Appendix A

Technical Notes

Most of these notes are derived from the *PIRLS Technical Report*. For a full discussion of these topics, see the *PIRLS Technical Report* at **www.pirls.org**.

Background

PIRLS 2001 formed the Reading Development Group (RDG) to help construct the PIRLS 2001 Framework and to review and endorse the final reading test. The RDG consisted of Marilyn Binkley, Karl Blueml, Sue Horner, Pirjo Linnakyla, Martine Remond, Keen See Tan, and William Tunmer provided valuable support in the design of the assessment. Jay Campbell of Educational Testing Service served as a technical advisor to the RDG.

The RDG concluded that at least 4 hours of assessment material and 2 hours for each reading purpose (literary and informational) were needed to provide a valid and reliable measure of reading achievement. Since it would not be possible to administer the entire test to any one child, PIRLS 2001 used a matrix sampling technique to distribute the assessment material among students, yet retain linkages necessary for scaling the achievement data.

Assessment Design

The reading material was divided into 40-minute "blocks," each comprising a story or article and items representing at least 15 score points. There were eight such blocks, four for each reading purpose: literary and informational. The eight assessment blocks were distributed across 10 test booklets, and each student completed one booklet in an 80-minute testing session. Each booklet contained two blocks.

One of the 10 booklets was the PIRLS 2001 Reader, a color booklet containing two reading passages, which appeared only in that booklet. The distribution of blocks across booklets linked the booklets to enable the achievement data to be scaled using item response theory methods.

Student Population Assessed

In 30 of the 35 PIRLS 2001 countries, including the United States, the students who completed the assessment had received 4 years of formal schooling. Five countries that assessed students who had received formal schooling for a different number of years were Belize (4 or 5 years), England (5 years), the Russian Federation (3 or 4 years), Scotland (5 years), and Slovenia (3 years). However, all participating PIRLS 2001 countries assessed the reading literacy of their students in the upper of two grades with the most 9-year-olds.

In the United States, the mean age of the students who completed the PIRLS 2001 assessment was 10.2 years. The average mean age of students in countries with significantly lower average scores than the United States on the combined reading literacy scale was also 10.2 years, and the mean age of participating students in countries with significantly higher average scores than the United States was 10.4 years.

Passages

The reading passages formed the foundation of the reading literacy test. In accordance with the framework, four assessment blocks contained literary texts and four contained informational texts. The passages were authentic texts drawn from children's storybook and informational sources. Submitted and reviewed by PIRLS 2001 participating countries, the literary passages included realistic stories and traditional tales. The informational texts included chronological and nonchronological articles, a biographical article, and an informational leaflet.

PIRLS 2001 examined all passages and test items for cultural bias. A large number of passages were initially provided by participating countries. Only those that were selected by all countries became a part of PIRLS 2001. Test items were examined for individual item statistics and item-by-country interactions. The analysis indicated that items were considered unbiased to a similar extent by all of the participating countries.

Item Development and Scoring

Two item formats were used to assess children's reading literacy: multiple choice and constructed response. Each type of item was used to assess both reading purposes and all four reading processes. Multiple-choice items provided students with four possible answers, one of which was correct. Each multiple-choice item was worth one point. Constructed-response items required students to construct their answers rather than select from among possible answers. These items were worth one, two, or three points, depending on the depth of understanding or extent of textual support the item required.

Each block of assessment material contained from 11 to 14 items that together represented at least 15 score points. Altogether, the PIRLS 2001 reading test included 98 items representing 119 score points–enough to estimate achievement reliably.

Scale anchoring is a way of describing students' performance at different points on the reading achievement scale in terms of the types of texts they were asked to read and the types of items they answered successfully. It involved an empirical component in which items that discriminate between successive points on the scale were identified, and a judgmental component in which reading experts examined the content of the texts and items and generalized to students' comprehension skills and strategies. Criteria were applied to the reading achievement scale results to identify the sets of items that students reaching each international benchmark were likely to answer correctly and that those at the next lower benchmark were unlikely to answer correctly.

Translation

The PIRLS 2001 reading test and questionnaires were prepared in English and translated into 31 other languages. Countries were responsible for translating the instruments into their local language or languages following internationally prescribed procedures. To ensure standardization of instruments across countries, PIRLS 2001 undertook an extensive verification process, whereby each country's data collection instrument was independently reviewed and verified by an external translation company engaged by the IEA. Instruments were verified twice, once before the field test and again before the main data collection. Also, statistical analyses of item data were conducted to check for any evidence of differences in student performance across countries that could indicate translation problems. More information about translation issues in the assessment is available in the *PIRLS Technical Report* at **www.pirls.org.**

Sampling

PIRLS 2001 used a three-stage stratified cluster sample design:

1. The first stage was a sample of primary sampling units (geographic units referred to as PSUs).

2. The second stage consisted of a sample of at least 150 schools using probability-proportional-to-size sampling. Schools were stratified by geographical characteristics (such as states or provinces), school type (such as public or private), and the level of urbanization (such as rural or urban). The United States selected 174 schools after substitution for non-responding schools and tested 3,763 fourth-grade students.

