8 Reporting Student Achievement in Mathematics and Science

Eugenio J. Gonzalez
Boston College

8.1 STANDARDIZING THE TIMSS INTERNATIONAL SCALE SCORES

The item response theory (IRT) scaling procedures described in the Chapter 7 yielded imputed scores or 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, which 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 methodology 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 averaged to get the best overall result. The variability among the five results is a measure of the error due to imputation and, where this is large, it may be combined with jackknife estimates of sampling error to give a more realistic indication of the total variability of a statistic. In TIMSS at Population 1 and 2 there was little variability between results from the five plausible values, and so it was decided to simplify the analytic procedures by ignoring this variability and using the first plausible value as the international student score in mathematics and science.

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 from all students from all countries at both grade levels were combined into a standardization sample. This sample consisted of student data from 40 countries, each country equally weighted. South Africa and the Philippines were not included in the sample. The means and standard deviations derived from this procedure are shown in Tables 8.1 through 8.4. These tables show the average logit for each of the five plausible values, and for the international student score (which is simply a copy of the first plausible value).

Table 8.1 Standardization Parameters of International Mathematics Scores Population 1

Variable	Mean Logit	Standard Deviation
International Mathematics Score	0.228345	1.070685
Mathematics Plausible Value #1	0.228345	1.070685
Mathematics Plausible Value #2	0.227183	1.069980
Mathematics Plausible Value #3	0.228378	1.069806
Mathematics Plausible Value #4	0.229702	1.070308
Mathematics Plausible Value #5	0.228632	1.072624
Average	0.228448	1.070681

Table 8.2 Standardization Parameters of International Science Scores Population 1

Variable	Mean Logit	Standard Deviation
International Science Score	0.288556	0.958956
Science Plausible Value #1	0.288556	0.958956
Science Plausible Value #2	0.283356	0.959373
Science Plausible Value #3	0.283130	0.959993
Science Plausible Value #4	0.286728	0.959670
Science Plausible Value #5	0.283406	0.960045
Average	0.285035	0.959607

Table 8.3 Standardization Parameters of International Mathematics Scores Population 2

Variable	Mean Logit	Standard Deviation
International Mathematics Score	0.214809	1.105079
Mathematics Plausible Value #1	0.214809	1.105079
Mathematics Plausible Value #2	0.215036	1.106252
Mathematics Plausible Value #3	0.215540	1.108284
Mathematics Plausible Value #4	0.215463	1.106881
Mathematics Plausible Value #5	0.213658	1.104365
Average	0.214901	1.106172

Table 8.4 Standardization Parameters of International Science Scores Population 2

Variable	Mean Logit	Standard Deviation
International Science Score	0.211454	0.770235
Science Plausible Value #1	0.211454	0.770235
Science Plausible Value #2	0.211574	0.770093
Science Plausible Value #3	0.211886	0.771142
Science Plausible Value #4	0.213772	0.769263
Science Plausible Value #5	0.210969	0.771090
Average	0.211931	0.770365

Each country was weighted to contribute equally to the calculation of the international mean and standard deviation, except for Kuwait and Israel, which tested only one grade at each population. These two countries were weighted to make only half the contribution of the countries with both grades. The contribution of the students from each grade within each country was proportional to the number of students at each grade level within the country. 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 50, the value was recoded to 50, therefore making 50 the lowest score on the transformed scale. This happened in very few cases across the countries. The highest transformed scores did not exceed 1000 points, so the transformed values were left untouched at the upper end of the distribution.

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 item difficulties for the item difficulty maps, the original item difficulties from the scaling were transformed in two ways. First, they were moved along the logit scale from the point where a student would have a 50 percent chance 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 onto the international achievement scale. The means and standard deviations for this transformation were the average of the plausible value means, and the average of the plausible value standard deviations from Table 8.1 through Table 8.4 above. This resulted in the following transformations for the mathematics and science items.

For the Populations 1 and 2 mathematics item difficulties, dm_j, the transformed item difficulty dm'_j was computed as follows:

$$\text{Population 1} \quad dm'_j = 500 + 100 \times \left(\frac{dm_j + \ln\left(\frac{0.65}{0.35}\right) - 0.228448}{1.07068} \right)$$

$$\text{Population 2} \quad dm'_j = 500 + 100 \times \left(\frac{dm_j + \ln\left(\frac{0.65}{0.35}\right) - 0.214901}{1.106172} \right)$$

For the Populations 1 and 2 science item difficulties, ds_j, the transformed item difficulty ds'_j was computed as follows:

$$\text{Population 1} \quad ds'_j = 500 + 100 \times \left(\frac{ds_j + \ln\left(\frac{0.65}{0.35}\right) - 0.285035}{0.959607} \right)$$

$$\text{Population 2} \quad ds'_j = 500 + 100 \times \left(\frac{ds_j + \ln\left(\frac{0.65}{0.35}\right) - 0.211931}{0.770365} \right)$$

The resulting values are the item difficulties presented in the item difficulty maps in the international reports.

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. In comparisons of student performance from two countries, the 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 reports 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 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 820 comparisons at the eighth-grade, 741 at the seventh-grade, 325 at the fourth-grade, and 276 at the third-grade. In decision-making terms this would be a very conservative procedure, however, and would run the risk of making an error of a different kind, i.e., 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 probably are interested in comparing a single country with all other countries and would not be making all possible between-country comparisons at any one time, a more realistic approach, which was adopted in TIMSS, seemed to be to adjust the significance level for a number of comparisons equal to the number of countries (minus one). From this perspective the number of simultaneous comparisons to be adjusted for at eighth grade, for example, is 40 rather than 820, and at seventh grade is 38 rather than 741. The number of comparisons is 25 for the fourth-grade table, and 23 for the third-grade table. As a consequence, we used the critical values shown in Table 8.5, given by the appropriate quantiles from the normal (Gaussian) distribution.

Table 8.5 Critical Values Used for the Multiple Comparison Figures in TIMSS International Reports

Grade Level	Alpha Level	Number of Comparisons	Critical Value
3rd Grade	0.05	23	3.0654
4th Grade	0.05	25	3.0902
7th Grade	0.05	38	3.2125
8th Grade	0.05	40	3.2273

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. Tables 8.6a and 8.6b show the means and standard errors used in the calculation of statistical significance between means for mathematics and science, at Population 2 and Population 1, respectively. 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 codence.

Table 8.6a Means and Standard Errors for Multiple Comparison Figures Mathematics and Science - Population 2

