. One approach would have been to adjust the significance level to compensate for all possible comparisons between the countries presented in the table. This would have meant adjusting the significance level for 420 comparisons for mathematics and science literacy. In decision-making terms this would have been a very conservative procedure, however, and would have run the risk of making an error of a different kind, that of concluding that a difference between sample means is not significant when in fact there is a difference between 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, a more realistic approach would be to adjust the significance level for a number of comparisons equal to the number of countries (minus one). This was the approach adopted in TIMSS. From this perspective, for mathematics and science literacy, the number of simultaneous comparisons to be adjusted for is 20 instead 420. The number of comparisons is 15 for mathematics and also 15 for physics. As a consequence, we used the critical values shown in Table 8.13, given by the appropriate quantiles from the normal (Gaussian) distribution.

Table 8.13 Critical Values Used for the Multiple Comparison Figures in TIMSS International Report

	Alpha Level	Number of Comparisons	Critical Value
Mathematics and Science Literacy	0.05	20	3.0233
Avanced Mathematics	0.05	15	2.9353
Physics	0.05	15	2.9353

Two means were considered significantly different from each other if the absolute differences between them was greater than the critical value multiplied by the standard error of the difference. The standard error of the difference between the two means was computed as the square root of the sum of the squared standard errors of the mean:

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

where se_1 and se_2 are the standard errors for each of the means being compared, respectively, computed using the jackknife method of variance estimation. Table 8.14 shows the means and standard errors used in the calculation of statistical significance between means for mathematics and science literacy, mathematics literacy, science literacy, advanced mathematics, and physics. By applying the Bonferroni correction, we were able to state that, for any given row or column of the multiple comparison chart, the differences between countries shown in the chart are statistically significant at the 95 percent level of confidence.

Table 8.14 Means and Standard Errors for Multiple Comparisons Figures

Country	Mathematics and Science Literacy		Mathematics Literacy		Science Literacy		Advanced Mathematics		Physics	
	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
Australia	525	9.5	522	9.3	527	9.8	525	11.6	518	6.2
Austria	519	5.4	518	5.3	520	5.6	436	7.2	435	6.4
Canada	526	2.6	519	2.8	532	2.6	509	4.3	485	3.3
Cyprus	447	2.5	446	2.5	448	3.0	518	4.3	494	5.8
Czech Republic	476	10.5	466	12.3	487	8.8	469	11.2	451	6.2
Denmark	528	3.2	547	3.3	509	3.6	522	3.4	534	4.2
France	505	4.9	523	5.1	487	5.1	557	3.9	466	3.8
Germany	496	5.4	495	5.9	497	5.1	465	5.6	522	11.9
Greece	-	-	-	-	-	-	513	6.0	486	5.6
Hungary	477	3.0	483	3.2	471	3.0	-	-	-	-
Iceland	541	1.6	534	2.0	549	1.5	-	-	-	-
Italy	475	5.3	476	5.5	475	5.3	474	9.6	.	.
Latvia (LSS)	-	-	-	-	-	-	-	-	488	21.5
Lithuania	465	5.8	469	6.1	461	5.7	516	2.6	-	-
Netherlands	559	4.9	560	4.7	558	5.3	-	-	-	-
New Zealand	525	4.7	522	4.5	529	5.2	-	-	-	-
Norway	536	4.0	528	4.1	544	4.1	-	-	581	6.5
Russian Federation	476	5.8	471	6.2	481	5.7	542	9.2	545	11.6
Slovenia	514	8.2	512	8.3	517	8.2	475	9.2	523	15.5
South Africa	352	9.3	356	8.3	349	10.5	-	-	-	-
Sweden	555	4.3	552	4.3	559	4.4	512	4.4	573	3.9
Switzerland	531	5.4	540	5.8	523	5.3	533	5.0	488	3.5
United States	471	3.1	461	3.2	480	3.3	442	5.9	423	3.3

A dash (–) indicates country did not participate in assessment.
S.E. = standard error.

