PIRLS

Appendix A

A

by the Reading Literacy Study, which served as a starting point for its development. PIRLS 2001 is intended to be the first in a continuing five-year cycle of trend studies in reading literacy, and has been designed from the outset to monitor progress in reading achievement into the future. The IEA scheduled the PIRLS data collection for 2001 to coincide with the 10th anniversary of its 1991 Reading Literacy Study. In order to provide countries that participated in the 1991 Reading Literacy Study an opportunity to measure changes from 1991 to 2001, PIRLS 2001 also included an option to re-administer the 1991 reading literacy test in 2001, at the same time as the main PIRLS assessment. The results of this study, known as the Trends in IEA's Reading Literacy Study, are presented in a separate report.[3]

Participants in PIRLS

Thirty-five countries took part in the 2001 PIRLS assessment. Of these countries, nine participated in the Trends in IEA's Reading Literacy Study to measure changes between 1991 and 2001 in student performance as measured by the 1991 reading literacy test (see Exhibit A.1).

Developing the PIRLS Tests

The assessment framework and specifications[4] for PIRLS was developed in collaboration with the PIRLS Reading Development Group (RDG) and with the assistance of the National Research Coordinators (NRCs) from the 35 countries participating in the study. The framework underwent several iterations in response to reviews and comments from the PIRLS countries and the reading research community, and embodies the ideas and interests of many individuals and organizations around the world. The IEA 1991 Reading Literacy Study served as the foundation for PIRLS, providing a basis for the PIRLS definition of reading literacy, and for establishing its framework and developing its assessment instruments. Although the 1991 study provided the groundwork for PIRLS, the PIRLS framework and instruments are new; reflecting the IEA's commitment to be forward-thinking, incorporating in PIRLS the latest approaches to measuring reading literacy.

3   ..., ..., ..., ..., & ..., ... (2003). *Trends in children's reading literacy achievement 1991-2001: IEA's repeat in nine countries of the 1991 Reading Literacy Study.* ...

4   ..., ..., ..., ..., ..., ..., & ..., ... (2001). *Framework and specifications for PIRLS assessment 2001* (2...). ...
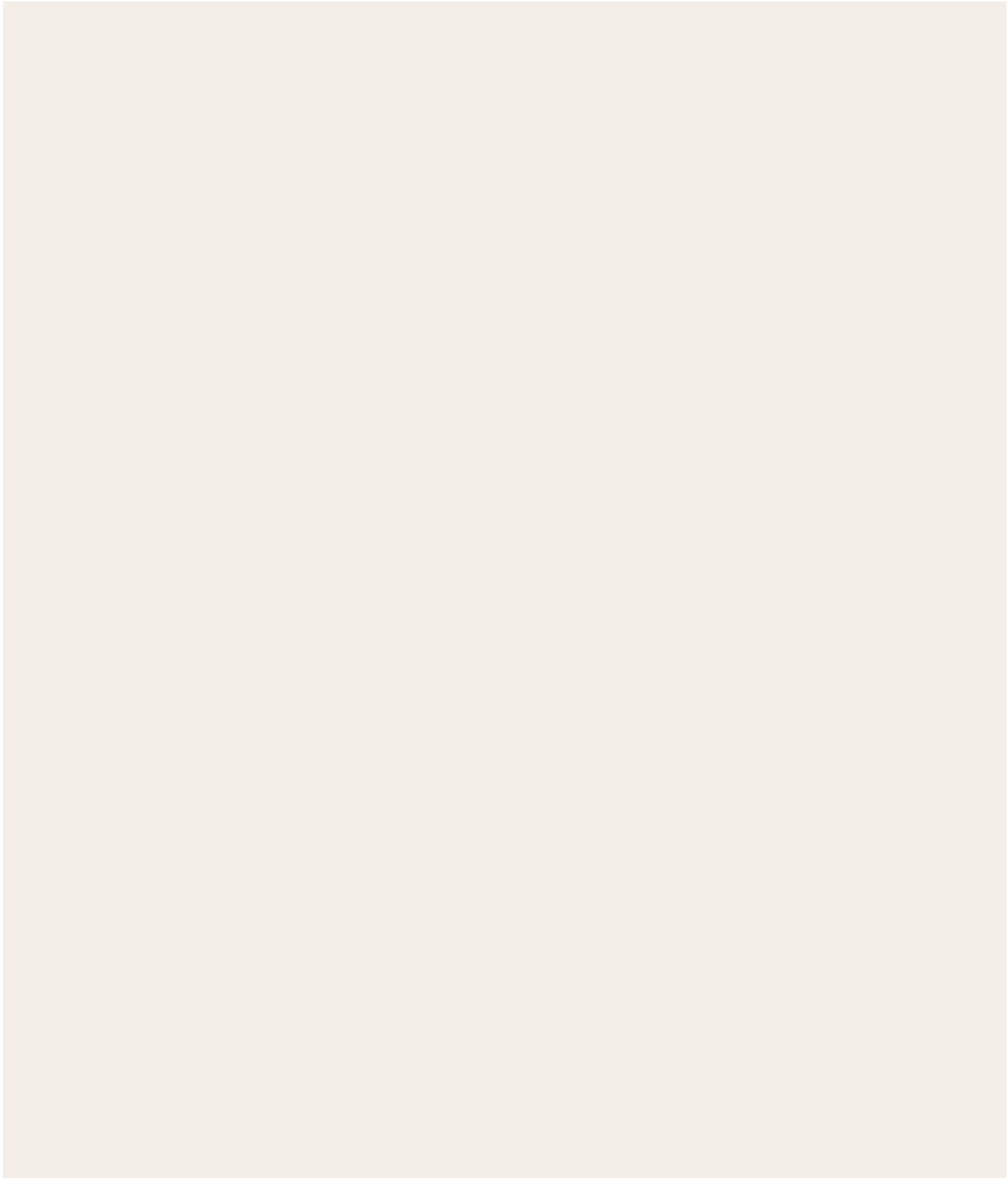
Exhibit A.2 portrays the interaction of the two: each process is assessed within each purpose for reading.

