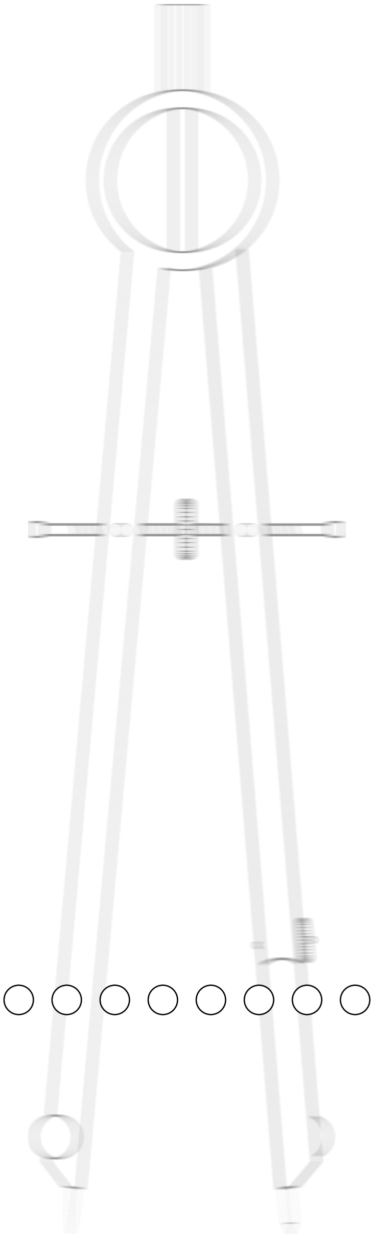
15

Reporting Student Achievement in Mathematics and Science for TIMSS 1999 Benchmarking

Eugenio J. Gonzalez
Kelvin D. Gregory



15 Reporting Student Achievement in Mathematics and Science for TIMSS 1999 Benchmarking[1]

Eugenio J. Gonzalez
Kelvin D. Gregory

15.1 Overview

As described in earlier chapters, the Benchmarking study makes extensive use of imputed student proficiency scores to report achievement in mathematics and science, both in the subjects overall and in the separate content areas. This chapter describes the procedures followed in computing the major statistics used to summarize achievement in the TIMSS 1999 Benchmarking Reports (Mullis et al., 2001; Martin et al., 2001), including average scores based on plausible values, Bonferroni adjustments for multiple comparisons, international benchmarks of achievement, and profiles of relative performance in subject-matter areas.

15.2 Computing Average Student Achievement

The item response theory (IRT) scaling procedure described in chapter 13 yields five imputed scores or plausible values in mathematics and science and in each of their content areas for each student. Average mathematics or science scores for countries or Benchmarking jurisdictions were computed by first taking the mean for each of the five plausible values, and then taking the mean of the five plausible-value means, as follows: The average for each plausible value was computed as the weighted mean

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

where

$\bar{X}_{pvl}$ is the country or jurisdiction mean for plausible value l

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

[1] This chapter is based on Gonzalez & Gregory (2000) from the TIMSS 1999 international technical report (Martin, Gregory, & Stemler, 2000).

$W^{i,j}$ is the weight associated with the j^{th} student in class i, described in chapters 5 and 6

N is the number of students in the sample.

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

The international average for mathematics and science was computed by taking the mean of the country means for each of the five plausible values and averaging across these five international means, as follows: The international average for each plausible value was computed as the average of that plausible value for each country:

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

where

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

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

and N is the number of countries.

The international average was the average of these five international means. The international averages were based on all TIMSS 1999 countries. Data from Benchmarking jurisdictions were not included in the computation of international averages.

15.3 Achievement Differences Across Benchmarking Jurisdictions

The TIMSS 1999 Benchmarking Reports aim to provide fair and accurate comparisons of student achievement across the participating jurisdictions. Most of the exhibits summarize achievement using a statistic such as a mean or percentage, and each statistic is accompanied by its standard error, which is a measure of the uncertainty due to student sampling and the imputation process. In comparisons of performance across jurisdictions, standard errors were used to assess the statistical significance of the difference between the summary statistics.

