Chapter 12

Reporting Student Achievement in Reading

Ann M. Kennedy and Kathleen L. Trong

12.1 Overview

The *PIRLS 2006 International Report* (Mullis, Martin, Kennedy, & Foy, 2007) presents a summary of reading achievement at the fourth grade in the 45 participating countries and provinces, as well as trends for those countries that also participated in PIRLS 2001. This chapter explains how the PIRLS International Benchmarks were established and the scale-anchoring process used to describe student achievement at each of these benchmarks. Additionally, the statistical procedures used to estimate the sampling and imputation variance that result from the PIRLS sampling and assessment design are described, as well as the methods used to calculate key statistics across countries.

12.2 PIRLS 2006 International Benchmarks of Student Achievement

As described in the previous chapter, substantial effort was put into creating the PIRLS reading achievement scale. To make full use of this information, it is essential that readers understand what scores on the scale mean. In other words, what skills did a student who scored 500 demonstrate? To facilitate this, the PIRLS International Benchmarks were created, and scale anchoring was used to describe student achievement at these points along the scale. The associated scale score for each benchmark is shown in Exhibit 12.1.

Exhibit 12.1 PIRLS 2006 International Benchmarks

Scale Score	International Benchmark
625	Advanced International Benchmark
550	High International Benchmark
475	Intermediate International Benchmark
400	Low International Benchmark

The PIRLS International Benchmarks are a set of unchanging points along the achievement scale that can be used to measure student achievement across countries and over time. It should be noted that the PIRLS 2006 International Benchmarks were established using procedures different from those in 2001. In PIRLS 2001, percentiles were used to determine benchmarks. That is, the points used to describe achievement were the Top 10 Percent (90^{th} percentile), Upper Quarter (75^{th} percentile), Median (50^{th} percentile), and Lower Quarter (25^{th} percentile). However, because benchmarks based on percentiles necessarily would be recalculated in each cycle according to the countries participating in that cycle, they would fluctuate as a greater range of countries participate in the future. To enable using the benchmarks to make comparisons across assessment cycles, the points need to be kept the same from cycle to cycle. Therefore, beginning in TIMSS 2003, permanent benchmarks were chosen for use with both IEA's TIMSS and PIRLS studies that were similar to those anchored in TIMSS 1999 for both mathematics and science (Gonzalez, Galia, Arora, Erberber, & Diaconu, 2004).

For reporting purposes, the 2006 benchmarks were applied to the 2001 data to allow for comparison across cycles. The permanent benchmarks are evenly distributed along the scale and are more dispersed than those in PIRLS 2001, with the 2006 benchmarks ranging from 400 (Low) to 625 (Advanced), whereas the 2001 benchmarks ranged from 435 (Lower Quarter) to 615 (Top 10 Percent). This greater breadth will be better able to capture the variance of achievement as more diverse countries participate in future assessments.

12.2.1 Identifying Students Achieving at Each Benchmark

Criteria were established for identifying students who scored at each of these International Benchmarks. As has been done in previous IEA studies, across all the PIRLS 2006 participants, all students scoring within +/- 5 score points of the benchmark were included in scale-anchoring analyses. This is done to create student groups that are large enough for analysis purposes, but small enough

that each benchmark remains clearly distinguished from the others. These ranges and the number of students scoring within each range in PIRLS 2006 are displayed in Exhibit 12.2.

Exhibit 12.2 Range Around Each International Benchmark and Number of Students Within Range

	Low International Benchmark 400	Intermediate International Benchmark 475	High International Benchmark 550	Advanced International Benchmark 625
Range of Scale Scores	395-405	470-480	545-555	620-630
Number of Students	2,681	6,484	10,360	4,844

12.2.2 Identifying Items Characterizing Achievement at Each Benchmark

Once the students achieving at each benchmark were identified, criteria were established to determine the items that these students were likely to answer correctly and that discriminate between the benchmarks (e.g., between the High and Advanced International Benchmarks). This allows for the development of descriptions of skills that students at each benchmark demonstrated through scale anchoring. To determine which items students at each anchor level were likely to answer correctly, the percent correct for those students was calculated for each item at each benchmark. For this analysis, students across the PIRLS 2006 participants were weighted so that students in each country contributed proportional to the size of the student population in that country.

For dichotomously scored items, the percent of students at each anchor point who answered each item correctly was computed. For constructed-response items with multiple score points (i.e., 2 or 3), each score level was treated separately because the different score levels may demonstrate different reading skills. For example, for a 2-point item, the percent of students at each anchor point earning only partial credit (1 point) was computed. In addition, the percent of students at each anchor point earning at least partial credit (1 or 2 points) was computed. This allowed the different score levels of an item to potentially anchor at different benchmarks.