The selection of the assessment passages and the development of the items and scoring guides were the result of an intensive process of collaboration, piloting, and review – spanning more than two years.[5] In selecting the passages for PIRLS, every effort was made to minimize cultural bias. Potential stimulus passages were collected from as many countries as possible, and the final selection was based, in part, on the national and cultural representation of the entire set of assessment passages. Everything possible was done to ensure that the PIRLS assessment represented the curricula of the participating countries, and that the items did not exhibit bias towards or against particular countries. Draft passages and items were subjected to full-scale field testing before the instruments for the main data collection were finalized.[6] The final version of the assessment was endorsed by the NRCs of the participating countries.

Exhibit A.3 shows the distribution of items by reading purpose and process category. There were 98 items in the assessment, approximately half of which were multiple-choice and half constructed-response. The constructed-response items required students to generate and write their own answers. Some items required short answers while others demanded a more elaborate response. In scoring the test, correct answers to most questions (including all those in multiple-choice format) were worth one point. However, responses to questions seeking more elaborate responses were evaluated for partial credit, with a fully-correct answer being awarded two or three points. Thus, the total number of score points available for analyses somewhat exceeds the number of items in the assessment. The student answer booklet provided an indication to the student of how many score points would be awarded for each answer, and how much writing was expected. Almost two-thirds of the score points came from constructed-response items.