The charts presented in the TIMSS 1999 Benchmarking Reports provide comparisons of average performance of a jurisdiction with that of the TIMSS 1999 countries as well as with other participating jurisdictions. The significance tests reported in these charts include a Bonferroni adjustment for multiple comparisons. The Bonferroni adjustment is necessary because the probability of finding a difference that is an artifact of chance greatly increases as the number of simultaneous comparisons increases.

15.3.1 Bonferroni Adjustments in TIMSS

If repeated samples were taken from two populations with the same mean and variance, and in each one the hypothesis that the two means are significantly different at the $\alpha = .05$ level (i.e., with 95% confidence) was tested, then it would be expected that in about 5% of the comparisons significant differences would be found between the sample means even though no difference exists in the populations. The probability of finding significant differences when none exist (the so-called Type I error) is given by α. Conversely, the probability of not making such an error is $1 - \alpha$, which in the case of a single test is .95. When $\alpha = .05$, comparing the means of three countries involves three tests (country A versus country B, country B versus country C, and country A versus country C). Since these are independent tests, the probability of avoiding a Type I error in any of the three is the product of the individual probabilities, which is $(1 - \alpha)(1 - \alpha)(1 - \alpha)$. With $\alpha= .05$, the overall probability of avoiding a Type I error is only .873. Stated differently, the probability of committing a Type I error rises from .05 for one comparisons to .127 with three comparisons, which is considerably less than the probability for a single test. As the number of tests increases, the probability of making a Type I error increases rapidly.

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

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

Most users of the multiple comparison exhibits in the international reports are likely to be interested in comparing a single country with all other countries, rather than in making all possible between-country comparisons at once; the more realistic approach of using the number of countries (minus one) to adjust the significance level was therefore adopted for the international reports. This meant that the number of simultaneous comparisons to be adjusted for was 37 instead of 703. The critical value for a 95% significance test adjusted for 37 simultaneous comparisons is 3.2049, from the appropriate quantiles from the normal (Gaussian) distribution.

In the multiple comparison exhibits of the TIMSS 1999 Benchmarking Reports (Martin et al., 2001; Mullis et al., 2001), it was decided to keep the same Bonferroni correction as in the international reports so that between-country significance tests in both sets of reports would have the same results. This decision was taken despite the fact that Benchmarking exhibits that included all 38 TIMSS countries as well as the 27 Benchmarking participants had more comparisons (65) than exhibits in the international reports, which involved just the 38 countries. Consequently, exhibits with all 65 comparisons, which are confined to the first chapter in each Benchmarking report, present significance tests that are slightly less conservative than they would otherwise be.

15.3.2 Standard Error of the Difference

Mean proficiencies were considered significantly different if the absolute difference between them, divided by the standard error of the difference, was greater than the Bonferroni-adjusted critical value. For differences between countries or Benchmarking jurisdictions, which can be considered as independent samples, the standard error of the difference in means was computed as the square root of the sum of the squared standard errors of each mean:

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

where se_1 and se_2 are the standard errors of the means. Exhibits 15.1 and 15.2 show the means and standard errors for mathematics and science used in the calculation of statistical significance for countries and Benchmarking jurisdictions, respectively.

Exhibit 15.1 Means and Standard Errors for Multiple-Comparisons Exhibits-Countries

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

Exhibit 15.2 Means and Standard Errors for Multiple-Comparisons Exhibits -States and Districts

States	Math		Science	
	Mean	S.E.	Mean	SE
Connecticut	512.389	9.075	529.485	10.436
Idaho	494.886	7.385	526.368	6.585
Illinois	509.478	6.730	520.515	6.546
Indiana	514.626	7.186	534.202	6.973
Maryland	494.610	6.245	506.110	7.689
Massachusetts	513.469	5.938	533.194	7.363
Michigan	516.630	7.452	544.142	8.624
Missouri	489.731	5.314	522.826	6.486
North Carolina	495.218	7.026	507.792	6.544
Oregon	514.110	5.953	536.094	6.051
Pennsylvania	507.452	6.299	528.951	6.475
South Carolina	501.610	7.393	510.958	6.693
Texas	516.445	9.066	508.698	10.427