8.4 ESTIMATING THE ACHIEVEMENT OF THE TOP 5 PERCENT, 10 PERCENT, AND 25 PERCENT OF STUDENTS IN THE SCHOOL-LEAVING AGE COHORT

As indicated by the test coverage indices, the samples of all final-year students in some countries represented nearly all of the students in the school-leaving age cohort, while in others it represented fewer and as low as only half of these students. For these latter countries, because of their target population, the physics and advanced mathematics samples represented a smaller fraction of the students in the school-leaving age cohort.

As described in Chapter 2, TIMSS computed an index quantifying the percentage of students in the school-leaving age cohort covered by the TIMSS samples. This index is called the TIMSS Coverage Index (TCI). Building on this index, the Mathematics TIMSS Coverage Index (MTCI) quantifies the percentage of students in the school-leaving age cohort covered by the advanced mathematics sample and the physics TIMSS Coverage Index (PTCI) quantifies the percentage of the school-leaving age cohort covered by the physics sample.

To take into account the different proportions of students in the school-leaving age cohort represented in the samples, TIMSS computed the performance in mathematics and science literacy for the top 25 percent of the students in the school-leaving age cohort, and the average performance in advanced mathematics and physics of the top 5 percent and top 10 percent of the students in the school-leaving age cohort. When computing each of these percentiles we assumed that students not tested in the subject

area would have scored below the percentile in question, primarily because they were not in school, in the case of the samples of all final-year students, or because they had not taken courses in advanced mathematics or physics, in the case of the advanced mathematics and physics samples.

When computing the average performance of the students above a certain percentile, the population of students covered by the TIMSS tests had to be adjusted as follows. We assumed that students not tested would score below the percentile. For example, in the United States the TCI was 63.1 percent. This means that the US school-leaving age cohort is approximately the population covered by TIMSS (2278258.19) plus the 36.9 percent that was not covered (2278258.19 * 100 / 63.1) or approximately 3,610,552 students. Now, if we had tested all students in the school-leaving age cohort (3.6 million), then the 75th percentile of those people would have been found easily. However, we did not test 1.3 million of these students, and we assume they would have performed below the 75th percentile of all the students. Then, to find the 75th percentile all we need to do is take away the top 25 percent of the 63.1 percent which corresponds to the 60.4th percentile of the tested sample, computed as $\left(1 - \frac{25}{63.1}\right) * 100$.

Table 8.15 shows, for each assessment, the percentile that was used to select the students in the sample above the percentile points.

Table 8.15 Percentiles of Performance

Country	TCI	MTCI	PTCI	Mathematics and Science Literacy	Advanced Mathematics		Physics	
				Percentile for Top 25%	Percentile for Top 10%	Percentile for Top 5%	Percentile for Top 10%	Percentile for Top 5%
Australia	68.1%	15.7%	12.6%	63.3	36.5	68.2	20.7	60.3
Austria	75.9%	33.3%	33.1%	67.1	70.0	85.0	69.7	84.9
Canada	70.3%	15.6%	13.7%	64.4	36.1	68.0	26.8	63.4
Cyprus	47.9%	8.8%	8.8%	47.8	–	43.5	–	43.5
Czech Republic	77.6%	11.0%	11.0%	67.8	8.9	54.4	8.8	54.4
Denmark	57.7%	20.6%	3.2%	56.6	51.4	75.7	–	–
France	83.9%	19.9%	19.9%	70.2	49.8	74.9	49.8	74.9
Germany	75.3%	26.3%	8.4%	66.8	62.0	81.0	–	40.5
Greece	–	10.0%	10.0%	–	0.0	50.0	0.0	50.0
Hungary	65.3%	–	–	61.7	–	–	–	–
Iceland	54.6%	–	–	54.2	–	–	–	–
Italy	51.5%	14.1%	8.6%	51.5	29.2	64.6	–	42.0
Lithuania	42.5%	2.6%	–	41.2	–	–	–	–
Netherlands	78.0%	–	–	67.9	–	–	–	–
New Zealand	70.5%	–	–	64.5	–	–	–	–
Norway	84.0%	–	8.4%	70.2	–	–	–	40.3
Russian Federation	48.1%	2.1%	1.8%	48.0	–	–	–	–
Slovenia	87.8%	75.4%	38.6%	71.5	86.7	93.4	74.1	87.0
South Africa	48.9%	–	–	48.9	–	–	–	–
Sweden	70.6%	16.2%	16.3%	64.6	38.4	69.2	38.6	69.3
Switzerland	81.9%	14.3%	14.2%	69.5	29.9	64.9	29.4	64.7
United States	63.1%	13.7%	14.5%	60.4	27.2	63.6	30.9	65.4