Exhibit A.2:   **Aspects of Reading Literacy**

PIRLS Reading Assessment – Reading Purposes and Processes

Purposes for Reading

**Processes of
Comprehension**
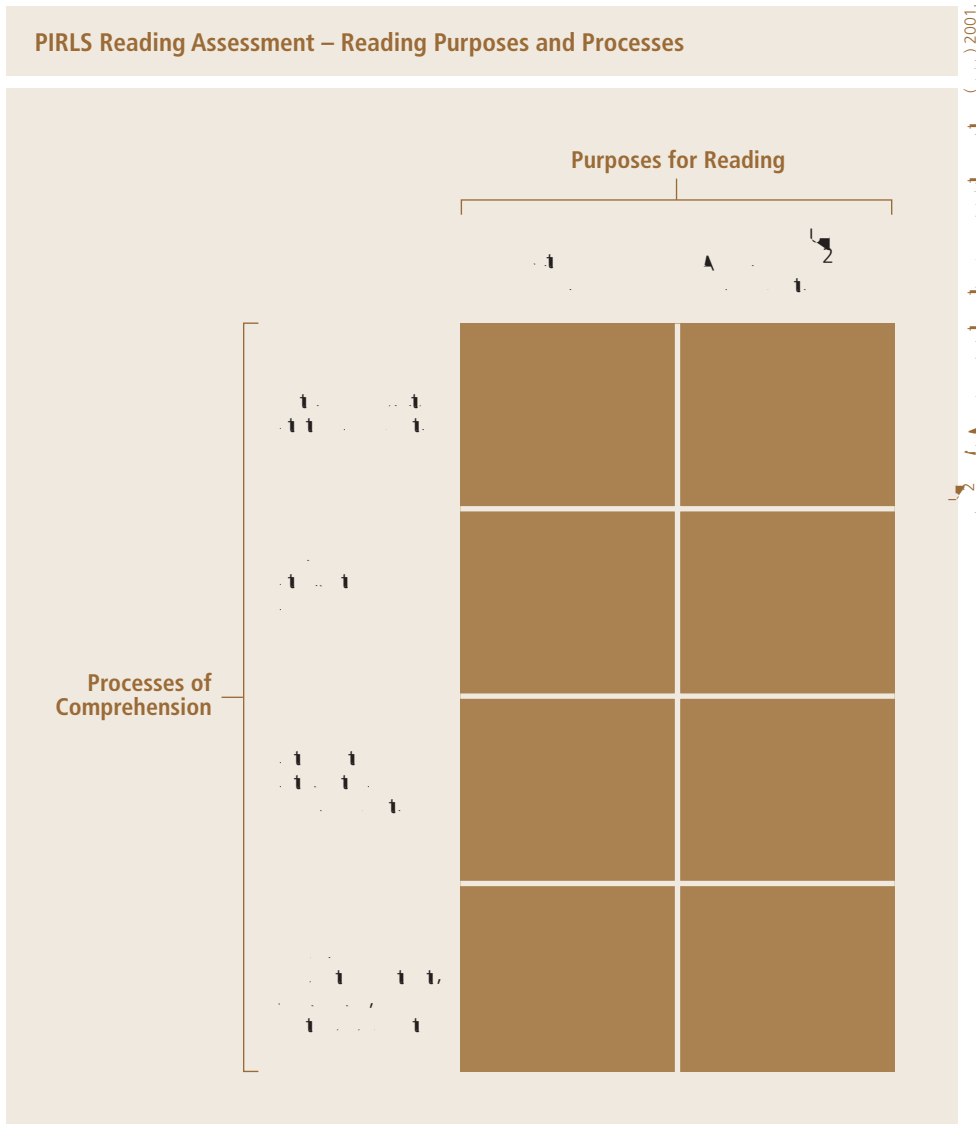
Exhibit A.3: Distribution of Items by Reading Purpose and Process Category

Reading Purpose	Percentage of Items	Total Number of Items	Number of Multiple-Choice Items	Number of Constructed-Response Items[1]	Number of Score Points[2]
Literary Experience	52	51	25	26	66
Acquire and Use Information	48	47	21	26	67
Total	100	98	46	52	133

Reading Process Involved					

PIRLS Test Design

Given the broad coverage goals of the PIRLS framework and its emphasis on the use of authentic texts, the specifications for the pool of reading passages and accompanying items included extensive testing time. The PIRLS RDG found that a valid assessment of two purposes for reading – reading for literary experience and reading to acquire and use information – required at least eight passages and items, four for each purpose. Students were given 40 minutes to complete a passage. With eight passages, a total assessment time would take up 320 minutes. While such an amount of assessment materials would provide good coverage of the reading material children meet in their everyday lives, it was an unreasonable expectation to administer the entire set of reading passages and test items to any one child. So as not to overburden the young chil-

that the corrections were made; and (6) a series of statistical checks after the testing to detect items that did not perform comparably across countries.

For the participating countries, the bulk of the translation effort took place prior to the field test. After the field test, countries needed only to make any changes to the items or passages that resulted from analysis of the field test data. The PIRLS data-collection instruments were verified twice – the field test versions before the field test and the final versions before the main data collection. Countries, therefore, had the benefit of two careful reviews of their translations. They also had the benefit of diagnostic item statistics from the field test data analysis, which helped to identify mistranslations that could be corrected before the main data collection.

Sample Implementation and Participation Rates

PIRLS 2001 had as its target population students enrolled in the upper of the two adjacent grades that contained the largest proportion of 9-year-old students at the time of testing.[9] Beyond the age criterion embedded in the above definition, the target grade should represent that point in the curriculum where students have essentially finished learning the basic reading skills and will focus more on "reading to learn" in the subsequent grades. Thus, the PIRLS 2001 target grade was expected to be fourth grade in most countries (some countries have students significantly older than nine years of age). Exhibit A.4 shows any differences in coverage between the international and national desired populations.

Selecting valid and efficient samples is critical to the quality and success of an international comparative study such as PIRLS. The accuracy of the survey results depends upon the quality of the sampling information available when planning the sample, and on the care with which the sampling activities are conducted. For PIRLS, NRCs worked on all phases of sampling in conjunction with staff from Statistics Canada. NRCs were trained in how to select the school and student samples, and in how to use the sampling software provided by the IEA Data Processing Center. In consultation with the PIRLS 2001 sampling referee (Keith Rust, Westat, Inc.), staff from Statistics Canada reviewed the national sampling plans, sampling data, sampling frames,

, , . (2003). , , & . (.), *PIRLS 2001 technical report.* , .

