12

Statistical Analysis and Reporting of the PIRLS Data

Eugenio J. Gonzalez

Ann M. Kennedy

12.1 Overview

The *PIRLS 2001 International Report* (Mullis, Martin, Gonzalez, and Kennedy, 2003) summarizes fourth-grade students' reading achievement in each country. This chapter provides information about how important statistics in the report were computed, including their standard errors; describes how international benchmarks of achievement were established to facilitate reporting achievement, outlines the scale-anchoring procedure followed to describe performance at these benchmarks; and describes briefly the reporting of the information collected by questionnaire from the students and their parents, teachers, and school principals.

12.2 Estimation of Sampling and Imputation Variance

To obtain estimates of students' reading proficiency that were both accurate and cost-effective, PIRLS 2001 made extensive use of probability sampling techniques to sample students from national fourth-grade student populations, and of matrix sampling methods to target individual students with a subset of the entire set of assessment materials. Statistics computed from these student samples were used to estimate population parameters. This approach made an efficient use of resources, in particular keeping student response burden to a minimum, but at a cost of some variance or uncertainty in the statistics. To quantify this uncertainty, each statistic in the *PIRLS 2001 International Report* (Mullis, Martin, Gonzalez, and Kennedy, 2003) and in the trend study report, in *Trends in Children's Reading Literacy Achievement 1991–2001* (Martin, Mullis, Gonzalez, and Kennedy, 2003) is accompanied by an estimate of its standard error. These standard errors incorporate components reflecting the uncertainty due to generalizing from student samples to the entire fourth-grade student population (sampling variance), and to inferring students' performance on the entire assessment from their performance on the subset of items that they took (imputation variance).

12.2.1 Estimating Sampling Variance

The PIRLS 2001 sampling design applied a stratified multistage cluster-sampling technique to the problem of selecting efficient and accurate samples of students while working with schools and classes. This design capitalized on the structure of the student population (i.e., students grouped in classes within schools) to derive student samples that permitted efficient and economical data collection. Unfortunately, however, such a complex sampling design complicated the task of computing standard errors to quantify sampling variability.

When, as in PIRLS, the sampling design involves multistage cluster sampling, there are several options for estimating sampling errors that avoid the assumption of simple random sampling (Wolter, 1985). The jackknife repeated replication technique (JRR) was chosen by PIRLS because it is computationally straightforward and provides approximately unbiased estimates of the sampling errors of means, totals, and percentages.

The variation on the JRR technique used in PIRLS 2001 is described in Johnson and Rust (1992). It assumes that the primary sampling units (PSUs) can be paired in a manner consistent with the sample design, with each pair regarded as members of a pseudo-stratum for variance estimation purposes. When used in this way, the JRR technique appropriately accounts for the combined effect of the between- and within-PSU contributions to the sampling variance. The general use of JRR entails systematically assigning pairs of schools to sampling zones, and randomly selecting one of these schools to have its contribution doubled and the other to have its contribution zeroed, so as to construct a number of "pseudo-replicates" of the original sample. The statistic of interest is computed once for all of the original sample, and once again for each pseudo-replicate sample. The variation between the estimates for each of the replicate samples and the original sample estimate is the jackknife estimate of the sampling error of the statistic.

Exhibit 12.1: Number of Sampling Zones Used in Each Country

Country	PIRLS 2001 Sampling Zones	Trends in IEA's Reading Literacy Study	
		2001 Sampling Zones	1991 Sampling Zones
Argentina	69	.	.
Belize	60	.	.
Bulgaria	75	.	.
Canada	75	.	.
Colombia	74	.	.
Cyprus	75	.	.
Czech Republic	71	.	.
England	66	.	.
France	73	.	.
Germany	75	.	.
Greece	73	35	75
Hong Kong	74	.	.
Hungary	75	75	72
Iceland	75	33	75
Iran, Islamic Rep.	75	.	.
Israel	74	.	.
Italy	75	46	75
Kuwait	75	.	.
Latvia	71	.	.
Lithuania	73	.	.
Macedonia, Rep. of	73	.	.
Moldova, Rep. of	75	.	.
Morocco	59	.	.
Netherlands	67	.	.
New Zealand	75	37	75
Norway	69	.	.
Romania	73	.	.
Russian Federation	61	.	.
Scotland	59	.	.
Singapore	75	49	75
Slovak Republic	75	.	.
Slovenia	75	38	70
Sweden	75	75	62
Turkey	75	.	.
United States	52	35	33

Construction of Sampling Zones

To apply the JRR technique used in PIRLS 2001, the sampled schools are paired and assigned to a series of groups known as sampling zones. This was done at Statistics Canada by working through the list of sampled schools in the order in which they were selected and assigning the first and second schools to the first sampling zone, the third and fourth schools to the second zone, and so on. 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 purposes of calculating the jackknife standard error. Each zone then consisted of a pair of schools or "quasi" schools. Exhibit 12.1 shows the number of sampling zones used in each country.

Computing Sampling Variance Using the JRR Method

The JRR algorithm used in PIRLS 2001 assumes that there are H sampling zones within each country, each containing two sampled schools selected independently. To compute a statistic t from the sample for a country, the formula for the JRR variance estimate of the statistic t is then given by the following equation:

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

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 hth jackknife replicate. This is computed using all cases except those in the hth zone of the sample; for those in the hth 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 is effectively 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 hth pair.

The computation of the JRR variance estimate for any statistic in PIRLS 2001 required the computation of the statistic up to 76 times for any given country: once to obtain the statistic for the full sample, and up to 75 times to obtain the statistics for each of the jackknife replicates (J_h). The number of times a statistic needed to be computed for a given country depended on the number of implicit strata or sampling zones defined for that country.

Doubling and zeroing the weights of the selected units within the sampling zones was accomplished effectively by creating replicate weights that were then used in the calculations. This approach required the user to temporarily create a new set of weights for each pseudo-replicate sample. Each replicate weight is equal to k times the overall sampling weight, where k can take values of 0, 1, or 2 depending on whether the case is to be removed from the computation, left as it is, or have its weight doubled. The value of k for an individual student record for a given replicate depends on the assignment of the record to the specific PSU and zone.

Within each zone, the members of the pair of schools are assigned an indicator (u_i), coded randomly to 1 or 0 so that one of them has 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 weights $\left(W_h^{g,i,j}\right)$ 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. In PIRLS 2001, the replicate weights were not permanent variables, but were created temporarily by the sampling variance estimation program as a useful computing device.

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 $W_o^{g,i,j}$ is the overall sampling weight, which is simply the product of the final school weight, the appropriate final classroom weight, and the appropriate final student weight, as described in Chapter 9.

The replicate weights for a single case were then computed as:

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

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.

In the PIRLS 2001 analysis, 75 replicate weights were computed for each country 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 number of zones within the country, were each the same as the overall sampling weight. Although this involved some redundant computation, having 75 replicate weights for each country had no effect on the size of the error variance computed using the jackknife formula, but it facilitated the computation of standard errors for a number of countries at a time.

Standard errors presented in the international reports were computed using SAS programs developed at the PIRLS International Study Center. As a quality control check, results were verified using the WesVarPC software (Westat, 1997).