Except at the Low International Benchmark, establishing criteria to identify items that were answered correctly by most students at the benchmark, but by fewer students at the next lower point, required considering achievement at adjacent benchmarks. For multiple-choice items, the criterion of 65 percent was used, since students would be likely (about two thirds of the time) to answer

the item correctly. The criterion of less than 50 percent was used for the next lower point, because this means that students were more likely to answer the item incorrectly than correctly. For example, if 65 percent of students scoring at the High International Benchmark answered a particular multiple-choice item correctly, but less than 50 percent of students at the Intermediate International Benchmark did so, this would be an anchor item for the High International Benchmark. For constructed-response items, a criterion of 50 percent was used, since there is no possibility of guessing to take into account, with no criterion for lower points.

Anchored Items

The criteria used to identify items that "anchored" at each of the four PIRLS 2006 International Benchmarks are outlined below.

An item anchored at the Low International Benchmark if:

- For a constructed-response item, at least 50 percent of students received either partial credit (e.g., at least 1 or at least 2 points, depending upon the maximum number of score points) or the full-credit score value (1, 2, or 3);
- For a multiple-choice item, at least 65 percent of students answered the item correctly. At the lowest level, only the 65 percent criterion is necessary, as there is no lower level from which to discriminate.

An item anchored at the Intermediate International Benchmark if:

- For a constructed-response item, at least 50 percent of students received at least partial or full credit;
- For a multiple-choice item at least 65 percent of students at the Intermediate International Benchmark, and less than 50 percent of students at the Low International Benchmark, answered the item correctly.

An item anchored at the High International Benchmark if:

- For a constructed-response item, at least 50 percent of students received at least partial or full credit;
- For a multiple-choice item, at least 65 percent of students at the High International Benchmark, and less than 50 percent of students at the Intermediate International Benchmark, answered the item correctly.

An item anchored at the Advanced International Benchmark if:

- For a constructed-response item, at least 50 percent of students received at least partial or full credit;
- For a multiple-choice item, at least 65 percent of students, and less than 50 percent of students at the High International Benchmark, answered the item correctly.

Almost Anchored Items

Not all items were assumed to be able to meet the anchoring criteria. Some items nearly met the 65 percent criterion, but did not discriminate between the anchor levels. Others discriminated well between anchor levels, but did not quite meet the 65 percent criterion.

The following criteria were established for those items nearly satisfying the anchoring criteria.

- An item "almost anchored" if more than 60 percent of students at a level answered an item correctly, and less than 50 percent of the students at the next lowest level answered correctly (the discrimination criterion is met).
- An item "anchored (only 60-65)" if more than 60 percent of students at a level answered an item correctly, but 50 percent or more students at the next lowest level answered correctly (the discrimination criterion is not met).

It is important to note that since there is no discrimination criterion for constructed-response items, the descriptions of the criteria for nearly meeting the anchoring requirements are for multiple-choice items only.

Items Too Difficult to Anchor

An item was too difficult to anchor if, for constructed-response items, less than 50 percent of students at the Advanced International Benchmark received at least partial or full credit, depending on the maximum score level for the item. For a multiple-choice item to be considered too difficult to anchor, less than 60 percent of students at the Advanced International Benchmark were able to answer correctly.

The results of the PIRLS 2006 scale anchoring of reading achievement are presented below in Exhibit 12.3. As this exhibit shows, considering items

that met the less stringent anchoring criteria added a substantial amount of information that could be used to describe student performance beyond what would have been available using only items that anchored.

Exhibit 12.3 Number of Items Anchoring at Each Benchmark

	Anchored	Almost Anchored	Met 60-65% Criterion	Total
Low (400)	9	4	0	13
Intermediate (475)	28	6	7	41
High (550)	53	5	9	67
Advanced (625)	28	0	6	34
			Too Difficult to Anchor	10
			Total	165

12.2.1 Expert Review of Anchor Items by Content Area

Once the empirical analysis identifying the items that anchored at each International Benchmark was completed, the items were reviewed by the PIRLS 2006 Reading Development Group (RDG), with the goal of developing descriptions of student performance. Members of the RDG were provided binders for each of the reading purposes, literary and informational, with their respective items grouped by benchmark and sorted by anchoring criteria. In other words, within the literary binder, there was a section for items that anchored at each benchmark, and in each section, the items that anchored appeared first, followed by those that almost anchored and those that met only the 60 to 65 percent criteria. For each item, the following information was displayed: item stem, answer key (for multiple-choice items), scoring guide for (constructed-response items), reading purpose, reading process, percent correct at each anchor point, overall international percent correct, and whether or not the item was released.

Using these materials, the descriptive portion of the scale anchoring analysis was conducted in Copenhagen, Denmark in April 2007. The task included developing a short description of the knowledge, understanding, or skills demonstrated by at least a partial-credit response for some constructed-response items, or by a full-credit response for a multiple-choice item or the maximum score level of a constructed-response item. Then, the item level descriptions for each International Benchmark were used to generalize and

draft a summary of the level of comprehension shown by students at each of the benchmarks. Following the meeting, the drafts were edited and presented in the international report. Additionally, example items that were selected to illustrate the benchmark descriptions were included in the international report.