Exhibit A.4:   Coverage of PIRLS Target Population

ISC | 4th Grade PIRLS 2001

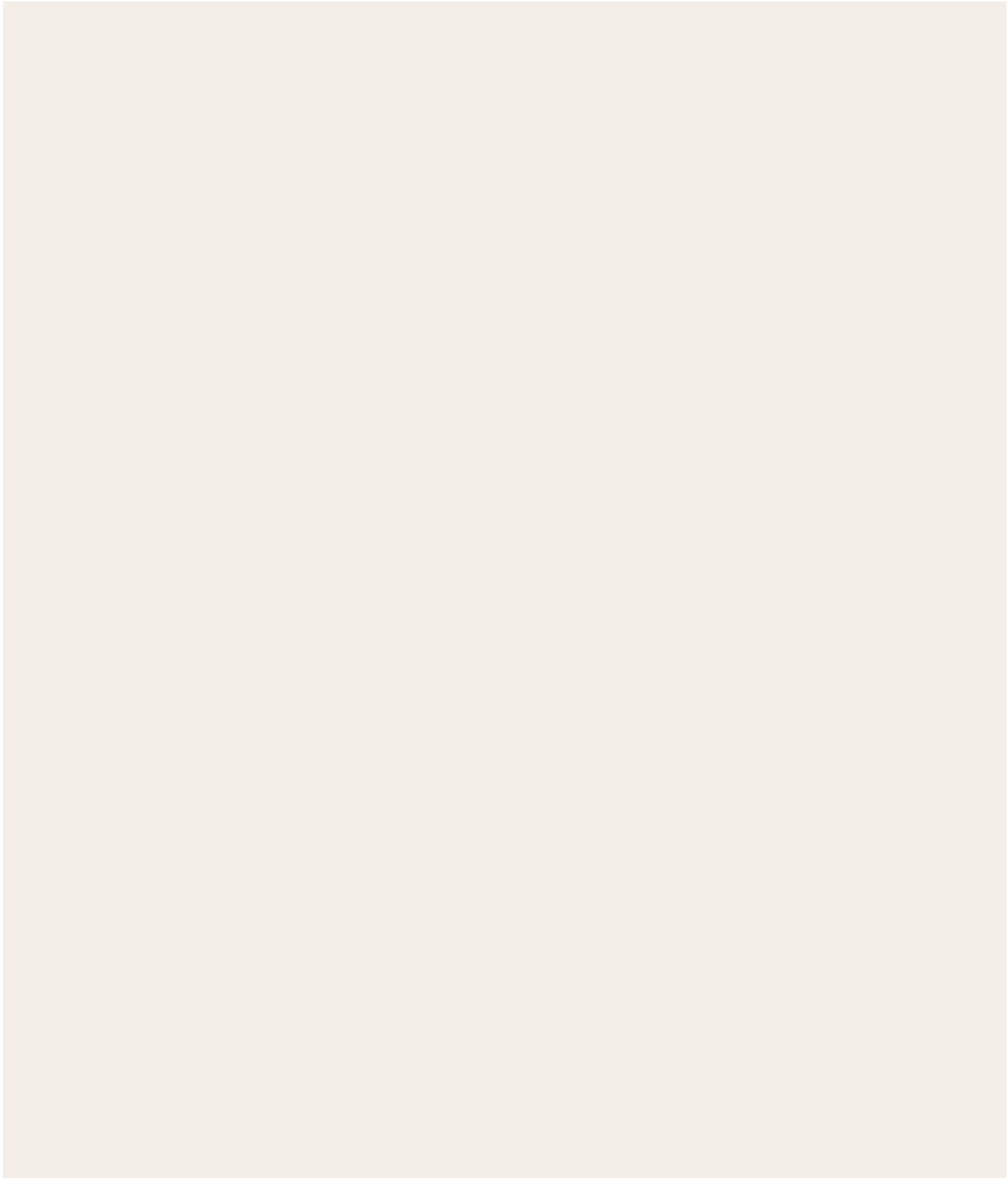| Countries	International Desired Population		National Desired Population		
	Country Coverage	Notes on Coverage	School-Level Exclusions	Within-Sample Exclusions	Overall Exclusions
Argentina	100%		3.7%	0.4%	4.1%
Belize	100%		0.8%	0.0%	0.8%
Bulgaria	100%		2.7%	0.0%	2.7%
Canada (O,Q)	60%	Provinces of Ontario and Quebec only	3.1%	2.2%	5.4%
Colombia	100%		3.2%	0.1%	3.3%
Cyprus	100%		0.0%	2.0%	2.0%
Czech Republic	100%		5.0%	0.0%	5.0%
England	100%		1.8%	3.9%	5.7%
France	100%		5.1%	0.3%	5.3%
Germany	100%		0.8%	1.0%	1.8%
Greece	100%		2.0%	5.3%	7.3%
Hong Kong, SAR	100%		2.8%	0.0%	2.8%
Hungary	100%		2.1%	0.0%	2.1%
Iceland	100%		1.8%	1.3%	3.1%
Iran, Islamic Rep. of	100%		0.5%	0.0%	0.5%
Israel	100%		16.5%	5.9%	22.4%
Italy	100%		0.0%	2.9%	2.9%
Kuwait	100%		0.0%	0.0%	0.0%
Latvia	100%		4.3%	0.3%	4.6%
Lithuania	90%	Lithuanian speaking students only	1.3%	2.5%	3.8%
Macedonia, Rep. of	100%		3.8%	0.4%	4.2%
Moldova, Rep. of	100%		0.5%	0.0%	0.5%
Morocco	100%		1.0%	0.0%	1.0%
Netherlands	100%		3.4%	0.3%	3.7%
New Zealand	100%		1.6%	1.7%	3.2%
Norway	100%		1.9%	0.8%	2.8%
Romania	100%		2.6%	1.9%	4.5%
Russian Federation	100%		2.8%	3.8%	6.6%
Scotland	100%		3.8%	0.8%	4.7%
Singapore	100%		1.3%	0.1%	1.4%
Slovak Republic	100%		1.4%	0.6%	2.0%
Slovenia	100%		0.0%	0.3%	0.3%
Sweden	100%		2.5%	2.5%	5.0%
Turkey	100%		3.9%	0.0%	3.9%
United States	100%		0.6%	4.7%	5.3%