Districts and Consortia	Math		Science	
	Mean	S.E.	Mean	SE
Academy School Dist. #20, CO	528.464	1.828	558.742	2.116
Chicago Public Schools, IL	462.500	6.102	449.447	9.505
Delaware Science Coalition, DE	479.483	8.928	500.446	8.379
First in the World Consort., IL	559.633	5.775	565.461	5.255
Fremont/Lincoln/WestSide PS, NE	488.142	8.215	511.302	5.780
Guilford County, NC	513.565	7.705	533.780	7.063
Jersey City Public Schools, NJ	474.814	8.610	439.666	9.756
Miami-Dade County PS, FL	421.330	9.449	425.956	10.937
Michigan Invitational Group, MI”	531.748	5.815	563.495	6.246
Montgomery County, MD	537.370	3.548	531.480	4.252
Naperville Sch. Dist. #203, IL	569.172	2.835	583.727	4.092
Project SMART Consortium, OH	520.593	7.507	539.223	8.370
Rochester City Sch. Dist., NY	444.404	6.462	451.669	7.372
SW Math/Sci. Collaborative, PA	516.719	7.547	543.249	7.429

15.4 Comparing Achievement with the International Mean

Many of the data exhibits in the TIMSS 1999 International Reports show countries’ and jurisdictions; mean achievement compared with the international mean. Since this resulted in 38 simultaneous comparisons, the critical value was adjusted to 3.2125 using the Dunn-Bonferroni procedure. In the Benchmarking Reports, the corresponding exhibits contained 40 comparisons (27 Benchmarking participants and 13 selected countries), but for consistency with the international reports, the critical value for 38 comparisons was used in Benchmarking exhibits also.

When comparing each country's mean with the international average, TIMSS took into account the fact that the country contributed to the international standard error. To correct for this contribution, TIMSS adjusted the standard error of the difference. The sampling component of the standard error of the difference for country *j* was

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

where

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

N is the number of countries

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

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

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

$$se_{i\_dif\_j} = \sqrt{\frac{6}{5} Var(d_{1\ldots} d_{l\ldots} d_5)}$$

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

Finally, the standard error of the difference was calculated as:

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

15.5 International Benchmarks of Achievement

In order to provide more information about student achievement, TIMSS identified four points on each of the mathematics and science scales for use as international benchmark as described in chapter 14. The top 10% benchmark was defined as the 90th percentile on the TIMSS scale, computed across all students in all participating countries, with countries weighted in proportion to the size of their eighth-grade population. This point on each scale (mathematics and science) is the point above which the top 10% of students in the 1999 TIMSS assessment scored. The upper quarter benchmark is the 75th percentile on the scale, above which the top 25% of students scored. The median benchmark is the 50th percentile, above which the top half of students scored. Finally, the lower quarter benchmark is the 25th percentile, the point reached by the top 75% of students. Comparing the percentage of students in Benchmarking jurisdictions that reached the achievement levels defined by these international benchmarks was a very useful way of describing student performance at various points of the ability distribution.

15.5.1 Establishing the International Benchmarks of Achievement

In computing of the international benchmarks of achievement, each country was weighted to contribute as many students as there were students in the target population. In other words, each country's contribution to setting the international benchmarks was proportional to the estimated population enrolled in the eighth grade. Exhibit 15.3 shows the contribution of each country to the estimation of the international benchmarks.

Exhibit 15.3 Estimated Enrollment at the Eighth Grade

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

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

Because of the imputation technology used to derive the student achievement scores, the international benchmarks had to be computed once for each of the five plausible values, and the results averaged to arrive at the final figure. The standard errors presented in the exhibits are computed by taking into account the sampling design as well as the variance due to imputation. The international benchmarks are presented in Exhibit 15.4 and 15.5 for mathematics and science, respectively.