Country	Mathematics				Science			
	7th Grade Mean	S.E.	8th Grade Mean	S.E.	7th Grade Mean	S.E.	8th Grade Mean	S.E.
Australia	497.9	3.8	529.6	4.0	504.4	3.6	544.6	3.9
Austria	509.2	3.0	539.4	3.0	518.8	3.1	557.7	3.7
Belgium (Fl)	557.6	3.5	565.2	5.7	528.7	2.6	550.3	4.2
Belgium (Fr)	507.1	3.5	526.3	3.4	442.0	3.0	470.6	2.8
Bulgaria	513.8	7.5	539.7	6.3	530.8	5.4	564.8	5.3
Canada	494.0	2.2	527.2	2.4	499.2	2.3	530.9	2.6
Colombia	368.5	2.7	384.8	3.4	387.5	3.2	411.1	4.1
Cyprus	445.7	1.9	473.6	1.9	419.9	1.8	462.6	1.9
Czech Republic	523.4	4.9	563.7	4.9	532.9	3.3	573.9	4.3
Denmark	464.8	2.1	502.3	2.8	439.0	2.1	478.3	3.1
England	476.2	3.7	505.7	2.6	512.0	3.5	552.1	3.3
France	492.2	3.1	537.8	2.9	451.5	2.6	497.7	2.5
Germany	484.4	4.1	509.2	4.5	499.5	4.1	531.3	4.8
Greece	439.9	2.8	483.9	3.1	448.6	2.6	497.3	2.2
Hong Kong	563.6	7.8	588.0	6.5	495.3	5.5	522.1	4.7
Hungary	501.8	3.7	537.3	3.2	517.9	3.2	553.7	2.8
Iceland	459.4	2.6	486.8	4.5	462.0	2.8	493.6	4.0
Iran, Islamic Rep.	400.9	2.0	428.3	2.2	436.3	2.6	469.7	2.4
Ireland	499.7	4.1	527.4	5.1	495.2	3.5	537.8	4.5
Israel	.	.	521.6	6.2	.	.	524.5	5.7
Japan	571.1	1.9	604.8	1.9	531.0	1.9	571.0	1.6
Korea	577.1	2.5	607.4	2.4	535.0	2.1	564.9	1.9
Kuwait	.	.	392.2	2.5	.	.	429.6	3.7
Latvia (LSS)	461.6	2.8	493.4	3.1	434.9	2.7	484.8	2.7
Lithuania	428.2	3.2	477.2	3.5	403.1	3.4	476.4	3.4
Netherlands	516.0	4.1	541.0	6.7	517.2	3.6	560.1	5.0
New Zealand	471.7	3.8	507.8	4.5	481.0	3.4	525.5	4.4
Norway	460.7	2.8	503.3	2.2	483.2	2.9	527.2	1.9
Portugal	423.1	2.2	454.4	2.5	427.9	2.1	479.6	2.3
Romania	454.4	3.4	481.6	4.0	451.6	4.4	486.1	4.7
Russian Federation	500.9	4.0	535.5	5.3	484.0	4.2	538.1	4.0
Scotland	462.9	3.7	498.5	5.5	468.1	3.8	517.2	5.1
Singapore	601.0	6.3	643.3	4.9	544.7	6.6	607.3	5.5
Slovak Republic	507.8	3.4	547.1	3.3	509.7	3.0	544.4	3.2
Slovenia	498.2	3.0	540.8	3.1	529.9	2.4	560.1	2.5
South Africa	347.5	3.8	354.1	4.4	317.1	5.3	325.9	6.6
Spain	448.0	2.2	487.3	2.0	477.2	2.1	517.0	1.7
Sweden	477.5	2.5	518.6	3.0	488.4	2.6	535.4	3.0
Switzerland	505.5	2.3	545.4	2.8	483.7	2.5	521.7	2.5
Thailand	494.7	4.8	522.5	5.7	492.8	3.0	525.5	3.7
United States	475.7	5.5	499.8	4.6	508.2	5.5	534.4	4.7

Table 8.6b Means and Standard Errors for Multiple Comparison Figures Mathematics and Science - Population 1

Country	Mathematics				Science			
	3th Grade Mean	S.E.	4th Grade Mean	S.E.	3rd Grade mean	S.E.	4th Grade Mean	S.E.
Australia	483.4	4.0	546.3	3.1	509.7	4.3	562.5	2.9
Austria	487.0	5.3	559.3	3.1	504.6	4.6	564.8	3.3
Canada	469.5	2.7	532.1	3.3	490.4	2.5	549.3	3.0
Cyprus	430.4	2.8	502.4	3.1	414.7	2.5	475.4	3.3
Czech Republic	497.2	3.3	567.1	3.3	493.7	3.4	556.5	3.1
Greece	428.1	4.0	491.9	4.4	445.9	3.9	497.2	4.1
Hong Kong	524.0	3.0	586.6	4.3	481.6	3.3	533.0	3.7
Hungary	476.1	4.2	548.4	3.7	464.4	4.1	531.6	3.4
Iceland	410.1	2.8	473.8	2.7	435.4	3.3	504.7	3.3
Iran, Islamic Rep.	378.0	3.5	428.5	4.0	356.2	4.2	416.5	3.9
Ireland	475.8	3.6	549.9	3.4	479.1	3.7	539.5	3.3
Israel	.	.	531.4	3.5	.	.	504.8	3.6
Japan	537.9	1.5	596.8	2.1	521.8	1.6	573.6	1.8
Korea	560.9	2.3	610.7	2.1	552.9	2.4	596.9	1.9
Kuwait	.	.	400.2	2.8	.	.	401.3	3.1
Latvia (LSS)	463.3	4.3	525.4	4.8	465.3	4.5	512.2	4.9
Netherlands	492.9	2.7	576.7	3.4	498.8	3.2	556.7	3.1
New Zealand	439.5	4.0	498.7	4.3	473.1	5.2	531.0	4.9
Norway	421.3	3.1	501.9	3.0	450.3	3.9	530.3	3.6
Portugal	425.3	3.8	475.4	3.5	423.0	4.3	479.8	4.0
Singapore	552.1	4.8	624.9	5.3	487.7	5.0	546.7	5.0
Thailand	444.3	5.1	490.2	4.7	432.6	6.6	472.9	4.9
England	456.5	3.0	512.7	3.2	499.2	3.5	551.5	3.3
Scotland	458.0	3.4	520.4	3.9	483.9	4.2	535.6	4.2
United States	479.8	3.4	544.6	3.0	511.2	3.2	565.5	3.1
Slovenia	487.6	2.9	552.4	3.2	486.9	2.8	545.7	3.3

8.4 INTERNATIONAL MARKER LEVELS OF ACHIEVEMENT

For both populations, international marker levels of achievement were computed at each grade level for mathematics and science. In order to compute the marker levels, all of the student data from all participating countries for a subject at a grade level were pooled, and then the pooled data were used to estimate the 90th, the 75th, and the 50th international percentiles. These percentiles were chosen as international markers because they have a ready interpretation. The 90th percentile in this instance corresponds to the "Top 10% Level," since it is the scale score above which the highest-scoring 10 percent of the students across all countries combined are to be found. Similarly, the 75th percentile corresponds to the "Top Quarter Level," since this is the score above which the top 25 percent of students are to be found, and the 50th percentile corresponds to the "Top Half Level," since this is the score above which the top 50 percent of students are to be found. If student proficiencies were distributed in the same way across countries we would expect about 10 percent of students in each country to score at or above the Top 10% Level, about 25 percent of students to score at or above the Top Quarter marker, and about 50 percent of students to score at or above the Top Half marker. In pooling the data, countries were weighted in accordance with their estimated enrollment size, as shown in Table 8.7.