Exhibit A.5:   **School Sample Sizes**

Exhibit A.6:   Student Sample Sizes (Unweighted)

ISC | 4th Grade PIRLS 2001

Countries	Within-School Student Participation (Weighted Percentage)	Number of Sampled Students in Participating Schools	Number of Students Withdrawn from Class/School	Number of Students Excluded	Number of Eligible Students	Number of Students Absent	Number of Students Assessed
Argentina	91%	3769	132	13	3624	324	3300
Belize	94%	3137	32	0	3105	196	2909
Bulgaria	97%	3633	53	0	3580	120	3460
Canada (O,Q)	94%	9151	99	228	8824	571	8253
Colombia	96%	5582	225	5	5352	221	5131
Cyprus	97%	3149	2	63	3084	83	3001
Czech Republic	94%	3220	10	0	3210	188	3022
England	94%	3528	46	122	3360	204	3156
France	97%	3673	20	11	3642	104	3538
Germany	88%	8997	27	58	8912	1186	7726
Greece	97%	2718	0	151	2567	73	2494
Hong Kong, SAR	99%	5192	69	0	5123	73	5050
Hungary	97%	4819	14	0	4805	139	4666
Iceland	87%	4320	29	58	4233	557	3676
Iran, Islamic Rep. of	98%	7703	104	0	7599	169	7430
Israel	96%	4400	33	214	4153	180	3973
Italy	98%	3703	15	103	3585	83	3502
Kuwait	91%	7874	0	0	7874	741	7133
Latvia	93%	3266	8	11	3247	228	3019
Lithuania	85%	3114	7	72	3035	468	2567
Macedonia, Rep. of	97%	3904	42	14	3848	137	3711
Moldova, Rep. of	96%	3679	9	0	3670	137	3533
Morocco	93%	3452	35	0	3417	264	3153
Netherlands	98%	4256	11	14	4231	119	4112
New Zealand	96%	2720	68	53	2599	111	2488
Norway	92%	3784	25	26	3733	274	3459
Romania	97%	3744	23	2	3719	94	3625
Russian Federation	97%	4281	24	42	4215	122	4093
Scotland	95%	2912	20	26	2866	149	2717
Singapore	98%	7162	46	4	7112	110	7002
Slovak Republic	96%	4034	33	18	3983	176	3807
Slovenia	95%	3112	10	8	3094	142	2952
Sweden	93%	6678	38	145	6495	451	6044
Turkey	97%	5390	123	0	5267	142	5125
United States	96%	4091	55	121	3915	152	3763

Exhibit A.7:   Participation Rates (Weighted)