Exhibit 12.4 presents the number of items (or point values, for multiple-point constructed-response items) that met one of the anchoring criteria for each benchmark, presented by reading purpose, as well as the number of items that were too difficult to anchor.

Exhibit 12.4 Number of Items Anchoring at Each Benchmark

	Low Benchmark	Intermediate Benchmark	High Benchmark	Advanced Benchmark	Too Difficult to Anchor	Total
Reading for Literary Purposes	5	24	37	15	3	84
Reading for Information	8	17	30	19	7	81

12.3 Capturing the Uncertainty in the PIRLS Student Achievement Measures

As discussed in previous chapters on sampling and scaling, PIRLS made extensive use of probability sampling techniques to sample students, and applied matrix sampling methods to administer a subset of the PIRLS 2006 assessment materials to each individual student. While this approach minimized the response burden to students, there is some variance or uncertainty in the statistics as a consequence. This uncertainty is measured and reported by providing an estimate of its standard error together with each statistic in the *PIRLS 2006 International Report* (Mullis, Martin, Kennedy, & Foy, 2007). For the achievement results, these standard errors reflect the uncertainty of the proficiency estimates due to two variance components—sampling variance and imputation variance.

12.3.1 Estimating Sampling Variance

There are several options for estimating sampling errors that take into account a complex sampling design, such as the stratified multistage cluster sampling applied in PIRLS 2006 (Brick, Morganstein, & Valliant, 2000). PIRLS uses a variation of the jackknife repeated replication (JRR) technique (Johnson & Rust, 1992) because it is computationally straightforward and provides approximately

unbiased estimates of the sampling errors of means, totals, and percentages. This technique assumes that the primary sampling units (PSUs) can be paired in a manner consistent with the sample design, with each pair belonging to a pseudo-stratum for variance estimation purposes. The JRR technique appropriately accounts for the combined effect of the between- and within-PSU contributions to the sampling variance.

The application of JRR involves pairing schools to sampling zones, and randomly selecting one of these schools to double its contribution and set the contribution of its paired school to zero, constructing a number of "pseudo-replicates" of the original sample. The statistic of interest is computed once for the original sample, and once again for each pseudo-replicate sample, with the variation between the estimates for each of the replicate samples and the original sample estimate being the jackknife estimate of the sampling error of the statistic.

12.3.2 Constructing Sampling Zones for Sampling Variance Estimation

Statistics Canada worked through the list of sampled schools for each PIRLS participating country and Canadian province to apply the JRR technique. Sampled schools were paired and assigned to a series of groups known as "sampling zones". Organized according to the order in which they were selected, the first and second sampled schools were assigned to the first sampling zone, the third and fourth schools to the second zone, and continuing through the list. In total, 75 zones were used, allowing for 150 schools per country. When more than 75 zones were constructed, they were collapsed to keep the total number to 75.

Sampling zones were constructed within design domains, or explicit strata. Where there was an odd number of schools in an explicit stratum, either by design or because of school nonresponse, the students in the remaining school were randomly divided to make up two "quasi" schools for the purpose of calculating the jackknife standard error. Each zone then consisted of a pair of schools or "quasi" schools. Exhibit 12.5 shows the range of sampling zones used in each country.

Exhibit 12.5 Number of Sampling Zones Used in PIRLS 2006 and PIRLS 2001

Countries	PIRLS 2006 Sampling Zones	PIRLS 2001 Sampling Zones
Austria	75	◊
Belgium (Flemish)	70	◊
Belgium (French)	75	◊
Bulgaria	74	75
Canada, Alberta	75	◊
Canada, British Columbia	74	◊
Canada, Nova Scotia	75	◊
Canada, Ontario	75	◊
Canada, Quebec	75	◊
Chinese Taipei	75	◊
Denmark	73	◊
England	75	66
France	75	73
Georgia	75	◊
Germany	75	75
Hong Kong SAR	74	74
Hungary	75	75
Iceland	75	75
Indonesia	75	◊
Iran, Islamic Rep. of	75	75
Israel	75	74
Italy	75	75
Kuwait	75	◊
Latvia	74	71
Lithuania	75	73
Luxembourg	75	◊
Macedonia, Rep. of	75	73
Moldova, Rep. of	75	75
Morocco	75	59
Netherlands	71	67
New Zealand	75	75
Norway	75	69
Poland	74	◊
Qatar	75	◊
Romania	75	73
Russian Federation	74	61
Scotland	66	59
Singapore	75	75
Slovak Republic	74	75
Slovenia	73	75
South Africa	75	◊
Spain	75	◊
Sweden	74	75
Trinidad and Tobago	75	◊
United States	47	52

A diamond (◊) indicates the country did not participate in the 2001 assessment.