12.2.2 Estimating Imputation Variance

The PIRLS 2001 item pool was far too extensive to be administered in its entirety to any one student, and so a matrix-sampling test design was developed whereby each student was given a single test booklet containing only a part of the entire assessment.[1] The results for all of the booklets were then aggregated using item response theory to provide results for the entire assessment. Since each student responded to a subset of the assessment items, multiple imputation (the generation of "plausible values") was used to derive reliable estimates of student performance on the assessment as a whole.[2] Since every student proficiency estimate incorporates some uncertainty, PIRLS followed the customary procedure of generating five estimates for each student and using the variability among them as a measure of this imputation uncertainty, or error. In the PIRLS 2001 international report the imputation error for each variable has been combined with the sampling error for that variable to provide a standard error incorporating both.

[1] Details of the PIRLS test design my be found in Chapter 3.

[2] See Chapter 11 for details of the methodology used in scaling the PIRLS 2001 data.

The general procedure for estimating the imputation variance using plausible values is the following (Mislevy et al., 1992). First compute the statistic (t) for each set of plausible values (M). The statistics $t_{m,}$ where m = 1, 2, …, 5, can be anything estimable from the data, such as a mean, the difference between means, percentiles, and so forth. Each of these statistics will be called t_m.

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

$$Var_{imp} = \left(1 + \frac{1}{M}\right) Var\left(t_m\right)$$

where M is the number of plausible values used in the calculation, and $Var(t_m)$ is the variance of the estimates computed using each plausible value.

12.2.3 Combining Sampling and Imputation Variance

The standard errors of the reading proficiency statistics reported by PIRLS include both sampling and imputation variance components. The standard errors were computed using the following formula:[3]

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

where $Var_{jrr}(t_1)$ is the sampling variance for the first plausible value and Var_{imp} in the imputation variance. The forthcoming User Guide for the PIRLS 2001 International Database contains programs in SAS and SPSS that compute each of these variance components for the PIRLS 2001 data.

Exhibits 12.2 through 12.4 show basic summary statistics for reading achievement in the PIRLS 2001 assessment, for reading overall, as well as for reading for literary and informational purposes. Each exhibit presents the student sample size, the mean 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. Exhibits 12.5 through 12.8 provide comparable statistics for the 1991 and 2001 data from IEA's trends in reading literacy study.

[3] Under ideal circumstances and with unlimited computing resources, the imputation variance for the plausible values and the JRR sampling variance for each of the plausible values would be computed. This would be equivalent to computing the same statistic up to 380 times (once overall for each of the five plausible values using the overall sampling weights, and then 75 times more for each plausible value using the complete set of replicate weights). An acceptable shortcut, however, is to compute the JRR variance component using one plausible value, and then the imputation variance using the five plausible values. Using this approach, a statistic needs to be computed only 80 times.

Exhibit 12.2: Summary Statistics and Standard Errors for PIRLS 2001 Overall Reading Achievement

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Argentina	3300	420	96	5.8	5.9
Belize	2909	327	106	4.6	4.7
Bulgaria	3460	550	83	3.8	3.8
Canada (O,Q)	8253	544	72	2.3	2.4
Colombia	5131	422	81	4.4	4.4
Cyprus	3001	494	81	2.8	3.0
Czech Republic	3022	537	65	2.3	2.3
England	3156	553	87	3.3	3.4
France	3538	525	70	2.3	2.4
Germany	7633	539	67	1.9	1.9
Greece	2494	524	73	3.5	3.5
Hong Kong, SAR	5050	528	63	3.1	3.1
Hungary	4666	543	66	2.1	2.2
Iceland	3676	512	75	1.1	1.2
Iran, Islamic Rep. of	7430	414	92	4.1	4.2
Israel	3973	509	94	2.7	2.8
Italy	3502	541	71	2.3	2.4
Kuwait	7126	396	89	4.2	4.3
Latvia	3019	545	62	2.1	2.3
Lithuania	2567	543	64	2.4	2.6
Macedonia, Rep. of	3711	442	103	4.6	4.6
Moldova, Rep. of	3533	492	75	4.0	4.0
Morocco	3153	350	115	9.6	9.7
Netherlands	4112	554	57	2.5	2.5
New Zealand	2488	529	93	3.5	3.6
Norway	3459	499	81	2.9	2.9
Romania	3625	512	90	4.6	4.6
Russian Federation	4093	528	66	4.4	4.4
Scotland	2717	528	84	3.6	3.6
Singapore	7002	528	92	5.1	5.2
Slovak Republic	3807	518	70	2.7	2.8
Slovenia	2952	502	72	1.9	2.0
Sweden	6044	561	66	2.1	2.2
Turkey	5125	449	86	3.5	3.5
United States	3763	542	83	3.8	3.8

Exhibit 12.3: Summary Statistics and Standard Errors for PIRLS 2001 Reading for Literary Experience

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Argentina	3300	419	97	5.8	5.8
Belize	2909	330	103	4.7	4.9
Bulgaria	3460	550	86	3.9	3.9
Canada (O,Q)	8253	545	75	2.5	2.6
Colombia	5131	425	79	4.2	4.2
Cyprus	3001	498	80	2.5	2.5
Czech Republic	3022	535	63	2.2	2.3
England	3156	559	94	3.6	3.9
France	3538	518	71	2.5	2.6
Germany	7633	537	66	1.9	1.9
Greece	2494	528	74	3.3	3.3
Hong Kong, SAR	5050	518	66	3.0	3.1
Hungary	4666	548	65	2.0	2.0
Iceland	3676	520	69	1.1	1.3
Iran, Islamic Rep. of	7430	421	91	4.4	4.5
Israel	3973	510	95	2.5	2.6
Italy	3502	543	76	2.4	2.7
Kuwait	7126	394	85	3.8	3.8
Latvia	3019	537	59	1.9	2.2
Lithuania	2567	546	68	2.6	3.1
Macedonia, Rep. of	3711	441	97	4.4	4.5
Moldova, Rep. of	3533	480	72	3.7	3.7
Morocco	3153	347	106	8.3	8.4
Netherlands	4112	552	58	2.4	2.5
New Zealand	2488	531	96	3.8	3.9
Norway	3459	506	84	2.6	2.8
Romania	3625	512	88	4.6	4.7
Russian Federation	4093	523	68	3.9	3.9
Scotland	2717	529	88	3.5	3.5
Singapore	7002	528	98	5.5	5.6
Slovak Republic	3807	512	68	2.4	2.6
Slovenia	2952	499	68	1.8	1.8
Sweden	6044	559	64	2.2	2.4
Turkey	5125	448	86	3.3	3.4
United States	3763	550	88	3.8	3.8

Exhibit 12.4: Summary Statistics and Standard Errors for PIRLS 2001 Reading to Acquire and Use Information