Countries	School Participation			Overall Participation	
	Before Replacement	After Replacement		Before Replacement	After Replacement
Argentina	89%	92%	91%	81%	84%
Belize	80%	80%	94%	75%	75%
Bulgaria	97%	97%	97%	93%	93%
Canada (O,Q)	90%	97%	94%	85%	91%
Colombia	80%	98%	96%	76%	94%
Cyprus	98%	100%	97%	95%	97%
Czech Republic	90%	95%	94%	85%	90%
England	57%	87%	94%	54%	82%
France	93%	97%	97%	90%	94%
Germany	98%	98%	88%	86%	86%
Greece	78%	85%	97%	76%	82%
Hong Kong, SAR	73%	98%	99%	72%	97%
Hungary	98%	98%	97%	95%	95%
Iceland	95%	95%	87%	82%	82%
Iran, Islamic Rep. of	97%	100%	98%	95%	98%
Israel	96%	98%	96%	92%	94%
Italy	90%	100%	98%	88%	98%
Kuwait	87%	89%	91%	80%	81%
Latvia	89%	96%	93%	83%	89%
Lithuania	56%	97%	85%	47%	83%
Macedonia, Rep. of	97%	97%	97%	94%	94%
Moldova, Rep. of	84%	100%	96%	81%	96%
Morocco	74%	74%	93%	69%	69%
Netherlands	53%	89%	98%	52%	87%
New Zealand	94%	100%	96%	90%	96%
Norway	82%	89%	92%	76%	82%
Romania	96%	96%	97%	93%	93%
Russian Federation	100%	100%	97%	97%	97%
Scotland	76%	79%	95%	72%	74%
Singapore	100%	100%	98%	98%	98%
Slovak Republic	88%	100%	96%	84%	96%
Slovenia	98%	99%	95%	94%	94%
Sweden	97%	99%	93%	90%	92%
Turkey	100%	100%	97%	97%	97%
United States	61%	86%	96%	59%	83%

Exhibit A.8:   Percentage of Students with Any Available Student, Parent, Teacher, and Principal Questionnaire Data
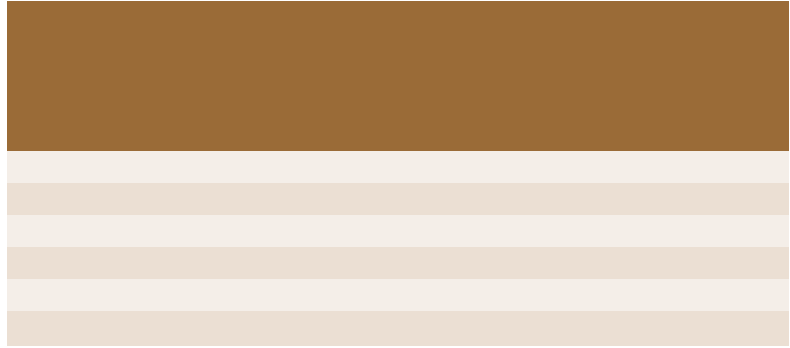
Each country was responsible for conducting quality control procedures and describing this effort in the Survey Activities report documenting procedures used in the study. In addition, the International Study Center considered it essential to independently monitor compliance with standardized procedures. NRCs were asked to nominate a person, unconnected with their national center, to serve as quality control monitors (QCMs) for their countries. The International Study Center developed manuals for the quality control monitors and, in a two-day training session, briefed them about PIRLS, the responsibilities of the national centers in conducting the study, and their own roles and responsibilities. Monitors from 33 countries attended the training session conducted by the International Study Center staff. In countries where the data collection schedule made it impossible for one quality control monitor to visit all the sampled schools, monitors who attended the training session were asked to recruit other monitors as necessary, in order to allow for efficiency in the coverage of the territory and testing timetable. In all, 71 quality control monitors participated.[11] They interviewed NRCs about data collection plans and procedures, and visited a sample of 15 schools in each country, where they observed testing sessions and interviewed school coordinators.[12] All together, quality control monitors visited observed testing sessions and interviewed school coordinators in 475 schools from 33 countries.

The results of the interviews indicate that, in general, NRCs had prepared well for data collection and – despite the heavy demands of the schedule and shortages of resources – were able to conduct the data collection efficiently and professionally. Similarly, the PIRLS test appeared to have been administered in compliance with international procedures – including the activities before the testing session, along with school-level activities related to receiving, distributing, and returning material from national centers.

Scoring the Constructed-Response Items

Because almost two-thirds of the score points came from constructed-response items, PIRLS needed to develop procedures for reliably evaluating student responses within and across countries. To ensure reliable scoring procedures based on the PIRLS rubrics, the International Study Center prepared detailed

11 ....