12.3.3 Computing Sampling Variance Using the JRR Method

The JRR algorithm assumes that there are *H* sampling zones within each country, each containing two sampled schools selected independently. The equation to compute the JRR variance estimate of a statistic *t* from the sample for a country is as follows:

$$Var_{jrr}(t) = \sum_{h=1}^{H} \left[t(J_h) - t(S) \right]^2$$

where *H* is the number of pairs in the sample for the country. The term $t(S)$ corresponds to the statistic for the whole sample (computed with any specific weights that may have been used to compensate for the unequal probability of selection of the different elements in the sample or any other post-stratification weight). The element $t(J_h)$ denotes the same statistic using the h^{th} jackknife replicate. This is computed using all cases except those in the h^{th} zone of the sample. For those in the h^{th} zone, all cases associated with one of the randomly selected units of the pair are removed, and the elements associated with the other unit in the zone are included twice. In practice, this process is accomplished by recoding to zero the weights for the cases of the element of the pair to be excluded from the replication, and multiplying by two the weights of the remaining element within the h^{th} pair.

Therefore, in PIRLS 2006, the computation of the JRR variance estimate for any statistic required the computation of the statistic up to 76 times for any given country: once to obtain the statistic for the whole sample, and as many as 75 times to obtain the statistics for each of the jackknife replicates (J_h). The number of jackknife replicates for a given country depended on the number of implicit strata or sampling zones defined for that country.

Replicate weights used in calculations of statistics were created by doubling and zeroing the weights of the selected units within the sampling zones. Within a zone, one of the schools was randomly assigned an indicator (u_i), code of 1 or 0 so that one member of the pair was assigned a value of 1 on the variable u_i, and the other a value of 0. This indicator determines whether the weights for the elements in the school in this zone are to be doubled or zeroed.

The replicate weight $W_h^{g,i,j}$ for the elements in a school assigned to zone *h* is computed as the product of k_h times their overall sampling weight, where k_h can take values of 0, 1, or 2 depending on whether the school is to be omitted, be

included with its usual weight, or have its weight doubled for the computation of the statistic of interest.

To create replicate weights, each sampled student was first assigned a vector of 75 weights, $W_h^{g,i,j}$, where *h* takes values from 1 to 75. The value of $W_0^{g,i,j}$ is the overall sampling weight, which is the product of the final school weight, classroom weight, and student weight.

The replicate weights for a single case were then computed as

$$W_h^{g,i,j} = W_0^{g,i,j} \cdot k_{hi}$$

where the variable k_h for an individual *i* takes the value $k_{hi} = 2 * u_i$ if the record belongs to zone *h*, and $k_{hi} = 1$ otherwise.

The replicate weights were not included as permanent variables in the PIRLS 2006 international database. Instead, they were created temporarily for each analysis by the sampling variance estimation program. For each country, PIRLS computed 75 replicate weights regardless of the number of actual zones within the country. If a country had fewer than 75 zones, then the replicate weights W_h, where *h* was greater than the total number of zones, were equal to the overall sampling weight. While computing 75 replicate weights for each country had no effect on the size of the error variance computed using the jackknife formula, the process facilitated the computation of standard errors for a number of countries simultaneously.

12.3.4 Estimating Imputation Variance

As described in Chapter 2, a matrix-sampling test design was used such that an individual student was administered a single test booklet containing only a portion of the PIRLS 2006 assessment. Using the scaling techniques described in Chapter 11, the results were aggregated across all booklets to provide results for the entire assessment, and plausible values were generated as estimates of student performance on the assessment as a whole. The variability among these estimates, or imputation error, for each variable was combined with the sampling error for that variable, providing an appropriate standard error that incorporates both error components.

To compute the imputation variance for any estimable statistic, t_m (e.g., mean, difference between means, or percentiles), the statistic must first be calculated for each set of *M* plausible values, where $m = 1, 2, \ldots, 5$.[1]

Once the statistics are computed, the imputation variance is computed as:

$$Var_{imp} = \left(1 + 1/M\right) Var\left(t_1, \ldots, t_M\right)$$

where *M* is the number of plausible values used in the calculation, and $Var\left(t_1, \ldots, t_M\right)$ is the variance of the *M* estimates computed using each plausible value.

12.3.5 Combining Sampling and Imputation Variance

In reporting reading proficiency statistics, PIRLS presented all calculated statistics with their standard errors, which incorporate both sampling and imputation variance components. The standard errors were computed using the following formula:[2]

$$Var\left(t_{pv}\right) = Var_{jrr}\left(t_1\right) + Var_{imp}$$

where $Var_{jrr}\left(t_1\right)$ is the sampling variance for the first plausible value and *Varimp* is the imputation variance. The *PIRLS 2006 User Guide for the International Database* (Foy & Kennedy, 2008) includes programs, for both SAS and SPSS statistical packages, that compute each of these variance components for the PIRLS 2006 data.