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Argentina	3300	422	99	5.4	5.4
Belize	2909	332	109	4.9	4.9
Bulgaria	3460	551	81	3.4	3.6
Canada (O,Q)	8253	541	71	2.3	2.4
Colombia	5131	424	83	4.2	4.3
Cyprus	3001	490	83	2.9	3.0
Czech Republic	3022	536	68	2.5	2.7
England	3156	546	82	3.4	3.6
France	3538	533	71	2.4	2.5
Germany	7633	538	68	1.8	1.9
Greece	2494	521	75	3.7	3.7
Hong Kong, SAR	5050	537	59	2.8	2.9
Hungary	4666	537	68	2.2	2.2
Iceland	3676	504	84	1.2	1.5
Iran, Islamic Rep. of	7430	408	97	4.6	4.6
Israel	3973	507	93	2.8	2.9
Italy	3502	536	69	2.3	2.4
Kuwait	7126	403	97	4.5	4.5
Latvia	3019	547	64	2.2	2.3
Lithuania	2567	540	64	2.5	2.7
Macedonia, Rep. of	3711	445	108	5.1	5.2
Moldova, Rep. of	3533	505	81	4.5	4.7
Morocco	3153	358	125	10.8	10.9
Netherlands	4112	553	58	2.4	2.6
New Zealand	2488	525	89	3.5	3.8
Norway	3459	492	81	2.8	2.8
Romania	3625	512	90	4.6	4.6
Russian Federation	4093	531	68	4.3	4.3
Scotland	2717	527	82	3.4	3.6
Singapore	7002	527	83	4.8	4.8
Slovak Republic	3807	522	71	2.7	2.7
Slovenia	2952	503	75	1.8	1.9
Sweden	6044	559	68	2.1	2.2
Turkey	5125	452	90	3.8	3.8
United States	3763	533	79	3.5	3.7

Exhibit 12.5: Summary Statistics and Standard Errors for IEA's Trends in Reading Literacy Study – Overall Reading Achievement

Country	2001				
	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Greece	1109	507	91	5.8	5.9
Hungary	4707	475	97	3.8	3.9
Iceland	1797	513	94	3.3	3.5
Italy	1590	513	92	4.4	4.4
New Zealand	1188	502	111	5.2	5.3
Singapore	3601	489	106	7.9	8.0
Slovenia	1502	493	91	3.7	3.7
Sweden	5361	498	115	3.8	3.9
United States	1826	511	94	6.3	6.3

Country	1991				
	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Greece	3516	466	96	4.5	4.5
Hungary	3009	459	93	3.9	4.0
Iceland	3961	486	104	1.4	1.5
Italy	2221	500	101	5.3	5.4
New Zealand	3016	498	110	4.1	4.1
Singapore	7326	481	88	3.5	3.6
Slovenia	3297	458	96	3.2	3.2
Sweden	4301	513	116	4.2	4.2
United States	6433	521	90	3.2	3.2

Exhibit 12.6: Summary Statistics and Standard Errors for IEA's Trends in Reading Literacy Study – Reading Narrative Texts

Country	2001				
	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Greece	1109	513	88	4.7	4.8
Hungary	4707	479	85	3.1	3.1
Iceland	1797	524	100	3.2	3.3
Italy	1590	517	88	3.9	4.1
New Zealand	1188	496	114	5.3	5.3
Singapore	3601	487	113	8.6	8.6
Slovenia	1502	490	88	3.4	3.7
Sweden	5361	496	104	3.2	3.6
United States	1826	498	105	6.6	6.8

Country	1991				
	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Greece	3516	479	87	3.6	3.7
Hungary	3009	467	81	3.1	3.2
Iceland	3961	493	98	1.4	1.6
Italy	2221	507	91	4.6	4.7
New Zealand	3016	500	111	4.2	4.3
Singapore	7326	486	94	3.5	3.5
Slovenia	3297	465	90	2.9	3.0
Sweden	4301	513	100	3.3	3.4
United States	6433	518	101	3.2	3.3

Exhibit 12.7: Summary Statistics and Standard Errors for IEA's Trends in Reading Literacy Study – Reading Expository Texts

Country	2001				
	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Greece	1109	509	91	5.1	5.2
Hungary	4707	464	111	4.3	4.4
Iceland	1797	502	97	3.1	3.3
Italy	1590	513	99	4.4	4.5
New Zealand	1188	510	101	5.2	5.3
Singapore	3601	495	91	6.5	6.6
Slovenia	1502	489	92	3.1	3.3
Sweden	5361	496	121	4.0	4.1
United States	1826	521	80	5.3	5.4

Country	1991				
	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Greece	3516	476	95	4.3	4.3
Hungary	3009	443	115	4.7	4.8
Iceland	3961	483	116	1.8	1.9
Italy	2221	507	103	5.3	5.5
New Zealand	3016	502	102	3.8	3.9
Singapore	7326	489	78	3.0	3.1
Slovenia	3297	455	101	3.6	3.6
Sweden	4301	519	130	4.3	4.4
United States	6433	516	82	3.1	3.2

Exhibit 12.8: Summary Statistics and Standard Errors for IEA's Trends in Reading Literacy Study – Reading Documents

Country	2001				
	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Greece	1109	490	92	5.1	5.2
Hungary	4707	486	102	3.7	3.7
Iceland	1797	506	89	3.2	3.4
Italy	1590	499	93	4.4	4.5
New Zealand	1188	506	113	4.9	5.2
Singapore	3601	484	96	6.8	6.8
Slovenia	1502	502	92	3.6	3.8
Sweden	5361	506	122	3.9	4.4
United States	1826	520	90	5.9	6.1

Country	1991				
	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Greece	3516	443	95	4.6	4.9
Hungary	3009	468	97	4.1	4.3
Iceland	3961	479	96	1.4	1.7
Italy	2221	482	104	5.3	5.4
New Zealand	3016	491	102	3.9	4.0
Singapore	7326	465	76	3.0	3.1
Slovenia	3297	456	94	2.8	3.0
Sweden	4301	504	120	4.4	4.5
United States	6433	527	82	2.8	3.2

12.3 Reporting Student Achievement in Reading

As described in earlier chapters, PIRLS made extensive use of imputed proficiency scores to report student achievement in reading, for each of the two reading purposes – reading for literary experience and to acquire and use information – and for reading overall. This section describes the procedures followed in computing the principal statistics used to summarize achievement in the *PIRLS 2001 International Report* (Mullis, Martin, Gonzalez, & Kennedy, 2003), including country means based on plausible values, international benchmarks of achievement, gender differences, and performance on example items. It also presents means and standard errors for the nine countries participating in the Trends in IEA's Reading Literacy Study (Martin, Mullis, Gonzalez, & Kennedy, 2003).

For each of the PIRLS reading scales, reading overall and literary and informational reading, the item response theory (IRT) scaling procedure described in Chapter 11 yields five imputed scores or plausible values for every student. The difference between the five values reflects the degree of uncertainty in the imputation process. Where the process yields consistent results, the differences between the five values is very small. To obtain the best estimate for each of the PIRLS statistics, each one was computed five times, using each of the five plausible values in turn, and the results averaged to derive the reported value. The standard errors that accompany each reported statistic include two components: one quantifying sampling error and the other quantifying imputation error, as described in the previous section.

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:

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

where:

$\overline{X}_{pvl}$ is the country mean for plausible value *l*

$pv_{l,j}$ 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*,

N is the number of students in the country's sample.