A dash (–) indicates country did not participate in assessment.

8.5 REPORTING GENDER DIFFERENCES WITHIN COUNTRIES

Gender differences were reported in overall student achievement in mathematics and science literacy, mathematics literacy, science literacy, advanced mathematics, and physics, as well as in the various subject matter content areas.

The analysis of overall gender differences focused on significant differences in achievement within each country in terms of the international scale scores. These results are presented in a table with an accompanying graph indicating whether the difference between male and female achievement was statistically significant. The significance of the difference was determined by comparing the absolute value of the standardized difference between the two means with a critical value of 1.96, corresponding to a 95 percent confidence level (two-tailed test; $\alpha = 0.05$, with infinite degrees of freedom). The standardized difference between the mean for males and females (*t)* was computed as

$$t_k = \frac{\bar{x}_{kb} - \bar{x}_{kg}}{\sqrt{se_{kb}^2 + se_{kg}^2}}$$

where t_k is the standardized difference between two means for country k, $\bar{x}_{kb}$ and $\bar{x}_{kg}$ are the means for males and females within country k, and se_{kb} and se_{kg} are the standard errors for the males' and females' means in country k computed using the jackknife error estimation method described earlier. The above formula assumes independent samples of males and females, and was used in TIMSS due to time constraints. However, since in most countries males and females attended the same schools, the samples of males and females are not completely independent. It would have been more correct to jackknife the difference between males and females. The appropriate test is then the difference between the mean for males and the mean for females divided by the jackknife standard error of the difference. Tables 8.16 through 8.20 show, for mathematics and science literacy, advanced mathematics, and physics, the standard errors of the differences computed under the assumption of independent sampling for males and females and computed using the jackknife technique for correlated samples. No corrections for multiple comparisons were made when comparing the achievement for males and females.

Table 8.16 Standard Error of the Gender Difference Mathematics and Science Literacy

Country	Males' Mean and (S.E.)	Females' Mean and (S.E.)	Males' and Females' Difference	JRR S.E. of Difference - Correlated Samples	JRR S.E. of Difference - Independent Samples
Australia	543 (10.7)	511 (9.3)	32.0	6.8	14.2
Austria	549 (7.8)	502 (5.5)	47.0	9.4	9.6
Canada	544 (3.4)	511 (3.4)	33.0	4.6	4.8
Cyprus	456 (4.9)	439 (3.0)	18.0	6.4	5.8
Czech Republic	500 (9.9)	452 (13.8)	48.0	14.7	17.0
Denmark	554 (4.5)	507 (3.7)	47.0	5.8	5.8
France	526 (5.9)	487 (4.8)	38.0	5.2	7.6
Germany	512 (8.2)	479 (8.5)	32.0	12.3	11.8
Hungary	485 (4.5)	468 (4.5)	17.0	6.9	6.3
Iceland	565 (2.9)	522 (1.9)	43.0	3.6	3.5
Italy	492 (6.9)	461 (5.7)	31.0	7.9	8.9
Lithuania	483 (6.7)	456 (7.4)	27.0	8.7	10.0
Netherlands	584 (5.5)	533 (5.9)	51.0	7.1	8.0
New Zealand	540 (5.7)	511 (5.5)	28.0	6.0	7.9
Norway	564 (5.0)	507 (4.5)	57.0	5.8	6.8
Russian Federation	499 (5.9)	462 (6.5)	37.0	5.0	8.8
Slovenia	538 (12.6)	492 (7.1)	46.0	12.2	14.4
South Africa	366 (10.3)	341 (11.8)	25.0	11.6	15.7
Sweden	579 (5.9)	533 (3.6)	46.0	6.1	6.9
Switzerland	547 (6.0)	511 (7.5)	37.0	8.7	9.6
United States	479 (4.2)	462 (3.5)	17.0	4.7	5.5