12 ...., ..., & ...., ... (2003). ....
   ...., ..., & .... (.), *PIRLS 2001 technical report.* ...., ...
   ....

and the PIRLS Reader. Median reliabilities ranged from 0.83 in The Netherlands to 0.91 in England, Israel, Macedonia, New Zealand, Romania, and Singapore. The international median (0.88) is the median of the reliability coefficients for all countries.

Data Processing

To ensure the availability of comparable, high-quality data for analysis, PIRLS took rigorous quality control steps to create the international database.[14] PIRLS prepared manuals and software for countries to use in creating and checking their data files, so that the information would be in a standardized international format before being forwarded to the IEA Data Processing Center in Hamburg for creation of the international database. Upon arrival at the Data Processing Center, the data underwent an exhaustive cleaning process. This involved several iterative steps and procedures designed to identify, document, and correct deviations from the international instruments, file structures, and coding schemes. The process also emphasized consistency of information within national data sets and appropriate linking among the student, parent, teacher, and school data files.

Throughout the process, the data were checked and double-checked by the IEA Data Processing Center, the International Study Center, and the national centers. The national centers were contacted regularly, and given multiple opportunities to review the data for their countries. In conjunction with the IEA Data Processing Center, the International Study Center reviewed item statistics for each cognitive item in each country to identify poorly performing items.[15] In general, the items exhibited very good psychometric properties in all countries. On only two occasions was an item deleted for a country; once because of a translation error in the student booklet, and once because of a misinterpretation of a scoring rubric.

IRT Scaling and Data Analysis

The general approach to reporting the PIRLS achievement data was based primarily on item response theory (IRT) scaling methods.[16] Student reading achievement was summarized using a family of 2- and 3-parameter IRT models

14   .   ꞏ   ꞏ   ꞏ   ꞏ   , 2ꞏ, &   ꞏ   ꞏ t, ꞏ. (2003).   ꞏ   ꞏ   ꞏ   ꞏ , ꞏ   , &
        ꞏ   ꞏ ꞏ   (.), *PIRLS 2001 technical report.*   ꞏ ꞏ ■ ꞏ, ꞏ ꞏ   ꞏ   ꞏ ꞏ .

15   ꞏ   , ꞏ ꞏ,   ꞏ   , ꞏ ꞏ, & ꞏ   , ꞏ. (2003).   ꞏ   ꞏ ꞏ ꞏ ꞏ   ꞏ , ꞏ   , & ꞏ ꞏ ꞏ
        (.), *PIRLS 2001 technical report.*   ꞏ ꞏ ■ ꞏ, ꞏ ꞏ   ꞏ   ꞏ ꞏ .

16   ꞏ   ꞏ   ꞏ   ꞏ   ,   , ꞏ ꞏ. (2003).   ꞏ   ꞏ   ꞏ ꞏ   ꞏ , ꞏ   ,
        ꞏꞏꞏ, & ꞏ ꞏ ꞏ   (.), *PIRLS 2001 technical report.*   ꞏ ꞏ ■ ꞏ, ꞏ ꞏ   ꞏ   ꞏ ꞏ .

Exhibit A.9:   PIRLS Within-Country Constructed-Response Scoring Reliability Data

	Minimum	Maximum
Argentina		
Belize		
Bulgaria		

Exhibit A.10:   PIRLS Cross-Country Constructed-Response Scoring Reliability   ISC | 4th Grade PIRLS 2001

Purpose	Item Label[1]	Total Valid Comparisons[2]	Exact Percent Agreement
Literary Experience	Unreleased C01	275496	99%
	Unreleased C02	275444	89%
	Unreleased C03	275548	93%
	Unreleased C06	275341	98%
	Unreleased C08	275496	92%
	Unreleased C10	275548	66%
	Unreleased C11	275444	72%
	Hare H03	275600	90%
	Hare H04	275393	93%
	Hare H07	275444	79%
	Hare H08	275086	84%
	Hare H09	275236	84%
	Hare H10	273661	73%
Acquire and Use Information	Unreleased A01	296892	96%
	Unreleased A03	296676	98%
	Unreleased A04	296676	90%
	Unreleased A07	296892	87%
	Unreleased A08	296623	80%
	Unreleased A09	296784	81%
	Unreleased A11	296191	80%
	Pufflings N07	274724	78%
	Pufflings N08	274724	83%
	Pufflings N10	273947	84%
	Pufflings N12	274673	76%
	Pufflings N13	274621	73%