These five weighted means were then averaged to obtain the national average for each country. To provide a reference point for comparison purposes, PIRLS presented the international average of many of the national statistics (means and percentages). International averages were calculated by first computing the national average for each plausible value for each country and then averaging across countries. These five estimates were then averaged to derive the international average presented in the PIRLS reports, as shown below:

$$\overline{X}_{pvl} = \frac{\sum_{k=1}^{N} \overline{X}_{pvl,k}}{K}$$

where

$\overline{X}_{pvl}$ is the international mean for plausible value *l*

$\overline{X}_{pvl,k}$ is the *k*-th country mean for plausible value *l*

K is the number of countries.

12.3.1 Achievement Differences Across Countries

A basic aim of the PIRLS 2001 international report is to provide fair and accurate comparisons of student achievement across the participating countries. Most of the exhibits in the PIRLS reports summarize student achievement by means of 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 countries, standard errors can be used to assess the statistical significance of the difference between the summary statistics.

The multiple comparison charts presented in the PIRLS 2001 international report facilitate the comparison of the average performance of a country with that of other participating countries. Reading achievement means were considered significantly different if the absolute difference between them, divided by the standard error of the difference, was greater than the critical value of 1.96, corresponding to a test of significance with 95% confidence. For differences between countries, 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. Exhibit 12.9 shows the PIRLS 2001 means and standard errors used in the calculation of statistical significance for the PIRLS international report.

The significance tests reported in these charts have NOT been adjusted for multiple comparisons. 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,[4] the results vary depending on the number of countries

[4] See Gonzalez and Gregory (2000) for a description of the Bonferroni procedure applied to IEA's TIMSS 1999 study.

Exhibit 12.9: Means and Standard Errors for International Comparisons – PIRLS 2001

Country	Overall Reading		Literary		Information	
	Mean	S.E.	Mean	S.E.	Mean	S.E.
Argentina	419.527	5.935	419.187	5.792	422.417	5.448
Belize	326.829	4.697	329.596	4.853	332.175	4.947
Bulgaria	550.498	3.847	549.542	3.866	551.310	3.573
Canada (O,Q)	544.146	2.377	544.567	2.609	541.300	2.449
Colombia	422.428	4.447	425.326	4.248	423.629	4.283
Cyprus	493.976	2.982	498.129	2.532	489.898	2.970
Czech Republic	536.883	2.321	535.287	2.335	536.399	2.680
England	552.878	3.394	559.177	3.883	545.624	3.557
France	525.170	2.367	518.149	2.642	533.133	2.537
Germany	539.090	1.935	536.515	1.942	538.181	1.949
Greece	524.167	3.487	527.640	3.345	520.986	3.707
Hong Kong, SAR	527.871	3.079	517.553	3.063	537.238	2.933
Hungary	543.226	2.199	548.462	2.031	537.273	2.199
Iceland	512.417	1.199	520.071	1.307	504.089	1.467
Iran, Islamic Rep. of	413.833	4.182	420.843	4.470	408.398	4.642
Israel	508.939	2.835	510.049	2.598	506.763	2.880
Italy	540.729	2.352	543.101	2.697	536.155	2.357
Kuwait	396.471	4.295	393.803	3.824	403.247	4.542
Latvia	544.607	2.284	537.206	2.177	546.946	2.345
Lithuania	543.387	2.589	545.518	3.086	539.544	2.677
Macedonia, Rep. of	441.586	4.610	441.477	4.457	445.321	5.200
Moldova, Rep. of	491.743	3.967	479.938	3.703	504.888	4.688
Morocco	349.511	9.650	347.148	8.352	358.014	10.855
Netherlands	554.209	2.497	552.285	2.494	552.834	2.621
New Zealand	528.824	3.563	531.368	3.880	524.857	3.825
Norway	499.179	2.922	505.703	2.750	492.133	2.836
Romania	511.710	4.589	511.822	4.727	512.424	4.598
Russian Federation	527.933	4.432	523.490	3.870	531.450	4.323
Scotland	528.176	3.601	529.097	3.543	527.033	3.605
Singapore	527.948	5.156	528.483	5.565	527.356	4.803
Slovak Republic	518.087	2.846	512.119	2.581	522.135	2.709
Slovenia	501.518	1.966	499.358	1.816	503.123	1.924
Sweden	561.014	2.218	559.403	2.383	558.605	2.212
Turkey	449.354	3.537	448.186	3.377	451.811	3.797
United States	542.149	3.817	550.408	3.812	533.325	3.655

included in the adjustment, leading to apparently conflicting results from comparisons using different numbers of countries.

12.3.2 Comparing Achievement with the International Mean

Many of the data exhibits in the PIRLS 2001 international reports show countries' mean achievement compared with the international mean, together with a test of the statistical significance of the difference between the two. These significance tests are based on the standard errors of the national and international means.

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

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

where

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

K 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

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

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

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

$$se_{dif\_j} = \sqrt{se_{i\_dif\_j}^2 + se_{s\_dif\_j}^2}$$

12.3.3 International Benchmarks of Reading Achievement

In order to provide information about the range of fourth-grade student reading achievement, PIRLS identified four points on the overall reading scale for use as international benchmarks, and reported the percentage of students reaching these benchmarks in each country. These four points correspond to the 90th, 75th, 50th, and 25th international percentiles of students achievement. The Top 10 percent Benchmark was defined as the 90th percentile on the PIRLS reading scale, computed across all students in all participating countries, with countries weighted in proportion to the size of their fourth-grade population. This point on the scale is the point above which the top 10 percent of students in the 2001 PIRLS assessment

Exhibit 12.10: Sample Size and Estimated Fourth-grade* Enrollment

Country	2001	
	Sample Size	Estimated Enrollment
Argentina	3300	709193
Belize	2909	7408
Bulgaria	3460	95702
Canada (O,Q)	8253	222012
Colombia	5131	975170
Cyprus	3001	10206
Czech Republic	3022	123831
England	3156	592787
France	3538	717378
Germany	7633	899014
Greece	2494	97288
Hong Kong, SAR	5050	88645
Hungary	4666	117238
Iceland	3676	4456
Iran, Islamic Rep. of	7430	1812810
Israel	3973	85802
Italy	3502	573318
Kuwait	7126	22318
Latvia	3019	34213
Lithuania	2567	43094
Macedonia, Rep. of	3711	27365
Moldova, Rep. of	3533	60634
Morocco	3153	554573
Netherlands	4112	181387
New Zealand	2488	58122
Norway	3459	58174
Romania	3625	283340
Russian Federation	4093	1823855
Scotland	2717	64375
Singapore	7002	49301
Slovak Republic	3807	71409
Slovenia	2952	21066
Sweden	6044	118134
Turkey	5125	977316
United States	3763	3802557

* Fourth-grade in most countries.

scored. If student reading achievement was distributed in the same way across all countries, approximately 10 percent of students within each country would be above the 90th percentile in the international distribution, regardless of the country's population size. Similarly, the upper quarter benchmark is the 75th percentile on the scale, above which the top 25 percent of students scored; the median benchmark is the 50th percentile, above which the top half of students scored; and the Lower Quarter Benchmark is the 25th percentile, the point reached by the top 75 percent of students.

In computing these benchmarks, the data were weighted so that each country contributed as many students to the analysis as it had students in the target population. In other words, each country's contribution to determining the international benchmarks was proportional to the estimated size of its fourth-grade population. Exhibit 12.10 shows the weighted number of students (estimated enrollment) each country contributed to the estimation of the international benchmarks.