JRR = jacknife repeated replicate method

S.E. = standard error

Table 8.17 Standard Error of the Gender Difference Mathematics Literacy

Country	Males' Mean and (S.E.)	Females' Mean and (S.E.)	Males' and Females' Difference	JRR S.E. of Difference - Correlated Samples	JRR S.E. of Difference - Independent Samples
Australia	540 (10.3)	510 (9.3)	30.0	6.7	13.9
Austria	545 (7.2)	503 (5.5)	41.0	8.5	9.0
Canada	537 (3.8)	504 (3.5)	34.0	4.9	5.2
Cyprus	454 (4.9)	439 (3.7)	15.0	7.0	6.1
Czech Republic	488 (11.3)	443 (16.8)	45.0	17.1	20.2
Denmark	575 (4.0)	523 (4.0)	52.0	5.7	5.7
France	544 (5.6)	506 (5.3)	38.0	5.1	7.7
Germany	509 (8.7)	480 (8.8)	29.0	12.3	12.4
Hungary	485 (4.9)	481 (4.8)	5.0	7.4	6.9
Iceland	558 (3.4)	514 (2.2)	44.0	3.9	4.1
Italy	490 (7.4)	464 (6.0)	26.0	8.5	9.5
Lithuania	485 (7.3)	461 (7.7)	23.0	9.3	10.6
Netherlands	585 (5.6)	533 (5.9)	53.0	7.6	8.2
New Zealand	536 (4.9)	507 (6.2)	29.0	6.4	7.9
Norway	555 (5.3)	501 (4.8)	54.0	6.2	7.1
Russian Federation	488 (6.5)	460 (6.6)	27.0	4.7	9.2
Slovenia	535 (12.7)	490 (8.0)	46.0	12.8	15.0
South Africa	365 (9.3)	348 (10.8)	17.0	11.0	14.3
Sweden	573 (5.9)	531 (3.9)	42.0	6.3	7.1
Switzerland	555 (6.4)	522 (7.4)	33.0	8.3	9.8
United States	466 (4.1)	456 (3.6)	11.0	4.4	5.5

JRR = jacknife repeated replicate method

S.E. = standard error

Table 8.18 Standard Error of the Gender Difference Science Literacy

Country	Males' Mean and (S.E.)	Females' Mean and (S.E.)	Males' and Females' Difference	JRR S.E. of Difference - Correlated Samples	JRR S.E. of Difference - Independent Samples
Australia	547 (11.5)	513 (9.4)	34.0	7.4	14.8
Austria	554 (8.7)	501 (5.8)	53.0	10.7	10.4
Canada	550 (3.6)	518 (3.8)	32.0	5.4	5.2
Cyprus	459 (5.8)	439 (3.0)	20.0	6.8	6.5
Czech Republic	512 (8.8)	460 (11.0)	51.0	12.6	14.0
Denmark	532 (5.4)	490 (4.1)	41.0	6.3	6.8
France	508 (6.7)	468 (4.8)	39.0	5.9	8.3
Germany	514 (7.9)	478 (8.5)	35.0	12.6	11.6
Hungary	484 (4.2)	455 (4.3)	29.0	6.6	6.0
Iceland	572 (2.7)	530 (2.1)	41.0	3.6	3.4
Italy	495 (6.7)	458 (5.6)	37.0	7.8	8.8
Lithuania	481 (6.4)	450 (7.3)	31.0	8.3	9.7
Netherlands	582 (5.7)	532 (6.2)	49.0	7.1	8.4
New Zealand	543 (7.1)	515 (5.2)	28.0	6.5	8.8
Norway	574 (5.1)	513 (4.5)	61.0	5.8	6.8
Russian Federation	510 (5.7)	463 (6.7)	47.0	5.8	8.8
South Africa	367 (11.5)	333 (13.0)	34.0	12.5	17.4
Sweden	585 (6.0)	534 (3.5)	51.0	6.1	6.9
Switzerland	540 (6.1)	500 (7.8)	40.0	9.4	9.9
United States	492 (4.5)	469 (3.9)	23.0	5.5	5.9
Slovenia	541 (12.7)	494 (6.4)	47.0	12.0	14.3