Table 8.7 Estimated Enrollment by Grade Level Within Country

Country	Third Grade	Fourth Grade	Seventh Grade	Eighth Grade
Australia	237828	245635	238294	231349
Austria	86044	91391	89593	86739
Belgium (Fl)	-	-	64177	75069
Belgium (Fr)	-	-	49898	59270
Bulgaria	-	-	140979	147094
Canada	371166	389160	377732	377426
Colombia	-	-	619462	527145
Cyprus	9740	9995	10033	9347
Czech Republic	116052	120406	152492	152494
Denmark	-	-	44980	54172
England	531682	534922	465457	485280
France	-	-	860657	815510
Germany	-	-	742346	726088
Greece	99000	106181	130222	121911
Hong Kong	83847	89901	88591	88574
Hungary	116779	117228	118727	112436
Iceland	3735	3739	4212	4234
Iran, Islamic Rep.	1391859	1433314	1052795	935093
Ireland	58503	60497	68477	67644
Israel	-	66967	-	60584
Japan	1388749	1438465	1562418	1641941
Korea	607007	615004	798409	810404
Kuwait	-	24071	-	13093
Latvia (LSS)	15121	18883	17041	15414
Lithuania	-	-	36551	39700
Netherlands	171561	173407	175419	191663
New Zealand	48386	52254	48508	51133
Norway	49036	49896	51165	50224
Portugal	114775	133186	146882	137459
Romania	-	-	295348	296534
Russian Federation	-	-	2168163	2004792
Scotland	59393	59054	61938	64638
Singapore	41904	41244	36181	36539
Slovak Republic	-	-	83074	79766
Slovenia	27453	27685	28049	26011
South Africa	-	-	649180	766334
Spain	-	-	549032	547114
Sweden	-	-	96494	98193
Switzerland	-	-	66681	69733
Thailand	883765	864525	680225	657748
United States	3643393	3563795	3156847	3188297

Having established the international marker levels, the next step was to compute the percentage of students in each country scoring at or above the marker levels. Countries with proportionately large numbers of high-achieving students had higher percentages of students scoring above the marker levels. For example, it was not unusual for high-achieving countries to have more than 30 percent of their students scoring at or above the Top 10% marker. Conversely, countries with lower achievement levels sometimes had very few students reaching that marker level.

Using these three marker levels, then, the students were classified into one of four groups: those below the international median or 50th percentile; those at or above the international median but below the third quartile or 75th percentile; those at or above the third quartile, but below the 90th percentile; and those at or above the 90th percentile. Standard errors for the percentage of students in each country were also computed using the jackknife method for sampling variance estimation. The international marker levels are presented in Table 8.8 below.

Table 8.8 International Marker Levels (Percentiles) of Achievement

Population 1

Mathematics			
Grade	**P50**	**P75**	**P90**
3	474	538	592
4	535	601	658
Science			
Grade	**P50**	**P75**	**P90**
3	488	554	610
4	541	607	660

Population 2

Mathematics			
Grade	**P50**	**P75**	**P90**
7	476	551	619
8	509	587	656
Science			
Grade	**P50**	**P75**	**P90**
7	483	553	615
8	521	592	655

8.5 REPORTING MEDIAN ACHIEVEMENT BY AGE

The target populations in TIMSS are defined in terms of adjacent grade levels (the two grades with the most 13-year-olds for Population 2 and the two grades with the most 9-year-olds for Population 1), and student achievement in the international reports is reported for the most part by grade. Since grades are primarily measures of years of schooling, they provide an appropriate basis on which to compare student achievement across countries. However, because of differences internationally in age of entry to formal schooling, and in promotion and retention practices through the grades, there is considerable variation across countries in the ages of students within comparable grade levels. Although TIMSS addressed this issue by using age as the primary basis for choosing the grades to be compared, there was still considerable variation between countries in the average age of their students within any given grade level.

Since TIMSS tested two adjacent grades at each of Populations 1 and 2, in many participating countries most or all 9-year-olds and 13-year-olds were included in the tested grades. Therefore, it was possible to extract just the students in these age groups from the total sample and make reasonable comparisons on the basis of age group. Although some countries had 100 percent of the age group in the grades tested, most countries had some, usually small, percentage of students in the age group outside of the tested grades. For example, in Population 2, some countries had a percentage of 13-year-olds below seventh grade, and a percentage above eighth grade. There was no way to estimate reliably the scores of the students missing from the age group, but it was possible to estimate how many students were involved by extrapolating from the distribution of ages within each of the tested grades.

Since the computation of the mean requires that all elements of the target group be present, it was not possible to compute the mean for 13-year-olds or for 9-year-olds without making assumptions about the scores of the students who were outside the tested grades. However, the median is a measure of the central tendency of a distribution which is less dependent on the values of the elements making up the distribution. In order to compute a median one need only be able to order the elements on the attribute of interest; it is not necessary to know their actual values. By capitalizing on this property of the median it was possible to estimate a median score for 9- and 13-year-olds while assuming only that those students who were in grades below the lower grade tested would score below the median, and those in grades above the upper grade tested would score above the median.

The first step was to estimate, from the age distribution within the tested grades, the percentages of students in the age group in grades below the lower grade tested and in grades above the upper grade tested. To do this it was assumed that the age distribution in the grades below the grades tested was similar to the age distribution in the lower grade lagged by one year for each grade below, and that the age distribution in the grades above the grades tested was similar to that of the upper grade increased by one year for each grade above. The next step was to adjust the median to compensate for the missing out-of-grade students. If there were no such missing students, that is, if the tested grades included all students in the age group, then the median would as usual be set to the 50th percentile, the score below which 50 percent of the student scores are found. However, when some percentage of the age group is outside the grades tested, the 50 percent refers to the entire age group, and not just to the tested students. In this case, the estimate for the number of out-of-grade students in the age group must be added to the number in the age group within the tested grades to estimate the size of the age group, and the percentage in the grades below the lower grade must be subtracted from 50 percent to find the percentile within the tested group that corresponds to 50 percent of the total age group.

8.5.1 Computational Example

Let us take for example a country in which the grades tested for Population 2 were the seventh and eighth. Table 8.9 shows the distribution of students by age in these two grades.[1] We can see that although the modal age of students in the grades tested is 13 at the time of testing, these are not the majority of the students. In fact, there are more students that are older or younger than the target age (53 percent).

Table 8.9 Observed Distribution of Age Groups Within Target Grades

Grade	Age					
	11	12	13	14	15	16
7	0	6506	28601	647	340	0
8	0	0	5121	25292	3702	2226
Total	0	6506	33722	25939	4042	2226

[1] The age of a student for the purpose of this analysis was considered to be the number of whole years between the date of birth of the student and the time of testing. For example, a student 13 years and 11 months old and a student 13 years and 1 month old were both considered to be 13 years old.

In Table 8.10 the age distribution of the seventh-grade students has been projected into the previous three grades with appropriate lags, and the figures from the eighth grade have been projected into the following school year with appropriate increases, until there are expected to be no more 13-year-old students. We notice from this table that the selection of grades to be tested in this country was right on target insofar as no other pair of adjacent grades would have more 13-year-olds. The two grades selected in this country included approximately 97 percent of the 13-year-olds in the country. Selecting the sixth and seventh grades would have yielded a coverage of only 84 percent of the 13-year-olds in the country, and selecting the eighth and ninth grades would have yielded a coverage of only 15 percent of the 13-year-olds.

Table 8.10 Observed and Estimated Distribution of Age Groups by Grade

	Age								
Grade	10	11	12	13	14	15	16	17	% of 13-Year-Olds
4	28601	647	340	0	.	.	.	.	0.00%
5	6506	28601	647	340	0	.	.	.	0.98%
6	0	6506	28601	647	340	0	.	.	1.86%
7	.	0	6506	28601	647	340	0	.	82.40%
8	.	.	0	5121	25292	3702	2226	0	14.75%
9	.	.	.	0	5121	25292	3702	2226	0.00%
Total of 13-Year-Olds:	.	.	.	34709	.	.	.	.	.