The percentiles corresponding to the international benchmarks were computed separately for each of the five plausible values and the results averaged to arrive at the final figure. The international benchmarks are presented in Exhibit 12.11.

Exhibit 12.11: International Benchmarks of Fourth-grade Reading Achievement

	90th Percentile	75th Percentile	50th Percentile	25th Percentile
Reading Scale Score	615	570	510	435

* Fourth-grade in most countries.

12.3.4 Gender Differences

PIRLS reported gender differences in student achievement in reading overall, as well as in the two reading purposes. Gender differences were presented in an exhibit showing the percentages of males and females and their mean reading achievement in each country, together with an indication of whether the male-female difference in reading achievement 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 for the purpose of statistical significance testing. Accordingly, PIRLS used a jackknife procedure applicable to correlated samples for estimating the standard error of the male-female difference. This involved computing the average difference between boys and girls once for each of the 75 replicate samples, and five more times, once for each plausible value, as described earlier in this chapter.

12.3.5 Reporting Student Performance on Individual Items

To portray student achievement as fully as possible, the PIRLS 2001 international report presents many examples of the items used in the assessment, together with the percentage of students in each country responding correctly to or earning partial credit on each item. The base of this percentage was the total number of students tested on an 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, the weighted percentage of students that achieved partial or full credit was reported. Omitted and not reached items were treated as incorrect.

When computing the percent correct for individual example items, 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 a particular score level of a constructed-response item, student responses to item j were classified as correct (C_j) when the corresponding number of points was obtained; incorrect (W_j) when the wrong answer or an answer worth less than the maximum points 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

Exhibit 12.12: Means and Standard Errors for International Comparisons – IEA's Trends in Reading Literacy Study 1991–2001

Country	Overall Reading				Narrative Text			
	2001		1991		2001		1991	
	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
Greece	507.020	5.875	466.270	4.501	512.667	4.795	478.800	3.684
Hungary	475.099	3.887	459.061	4.005	479.152	3.130	467.334	3.228
Iceland	512.898	3.523	485.921	1.534	523.860	3.337	492.789	1.627
Italy	512.607	4.417	500.461	5.368	517.126	4.084	507.407	4.650
New Zealand	502.130	5.322	498.397	4.144	495.541	5.341	500.226	4.309
Singapore	488.500	7.950	480.629	3.565	487.209	8.631	486.334	3.525
Slovenia	493.407	3.702	457.673	3.191	490.279	3.661	465.302	3.023
Sweden	497.703	3.879	512.965	4.204	496.234	3.574	513.344	3.409
United States	510.636	6.320	520.839	3.249	497.934	6.834	517.813	3.317

Country	Expository Text				Documents			
	2001		1991		2001		1991	
	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
Greece	509.115	5.171	476.493	4.307	490.396	5.220	442.707	4.853
Hungary	464.450	4.357	443.036	4.807	486.078	3.709	467.601	4.281
Iceland	501.637	3.301	483.479	1.889	506.258	3.411	478.690	1.698
Italy	513.056	4.487	506.798	5.530	498.935	4.458	481.980	5.392
New Zealand	510.497	5.256	502.431	3.903	506.243	5.168	490.688	4.034
Singapore	495.489	6.550	489.406	3.111	483.681	6.798	465.439	3.100
Slovenia	488.913	3.285	455.105	3.635	502.402	3.794	455.651	2.968
Sweden	496.238	4.063	518.965	4.436	506.343	4.354	504.007	4.532
United States	520.605	5.398	515.738	3.211	519.664	6.122	526.849	3.157

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.

12.3.6 Trends in Achievement on the IEA Reading Literacy Test 1991–2001

The Trends in IEA's Reading Literacy Study was designed to describe changes in performance from 1991 to 2001 on IEA's 1991 reading literacy test. Nine of the PIRLS countries that participated in 1991 took part in the study. Exhibit 12.12 presents average achievement for the nine participating countries in 1991 and 2001 for overall reading literacy and for narrative texts, expository texts, and documents.

12.4 Describing International Benchmarks of Student Achievement[5]

To describe the level of comprehension of students scoring at the international benchmarks, PIRLS used scale anchoring to summarize and describe student achievement at these four points on the reading scale – Top 10% Benchmark, Upper Quarter Benchmark, Median Benchmark, and Lower Quarter Benchmark. Scale anchoring involves identifying items that students scoring at the anchor points (the international benchmarks) can answer correctly and having reading experts review the items, delineate the kind of comprehension they require, and summarize this in a brief description for each anchor point.

12.4.1 Identifying the Anchor Items

The first step in the scale-anchoring procedure is to establish criteria for identifying those students scoring at the anchor points – the international benchmarks in the case of PIRLS. Following the procedure used in previous IEA studies, a student scoring within five scale score points of a benchmark was deemed to be scoring at that benchmark. The score ranges around each benchmark and the number of students scoring in each range are shown in Exhibit 12.13. The range of plus and minus five points around a benchmark is intended to provide an adequate sample in each group, yet be small enough so each benchmark anchor point is still distinguishable from the next. The data analysis for the scale anchoring was based on these students scoring at each anchor point.

Having identified the students scoring at each benchmark anchor point, the next step is to choose criteria for determining whether particular items anchor at each of the anchor points. An important feature of the scale anchoring method is that it yields

[5] The description of the scale anchoring procedure was adapted from Kelly (1999) and Gregory & Mullis (2000).

Exhibit 12.13: Range Around Each Anchor Point and Number of Observations within Ranges

	25th Percentile	50th Percentile	75th Percentile	90th Percentile
Scale Score	430 to 440	505 to 515	565 to 575	610 to 620
Students	3642	6259	6210	3480

descriptions of the comprehension of students reaching certain performance levels on a scale, and that these descriptions reflect demonstrably different accomplishments from anchor point to anchor point. The process entails the delineation of sets of items that students at each benchmark anchor point are very likely to answer correctly and that discriminate between performance at the various benchmarks. Criteria are applied to identify the items that are answered correctly by most of the students at the anchor point, but by fewer students at the next lower point.

Anchoring Criteria

In scale anchoring, the anchor items for each point are intended to be those that differentiate between adjacent anchor points, e.g., between the Top 10% and the Upper Quarter international benchmarks. To meet this goal, the criteria for identifying the items must take into consideration performance at more than one anchor point. Therefore, in addition to a criterion for the percentage of students at a particular anchor point correctly answering an item, it is necessary to use a criterion for the percentage of students scoring at the next lower anchor point who correctly answer an item. For multiple choice items, the criterion of 65 percent was used for the anchor point, 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 with this response probability, students were more likely to have answered the item incorrectly than correctly. For constructed response items the criterion of 50% was used for the anchor point and no criterion was used for the lower points.