JRR = jacknife repeated replicate method

S.E. = standard error

Table 8.19 Standard Error of the Gender Difference Advanced Mathematics

Country	Males' Mean and (S.E.)	Females' Mean and (S.E.)	Males' and Females' Difference	JRR S.E. of Difference - Correlated Samples	JRR S.E. of Difference - Independent Samples
Australia	531 (11.4)	517 (15.1)	14.0	12.5	18.9
Austria	486 (7.3)	406 (8.6)	80.0	11.5	11.2
Canada	528 (6.4)	489 (4.4)	39.0	7.5	7.7
Cyprus	524 (4.4)	509 (6.4)	15.0	6.0	7.8
Czech Republic	524 (13.0)	432 (8.9)	92.0	10.0	15.7
Denmark	529 (4.4)	510 (4.6)	19.0	5.8	6.3
France	567 (5.1)	543 (5.1)	23.0	7.0	7.2
Germany	484 (6.5)	452 (6.6)	32.0	7.2	9.2
Greece	516 (6.6)	505 (10.2)	11.0	11.3	12.1
Italy	484 (10.6)	460 (14.1)	24.0	15.1	17.7
Lithuania	542 (3.7)	490 (5.6)	51.0	8.1	6.7
Russian Federation	568 (9.7)	515 (10.2)	53.0	10.5	14.1
Slovenia	484 (11.5)	464 (11.0)	20.0	13.5	15.9
Sweden	519 (5.9)	496 (5.2)	23.0	8.2	7.9
Switzerland	559 (5.6)	503 (5.7)	56.0	6.0	8.0
United States	457 (7.8)	426 (7.1)	31.0	8.7	10.5

JRR = jacknife repeated replicate method

S.E. = standard error

Table 8.20 Standard Error of the Gender Difference Physics

Country	Males' Mean and (S.E.)	Females' Mean and (S.E.)	Males' and Females' Difference	JRR S.E. of Difference - Correlated Samples	JRR S.E. of Difference - Independent Samples
Australia	532 (6.7)	490 (8.4)	42.0	8.2	10.8
Austria	479 (8.1)	408 (7.4)	71.0	10.4	11.0
Canada	506 (6.0)	459 (6.3)	47.0	10.5	8.7
Cyprus	509 (8.9)	470 (7.1)	40.0	12.6	11.4
Czech Republic	503 (8.8)	419 (3.9)	83.0	8.2	9.7
Denmark	542 (5.2)	500 (8.1)	42.0	10.1	9.6
France	478 (4.2)	450 (5.6)	28.0	5.8	7.0
Germany	542 (14.3)	479 (9.1)	64.0	13.5	17.0
Greece	495 (6.1)	468 (8.1)	28.0	8.2	10.1
Italy	. .	. .	.	.	.
Latvia (LSS)	509 (19.0)	467 (22.6)	42.0	7.6	29.5
Norway	594 (6.3)	544 (9.3)	51.0	8.0	11.2
Russian Federation	575 (9.9)	509 (15.3)	66.0	10.7	18.2
Slovenia	546 (16.3)	455 (18.7)	91.0	20.1	24.8
Sweden	589 (5.1)	540 (5.3)	49.0	7.3	7.4
Switzerland	529 (5.2)	446 (3.6)	83.0	5.7	6.3
United States	439 (4.3)	405 (3.1)	33.0	4.9	5.3

JRR = jacknife repeated replicate method

S.E. = standard error

8.6 PERCENT CORRECT FOR INDIVIDUAL ITEMS

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

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 was selected, invalid (I_j) when two or more choices were made on the same question, 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. For free-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 students' response was not legible or interpretable, 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.