After the corresponding lags and increases are projected to the grades adjacent to the grades tested, we estimate that there are approximately 34,709 13-year-olds in the country (340 + 647 + 28601 + 5121). Of those 34,709 13-year-olds, about 3 percent are in grades below the lower grade, there are none in grades above the upper grade, and about 97 percent are in the two grades tested. With this information we can estimate the median achievement of the 13-year-olds, but we need to make one further assumption. We know that, in general, as the students move along the educational system their performance on the test improves. So it is reasonable to assume that those 13-year-olds who are in grades below the lower grade will perform below the median of all 13-year-olds, and those above the target grades will perform above the median of all 13-year-olds. Based on this assumption we can then compute the median of the 13-year-olds by looking at the percentile (Px) from the 13-year-olds in the target grades given by the following formula:

$$Px = \left(\frac{(50 - PBTG) * 100}{PITG}\right)$$

where $PBTG$ is the estimated percent of 13-year-old students below the target grades, and $PITG$ is the percent of students in the target grades. To complete our example, we would then look up the Px percentile in the distribution of 13-year-olds within the country. This works out to be

$$48.54 = \left(\frac{(50 - 2.84) * 100}{97.15}\right)$$

The median for the 13-year-olds in this particular country corresponds to the 48.54th percentile in the distribution of 13-year-olds in the tested grades. For the purpose of the tables presented in the international reports, the median for the students in the age group was computed only if both grades were tested within the country, the appropriate target grades were selected for the testing, and at least an estimated 75 percent of the 13-year-olds were in the target grades. The distribution of students by age across the grades tested is presented in Tables 8.11 and 8.12.

Table 8.11 Coverage of 9-Year-Olds in the Population 1 Sample

Country	Coverage of 9-Year-Olds					
	% Below Lower Grade*	% in Lower Grade	% in Upper Grade	% Above Upper Grade*	Percent of 9-Year-Olds Tested	Percentile in 9-Year-Olds Sample Representing Median for 9-Year-Olds in Country
Australia	5.8%	64.9%	28.9%	0.4%	93.8%	47.1
Austria	13.2%	71.5%	15.2%	0.0%	86.8%	42.4
Canada	4.8%	46.3%	47.5%	1.3%	93.8%	48.1
Cyprus	1.4%	35.1%	62.5%	0.9%	97.7%	49.8
Czech Republic	9.2%	75.5%	15.4%	0.0%	90.8%	45.0
England	0.9%	57.8%	41.2%	0.1%	99.0%	49.6
Greece	0.8%	10.9%	87.6%	0.7%	98.6%	50.0
Hong Kong	6.2%	43.2%	50.0%	0.7%	93.1%	47.0
Hungary	10.5%	70.2%	19.0%	0.3%	89.2%	44.3
Iceland	0.4%	14.8%	84.4%	0.4%	99.2%	50.0
Iran, Islamic Rep.	16.9%	50.7%	32.0%	0.4%	82.7%	40.0
Ireland	8.4%	68.4%	23.2%	0.0%	91.6%	45.4
Israel	.	.	.	.	.	.
Japan	0.5%	90.8%	8.7%	0.0%	99.5%	49.7
Korea	7.9%	67.2%	24.3%	0.7%	91.5%	46.1
Kuwait	.	.	.	.	.	.
Latvia	23.8%	54.7%	21.2%	0.3%	75.9%	34.5
Netherlands	6.9%	63.0%	30.1%	0.0%	93.1%	46.3
New Zealand	0.3%	50.2%	49.1%	0.3%	99.4%	50.0
Norway	0.1%	38.1%	61.7%	0.1%	99.9%	50.0
Portugal	6.7%	45.0%	47.9%	0.4%	92.9%	46.6
Scotland	0.3%	22.9%	75.7%	1.1%	98.6%	50.4
Singapore	2.1%	80.5%	17.4%	0.1%	97.8%	49.0
Slovenia	40.0%	59.6%	0.4%	0.0%	60.0%	.
Thailand	29.2%	60.1%	10.6%	0.2%	70.6%	.
United States	4.5%	61.1%	34.2%	0.2%	95.3%	47.8

*Data are estimated; students below the lower grade and above the upper grade were not included in the sample.

Table 8.12 Coverage of 13-Year-Olds in the Population 2 Sample

Country	Coverage of 13-Year-Olds					
	% Below Lower Grade*	% in Lower Grade	% in Upper Grade	% Above Upper Grade*	Percent or 13-Year-Olds Tested	Percentile in 13-Year-Olds Sample Representing Median for 13-Year-Olds in Country
Australia	7.5%	63.9%	28.2%	0.4%	92.1%	46.2
Austria	10.4%	62.5%	27.1%	0.0%	89.6%	44.2
Belgium (Fl)	5.4%	45.4%	48.8%	0.4%	94.2%	47.3
Belgium (Fr)	13.3%	40.5%	46.0%	0.2%	86.5%	42.4
Bulgaria	2.9%	58.6%	36.6%	1.9%	95.2%	49.4
Canada	8.0%	48.5%	42.9%	0.6%	91.4%	45.9
Colombia	51.5%	30.5%	15.8%	2.2%	46.3%	.
Cyprus	1.6%	27.3%	70.4%	0.8%	97.7%	49.6
Czech Republic	9.7%	72.9%	17.3%	0.0%	90.3%	44.6
Denmark	1.0%	33.9%	64.2%	0.9%	98.1%	49.9
England	0.6%	57.2%	41.7%	0.5%	98.9%	50.0
France	20.2%	43.6%	34.7%	1.5%	78.3%	38.1
Germany	26.1%	71.5%	2.2%	0.2%	73.7%	.
Greece	2.9%	10.3%	85.6%	1.2%	95.9%	49.1
Hong Kong	10.0%	44.2%	45.5%	0.3%	89.6%	44.6
Hungary	10.2%	65.2%	24.3%	0.3%	89.5%	44.5
Iceland	0.0%	16.6%	82.9%	0.5%	99.5%	50.3
Indonesia	10.2%	58.3%	27.5%	4.0%	85.8%	46.3
Iran	28.1%	47.0%	24.9%	0.1%	71.9%	
Ireland	13.9%	68.8%	17.3%	0.1%	86.1%	42.0
Israel	.	.	.	.	.	.
Japan	0.3%	90.9%	8.8%	0.0%	99.7%	49.8
Korea	1.5%	69.9%	28.2%	0.4%	98.1%	49.4
Kuwait	.	.	.	.	.	.
Latvia	10.7%	60.0%	29.1%	0.1%	89.1%	44.0
Lithuania	10.2%	64.2%	25.5%	0.2%	89.6%	44.4
Mexico	18.9%	40.4%	37.0%	3.7%	77.4%	40.2
Netherlands	9.8%	58.7%	31.2%	0.4%	89.8%	44.8
New Zealand	0.5%	51.7%	47.4%	0.4%	99.1%	50.0
Norway	0.2%	42.4%	57.2%	0.1%	99.7%	49.9
Philippines	.	.	.	.	.	.
Portugal	22.9%	43.7%	33.1%	0.3%	76.8%	35.3
Romania	24.4%	66.2%	9.4%	0.0%	75.6%	33.8
Russian Federation	4.5%	50.5%	44.3%	0.7%	94.8%	48.0
Scotland	0.2%	22.7%	76.8%	0.3%	99.5%	50.0
Singapore	3.1%	82.2%	14.7%	0.0%	96.9%	48.4
Slovak Republic	4.4%	73.2%	22.4%	0.0%	95.6%	47.7
Slovenia	33.1%	65.3%	1.6%	0.1%	66.9%	.
South Africa	40.6%	35.1%	21.1%	3.2%	56.2%	.
Spain	14.9%	45.8%	39.0%	0.3%	84.7%	41.4
Sweden	0.8%	45.0%	54.1%	0.1%	99.1%	49.6
Switzerland	8.3%	47.5%	44.0%	0.2%	91.5%	45.6
Thailand	18.0%	58.4%	19.6%	4.0%	78.0%	41.0
United States	8.7%	57.5%	33.5%	0.3%	91.0%	45.4

*Data are estimated; Students below the lower grade and above the upper grade were not included in the sample.