Exhibit 15.4 International Benchmarks of Achievement for Eighth Grade—Mathematics

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

Exhibit 15.5 International Benchmarks of Achievement for Eighth Grade—Science

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

Exhibit 15.6 Percentages of Students Reaching TIMSS 1999 International Benchmarks of Mathematics Achievement

States, Districts and Consortia	Top 10%	Upper Quarter	Median	Lower Quarter
Connecticut	11 (2.5)	31 (3.9)	67 (4.4)	91 (1.9)
Idaho	5 (1.1)	24 (2.9)	61 (3.5)	88 (2.2)
Illinois	10 (1.6)	29 (2.9)	65 (3.3)	92 (1.5)
Indiana †	9 (1.9)	30 (3.9)	69 (3.6)	94 (1.2)
Maryland	8 (1.4)	27 (2.5)	57 (3.2)	87 (2.0)
Massachusetts	10 (1.6)	31 (2.6)	68 (3.0)	92 (1.6)
Michigan	10 (2.0)	33 (3.7)	70 (3.3)	92 (1.7)
Missouri	4 (0.9)	20 (2.4)	58 (2.9)	89 (1.5)
North Carolina	7 (1.6)	25 (3.1)	57 (3.3)	88 (2.0)
Oregon	10 (1.8)	32 (2.8)	69 (2.8)	91 (1.4)
Pennsylvania	9 (1.3)	28 (2.6)	65 (3.0)	91 (1.8)
South Carolina	10 (2.0)	30 (3.2)	60 (3.5)	88 (1.8)
Texas	13 (2.2)	37 (3.8)	66 (4.3)	90 (2.1)

States, Districts and Consortia	Top 10%	Upper Quarter	Median	Lower Quarter
Academy School Dist. #20, CO	12 (0.8)	38 (1.5)	75 (1.5)	95 (0.7)
Chicago Public Schools, IL	2 (0.9)	12 (1.7)	41 (4.3)	81 (2.5)
Delaware Science Coalition, DE	5 (1.8)	22 (4.1)	51 (4.5)	83 (2.4)
First in the World Consort., IL	22 (3.2)	56 (3.3)	87 (2.1)	98 (0.6)
Fremont/Lincoln/WestSide PS, NE	6 (2.3)	23 (4.1)	58 (4.0)	84 (2.7)
Guilford County, NC 1	10 (2.2)	33 (3.5)	66 (4.1)	91 (1.6)
Jersey City Public Schools, NJ	6 (1.9)	17 (3.4)	48 (3.9)	82 (2.9)
Miami-Dade County PS, FL	2 (0.9)	9 (2.4)	29 (3.6)	61 (3.5)
Michigan Invitational Group, MI	12 (2.4)	39 (3.4)	77 (3.0)	96 (1.3)
Montgomery County, MD 1	17 (2.2)	45 (1.8)	77 (1.4)	95 (1.1)
Naperville Sch. Dist. #203, IL	24 (1.7)	59 (2.2)	91 (1.1)	99 (0.4)
Project SMART Consortium, OH	11 (2.9)	34 (4.7)	70 (3.1)	95 (1.0)
Rochester City Sch. Dist., NY	2 (0.9)	9 (2.5)	32 (3.2)	73 (2.9)
SW Math/Sci. Collaborative, PA	11 (2.7)	32 (3.9)	68 (3.1)	93 (1.6)

Top 10% Benchmark (90th Percentile)	616
Upper Quarter Benchmark (75th Percentile)	555
Median Benchmark (50th Percentile)	479
Lower Quarter Benchmark (25th Percentile)	396

States in *italics* did not fully satisfy guidelines for sample participation rates (see Appendix A for details).

† Met guidelines for sample participation rates only after replacement schools were included (see Exhibit A.6).

1 National Defined Population covers less than 90 percent of National Desired Population (see Exhibit A.3).

( ) Standard errors appear in parentheses. Because results are rounded to the nearest whole number, some totals may appear inconsistent.