Note that although the not-reached responses were treated as missing for the purpose of estimating the item parameters in the international IRT scaling, they were considered to be wrong answers for a student when percents correct for an item were computed.

8.7 THE TEST-CURRICULUM MATCHING ANALYSIS

TIMSS developed international tests of advanced mathematics and physics 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 Subject Matter Advisory Committee, which consists of mathematics and physics 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 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 asked each country to indicate which items on the tests, if any, were inappropriate to its curriculum. For each country, in turn, TIMSS 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. 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 not only to see the performance of all countries on the items appropriate for its curriculum, but also to see the performance of its students on items judged appropriate for the curriculum in other countries.

Each NRC was given a questionnaire with all the items included in the TIMSS advanced mathematics and physics tests and was asked to indicate, for each item, whether it was considered an appropriate item for their curriculum. The results from these questionnaires were then used to assess the curricular coverage of the items in the tests, and what effect omitting items identified by each NRC had on the test results of all countries. It must be stressed that this analysis was not intended to replace the carefully constructed and agreed-upon tests that TIMSS used for its international comparisons and research analyses. The IRT scaling and research analyses used all items that were included in the tests and that met psychometric standards. In the TCMA analysis, items identified by NRCs were omitted from test results only in the analyses designed to illuminate and explain the international comparisons based on the entire test.

8.7.1 The Analytical Method of the TCMA[1]

The TCMA makes use of the average proportion-correct technology. The basic item-level data for a participating country were represented by the matrix D_{ikj}. This matrix contains elements d_{ikj}, which represent the scored response of student i in country k to item j. The possible values for item j are 0 or 1 for multiple-choice items, and between 0 and 3 for multiple-score items. Most of the elements of D are missing since each student took only one of four possible booklets administered in each subject. Depending on the booklet, each student took between one-seventh and three-sevenths of the total item pool (Adams and Gonzalez, 1996).

The information provided by the NRC as to whether or not an item should be omitted from these analyses was summarized in a matrix T_{kj}, where the elements t_{kj} represent the information that the NRC in country k submitted about item j (for a particular grade). The actual responses of the NRCs for an item were 0 (meaning omit this item for my country) or 1 (meaning include it). Given that multiple-score items were included

[1] The analytic method of the TCMA was developed by Albert E. Beaton, TIMSS International Study Director.

in the TIMSS tests, both matrices D_{ikj} and T_{kj} were then converted to $D_{ikj'}$ and $T_{kj'}$ matrices as described in the previous chapter. In that conversion, the score points on each item in the matrix $D_{ikj'}$ were transformed into their binary representation, and the item selection by the NRC, contained in the matrix $T_{kj'}$ was transformed into a matrix that matched the $D_{ikj'}$.

Although the procedure described here will work generally for any item selection proportion from 0 to 1, the TCMA analysis in TIMSS was limited to a binary choice of either including or excluding the item at the specific grade level. The computational procedure used for the TCMA analysis was as follows. First form the $P'_{kj'}$ matrix. The elements in matrix $P'_{kj'}$ are computed from the D_{ikj} matrix after the transformations and estimation outlined in Chapter 9 in the *TIMSS Technical Report, Volume II* (Martin and Kelly, 1997) are applied to the data. The elements of $P'_{kj'}$ are the weighted averages of the student responses in country k to item j', that is, the average of the student responses $d_{ikj'}$, estimated for some elements. Under the TIMSS design, students not administered particular items may be considered missing at random and treated as not having taken the item. Item responses coded as not reached or omitted are treated as incorrect responses.

The next step is to compute an index of text coverage. A reasonable index is the percentage of the total possible test points that were deemed appropriate by each country. This index should not be confused with the TIMSS Coverage Index (TCI) discussed in Chapter 2 and earlier in this chapter. The total possible test points in a TIMSS test are equal to C_t, and the total possible score on the items deemed appropriate in country k is computed as

$$C_k = \sum_{j'} t_{kj'}$$

The index can then be computed as the ratio of the total possible score on the items deemed appropriate in country k to the total possible test points in the TIMSS test:

$$\frac{C_k}{C_t}$$

This index indicates the proportion of score points of the test that was considered appropriate to the curriculum in the country. The index for each country is presented in Table 8.21.