8.6 REPORTING GENDER DIFFERENCES WITHIN COUNTRIES

Gender differences were reported in overall student achievement in mathematics and science, as well as in several subject matter content areas. The computational procedures differed in several ways because of the different approaches to summarizing student performance: IRT scaling for the overall mathematics and science scores, and average percent correct for the subject matter content areas. This chapter describes the procedure for computing gender differences for the overall scores. The procedure for reporting gender differences in content areas is described in Chapter 9.

The analysis of overall gender differences focused on significant differences in mathematics and science achievement within each country using the international scale scores. These results are presented for each country in a table with an accompanying graph indicating where the difference between the boys' achievement and the girls' 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 same critical value was used for the third, fourth, seventh, and eighth grade results. The standardized difference between the mean for boys and girls (*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 boys and girls within country k, and se_{kb} and se_{kg} are the standard errors for the boys' and girls' means in country k computed using the jackknife error estimation method described earlier. The above formula assumes independent samples of boys and girls, and was used in TIMSS due to time constraints. However, since in most countries boys and girls attended the same schools, in fact the samples of boys and girls are not completely independent. It would have been more correct to jackknife the difference between boys and girls. The appropriate test is then the difference between the mean for boys and the mean for girls divided by the jackknife standard error of the difference. Tables 8.13 through 8.20 show the standard errors of the differences computed under the assumption of independent sampling for boys and girls and computed using the jackknife technique for correlated samples. No corrections for multiple comparisons were made when comparing the achievement for boys and girls.

Table 8.13 Standard Error of the Gender Difference Mathematics - Third Grade

Country	Boys Mean and s.e.	Girls Mean and s.e.	S.E. of the Difference Using JRR	S.E. of the Difference Assuming SRS
Australia	487.0 (4.5)	479.8 (4.4)	4.0	6.3
Austria	493.6 (9.2)	481.3 (3.8)	9.6	10.0
Canada	476.7 (3.2)	462.9 (3.0)	3.4	4.4
Cyprus	433.3 (3.3)	428.0 (3.1)	3.2	4.5
Czech Republic	502.0 (3.7)	492.5 (3.8)	3.4	5.3
Greece	432.2 (4.4)	423.9 (4.2)	3.4	6.0
Hong Kong	528.5 (3.2)	518.4 (3.5)	2.9	4.8
Hungary	479.0 (4.9)	476.2 (4.4)	3.7	6.6
Iceland	417.9 (3.5)	402.5 (3.0)	3.4	4.7
Iran, Islamic Rep.	384.2 (4.4)	372.7 (4.9)	6.2	6.6
Ireland	473.2 (4.3)	478.7 (4.5)	5.2	6.3
Japan	539.5 (2.0)	536.3 (1.7)	2.2	2.7
Korea	566.9 (2.8)	554.3 (2.5)	2.7	3.8
Latvia (LSS)	462.4 (5.3)	464.1 (4.5)	4.9	7.0
Netherlands	496.7 (2.9)	488.9 (3.2)	2.8	4.3
New Zealand	435.8 (4.4)	443.0 (4.5)	3.9	6.3
Norway	429.9 (3.5)	411.4 (3.8)	4.0	5.2
Portugal	430.0 (3.5)	420.4 (5.0)	4.1	6.1
Singapore	550.8 (5.4)	553.5 (5.0)	4.1	7.4
Thailand	440.2 (5.0)	448.3 (5.6)	3.2	7.5
England	460.7 (3.5)	452.3 (3.4)	3.2	4.8
Scotland	461.9 (3.8)	453.7 (3.5)	3.0	5.2
United States	480.2 (3.1)	479.3 (4.4)	3.3	5.4
Slovenia	492.4 (3.1)	482.6 (3.5)	3.0	4.7

JRR = jackknife repeated replicate method

SRS = simple random sample

Table 8.14 Standard Error of the Gender Difference Mathematics - Fourth Grade

Country	Boys Mean and s.e.	Girls Mean and s.e.	S.E. of the Difference Using JRR	S.E. of the Difference Assuming SRS
Australia	547.2 (3.5)	545.5 (3.7)	3.7	5.1
Austria	563.2 (3.6)	555.5 (3.6)	3.6	5.1
Canada	533.5 (3.4)	530.9 (3.9)	2.9	5.2
Cyprus	506.4 (3.5)	498.7 (3.3)	2.7	4.8
Czech Republic	568.5 (3.4)	565.8 (3.6)	2.7	5.0
Greece	491.0 (5.0)	492.7 (4.5)	3.9	6.8
Hong Kong	586.5 (4.7)	587.3 (4.2)	2.6	6.3
Hungary	551.6 (4.2)	546.4 (3.9)	3.6	5.8
Iceland	474.3 (3.3)	473.3 (3.0)	3.4	4.5
Iran, Islamic Rep.	432.9 (6.0)	423.8 (5.0)	7.8	7.8
Ireland	548.5 (3.9)	551.4 (4.3)	4.6	5.8
Israel	537.2 (4.4)	528.0 (4.1)	4.5	6.0
Japan	600.6 (2.5)	593.1 (2.2)	2.3	3.3
Korea	618.2 (2.5)	603.0 (2.6)	2.9	3.6
Kuwait	398.8 (4.6)	401.6 (2.5)	5.1	5.3
Latvia (LSS)	520.7 (5.5)	530.2 (5.2)	4.5	7.5
Netherlands	584.7 (3.8)	569.5 (3.4)	2.6	5.1
New Zealand	493.8 (5.7)	503.5 (4.3)	5.3	7.1
Norway	504.2 (3.5)	499.1 (3.6)	3.5	5.0
Portugal	477.6 (3.8)	473.1 (3.7)	2.6	5.3
Singapore	620.2 (5.5)	630.2 (6.4)	5.4	8.4
Thailand	484.8 (5.8)	495.6 (4.2)	3.9	7.1
England	515.1 (3.4)	510.3 (4.4)	4.4	5.5
Scotland	520.3 (4.3)	520.2 (3.8)	2.6	5.8
United States	545.4 (3.1)	543.8 (3.3)	1.9	4.5
Slovenia	551.1 (3.4)	553.9 (4.0)	3.6	5.2

JRR = jackknife repeated replicate method

SRS = simple random sample

Table 8.15 Standard Error of the Gender Difference Science - Third Grade

Country	Boys Mean and s.e.	Girls Mean and s.e.	S.E. of the Difference Using JRR	S.E. of the Difference Assuming SRS
Australia	509.8 (5.6)	509.6 (4.3)	5.1	7.1
Austria	508.3 (6.9)	501.2 (4.0)	6.7	7.9
Canada	496.0 (3.2)	485.9 (2.9)	3.3	4.3
Cyprus	417.6 (2.7)	412.0 (3.0)	2.6	4.0
Czech Republic	503.0 (4.1)	484.7 (3.9)	3.9	5.6
Greece	452.7 (4.6)	438.8 (3.9)	3.6	6.0
Hong Kong	488.3 (3.4)	473.5 (3.8)	3.1	5.1
Hungary	472.0 (4.2)	459.6 (4.7)	3.4	6.3
Iceland	439.9 (4.0)	431.0 (3.9)	4.5	5.6
Iran, Islamic Rep.	358.7 (5.7)	354.0 (5.7)	7.8	8.1
Ireland	481.2 (4.6)	476.8 (4.4)	5.3	6.4
Japan	523.0 (2.1)	520.6 (2.0)	2.5	2.8
Korea	561.8 (2.8)	543.1 (2.7)	2.7	3.9
Latvia (LSS)	461.7 (5.2)	468.7 (4.8)	4.2	7.1
Netherlands	504.4 (3.8)	493.4 (3.1)	2.4	4.9
New Zealand	469.6 (5.9)	476.3 (5.7)	5.2	8.2
Norway	456.8 (4.6)	444.0 (4.5)	4.6	6.4
Portugal	430.8 (4.3)	415.0 (5.4)	4.7	6.9
Singapore	490.8 (5.8)	484.5 (5.2)	4.3	7.7
Thailand	428.4 (6.5)	436.6 (7.1)	3.8	9.6
England	503.3 (4.8)	495.3 (3.4)	4.7	5.9
Scotland	485.3 (4.4)	482.0 (4.7)	3.5	6.5
United States	514.2 (4.2)	508.1 (3.2)	3.8	5.2
Slovenia	495.7 (3.4)	477.7 (3.4)	3.7	4.8