Exhibit 15.7 Percentages of Students Reaching TIMSS 1999 International Benchmarks of Science Achievement

States	Top 10%	Upper Quarter	Median	Lower Quarter
Connecticut	17 (3.0)	39 (4.4)	69 (4.6)	90 (2.5)
Idaho	13 (1.8)	37 (3.2)	70 (3.3)	91 (1.8)
Illinois	14 (1.9)	36 (3.0)	66 (3.0)	88 (1.5)
Indiana †	18 (2.5)	41 (3.6)	72 (2.8)	92 (1.4)
Maryland	12 (1.3)	31 (3.0)	59 (3.5)	84 (2.5)
Massachusetts	17 (2.4)	40 (3.0)	71 (3.4)	92 (1.7)
Michigan	22 (2.6)	47 (3.6)	75 (3.4)	91 (2.2)
Missouri	14 (2.3)	36 (3.0)	67 (2.8)	89 (1.8)
North Carolina	11 (1.4)	30 (2.9)	60 (3.4)	85 (2.1)
Oregon	19 (2.3)	43 (2.7)	73 (2.6)	91 (1.9)
Pennsylvania	15 (1.5)	38 (2.5)	70 (3.2)	91 (1.6)
South Carolina	13 (1.8)	34 (2.7)	60 (3.4)	85 (1.7)
Texas	15 (2.1)	35 (3.6)	61 (4.5)	83 (3.3)

Districts and Consortia	Top 10%	Upper Quarter	Median	Lower Quarter
Academy School Dist. #20, CO	23 (1.6)	52 (1.5)	84 (1.2)	97 (0.6)
Chicago Public Schools, IL	3 (1.1)	11 (2.4)	34 (3.9)	67 (3.8)
Delaware Science Coalition, DE	10 (1.8)	29 (4.0)	56 (4.2)	83 (2.1)
First in the World Consort., IL	27 (3.7)	54 (3.6)	85 (2.0)	97 (0.9)
Fremont/Lincoln/WestSide PS, NE	11 (1.7)	32 (3.1)	63 (3.2)	86 (2.1)
Guilford County, NC 1	19 (2.5)	43 (3.6)	69 (3.5)	90 (2.0)
Jersey City Public Schools, NJ	3 (1.5)	11 (3.1)	31 (3.6)	64 (3.5)
Miami-Dade County PS, FL	4 (1.4)	10 (2.4)	28 (3.0)	58 (3.7)
Michigan Invitational Group, MI	25 (3.1)	54 (3.0)	84 (2.1)	96 (1.1)
Montgomery County, MD 1	17 (1.1)	40 (2.5)	70 (2.3)	91 (1.3)
Naperville Sch. Dist. #203, IL	33 (2.5)	64 (2.2)	90 (1.2)	98 (0.6)
Project SMART Consortium, OH	19 (3.6)	43 (5.0)	73 (3.3)	93 (1.1)
Rochester City Sch. Dist., NY	3 (1.3)	12 (2.5)	33 (3.7)	68 (3.0)
SW Math/Sci. Collaborative, PA	19 (3.1)	45 (3.6)	75 (3.5)	94 (1.7)

Top 10% Benchmark (90th Percentile)	616
Upper Quarter Benchmark (75th Percentile)	558
Median Benchmark (50th Percentile)	488
Lower Quarter Benchmark (25th Percentile)	410

States in *italics* did not fully satisfy guidelines for sample participation rates (see Appendix A for details).

† Met guidelines for sample participation rates only after replacement schools were included (see Exhibit A.6).

1 National Defined Population covers less than 90 percent of National Desired Population (see Exhibit A.3).

( ) Standard errors appear in parentheses. Because results are rounded to the nearest whole number, some totals may appear inconsistent.

15.5.2 Reporting Student Achievement at the International Benchmarks

To compare student performance at the international benchmarks, TIMSS computed the percentage of students in each Benchmarking jurisdiction reaching each international benchmark. These percentages and their standard errors are presented in Exhibit 15.6 for mathematics and in Exhibit 15.7 for science.

15.6 Reporting Gender Differences

TIMSS reported gender differences in student achievement in mathematics and science overall, as well as in content areas. Gender differences in countries and Benchmarking jurisdictions were presented in an exhibit showing mean achievement for males and females, the differences between them, and an accompanying graph indicating whether the difference was statistically significant. The significance test was adjusted for multiple comparisons, based on the number of countries presented.