3. The third stage consisted of sampling of one or more classrooms from the target grade in sampled schools. The target grade in each country was the upper of the two grades with the most 9-year-olds. In the United States and the majority of other countries, the target grade was the fourth grade. Each fourth-grade classroom in all selected schools had an equal likelihood of being selected. This resulted in a sample size of at least 3,750 students in each country. For more information about the grade levels that were assessed in each country, see the *PIRLS Technical Report* at **www.pirls.org**.

Exclusions in the PIRLS Sample

A major objective of PIRLS was that the target population, the population actually sampled by PIRLS, be as close as possible to the international desired population. Consequently, each country had to account for any exclusion of eligible students from the international desired population. This applied to school-level exclusions as well as within-school exclusions. Within-school exclusions included the following three groups:

Educable mentally disabled students. These are students who were considered, in the professional opinion of the school principal or other qualified staff members, to be educable mentally disabled, or who had been so diagnosed in psychological tests. This

category included students who were emotionally or mentally unable to follow even the general instructions of the PIRLS test. It did not include students who merely exhibited poor academic performance or discipline problems.

Functionally disabled students. These are students who were permanently physically disabled in such a way that they could not perform in the PIRLS tests. Functionally disabled students who could perform were included in the testing.

Non-native-language speakers. These are students who could not read or speak the language of the test and so could not overcome the language barrier of testing. Typically, a student who had received less than 1 year of instruction in the language of the test was excluded, but this definition was adapted in different countries.

School-level exclusions consisted of students in special education schools, students in vocational/technical schools, and students in alternative schools.

The United States produced a within-school exclusion rate of 4.7 percent and a school level exclusion rate of 0.6 percent, for a combined exclusion rate of 5.3 percent. Internationally, combined exclusion rates ranged from no exclusions in Kuwait to 22.4 percent in Israel, with an average rate of 3.8 percent.

Data Collection

Each country was responsible for carrying out all aspects of the data collection, using standardized procedures developed for the study by IEA. Manuals provided explicit instructions on all aspects of the data collection, from contacting sampled schools to packing and shipping materials to the IEA Data Processing Center in Hamburg, Germany. Manuals were also prepared for test administrators and for cooperating school officials. In all participating PIRLS 2001 countries, data were collected in the final months of the 2000–01 school year.

Quality Control

PIRLS 2001 also implemented an international program of site visits, whereby international quality control monitors visited a sample of 15 schools in each country and observed test administrations. PIRLS 2001 National Research Coordinators in each country were also expected to organize national quality control programs based on the international model, to ensure that data across countries were comparable. The national quality control monitors visited random samples of 10 percent of the schools (in addition to those visited by the international quality control monitors) and monitored testing sessions, recording their observations for later analysis.

Statistical Comparisons in This Report

Comparisons made in this highlights report have been tested for statistical significance. For example, in the commonly made comparison of country averages against the average of the United States, tests of statistical significance were used to establish whether or not the observed differences from the U.S. average were statistically significant.

In almost all instances the tests used were standard t-tests. These fell into two categories according to the nature of the comparison being made. In simple comparisons of country averages against the U.S. average or against the international average, the following formula was used to compute the t statistic:

$$t = (Est_1 - Est_2) \,/\, SQRT[(se_1)^2 + (se_2)^2]$$

Est_1 and Est_2 are the estimates being compared (e.g., average of country A and the U.S. average) and se_1 and se_2 are the corresponding standard errors of these averages.

In several places, between-country comparisons of group differences within countries were made. Comparisons of sex differences in other PIRLS 2001 countries against sex differences in the United States is an example. In these instances the following formula was used:

$$t = \frac{[(Est_{11} - Est_{21}) - (Est_{12} - Est_{22})]}{SQRT[(se_{11}^2 + se_{21}^2) + (se_{12}^2 + se_{22}^2)]}$$

Est_{11} and Est_{21} are the estimates being compared within country A (e.g., girls' reading average and boys' reading average), Est_{12} and Est_{22} are the corresponding estimates for the United States, and se_{11}, se_{21}, se_{12}, and se_{22} are their corresponding standard errors.