JRR = jackknife repeated replicate method

SRS = simple random sample

Table 8.16 Standard Error of the Gender Difference Science - Fourth Grade

Country	Boys Mean and s.e.	Girls Mean and s.e.	S. E. of the Difference Using JRR	S. E. of the Difference Assuming SRS
Australia	568.9 (3.3)	555.8 (3.2)	3.0	4.6
Austria	571.8 (3.9)	556.4 (3.7)	3.7	5.3
Canada	552.7 (3.7)	545.0 (3.2)	3.0	4.9
Cyprus	480.3 (4.0)	470.6 (3.1)	2.9	5.1
Czech Republic	565.5 (3.4)	548.3 (3.6)	3.3	5.0
Greece	500.7 (4.5)	493.8 (4.3)	3.2	6.2
Hong Kong	539.7 (4.1)	525.7 (3.8)	2.9	5.6
Hungary	539.3 (3.8)	525.1 (3.9)	3.5	5.4
Iceland	513.8 (4.3)	496.2 (3.3)	3.8	5.4
Iran, Islamic Rep.	420.7 (5.9)	412.0 (4.7)	7.5	7.6
Ireland	542.8 (3.5)	536.2 (4.5)	4.5	5.7
Israel	512.2 (4.5)	501.1 (3.8)	4.0	5.9
Japan	580.4 (2.0)	566.8 (2.0)	2.0	2.9
Korea	603.8 (2.2)	589.9 (2.5)	2.9	3.3
Kuwait	389.1 (5.8)	414.3 (3.1)	7.0	6.6
Latvia (LSS)	511.7 (5.4)	512.7 (5.5)	4.7	7.7
Netherlands	569.8 (3.6)	544.3 (3.5)	3.3	5.0
New Zealand	527.0 (6.1)	535.0 (4.8)	4.9	7.7
Norway	533.6 (4.7)	525.7 (3.7)	4.4	5.9
Portugal	481.3 (4.5)	478.2 (4.2)	3.3	6.2
Singapore	548.5 (5.4)	544.5 (6.3)	5.8	8.3
Thailand	471.2 (5.9)	474.5 (4.3)	3.3	7.3
England	555.0 (4.0)	548.1 (3.4)	3.6	5.3
Scotland	537.6 (4.5)	533.4 (4.3)	2.9	6.2
United States	571.5 (3.3)	559.6 (3.3)	2.4	4.6
Slovenia	547.9 (3.3)	544.1 (4.0)	3.0	5.2

JRR = jackknife repeated replicate method

SRS = simple random sample

Table 8.17 Standard Error of the Gender Difference Mathematics - Seventh Grade

Country	Boys Mean and s.e.	Girls Mean and s.e.	S.E. of the Difference Using JRR	S.E. of the Difference Assuming SRS
Australia	495.1 (5.2)	500.5 (4.3)	5.8	6.8
Austria	510.0 (4.6)	508.6 (3.3)	4.8	5.6
Belgium (Fl)	556.7 (4.5)	558.5 (4.7)	5.9	6.5
Belgium (Fr)	513.8 (4.1)	501.1 (4.2)	4.1	5.9
Bulgaria	508.0 (6.9)	518.3 (8.7)	5.1	11.1
Canada	495.1 (2.7)	493.4 (2.6)	3.1	3.8
Colombia	371.7 (3.8)	365.0 (3.9)	5.3	5.4
Cyprus	445.9 (2.5)	445.6 (2.6)	3.4	3.6
Czech Republic	526.6 (4.8)	520.3 (5.6)	3.6	7.4
Slovak Republic	510.9 (4.4)	504.9 (3.3)	3.9	5.5
Denmark	468.5 (2.8)	461.8 (2.9)	3.7	4.0
France	497.0 (3.6)	488.8 (3.3)	2.7	4.9
Germany	486.3 (4.8)	483.8 (4.5)	4.3	6.6
Greece	439.5 (3.2)	440.4 (3.0)	2.7	4.4
Hong Kong	569.7 (9.7)	555.8 (8.3)	9.6	12.8
Hungary	502.5 (3.8)	501.1 (4.4)	3.8	5.8
Iceland	460.5 (2.7)	458.3 (3.2)	2.9	4.2
Iran, Islamic Rep.	407.1 (2.7)	393.1 (2.3)	3.7	3.5
Ireland	506.7 (6.0)	493.7 (4.8)	6.9	7.7
Japan	576.4 (2.7)	565.4 (2.0)	3.0	3.4
Korea	584.4 (3.7)	567.1 (4.4)	6.2	5.7
Latvia (LSS)	463.3 (3.5)	459.6 (3.3)	3.8	4.8
Lithuania	423.3 (3.6)	433.1 (3.5)	3.2	5.0
Netherlands	517.5 (5.2)	514.6 (4.3)	4.8	6.7
New Zealand	473.1 (4.6)	470.1 (3.8)	3.7	5.9
Norway	462.4 (3.3)	458.8 (3.2)	3.2	4.6
Portugal	426.3 (2.7)	420.2 (2.2)	2.2	3.5
Romania	456.6 (3.7)	452.4 (3.7)	2.9	5.2
Russian Federation	502.4 (5.1)	499.5 (3.5)	3.5	6.1
Singapore	601.3 (7.1)	600.8 (8.0)	8.2	10.7
South Africa	351.8 (5.3)	344.2 (3.3)	4.1	6.2
Spain	450.7 (2.7)	445.2 (2.7)	3.1	3.8
Sweden	480.1 (2.8)	474.8 (3.2)	3.4	4.2
Switzerland	512.5 (2.9)	498.5 (2.6)	2.9	3.9
Thailand	494.3 (4.8)	495.4 (5.7)	4.4	7.5
England	483.9 (6.2)	467.0 (4.3)	8.3	7.5
Scotland	464.5 (4.6)	461.7 (3.8)	3.8	5.9
United States	478.1 (5.7)	473.3 (5.7)	3.2	8.1
Slovenia	500.6 (3.5)	495.8 (3.2)	3.2	4.7