The criteria used to identify multiple choice items that "anchored" are outlined below:

For the 25th percentile (the Lower Quarter Benchmark), an item anchored if:

- At least 65 percent of students scoring in the range answered the item correctly

- Because the 25th percentile is the lowest point, items were not identified in terms of performance at a lower point

For the 50th percentile (the Median Benchmark), an item anchored if:

- At least 65 percent of students scoring in the range answered the item correctly and

- Less than 50 percent of students at the 25th percentile answered the item correctly

For the 75th percentile (the Upper Quarter Benchmark), an item anchored if:

- At least 65 percent of students scoring in the range answered the item correctly and

- Less than 50 percent of students at the 50th percentile answered the item correctly

For the 90th percentile (the Top 10% Benchmark), an item anchored if:

- At least 65 percent of students scoring in the range answered the item correctly and

- Less than 50 percent of students at the 75th percentile answered the item correctly

To supplement the pool of anchor items, items that met a slightly less stringent set of criteria were also identified. The criteria to identify items that "almost anchored" were the following:

For the 25th percentile, an item almost anchored if:

- At least 60 percent of students scoring in the range answered the item correctly

- Because the 25th percentile is the lowest point, items were not identified in terms of performance at a lower point

For the 50th percentile, an item almost anchored if:

- At least 60 percent of students scoring in the range answered the item correctly and

- Less than 50 percent of students at the 25th percentile answered the item correctly

For the 75th percentile, an item almost anchored if:

- At least 60 percent of students scoring in the range answered the item correctly and

- Less than 50 percent of students at the 50th percentile answered the item correctly

For the 90th percentile, an item almost anchored if:

- At least 60 percent of students scoring in the range answered the item correctly and

- Less than 50 percent of students at the 75th percentile answered the item correctly

To further supplement the pool of items, items that met only the criterion that at least 60 percent of the students answered correctly (regardless of the performance of students at the next lower point) were identified. The three categories of items were mutually exclusive, and ensured that all of the items were available to inform the descriptions of student achievement at the anchor levels.

Computing the Item Percent Correct at Each Level

The percentage of students scoring in the range around each anchor point that answered the item correctly was computed. To that end, students were weighted to contribute proportionally to the size of the student population in a country. About half of the PIRLS 2001 items are scored dichotomously. For these items, the percentage of students at each anchor point who answered each item correctly was computed. Some of the open-ended items, however, are scored on a partial-credit basis (one, two, or three points); these were transformed into a series of dichotomously scored items, as follows. Consider an item that was scored 0, 1, or 2. Two variables were created:

- $v_1 = 1$ if the student receives a 1, or 2, and 0 otherwise

- $v_2 = 1$ if the student receives a 2 and 0 otherwise.

The percent of students receiving a 1 on v_1 and the percentage of those receiving a 1 on v_2 were computed. This yielded the percent of students receiving at least one point, and full credit.

Identifying Anchor Items

For the PIRLS 2001 reading scale, the criteria described above were applied to identify the items that anchored, almost anchored, and met only the 60 to 65 percent criterion. Exhibits 12.14 and 12.15 present the number of these items at each anchor point.

Including items meeting the less stringent anchoring criteria substantially increased the number of items that could be used to characterize performance at each anchor point, beyond what would have been available if only the items that met the 65%–50% criteria were included. Even though these items did not meet the 65%–50% anchoring criteria, they were still items that students scoring at the anchor points had a high probability of answering correctly.

Exhibit 12.14: Number of Multiple-Choice Items Anchoring at Each Anchor Level

	Anchored	Almost Anchored	Met 60–65% Criterion	Total
25th Percentile	11	3	0	14
50th Percentile	6	1	6	13
75th Percentile	5	4	7	16
90th Percentile	0	0	3	3
Total	22	8	16	46

Exhibit 12.15: Number of Constructed-Response Point Values Anchoring at Each Anchor Level

	Anchored
25th Percentile	15
50th Percentile	31
75th Percentile	17
90th Percentile	11
Too difficult for 90th	13

Exhibit 12.16 presents, by reading purpose, the number of items that met the anchoring criteria discussed above, at each international percentile, and the number of items that were too difficult for the 90th percentile.

12.4.2 Review of Anchor Items Development of Anchor Level Descriptions

Having identified the items that anchored at each of the international benchmarks, the next step was to have the items reviewed by reading experts to develop descriptions of the level of reading comprehension the items demand. In view of their extensive experience in reading and their thorough knowledge of the PIRLS frameworks and achievement tests, the PIRLS Reading Development Group (RDG) was asked to perform this task. In preparation for the review by the RDG, the items were organized in binders grouped by benchmark anchor point and within anchor point by reading purpose, each binder having four sections, corresponding to the four anchor points. Within each section, the items were sorted by reading purpose and then by the anchoring criteria they met – items that anchored, followed by items that almost anchored, followed by items that met only the 60 to 65 percent criteria. The following information was included for each item: its PIRLS 2001 reading purpose and reading process categories; its answer key; percent correct at each anchor point; and overall international percent correct. For constructed-response items, the scoring guides were included.

The PIRLS International Study Center convened the RDG for a three-day meeting. The assignment consisted of three tasks: (1) work through each item in each binder and arrive at a short description of the knowledge, understanding, and/or skills demonstrated by students answering the item correctly; (2) based on the items that anchored, almost anchored, and met only the 60–65 percent criterion, draft a description of the level of comprehension demonstrated by students at each of the four

Exhibit 12.16: Number of Point Values Anchoring* at Each Anchor Level, by Reading Purpose

	25th Percentile	50th Percentile	75th Percentile	90th Percentile	Too Difficult for 90th Percentile	Total
Information Purpose	12	19	20	7	9	67
Literary Purpose	17	25	13	7	4	66

* The numbers in each column include those point values that met or nearly met the anchoring criteria.

benchmark anchor point; and (3) select example items to support and illustrate the anchor point descriptions. Following the meeting, these drafts were edited and revised as necessary for use in the PIRLS 2001 International Report.

12.5 Reporting Questionnaire Data

As described in chapter 3, PIRLS 2001 used four questionnaires to gather information about students' home and school environments and their experiences in learning to read:

1. Students answered questions pertaining to their home and school experiences in learning to read, including instructional experiences, self-perception and attitudes towards reading, out-of-school reading habits, computer use, home literacy resources, and basic demographic information.

2. Parents or caregivers of the sampled students responded to questions about the students' early reading experiences, child-parent literacy interactions, parents' reading habits and attitudes, home-school connections, and demographic and socioeconomic indicators.

3. The teachers of the sampled students were asked about characteristics of the class tested, instructional activities for teaching reading, classroom resources, assessment practices, and about their education, training, and opportunities for professional development.

4. The principals of schools reported on enrollment and school characteristics, school organization for reading instruction, school staffing and resources, home-school connections, and the school environment.

The *PIRLS 2001 International Report* devotes five chapters to the questionnaire data, dealing with literacy-related activities in the home, the school curriculum and organization for teaching reading, teachers and reading instruction, school contexts, and students' reading attitudes, self-concept, and out-of-school activities.