Table 8.21 Index of Test Coverage
Advanced Mathematics and Physics

Country	Advanced Mathematics	Physics
Australia	0.87	0.96
Austria	1.00	1.00
Canada	0.85	0.73
Cyprus	0.93	0.96
Czech Republic	0.98	0.95
Denmark	0.79	0.90
France	0.98	0.74
Germany	0.79	0.96
Russian Federation	0.82	0.47
Slovenia	0.99	0.96
Sweden	0.76	-
Switzerland	0.88	0.53
United States	1.00	1.00

After computing the index of test coverage, the next step was to compute the normalized weight matrix. To facilitate cross-national comparisons, it is useful to anchor the various national proficiency estimates in a common manner. The national proficiency estimates described in the next section have the property that, if the students in a country correctly answer all of the items deemed appropriate for that country, then the country will receive a value of 100; if the students answer all of those items incorrectly, then the country will receive a value of 0. Items not deemed appropriate to the curriculum of a country are not used in computing these values. In situations where the information in T is either 1 (include) or 0 (omit), the country values may be considered percentages of possible points attained on included items. If T contains proportions other than 0 and 1, then the country values may be greater than 100, in which case the students answered more items correctly than was expected from the values in T.

To compute such country estimates, it is necessary to construct the matrix $W_{kj'}$, with the elements $w_{kj'}$, where the matrix elements are computed as follows:

$$w_{kj} = \frac{t_{kj'}}{\sum_{j'} t_{kj'}^2}$$

where the denominator of this equation is the sum of the squares of the NRCs' judgments of the items.

The Country Comparison Matrix can be computed from $P_{kj'}$ and $W_{kj'}$ by the matrix multiplication

$$C_{kk'} = 100 * (W_{kj'} * P'_{kj'})$$

where the elements of $C_{kk'}$ indicate how the students in country k' scored on the items deemed appropriate in country k.

Another way to estimate the $C_{kk'}$ matrix directly without going through the intermediate step of computing the w_{kj} matrix is as follows:

$$C_{kk'} = \frac{\sum_{j'} t_{kj'} * p_{kj'}}{\sum_{j'} t_{kj'}^2} * 100 .$$

The estimates in the resulting Country Comparison Matrix are unbiased estimators of average student performance based on the items selected by each country for inclusion in the TCMA. The precision of estimates varies as a result of the test booklet rotation as well as the different school and student sampling plans.

8.7.2 Computing Standard Errors

The computation of standard errors for the TCMA is a continuation of the procedure for computing the standard error for the average percent correct as described in Chapter 9 of the *TIMSS Technical Report, Volume II* (Martin and Kelly, 1997). Once the $P_{kj}'^{h'}$ matrices are obtained, we then continue to compute each of the $C_{kk'}^{h'}$ matrices, which can be computed with each of the different $P_{kj}'^{h'}$ replicate matrices. This is accomplished in a straightforward manner by use of the following multiplication:

$$C_{kk'}^{h'} = \frac{\sum_{j'} t_{kj'} * p_{kj'}^{h'}}{\sum_{j'} t_{kj'}^2} * 100 .$$

The jackknifed standard errors for each of the elements in the $C_{kk'}$ matrix are then computed by applying the following formula:

$$jse_{C_{kk'}} = \sqrt{\sum_{h'} (c_{kk'} - c'^{h'}_{kk'})^2} .$$

REFERENCES

Adams, R. J. and Gonzalez, E. J. (1996). The TIMSS test design. In Martin, M.O. and Kelly, D.L. (Eds.), *TIMSS technical report, volume I: Design and development*. 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., and Kelly, D.L., Eds. (1997). *TIMSS technical report, volume II: Implementation and analysis- primary and middle school years*. Chestnut Hill, MA: Boston College.

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