Because in most countries males and females attend the same schools, the two samples cannot be treated as independent for the purpose of statistical tests. Accordingly, TIMSS used a jackknife procedure applicable to correlated samples for estimating the standard error of the male-female difference. This involves computing the differences between boys and girls once for each of the 75 replicate samples, and five more times, once for each plausible value, as described in chapter 11.

15.7 Relative Performance by Content Areas

In addition to performance in mathematics and science overall, it was of interest to see how Benchmarking participants and countries performed on the content areas relative to performance on the subject overall. Five content areas in mathematics and six in science were used in this analysis. Relative performance on the content areas was examined separately for the two subjects. The average across content area scores was computed for each jurisdiction, and then performance in each content area was shown as the difference between that average and the overall average. Confidence intervals were estimated for each difference.

In order to do this, TIMSS computed the vector of average proficiencies for each of the content areas on the test, and joined each vector to form a matrix called R_{ks}, where a row contained the average proficiency score for jurisdiction k on scale s for a specific subject. This R_{ks} matrix had also a "zeroth" row and column. The elements in r_{k0} contained the average of the elements on the k^{th}

row of the R_{ks} matrix. These were the jurisdiction averages across the content areas. The elements in r_{0s} contained the average of the elements of the s^{th} column of the R_{ks} matrix. These were the content area averages across all jurisdictions. The element r_{00} contained the overall average for the elements in vector r_{0j} or r_{k0}. Based on this information, the matrix I_{ks} was constructed in which the elements are computed as

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

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

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

To compute the sampling portion of the standard error, the vector of average proficiency was computed for each of the jurisdiction replicates for each content area in the test. For each jurisdiction and each content area, 75 replicates were created.[2] Each replicate was randomly reassigned to one of 75 sampling zones or replicates (h). These column vectors were then joined to form a new set of matrices each called R_{ks}^{h}, where a row contains the average proficiency for jurisdiction k in content area s for a specific subject, for the h^{th} international set of replicates. Each of these R_{ks}^{h} matrices has also a zeroth row and

[2] In countries and jurisdictions where there were fewer than 75 jackknife zones, 75 replicates were also created by assigning the overall mean to as many replicates as were necessary to have 75.

column. The elements in r^h_{k0} contain the average of the elements on the k^{th} row of the R^h_{ks} matrix. These are the jurisdiction averages across the content areas. The elements in r^h_{0s} contain the average of the elements of the s^{th} column of the R^h_{ks} matrix. These are the content area averages across all countries. The element r^h_{00} contains the overall average for the elements in vector r^h_{0j} or r^h_{k0}. Based on this information the set of matrices $R^h_{ks'}$ were constructed, in which the elements were computed as

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

The jackknife repeated replication (JRR) standard error is then given by the formula

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

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

The critical values were determined by adjusting the critical value for a two-tailed test, at the $\alpha = .05$ level of significance for multiple comparisons according the Dunn-Bonferroni procedure. The critical value for mathematics, with five content scales, was 2.5758, and for science, with six content scales, was 2.6383.

15.8 Percent Correct for Individual Items

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

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

References

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

Gonzalez, E.G., & Gregory, K.D. (2000). Reporting student achievement in mathematics and science. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS 1999 technical report.* Chestnut Hill, MA: Boston College.

Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., O'Connor, K.M., Chrostowski, S.J., Gregory, K.D., Smith, T.A., & Garden, R.A. (2001). *Science benchmarking report: TIMSS 1999—Eighth grade.* Chestnut Hill, MA: Boston College.

Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., Gregory, K.D., Smith, T.A., Chrostowski, S.J., Garden, R.A., & O'Connor, K.M. (2000). *TIMSS 1999 international science report.* Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., O'Connor, K.M., Chrostowski, S.J., Gregory, K.D., Garden, R.A., & Smith, T.A. (2001). *Mathematics benchmarking report: TIMSS 1999—Eighth grade.*Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., Gregory, K.D., Garden, R.A., O'Connor, K.M., Chrostowski, S.J., & Smith, T.A. (2000). *TIMSS 1999 international mathematics report.* Chestnut Hill, MA: Boston College.

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