12.4 Calculating National and International Statistics for Student Achievement

This section describes the procedures for computing the statistics used to summarize reading achievement in the *PIRLS 2006 International Report*, including mean achievement scale scores based on plausible values, gender differences in average achievement, and performance on example items.

1 The general procedure for estimating the imputation variance using plausible values is described in Mislevy, R.J., Beaton, A.E., Kaplan, B., & Sheehan, K.M. (1992).

2 With unlimited computing resources, computing the imputation variance for the plausible values and the JRR sampling variance for each of the plausible values (pv) (i.e., computing the same statistic as many as 380 times: once for each pv using the overall sampling weight, and then 75 times for each pv using the complete set of replicate weights) is ideal. An acceptable shortcut, however, is to compute the JRR variance component using one pv, and then the imputation variance using the five pv. Using this approach, a statistic would be computed only 80 times.

National averages were computed as the average of the weighted means for each of the five plausible values. The weighted mean for each plausible value was computed as follows:

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

where

- $\bar{X}_{pvl}$ is the country mean for plausible value l
- pv_{lj} is the l^{th} plausible value for the j^{th} student
- $W^{i,j}$ is the weight associated with the j^{th} student in class i, and
- N is the number of students in the country's sample.

Exhibits 12.6 through 12.10 provide basic summary statistics for reading achievement overall, as well as by purposes and processes. Each exhibit presents the student sample size, the mean achievement scale score and standard deviation, averaged across the five plausible values, the jackknife standard error for the mean, and the overall standard errors for the mean including imputation error.

12.4.1 Comparing Achievement Differences Across Countries

In reporting student achievement in the international report, PIRLS compares average performance of a participant with that of the other participants. Differences in mean achievement between countries are considered statistically significant if the absolute difference between them, divided by the standard error of the difference, is greater than the critical value. For differences between countries, which can be considered as independent samples, the standard error of the difference between means is 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. The means and standard errors used in the calculation of statistical significance for

reading achievement overall and by purposes and processes are presented in Exhibits 12.6-12.9.

The significance tests presented were not adjusted for multiple comparisons among countries. Although adjustments such as the Bonferroni procedure guard against misinterpreting the outcome of multiple simultaneous significance tests, and have been used in previous IEA studies, the results vary depending on the number of countries included in the adjustment, leading to apparently conflicting results from comparisons using different combinations of countries.

12.4.2 Comparing National Average Achievement to the PIRLS Scale Average

Several exhibits in the international report compare the mean achievement for a country with the PIRLS scale average (500, with no standard error), together with a test of the statistical significance of the difference. The standard error of the difference is equal to the standard error of the mean achievement score for the country.

12.4.3 Reporting Gender Differences Within Countries

Gender differences were reported in overall student achievement in reading, as well as in the reading purposes and processes scales. Gender differences were presented in an exhibit showing mean achievement for girls and boys and their differences, with an accompanying graph indicating whether the difference was statistically significant. Because in most countries males and females attend the same schools, the samples of males and females cannot be treated as independent samples for the purpose of statistical tests. Accordingly, PIRLS applied a jackknife procedure for correlated samples to estimate the standard errors of the differences. This procedure involved computing the average difference between boys and girls in each country once for each of the 75 replicate samples, and five more times, once for each plausible value, as described in the earlier section on estimating imputation variance.

Exhibit 12.6 Summary Statistics and Standard Errors in Overall Reading Achievement