Summary Indices from Background Data
To summarize the information obtained from the background questionnaires concisely, and focus attention on educationally relevant support and practice, PIRLS sometimes combined information from a number of questions to form an index that was more global and reliable than the component questions. According to the responses of students, their parents, teachers or school principals, students were placed in a "high," "medium," or "low" category for each index, with the high level being set so that it corresponds to conditions or activities generally associated with higher academic achievement. For example, a three-level index of home educational resources was constructed from students' responses to two questions about home educational resources: number of books in the home and educational aids in the home (computer, study desk/table for own use, books of their own, access to a daily newspaper); and parents' responses to two questions: number of children's books in the

Exhibit 12.17: Summary Indices from Background Data in the PIRLS 2001 International Report

Name of Index	Label	Exhibit[a]	Analysis Method
Index of Early Home Literacy Activities	EHLA	4.10	Index based on parents' responses to the frequency of the following activities they engaged in with their child prior to entry into primary school: read books; tell stories; sing songs; play with alphabet toys (e.g., blocks with letters of the alphabet); play word games; or read aloud signs and labels. Average is computed across the 6 items based on a 3-point scale: Never or almost never = 1, Sometimes = 2, and Often = 3. High level indicates an average of greater than 2.33 through 3. Medium level indicates an average of 1.67 through 2.33. Low level indicates an average of 1 to less than 1.67.
Index of Home Educational Resources	HER	4.60	Index based on students' responses to two questions about home educational resources: number of books in the home, and educational aids in the home (computer, study desk/table for own use, books of their own, access to a daily newspaper); and parents' responses to two questions: number of children's books in the home, and parents' education. High level indicates more than 100 books in the home; more than 25 children's books; 3 or 4 educational aids; and highest level of education for either parent is finished university. Low level indicates 25 or fewer books in the home; 25 or fewer children's books; 2 or fewer educational aids; and highest level of education for either parent is some secondary or less. Medium level includes all other combinations of responses.
Index of Parents' Attitudes Toward Reading	PATR	4.17	Index based on parents' agreement with the following: I read only if I have to; I like talking about books with other people; I like to spend my spare time reading; I read only if I need information; and Reading is an important activity in my home. Average is computed across the 5 items based on a 4-point scale: Disagree a lot = 1, Disagree a little = 2, Agree a little = 3, and Agree a lot = 4. Responses for negative statements were reverse-coded. High level indicates an average of greater than 3 through 4, Medium level indicates an average of 2 through 3, and Low level indicates an average of 1 to less than 2.
Index of Reading for Homework	RFH	6.34	Index based on teachers' responses to two questions: How often do you assign reading as part of homework (for any subject)? In general, how much time do you expect students to spend on homework involving reading (for any subject) each time you assign it? High level indicates students are expected to spend more than 3 minutes at least 1-2 times a week. Low level indicates students are never assigned homework or are expected to spend no more than 30 minutes less than once a week. Medium level indicates all other combinations of frequencies.
Index of Home-School Involvement	HSI	7.90	Index based on principals' responses to how often and what percentage of students' parents participate in the following provided by the school: teacher-parent conferences; letters, calendars, newsletters, etc., sent home to provide information about school; written reports (report cards) of child's performance sent home; and events at school to which parents are invited. High level indicates that 4 or more times a year schools hold teacher-parent conferences and events at school attended by more than half of the parents; send home letters, calendars, newsletters, etc., with information about the school 7 or more times a year; and send written reports (report cards) of child's performance 4 or more times a year. Low level indicates schools never hold teacher-parent conferences, or if they do, only 0-25% of parents attend; schools never hold events, or do so only yearly, attended by 0-25% of parents; send home letters, calendars, newsletters, etc., no more than 3 times a year; and send home written reports of children's performance never or only once a year. Medium level indicates all other combinations.

[a] Exhibit number in the international report where data based on the index were presented.

Exhibit 12.17: Summary Indices from Background Data in the PIRLS 2001 International Report (continued)

Name of Index	Label	Exhibit[a]	Analysis Method
Index of Principals' Perceptions of School Climate	PPSC	7.14	Index based on principals' characterization in their school: teachers' job satisfaction; teachers' expectations for student achievement; parental support for student achievement; students' regard for school property; and students' desire to do well in school. Average is computed on a 5-point scale: Very high = 1, High = 2, Medium = 3, Low = 4, and Very low = 5. High level indicates an average of 1 to less than 2.33. Medium level indicates an average of 2.33 through 3.67. Low level indicates an average of greater than 3.67 through 5.
Index of Principals' Perceptions of School Safety	PPSS	7.17	Index based on principals' responses about the degree each was a school problem: classroom disturbances; cheating; profanity; vandalism; theft; intimidation or verbal abuse of other students; and physical conflicts among students. Average is computed on a 4-point scale: Not a problem = 1, Minor problem = 2, Moderate problem = 3, and Serious problem = 4. High level indicates an average of 1 to less than 2. Medium level indicates an average of 2 through 3. Low level indicates an average of greater than 3 through 4.
Index of Availability of School Resources	ASR	7.18	Index based on principals' responses to how much the school's capacity to provide instruction is affected by a shortage or inadequacy of the following: instructional staff; teachers quali- fied to teach reading; instructional materials; supplies (e.g., paper, pencils); school buildings and grounds; heating/cooling and lighting systems; instructional space (e.g., classrooms); special equipment for physically disabled students; computers for instructional purposes; computer software for instructional purposes; computer support staff; library books; and audiovisual resources. Average is computed on a 4-point scale: Not at all = 1, A little = 2, Some = 3, and A lot = 4. High level indicates an average of 1 to less than 2. Medium level indicates an average of 2 through 3. Low level indicates an average of greater than 3 through 4.
Index of Students' Attitudes Toward Reading	SATR	8.1 and 8.2	Index based on students' agreement with the following: I read only if I have to; I like talking about books with other people; I would be happy if someone gave me a book as a present; I think reading is boring; and I enjoy reading. Average is computed on a 4-point scale: Disagree a lot = 1, Disagree a little = 2, Agree a little = 3, and Agree a lot = 4. Responses for negative statement were reverse-coded. High level indicates an average greater than 3 through 4. Medium level indicates an average of 2 through 3. Low level indicates an average of 1 to less than 2.
Index of Students' Reading Self Concept	SRSC	8.3 and 8.4	Index based on students' agreement with the following: reading is very easy for me; I do not read as well as other students in my class; and reading aloud is very hard for me. Average is computed on a 4-point scale: Disagree a lot = 1, Disagree a little = 2, Agree a little = 3, and Agree a lot = 4. Responses for negative statement were reverse-coded. High indicates an average of greater than 3 through 4. Medium indicates an average of 2 through 3. Low indicates an average of 1 to less than 2.

[a] Exhibit number in the international report where data based on the index were presented.

home, and parents' education. Students were assigned to the high level if there were more than 100 books, more than 25 children's books, and at least three of the educational aids in the home, and at least one parent finished university. Students at the low level had 25 or fewer books, 25 or fewer children's books, no more than two educational aids, and the highest level of education for either parent was some secondary or less. Students with all other response combinations were assigned to the middle category.

The 10 indices constructed for the PIRLS 2001 international report are listed in Exhibit 12.17.

The exhibit that displays each index shows the percentages of students at each level of the index, together with their reading achievement. In addition, the percentage at the high level was displayed graphically, with the countries ranked in order.

12.5.1 Reporting Student Questionnaire Data

Reporting the data from the student questionnaire was fairly straightforward. Most of the exhibits in the international report that include data from the student questionnaire present weighted percentages of students in each country for each response category, together with the mean reading achievement of those students. International averages are also displayed for each category. In general, jackknife standard errors accompany the statistics reported. In addition to the exhibits showing percentages of students overall, the international report include some information separately by gender. For gender-based exhibits, the percentages of boys and girls in each category were displayed, and the statistical significance of the difference between genders was indicated.

