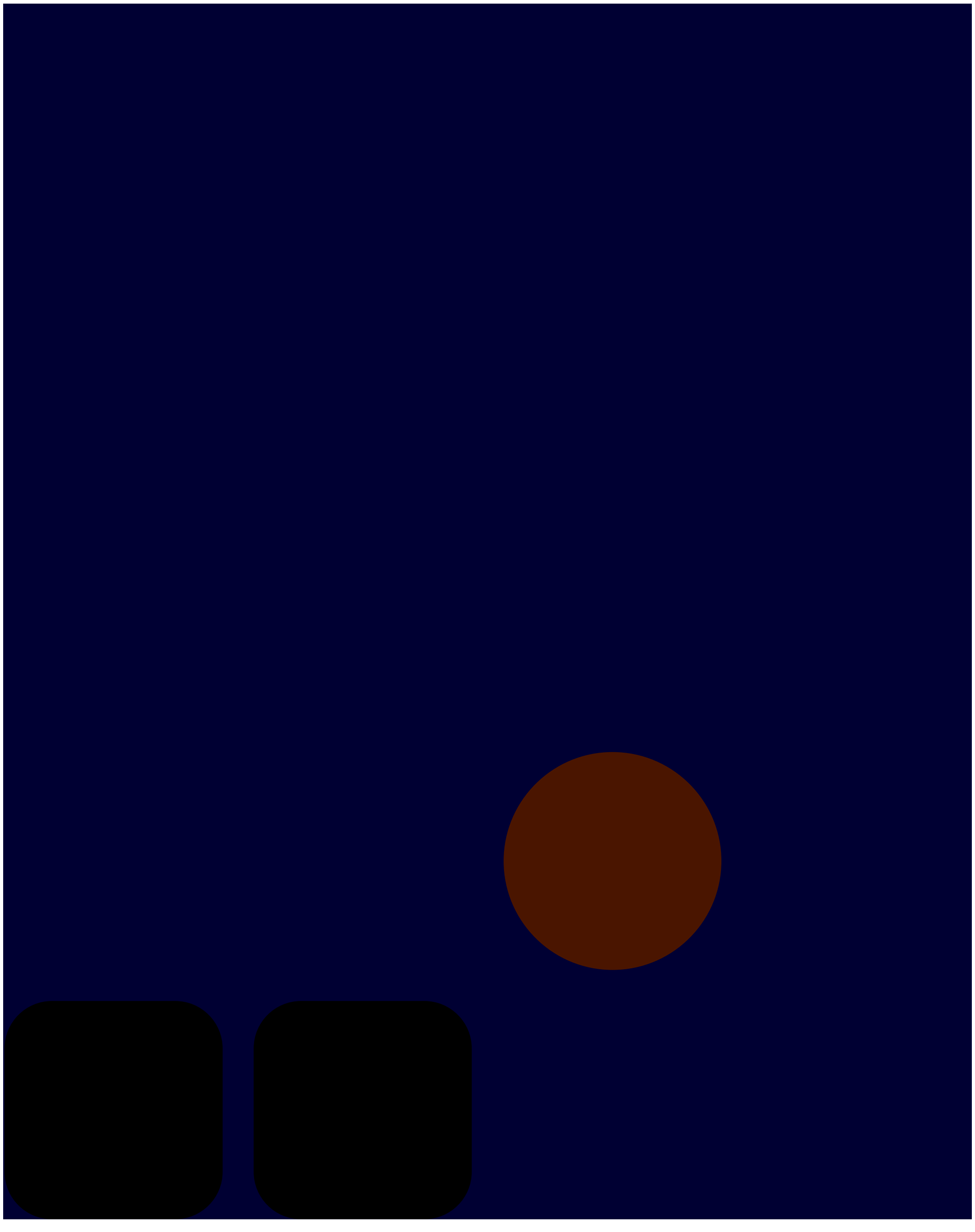

15 Reporting Student Achievement in Mathematics and Science for TIMSS 1999 Benchmarking[1]

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

$$X_{pvl}$$

where

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 α=

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		

ter 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.

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

[illegible]

[illegible]

1 [illegible]

( ) [illegible]

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

States	Top 10%	Upper Quarter

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 for 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

.

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 1s5 1 Tf0.6609 0 TD( )Tj/F4 1 Tfcolumn. The elements in contain the average of the elements on the k^{th}

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*

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*