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Austria	5,067	538.296	63.654	2.105	2.200
Belgium (Flemish)	4,479	547.044	55.622	1.866	1.964
Belgium (French)	4,552	499.666	68.585	2.590	2.640
Bulgaria	3,863	547.032	82.682	4.341	4.366
Chinese Taipei	4,589	535.371	64.143	1.928	2.040
Denmark	4,001	546.346	69.712	2.257	2.266
England	4,036	539.483	86.845	2.464	2.560
France	4,404	521.593	66.584	2.061	2.066
Georgia	4,402	470.836	74.877	3.075	3.138
Germany	7,899	547.591	66.977	2.094	2.175
Hong Kong SAR	4,712	563.911	59.327	2.337	2.354
Hungary	4,068	550.889	70.238	2.931	2.976
Iceland	3,673	510.597	68.107	1.125	1.289
Indonesia	4,774	404.737	78.616	4.039	4.074
Iran, Islamic Rep. of	5,411	420.933	94.685	3.044	3.088
Israel	3,908	512.462	98.825	3.345	3.348
Italy	3,581	551.468	67.854	2.882	2.932
Kuwait	3,958	330.300	110.751	3.632	4.216
Latvia	4,162	540.912	62.635	2.210	2.335
Lithuania	4,701	537.033	56.895	1.610	1.640
Luxembourg	5,101	557.195	66.405	0.873	1.084
Macedonia, Rep. of	4,002	442.395	101.330	3.940	4.089
Moldova, Rep. of	4,036	499.884	69.038	3.025	3.037
Morocco	3,249	322.580	109.139	5.797	5.938
Netherlands	4,156	547.152	53.026	1.458	1.520
New Zealand	6,256	531.715	86.948	1.974	2.016
Norway	3,837	498.008	66.601	2.442	2.553
Poland	4,854	519.389	75.250	2.205	2.356
Qatar	6,680	353.436	95.575	1.070	1.090
Romania	4,273	489.473	91.463	4.998	5.012
Russian Federation	4,720	564.744	68.744	3.301	3.355
Scotland	3,775	527.355	79.862	2.755	2.791
Singapore	6,390	558.273	76.658	2.835	2.883
Slovak Republic	5,380	530.815	74.164	2.732	2.755
Slovenia	5,337	521.531	70.721	2.072	2.087
South Africa	14,657	301.613	136.181	5.467	5.555
Spain	4,094	512.504	70.965	2.394	2.482
Sweden	4,394	549.282	63.642	2.168	2.280
Trinidad and Tobago	3,951	435.588	103.316	4.863	4.885
United States	5,190	539.925	74.063	3.541	3.549

Exhibit 12.7 Summary Statistics and Standard Errors in Reading Achievement for Literary Purposes

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Austria	5,067	537.074	62.275	1.999	2.112
Belgium (Flemish)	4,479	543.807	57.634	1.878	1.908
Belgium (French)	4,552	499.482	67.463	2.401	2.419
Bulgaria	3,863	542.150	83.832	4.483	4.513
Chinese Taipei	4,589	530.438	69.442	1.946	1.994
Denmark	4,001	547.387	68.435	2.212	2.626
England	4,036	538.707	89.363	2.493	2.605
France	4,404	516.297	65.632	2.000	2.405
Georgia	4,402	476.456	75.489	3.130	3.238
Germany	7,899	548.768	66.452	1.992	2.161
Hong Kong SAR	4,712	556.926	64.015	2.538	2.607
Hungary	4,068	556.761	70.087	2.861	2.928
Iceland	3,673	514.476	65.901	1.026	1.660
Indonesia	4,774	397.186	78.412	3.889	3.922
Iran, Islamic Rep. of	5,411	426.209	96.459	3.076	3.147
Israel	3,908	516.439	97.702	3.203	3.429
Italy	3,581	551.490	73.744	3.147	3.269
Kuwait	3,958	340.428	108.051	3.509	3.659
Latvia	4,162	539.283	63.419	2.085	2.386
Lithuania	4,701	541.633	58.441	1.771	1.933
Luxembourg	5,101	554.897	68.090	0.802	0.954
Macedonia, Rep. of	4,002	438.603	97.225	3.574	3.735
Moldova, Rep. of	4,036	492.228	68.133	2.621	2.814
Morocco	3,249	317.357	116.430	6.240	6.452
Netherlands	4,156	544.552	56.522	1.636	1.837
New Zealand	6,256	527.324	86.488	2.017	2.059
Norway	3,837	501.131	66.508	2.464	2.508
Poland	4,854	523.138	77.809	2.263	2.482
Qatar	6,680	358.373	96.300	1.026	1.255
Romania	4,273	493.009	91.085	4.806	4.840
Russian Federation	4,720	561.032	69.422	3.192	3.297
Scotland	3,775	526.900	81.191	2.464	2.575
Singapore	6,390	551.518	80.283	2.904	2.915
Slovak Republic	5,380	533.326	74.230	2.773	2.864
Slovenia	5,337	519.435	68.958	1.977	2.032
South Africa	14,657	299.431	134.651	5.150	5.249
Spain	4,094	516.423	75.241	2.632	2.694
Sweden	4,394	546.026	61.406	2.169	2.256
Trinidad and Tobago	3,951	434.137	103.793	4.586	4.631
United States	5,190	540.658	77.645	3.434	3.571

Exhibit 12.8 Summary Statistics and Standard Errors in Reading Achievement for Informational Purposes