Average Percent Agreement	85%

Exhibit A.11 :   Cronbach's Alpha Reliability Coefficient – PIRLS 2001

ISC 4th Grade
PIRLS 2001

Countries	Reliability Coefficient[1]
Argentina	0.90
Belize	0.87
Bulgaria	0.89
Canada (O,Q)	0.87
Colombia	0.87
Cyprus	0.90
Czech Republic	0.85
England	0.91
France	0.87
Germany	0.87
Greece	0.88
Hong Kong, SAR	0.85
Hungary	0.87
Iceland	0.89
Iran, Islamic Rep.	0.89
Israel	0.91
Italy	0.87
Kuwait	0.86
Latvia	0.85
Lithuania	0.85
Macedonia, Rep. of	0.91
Moldova, Rep. of	0.87
Morocco	0.90
Netherlands	0.83
New Zealand	0.91
Norway	0.89
Romania	0.91
Russian Federation	0.86
Scotland	0.90
Singapore	0.91
Slovak Republic	0.88
Slovenia	0.88
Sweden	0.85
Turkey	0.89
United States	0.90
International Median	0.88

for dichotomously-scored items (right or wrong), and generalized partial credit models for items with two or three available score points. The IRT scaling method produces a score by averaging the responses of each student to the items that he or she took which takes into account the difficulty and discriminating power of each item. The methodology used in PIRLS included refinements enabling reliable scores to be produced even though individual students responded to just two of the eight assessment passages. Achievement scales were produced for each of the two reading purposes (reading for literary experience and reading for information), as well as for reading overall. Exhibit A.12 presents the Pearson correlation coefficient indicating the linear relationship between the two reading purposes in each of the PIRLS countries.

The IRT methodology was preferred for developing comparable estimates of performance for all students, since students responded to different passages and items depending upon which of the test booklets they received (Booklets 1 through 9, or the PIRLS Reader). The IRT analysis provides a common scale on which performance can be compared across countries. In addition to providing a basis for estimating mean achievement, scale scores permit estimates of how students within countries vary and provide information on percentiles of performance. Treating all participating countries equally, the PIRLS scale average across countries was set to 500, and the standard deviation was set at 100. Since the countries varied in size, each country was weighted to contribute equally to the mean and standard deviation of the scale. The average and standard deviation of the scale scores are arbitrary and do not affect scale interpretation.

To allow more accurate estimation of summary statistics for student subpopulations, the PIRLS scaling made use of plausible-value technology, whereby five separate estimates of each student's score were generated on each scale – based on the student's responses to the items in the student's booklet, and on the student's background characteristics. The five score estimates are known as "plausible values," and the variability between them encapsulates the uncertainty inherent in the score estimation process.

Exhibit A.12:   **Correlation Between Reading for Literary Purposes and Reading for Informational Purposes**

Argentina	0.81
Belize	0.85
Bulgaria	0.85
Canada (O,Q)	0.82
Colombia	0.83
Cyprus	0.87
	0.87
Czech Republic10 1 TGQq1 Gr.mae1.mae1	
Bulgaria	0.85
	Bulgaria

Estimating Sampling Error

Because the statistics presented in this report are estimates of national performance based on samples of students – rather than on the values that could be calculated if every student in every country had answered every question – it is important to have measures for the degree of uncertainty of the estimates. The jackknife procedure was used to estimate the standard error associated with each statistic presented in this report.[17] The jackknife standard errors also include an error component due to variation between the five plausible values generated for each student. The use of confidence intervals (based on the standard errors) provides a way to make inferences about the population means and proportions in a manner that reflects the uncertainty associated with the sample estimates. An estimated sample statistic plus or minus two standard errors represents a 95 percent confidence interval for the corresponding population result.

Setting International Benchmarks of Student Achievement

To facilitate reporting of student reading achievement at a variety of performance levels, PIRLS identified four international benchmarks of student achievement. These benchmarks are the points on the PIRLS reading scale that separate the 10 percent of students located on top of the distribution, the top 25 percent of students, the top 50 percent, and the bottom 25 percent. The percentage of students in each country meeting or exceeding the international benchmarks is reported. The benchmarks correspond to the 90th, 75th, 50th, and 25th percentiles of the international distribution of achievement. When computing these percentiles, sampling weights were applied so that each country contributed as many students to the distribution as there were students in the target population in the country. That is, each country's contribution to setting the international benchmarks was proportional to the estimated population enrolled at the fourth grade.

In order to interpret the PIRLS scale scores and analyze achievement at the international benchmarks, PIRLS conducted a scale anchoring analysis to describe achievement of students at those four points on the scale.

17   t   t   t   , , , & ʌ   , ʌ   . (2003). t t t   t-
.   t   t   t, ,   , & ʌ ʌ   (.), *PIRLS 2001 technical report.*   t   t ▪   , ʌ l t   . .

Scale anchoring is a way of describing students' performance at different points