JRR = jackknife repeated replicate method

SRS = simple random sample

Table 8.18 Standard Error of the Gender Difference Mathematics - Eighth Grade

Country	Boys Mean and s.e.	Girls Mean and s.e.	S.E. of the Difference Using JRR	S.E. of the Difference Assuming SRS
Australia	527.4 (5.1)	532.0 (4.6)	5.4	6.9
Austria	543.6 (3.2)	535.6 (4.5)	4.9	5.6
Belgium (Fl)	563.1 (8.8)	567.2 (7.4)	11.7	11.5
Belgium (Fr)	530.0 (4.7)	523.5 (3.7)	5.1	6.0
Bulgaria	533.2 (7.0)	546.2 (6.7)	5.2	9.6
Canada	526.0 (3.2)	529.6 (2.7)	3.4	4.2
Colombia	385.7 (6.9)	384.0 (3.6)	8.2	7.7
Cyprus	472.2 (2.8)	475.3 (2.5)	3.7	3.7
Czech Republic	569.0 (4.5)	558.4 (6.3)	4.5	7.7
Slovak Republic	549.0 (3.7)	545.3 (3.6)	3.2	5.2
Denmark	511.5 (3.2)	494.3 (3.4)	3.4	4.7
France	541.9 (3.1)	535.7 (3.8)	3.2	4.9
Germany	511.6 (5.1)	509.1 (5.0)	4.7	7.1
Greece	489.7 (3.7)	477.8 (3.1)	2.9	4.8
Hong Kong	597.2 (7.7)	577.2 (7.7)	8.6	10.9
Hungary	537.3 (3.6)	537.2 (3.6)	3.3	5.1
Iceland	487.6 (5.5)	485.9 (5.6)	6.3	7.8
Iran, Islamic Rep.	434.1 (2.9)	420.8 (3.3)	4.5	4.4
Ireland	534.6 (7.2)	520.3 (6.0)	8.2	9.3
Israel	538.7 (6.6)	509.4 (6.9)	5.8	9.6
Japan	609.2 (2.6)	600.0 (2.1)	2.9	3.3
Korea	615.2 (3.2)	597.9 (3.4)	4.8	4.7
Kuwait	389.0 (4.3)	395.5 (2.6)	5.0	5.0
Latvia (LSS)	495.6 (3.8)	491.2 (3.5)	3.7	5.2
Lithuania	476.8 (4.0)	477.6 (4.1)	4.0	5.7
Netherlands	544.8 (7.8)	536.4 (6.4)	4.4	10.1
New Zealand	512.2 (5.9)	503.0 (5.3)	6.7	7.9
Norway	505.3 (2.8)	501.3 (2.7)	3.3	3.9
Portugal	459.8 (2.8)	448.9 (2.7)	2.4	3.9
Romania	482.9 (4.8)	480.2 (4.0)	3.4	6.2
Russian Federation	534.8 (6.3)	536.0 (5.0)	3.7	8.0
Singapore	642.2 (6.3)	644.6 (5.4)	6.5	8.3
South Africa	359.8 (6.3)	349.2 (4.1)	5.7	7.5
Spain	492.2 (2.5)	482.7 (2.6)	3.2	3.6
Sweden	519.5 (3.6)	517.7 (3.1)	3.1	4.7
Switzerland	547.8 (3.5)	543.0 (3.1)	3.7	4.7
Thailand	517.0 (5.6)	526.2 (7.0)	6.5	9.0
England	507.7 (5.1)	503.5 (3.5)	7.1	6.2
Scotland	506.2 (6.6)	490.3 (5.2)	4.9	8.4
United States	502.0 (5.2)	497.5 (4.5)	2.9	6.9
Slovenia	544.9 (3.8)	536.9 (3.3)	3.4	5.0

JRR = jackknife repeated replicate method

SRS = simple random sample

Table 8.19 Standard Error of the Gender Difference Science - Eighth Grade

Country	Boys Mean and s.e.	Girls Mean and s.e.	S.E. of the Difference Using JRR	S.E. of the Difference Assuming SRS
Australia	506.6 (5.2)	502.3 (4.0)	5.9	6.6
Austria	522.3 (4.3)	515.5 (4.1)	5.2	6.0
Belgium (Fl)	535.8 (3.3)	521.4 (3.1)	3.9	4.5
Belgium (Fr)	452.7 (3.6)	432.1 (3.5)	3.5	5.0
Bulgaria	529.2 (5.5)	532.0 (6.7)	5.7	8.7
Canada	505.5 (2.9)	493.1 (2.5)	2.9	3.8
Colombia	396.4 (3.8)	378.5 (4.4)	4.8	5.8
Cyprus	420.1 (2.8)	420.1 (2.6)	4.1	3.9
Czech Republic	543.2 (3.2)	522.9 (4.1)	3.2	5.2
Slovak Republic	520.3 (4.0)	499.4 (3.1)	3.9	5.1
Denmark	452.0 (3.0)	427.4 (2.8)	3.9	4.1
France	460.8 (3.1)	442.7 (3.0)	3.1	4.3
Germany	504.9 (4.9)	495.4 (4.5)	4.5	6.6
Greece	451.7 (3.2)	445.5 (2.8)	3.1	4.2
Hong Kong	503.5 (6.6)	485.0 (5.8)	6.3	8.7
Hungary	525.3 (3.9)	510.5 (3.4)	3.4	5.1
Iceland	467.7 (4.4)	455.9 (2.4)	4.5	5.0
Iran, Islamic Rep.	443.0 (2.9)	427.8 (4.1)	4.9	5.0
Ireland	504.4 (4.6)	487.3 (4.5)	5.9	6.4
Japan	536.0 (2.6)	525.8 (1.9)	2.7	3.2
Korea	545.4 (2.8)	520.8 (3.2)	4.4	4.2
Latvia (LSS)	439.6 (3.6)	430.1 (3.0)	3.8	4.7
Lithuania	405.4 (3.5)	400.7 (4.2)	3.8	5.5
Netherlands	522.8 (4.0)	512.2 (4.4)	4.4	5.9
New Zealand	489.1 (4.3)	471.7 (3.7)	4.3	5.7
Norway	488.9 (3.6)	477.2 (3.6)	4.3	5.1
Portugal	436.3 (2.4)	420.1 (2.4)	2.4	3.4
Romania	455.8 (4.7)	447.7 (4.9)	3.5	6.7
Russian Federation	492.9 (5.3)	475.4 (3.8)	3.8	6.5
Singapore	548.1 (7.9)	541.3 (8.2)	9.2	11.4
South Africa	323.8 (6.4)	312.5 (5.2)	4.9	8.3
Spain	487.5 (2.9)	466.7 (2.3)	3.3	3.7
Sweden	493.3 (2.9)	483.6 (3.3)	3.5	4.4
Switzerland	492.4 (2.9)	474.8 (2.9)	3.0	4.1
Thailand	494.8 (3.3)	491.7 (3.5)	3.1	4.8
England	522.2 (5.6)	499.9 (4.6)	8.0	7.3
Scotland	477.0 (4.4)	459.2 (4.1)	3.8	6.0
United States	514.4 (6.3)	502.2 (5.8)	5.2	8.6
Slovenia	539.2 (3.0)	521.2 (2.8)	3.2	4.1