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Austria	5,067	536.131	64.668	2.216	2.309
Belgium (Flemish)	4,479	547.126	53.133	1.763	2.036
Belgium (French)	4,552	497.958	67.894	2.617	2.785
Bulgaria	3,863	549.828	82.797	4.272	4.355
Chinese Taipei	4,589	538.261	58.521	1.693	1.815
Denmark	4,001	541.709	71.717	2.298	2.407
England	4,036	537.069	84.067	2.383	2.530
France	4,404	526.076	66.505	1.985	2.110
Georgia	4,402	465.178	77.053	3.324	3.552
Germany	7,899	544.445	66.448	2.142	2.265
Hong Kong SAR	4,712	568.232	55.924	2.215	2.250
Hungary	4,068	541.154	70.292	2.953	3.081
Iceland	3,673	505.181	71.493	1.194	1.383
Indonesia	4,774	417.685	82.163	4.151	4.165
Iran, Islamic Rep. of	5,411	419.796	90.683	3.023	3.122
Israel	3,908	507.409	98.601	3.437	3.619
Italy	3,581	548.937	64.080	2.727	2.934
Kuwait	3,958	326.510	117.862	4.044	4.296
Latvia	4,162	539.895	62.530	2.247	2.390
Lithuania	4,701	529.879	54.480	1.597	1.628
Luxembourg	5,101	556.644	63.982	0.794	0.971
Macedonia, Rep. of	4,002	449.857	102.559	3.996	4.174
Moldova, Rep. of	4,036	508.045	70.362	3.022	3.042
Morocco	3,249	334.506	104.921	5.869	6.020
Netherlands	4,156	547.557	49.555	1.323	1.594
New Zealand	6,256	533.516	83.737	2.083	2.234
Norway	3,837	494.263	68.335	2.642	2.754
Poland	4,854	515.055	72.322	1.992	2.191
Qatar	6,680	356.046	93.687	0.968	1.621
Romania	4,273	487.202	88.408	4.929	4.943
Russian Federation	4,720	563.774	65.985	3.270	3.346
Scotland	3,775	526.952	77.890	2.448	2.556
Singapore	6,390	563.166	70.399	2.665	2.832
Slovak Republic	5,380	526.803	72.807	2.513	2.644
Slovenia	5,337	522.956	70.774	2.175	2.390
South Africa	14,657	315.626	131.904	5.083	5.150
Spain	4,094	508.187	67.542	2.415	2.889
Sweden	4,394	548.617	67.171	2.229	2.351
Trinidad and Tobago	3,951	440.119	99.288	4.341	4.586
United States	5,190	537.164	69.909	3.298	3.440

Exhibit 12.9 Summary Statistics and Standard Errors in Reading Achievement for Retrieving and Straightforward Inferencing Processes

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Austria	5,067	544.012	65.937	2.002	2.087
Belgium (Flemish)	4,479	544.562	59.380	1.671	1.920
Belgium (French)	4,552	501.166	70.840	2.488	2.637
Bulgaria	3,863	537.648	78.864	4.144	4.233
Chinese Taipei	4,589	540.923	67.730	1.915	1.961
Denmark	4,001	550.980	78.411	2.532	2.691
England	4,036	533.309	90.704	2.477	2.841
France	4,404	523.467	67.235	1.964	2.098
Georgia	4,402	477.963	73.262	3.267	3.320
Germany	7,899	554.563	71.815	2.110	2.624
Hong Kong SAR	4,712	557.528	59.249	2.432	2.515
Hungary	4,068	543.514	69.262	2.623	2.781
Iceland	3,673	516.355	72.863	1.125	1.227
Indonesia	4,774	409.457	77.640	3.854	3.927
Iran, Islamic Rep. of	5,411	427.870	96.146	3.204	3.294
Israel	3,908	507.349	94.658	3.000	3.216
Italy	3,581	544.103	69.523	2.768	2.816
Kuwait	3,958	336.978	106.949	3.303	3.865
Latvia	4,162	534.034	64.956	2.317	2.462
Lithuania	4,701	531.073	60.179	1.715	1.899
Luxembourg	5,101	565.086	72.780	0.849	1.205
Macedonia, Rep. of	4,002	445.981	97.680	3.787	3.830
Moldova, Rep. of	4,036	485.985	68.782	2.820	2.870
Morocco	3,249	336.209	103.833	6.014	6.170
Netherlands	4,156	551.212	60.907	1.619	2.036
New Zealand	6,256	523.595	86.261	2.148	2.269
Norway	3,837	501.977	71.976	2.246	2.291
Poland	4,854	515.977	75.800	2.198	2.356
Qatar	6,680	360.581	94.470	0.963	1.202
Romania	4,273	488.843	88.819	5.114	5.203
Russian Federation	4,720	562.323	70.091	3.223	3.438
Scotland	3,775	524.682	81.700	2.569	2.810
Singapore	6,390	560.224	84.587	3.227	3.293
Slovak Republic	5,380	529.011	74.858	2.697	2.754
Slovenia	5,337	518.658	72.106	1.972	2.063
South Africa	14,657	306.569	130.940	5.163	5.322
Spain	4,094	508.235	69.074	2.484	2.515
Sweden	4,394	550.238	68.608	2.226	2.360
Trinidad and Tobago	3,951	438.496	102.661	4.596	4.708
United States	5,190	532.155	78.000	3.312	3.339