12.5.2 Reporting Teacher Questionnaire Data

The teacher of each PIRLS fourth-grade class was asked to complete a questionnaire to provide information about the students in the class, reading instruction for those students, computer use and library facilities, homework and assessment, and about the teacher's own education and professional training and development. Because the sampling for the teacher questionnaires was based on participating students, the teachers that responded do not necessarily represent all of the teachers of the target grade in each of the PIRLS countries. Rather, they represent teachers of the representative samples of students assessed. It is important to note that in the international report, the student was always the unit of analysis, even when information from the teacher questionnaires was being reported. That is, the data presented are the percentages of *students* whose teachers reported various characteristics or instructional strategies. Using the student as the unit of analysis makes it possible to describe the instruction received by representative samples of students. Although this approach may provide a different perspective from that obtained by simply collecting information from teachers, it is consistent with the PIRLS goals of illuminating students' educational contexts and performance.

Although the vast majority of the PIRLS classes were taught by a single teacher, in Sweden each class had two teachers, each of which completed a teacher questionnaire. For reporting in these cases, the student's sampling weight was divided between the teachers, so that the student's contribution to student population estimates thus remained constant regardless of the number of teachers. This was consistent with the policy of reporting attributes of teachers and their classrooms in terms of the percentages of students taught by teachers with these attributes.

12.5.3 Reporting Parents' Questionnaire Data

The PIRLS *Learning to Read Survey* was completed by the parents or primary caregivers of the students participating in the study. Like the teacher questionnaire, the data from the parents' questionnaire were linked to the student, who was always the unit of analysis, even when information from the parents' questionnaires was being reported. That is, the data presented are the percentages of students whose parents reported various characteristics or instructional strategies.

12.5.4 Reporting School Questionnaire Data

The principals of the selected schools in PIRLS completed questionnaires on the school contexts in which the learning and teaching of reading occur. Although schools constituted the first stage of sampling, the PIRLS school sample was designed to optimize the student sample, not to provide an optimal sample of schools.[6] Therefore, like the teacher data, the school-level data were reported using the student as the unit of analysis to describe the school contexts for the representative samples of students. In general, the exhibits based on the school data present percentages of students in schools with different characteristics for each country and for the international average.

12.5.5 Reporting Response Rates for Background Questionnaire Data

While it is desirable that all questions included in a data collection instrument be answered by all intended respondents, a certain percentage of non-response is inevitable. Not only do some questions remain unanswered; sometimes entire questionnaires are not completed or not returned. In PIRLS 2001, since students, parents, teachers, or principals sometimes did not complete the questionnaire assigned to them or some questions within it, certain variables had less than a 100 percent response rate.

The handling of non-responses varied depending on how the data were to be reported. For background variables that were reported directly, the non-response rates indicate the percentage of students for whom no response was available for a given question. In general, derived variables based on more than one background question were coded as missing if data for any

[6] See Chapter 5 for a description of the PIRLS sampling design.

of the required background variables were missing. However, for the 10 indices described earlier in this chapter, cases were coded as missing only if there was no response for more the one-third of the questions used to compute the index; index values were be computed if there were valid data for at least two-thirds of the required variables.

The tables in the PIRLS international reports contain special notations on response rates for the background variables. Although in general the response rates for background variables were high, some variables and some countries exhibited less than acceptable rates. Since the student is the unit of analysis, the non-response rates given in the international report always reflect the percentage of students for whom the required responses from students, parents, teachers, or schools were not available. The following special notations were used to convey information about response rates in exhibits in the international report·

- For a country where student, parent, teacher or school responses were available for 70 percent to 84 percent of the students, an "r" appears next to the data for that country.

- When student, parent, teacher or school responses were available for 50 to 69 percent of the students, an "s" appears next to the data for that country.

- When student, parent, teacher or school responses were available for fewer than 50 percent of the students, an "x" replaces the data.

- When the percentage of students in a particular category fell below 2 percent, achievement data were not reported in that category. The data were replaced by a tilde (~).

- When data were unavailable for all respondents in a country, dashes (–) were used in place of data in all of the affected columns.

12.5.6 Development of the PIRLS International Report

The goal of the PIRLS international report was to describe fourth-grade students' reading achievement in participating countries and present as much information as possible about the contexts for learning to read. Beginning in September 2001, staff at the PIRLS International Study Center drafted an outline of the report, and, following a careful review of the questionnaires, developed specifications for the variables and indices to be included. Staff also prepared detailed analysis plans specifying how the analyses underlying each proposed exhibit in the draft report outline should be conducted, and began work developing the programs to implement the plans. Analysis plans included detailed documentation of the variables and response categories involved, and the specification for any country-specific modifications to analyses necessitated by national adaptations to questions. These plans were incorporated in analysis notes for each proposed exhibit. The analyses required to produce the proposed exhibits were planned, and prototype exhibits prepared.

The analysis plans, report outlines, and prototype exhibits underwent a lengthy review involving the National Research Coordinators and project staff, following which consensus was achieved as to the contents of the international report, including the indices and variables to be reported. The analysis plans, outlines, and prototype exhibits were reviewed at the seventh meeting of the PIRLS 2001 National Research Coordinators in Athens, Greece, in March 2002. Following this meeting, the material was revised and updated to reflect the ideas and suggestions that were made. Some exhibits were deleted or added, and some of the analyses or presentational modes were modified.

After the data for all countries became available for analysis in mid-2002, the International Study Center conducted the psychometric scaling of the reading achievement data[7] and implemented the analyses documented in the analysis notes. In September 2002, staff met with the PIRLS Reading Development Group to conduct scale-anchoring. Analyses were completed and the text of the report drafted in November 2002, after which draft reports were circulated by mail to NRCs for review. The draft report was reviewed in detail by NRCs at the eighth and final PIRLS NRC meeting in Istanbul, Turkey, in December 2002. Comments and suggestions from NRCs were incorporated into the final version of the report. Final revisions were made in January 2003, and the report was published in April 2003 (Mullis et al., 2003).

[7] The scaling of the PIRLS achievement data is described in Chapter 11.

References

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.

Gregory, K.D. & Mullis, I.V.S. (2000). Describing international benchmarks of student achievement. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS 1999 technical report*. 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.

Kelly, D. L. (1999). *Interpreting the Third International Mathematics and Science Study (TIMSS) achievement scales using scale anchoring*. Unpublished doctoral dissertation, Boston College.

Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., & Kennedy, A.M. (2003). *Trends in children's reading literacy achievement 1991–2002: IEA's repeat in nine countries of the 1991 Reading Literacy Study*. Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., & Kennedy, A.M. (2003). *PIRLS 2001 International Report: IEA's study of reading literacy achievement in primary schools in 35 countries*. Chestnut Hill, MA: Boston College.

Westat, Inc. (1997). *A user's guide to WesVarPC*. Rockville, MD: Westat, Inc.

Wolter, K.M. (1985). *Introduction to variance estimation*. New York: Springer-Verlag.