JRR = jackknife repeated replicate method

SRS = simple random sample

Table 8.20 Standard Error of the Gender Difference Science - Eighth Grade

Country	Boys Mean and s.e.	Girls Mean and s.e.	S.E. of the Difference Using JRR	S.E. of the Difference Assuming SRS
Australia	549.6 (5.2)	539.5 (4.1)	5.3	6.6
Austria	566.4 (4.0)	548.7 (4.6)	4.3	6.1
Belgium (Fl)	557.6 (6.0)	542.9 (5.8)	8.7	8.4
Belgium (Fr)	478.9 (4.8)	463.0 (2.9)	5.5	5.6
Bulgaria	563.2 (5.7)	566.8 (6.6)	6.3	8.7
Canada	537.4 (3.1)	525.4 (3.7)	4.3	4.8
Colombia	417.6 (7.3)	404.9 (4.6)	8.4	8.6
Cyprus	461.0 (2.2)	464.8 (2.7)	3.0	3.4
Czech Republic	585.9 (4.2)	561.6 (5.8)	4.5	7.2
Slovak Republic	552.2 (3.5)	536.9 (3.9)	3.6	5.2
Denmark	494.2 (3.6)	463.3 (3.9)	4.5	5.3
France	505.9 (2.7)	490.1 (3.3)	3.1	4.3
Germany	541.7 (5.9)	523.9 (4.9)	4.8	7.6
Greece	504.8 (2.6)	489.3 (3.1)	3.3	4.0
Hong Kong	534.7 (5.5)	507.3 (5.1)	5.8	7.5
Hungary	563.0 (3.1)	544.6 (3.4)	3.6	4.7
Iceland	501.1 (5.1)	485.5 (4.6)	5.2	6.9
Iran, Islamic Rep.	477.3 (3.8)	460.5 (3.2)	5.2	4.9
Ireland	543.6 (6.6)	532.0 (5.2)	7.6	8.4
Israel	544.8 (6.4)	512.2 (6.1)	7.1	8.9
Japan	579.0 (2.4)	562.4 (2.0)	3.0	3.1
Korea	575.9 (2.7)	551.5 (2.3)	3.8	3.6
Kuwait	416.0 (6.6)	443.5 (3.3)	7.4	7.4
Latvia (LSS)	492.4 (3.3)	477.8 (3.2)	3.5	4.6
Lithuania	483.9 (3.8)	470.3 (4.0)	3.9	5.5
Netherlands	570.2 (6.4)	549.8 (4.9)	5.1	8.1
New Zealand	537.6 (5.4)	512.3 (5.2)	6.2	7.6
Norway	534.0 (3.2)	520.5 (2.0)	3.7	3.8
Portugal	490.5 (2.8)	468.4 (2.7)	2.8	3.9
Romania	492.0 (5.3)	480.1 (5.0)	3.8	7.3
Russian Federation	544.0 (4.9)	532.9 (3.7)	3.4	6.2
Singapore	611.9 (6.7)	602.7 (7.0)	8.1	9.7
South Africa	336.6 (9.5)	315.4 (6.0)	8.6	11.3
Spain	526.4 (2.1)	508.1 (2.3)	2.9	3.1
Sweden	542.5 (3.4)	528.0 (3.4)	3.4	4.8
Switzerland	529.0 (3.2)	514.0 (3.0)	3.7	4.4
Thailand	524.4 (3.9)	526.3 (4.3)	3.6	5.8
England	561.6 (5.6)	541.6 (4.2)	7.7	7.1
Scotland	527.3 (6.4)	506.9 (4.7)	5.1	7.9
United States	538.8 (4.9)	530.0 (5.2)	3.6	7.2
Slovenia	573.2 (3.2)	547.8 (3.2)	4.1	4.5

JRR = jackknife repeated replicate method

SRS = simple random sample

8.7 REPORTING POPULATION 1 ACHIEVEMENT ON THE POPULATION 2 SCALE

In order to establish a link between the reporting scales for Population 1 and Population 2, a number of items in the TIMSS tests were administered to students in both populations. A total of 15 mathematics and 18 science items were administered in both populations, at grades three and four and at grades seven and eight. The 15 mathematics items were exclusively multiple choice, while the 18 science items consisted of 10 multiple-choice items and 8 free-response items. All of these items were dichotomously scored, and were worth one score point each. Because of the existence of these "link items," it was possible to link the Population 1 results to those of Population 2.

8.7.1 Estimating the Shift in Item Difficulties

The scaling of the student achievement data for Population 2 and the reporting of results on that scale were completed before those for Population 1. Because of this, the scales from the two populations were linked by reporting the Population 1 results on the Population 2 scale. In order to achieve this, the item difficulties were first calibrated separately in each population. The link items were then identified and the average of the differences between the item difficulties from each of the calibrations was computed, separately for mathematics and science. This average is an estimate of the shift in item difficulty that would have to be made in order to report the results from the Population 1 scaling on a scale based on the calibration of the Population 2 items. Table 8.21 and 8.22 present the mathematics and science link items with their item difficulties calibrated separately for Population 1 and Population 2, the difference between them, and the average of the difference (the shift) calculated as

$$Shift(s) = \frac{\sum_{L} (d_l^{pop1} - d_l^{pop2})}{L}$$

where L is the number of link items, d_l^{pop1} is the item difficulty of item L at Population 1, d_l^{pop2} is the item difficulty of item L at Population 2. This shift is applied to the logit metric in which the Population 1 scores are first computed.

Table 8.21 Mathematics Link Items

Item Name in Population 1	Item Name in Population 2	Difficulty at Population 1	Difficulty at Population 2	Difference in Difficulty Between Populations
F08	D11	-0.227	-1.906	-1.679
K07	E06	1.852	0.693	-1.159
C03	H08	0.352	-1.047	-1.399
G04	H12	0.414	-0.996	-1.410
U02	I06	0.297	-1.216	-1.513
G01	J17	0.984	-0.585	-1.569
H05	K03	0.461	-0.644	-1.105
F05	L10	-0.516	-2.025	-1.509
L08	L12	1.461	-0.977	-2.438
L04	L13	-0.516	-2.332	-1.816
L02	M03	0.516	-1.143	-1.659
B07	P12	0.758	-0.809	-1.567
F06	P14	-0.164	-1.262	-1.098
B05	Q04	-0.234	-1.777	-1.543
I09	R12	-0.250	-1.953	-1.703
Average		0.346	-1.199	-1.544
Standard Error				0.085

Table 8.22 Science Link Items

Item Name in Population 1	Item Name in Population 2	Difficulty at Population 1	Difficulty at Population 2	Difference in Difficulty Between Populations
D04	B01	-0.773	-1.845	-1.072
E07	B04	0.000	-1.998	-1.998
N08	C10	-0.008	-1.102	-1.094
P05	D02	0.633	-0.914	-1.547
P09	D06	0.805	-0.877	-1.682
B04	F03	0.031	-0.515	-0.546
O04	H03	-0.477	-1.233	-0.756
Q02	I10	-0.211	-0.921	-0.710
O06	K19	0.156	-0.852	-1.008
Q08	M14	0.617	-0.764	-1.381
Q04	N07	0.047	-2.016	-2.063
O01	N08	0.680	-0.727	-1.407
R01	N10	2.109	0.123	-1.986
Y01	O14	1.516	0.053	-1.463
W03	O16	1.656	-0.112	-1.768
O05	R01	0.156	-0.654	-0.810
Z01A	W01A	0.023	-1.306	-1.329
Z01B	W01B	2.000	0.606	-1.394
Average		0.498	-0.836	-1.334
Standard Error				0.109

8.7.2 Estimating the Variance of the Shift

Because the student responses from which the item difficulty parameters are estimated are derived from random samples of students, the estimates of the relative item difficulty of the items in the two samples are subject to sampling variation. It is important to take this variation into account when reporting results on a scale that has been constructed by means of a shift from another scale. This variance component, known as the variance of the shift, is computed as the variance of the differences in item difficulty with respect to the mean difference in item difficulty. The formula for this calculation is as follows

$$Var_{Shift(s)} = \frac{\sum_{L}((d_l^{pop1} - d_l^{pop2}) - Shift(s))^2}{L^2}$$

where L is the number of link items, d_l^{pop1} is the item difficulty of item L at Population 1, d_l^{pop2} is the item difficulty of item L at Population 2, and *Shift (s)* is the average difference between two calibrations. The variance of the shift is used only when reporting the scores from one scale onto another scale. This variance component is added to the standard variance estimate.

REFERENCES

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

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