Exhibit 12.10 Summary Statistics and Standard Errors in Reading Achievement for Interpreting, Integrating, and Evaluating Processes

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Austria	5,067	530.060	64.056	2.115	2.213
Belgium (Flemish)	4,479	547.098	53.047	1.770	1.819
Belgium (French)	4,552	496.888	66.771	2.445	2.460
Bulgaria	3,863	552.640	83.558	4.409	4.428
Chinese Taipei	4,589	529.729	62.136	1.775	1.858
Denmark	4,001	542.249	62.359	2.004	2.326
England	4,036	543.082	81.437	2.248	2.450
France	4,404	517.834	66.101	2.135	2.291
Georgia	4,402	461.308	80.216	3.403	3.539
Germany	7,899	540.149	65.042	2.096	2.162
Hong Kong SAR	4,712	565.539	58.882	2.308	2.436
Hungary	4,068	553.827	67.512	2.718	2.990
Iceland	3,673	503.032	65.802	1.062	1.267
Indonesia	4,774	404.170	80.212	3.956	4.133
Iran, Islamic Rep. of	5,411	417.701	92.501	3.200	3.280
Israel	3,908	516.149	97.497	3.357	3.569
Italy	3,581	555.668	64.610	2.775	2.852
Kuwait	3,958	329.859	113.084	3.761	3.953
Latvia	4,162	545.200	58.113	1.863	1.882
Lithuania	4,701	540.190	53.081	1.590	1.635
Luxembourg	5,101	548.282	62.905	0.754	0.888
Macedonia, Rep. of	4,002	439.069	104.679	3.914	4.028
Moldova, Rep. of	4,036	515.334	67.212	2.789	2.919
Morocco	3,249	312.993	116.086	6.430	6.552
Netherlands	4,156	542.283	50.528	1.328	1.496
New Zealand	6,256	537.930	81.422	2.072	2.182
Norway	3,837	494.934	65.567	2.165	2.415
Poland	4,854	521.798	72.491	2.150	2.290
Qatar	6,680	355.309	92.402	0.908	1.553
Romania	4,273	490.000	90.988	5.228	5.321
Russian Federation	4,720	562.554	66.192	3.205	3.248
Scotland	3,775	528.473	76.794	2.455	2.561
Singapore	6,390	555.562	69.400	2.672	2.705
Slovak Republic	5,380	531.238	71.335	2.755	2.791
Slovenia	5,337	523.322	66.245	1.916	1.959
South Africa	14,657	313.039	130.433	5.143	5.284
Spain	4,094	515.320	71.554	2.571	2.615
Sweden	4,394	546.476	62.141	2.040	2.187
Trinidad and Tobago	3,951	436.522	100.312	4.720	5.032
United States	5,190	545.830	67.134	3.204	3.331

12.4.4 Reporting Student Performance on Individual Items

To describe the PIRLS International Benchmarks, PIRLS provides several examples of achievement items from the assessment together with the percentages of students in each country responding correctly to or earning partial or full credit on the items. The basis for calculating these percentages was the total number of students that were administered the item. For multiple-choice items, the weighted percentage of students that answered the item correctly was reported. For constructed-response items with more than one score level, it was the weighted percentage of students that achieved at least partial credit or full credit on the item. Omitted and not-reached items were treated as incorrect.

When the percent correct for example items was computed, student responses were classified in the following way.

For multiple-choice items, the responses to item *j* were classified as:

- Correct (C_j) when the correct option for an item was selected,
- Incorrect (W_j) when the incorrect option or no option at all was selected,
- Invalid (I_j) when two or more choices were made on the same question,
- 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 constructed-response items, student responses to item *j* were classified as:

- Correct (C_j) when the maximum number of points was obtained on the question,
- Incorrect (W_j) when the wrong answer or an answer not worth all the points in the question was given,
- Invalid (N_j) when the student's response was not legible or interpretable, or 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

Brick, J.M., Morganstein, D., & Valliant, R. (2000). *Analysis of complex sample data using replication.* Rockville, MD: Westat.

Foy, P. & Kennedy, A.M. (Eds). (2008). *PIRLS 2006 user guide for the international database.* Chestnut Hill, MA: Boston College.

Gonzalez, E.J., Galia, J., Arora, A, Erberber, E., & Diaconu, D. (2004). Reporting student achievement in mathematics and science. In M.O. Martin, I.V.S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 technical report* (pp. 275-307). Chestnut Hill, MA: Boston College.

Johnson, E.G., & Rust, K F. (1992). Population references and variance estimation for NAEP data. *Journal of Educational Statistics, 17,* 175-190.

Mislevy, R.J., Beaton, A.E., Kaplan, B., & Sheehan, K.M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29,* 133-161.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., & Foy, P. (2007). *PIRLS 2006 international report: IEA's progress in international reading literacy study in primary schools in 40 countries*. Chestnut Hill, MA: Boston College.