Highlights From PIRLS 2011

Reading Achievement of U.S. Fourth-Grade Students in an International Context

NCES 2013-010

U.S. DEPARTMENT OF EDUCATION

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

Page intentionally left blank

Highlights From PIRLS 2011

Reading Achievement of U.S. Fourth-Grade Students in an International Context

DECEMBER 2012

Sheila Thompson
Project Officer
National Center for Education Statistics

Stephen Provasnik
National Center for Education Statistics

David Kastberg
David Ferraro
Nita Lemanski
Stephen Roey
Frank Jenkins
Westat

NCES 2013-010
U.S. DEPARTMENT OF EDUCATION

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

U.S. Department of Education
Arne Duncan
Secretary

Institute of Education Sciences
John Q. Easton
Director

National Center for Education Statistics
Jack Buckley
Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high-priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high-quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public. Unless specifically noted, all information contained herein is in the public domain.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to

NCES, IES, U.S. Department of Education
1990 K Street NW
Washington, DC 20006-5651

December 2012

The NCES Home Page address is http://nces.ed.gov.
The NCES Publications and Products address is http://nces.ed.gov/pubsearch.

This publication is only available online. To download, view, and print the report as a PDF file, go to the NCES Publications and Products address shown above.

This report was prepared in part under Contract No. ED-04-CO-0059/0026 with Westat. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

Suggested Citation

Thompson, S., Provasnik, S., Kastberg, D., Ferraro, D., Lemanski, N., Roey, S., and Jenkins, F. (2012). *Highlights From PIRLS 2011: Reading Achievement of U.S. Fourth-Grade Students in an International Context* (NCES 2013–010). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. Government Printing Office.

Content Contact
Sheila Thompson
(202) 502-7425
Sheila.Thompson@ed.gov

Executive Summary

The Progress in International Reading Literacy Study (PIRLS) is an international comparative study of student achievement. In 2011, PIRLS was administered to nationally representative samples of 4th-grade students in 53 education systems around the world.[1] The PIRLS assessment measures student performance on a combined reading literacy scale, as well as two subscales of purposes of reading: reading for literary experience and reading to acquire and use information.[2]

This report compares the performance of U.S. students with their peers around the world and also examines how the reading literacy of U.S. 4th-grade students has changed since the first administration of PIRLS in 2001 and the previous administration in 2006.[3]

Results are presented by two student characteristics (sex and race/ethnicity) and by one measure of school poverty (percent of students in the school eligible for free or reduced-price lunch).

In the United States, one state, Florida, participated as a separate education system and is included in international comparisons. Total counts of education systems include Florida, not only as part of the U.S. national sample of public and private schools, but also individually with the state level public school sample. Note that because all education systems participating in PIRLS are treated equally, Florida is compared with the United States (national sample) throughout this report. All differences described in this report are statistically significant at the .05 level. No statistical adjustments to account for multiple comparisons were used.

Key findings for the reading literacy scale in 2011 include the following:

- The average score for U.S. students (556) was higher than the international PIRLS scale average, which is set to 500.[4]
- In 2011 the United States was among the top 13 education systems (5 education systems had higher averages and 7 were not measurably different). The United States average was higher than 40 education systems.
- The 5 education systems with average scores above the U.S. average were Hong Kong-CHN, Florida-USA, the Russian Federation, Finland, and Singapore.
- Compared with 2001, the U.S. average score was 14 points higher in 2011 (542 in 2001 vs. 556 in 2011).
- Compared with 2006, the U.S. average score was 16 points higher in 2011 (540 in 2006 vs. 556 in 2011).
- Considering the percentage of 4th-graders performing at or above the *Advanced* international reading benchmark: two education systems had a percentage that was higher than the United States, 7 education systems had percentages that were not measurably different than the United States, and 43 education systems had percentages lower than the United States.[5]
- The average score for girls was higher than the average scores for boys in the United States (562 vs. 551) and in the one education system separately assessed in the United States, Florida (576 vs. 561).
- Compared to the U.S. national average reading score: White, Asian, and multiracial students scored higher on average, while Black and Hispanic 4th-graders scored lower on average than the U.S. average.[6]
- In the United States, schools were classified into five categories on the basis of the percentage of students in the school eligible for free or reduced-price lunch. The percentage of students eligible and the average reading score in each category are as follows: less than 10 percent (605), 10 to 24.9 percent (584), 25 to 49.9 percent (568), 50 to 74.9 percent (544), and 75 percent or more (520). In all cases, children from schools with a lower level of free lunch eligibility had a higher average score than children from schools with a higher level of free lunch eligibility.

[1]For the purposes of this report "countries" are complete, independent political entities, whereas "other education systems" represent a portion of a country, nation, kingdom, or emirate or are other non-national entities (e.g., U.S. states, Canadian provinces, and Northern Ireland). The total number of education systems reported here differs from the total number reported in the international PIRLS reports (Mullis et al. 2012; Martin et al. forthcoming) because four education systems administered the PIRLS grade 4 assessment only to 5th- and 6th-grade students. Education systems that did not assess students at the target grade level are not counted or included in this report.
[2]The PIRLS 2011 International Report also presents results for two subscales of processes of comprehension: Retrieving and Straightforward Inferencing and Interpreting, Integrating, and Evaluation. In the interest of space, these results are not included here.
[3]In the United States, a total of 370 schools and 12,726 4th-grade students participated in 2011. The final weighted student response rate was 96 percent. The overall weighted school response rate before the use of substitute schools was 80 percent. The final weighted school response rate was 85 percent.
[4]The scores are reported on a scale from 0 to 1,000, with the PIRLS scale average set at 500 (the 2001 mean) and standard deviation set at 100.
[5]PIRLS reports on four benchmarks to describe student performance in reading. Each benchmark is associated with a score on the achievement scale and a description of the knowledge and skills demonstrated by students at that level of achievement. The *Advanced* international benchmark indicates that students scored 625 or higher. More information on the benchmarks can be found in the main body of the report and appendix A.
[6]The White, Asian, and Black categories are exclusive of Hispanics.

Page intentionally left blank

Acknowledgments

The authors wish to thank the students, teachers, and school officials who participated in PIRLS 2011. Without their assistance and cooperation, this study would not be possible. The authors also wish to thank all those who contributed to the PIRLS design, implementation, and data collection as well as the writing, production, and review of this report.

Page intentionally left blank

Contents

List of Tables

List of Figures

Introduction

PIRLS in brief

The Progress in International Reading Literacy Study (PIRLS) is an international comparative study of student achievement. PIRLS 2011 represents the third such study since PIRLS was first conducted in 2001. Developed and implemented by the International Association for the Evaluation of Educational Achievement (IEA), an international organization of national research institutions and governmental research agencies, PIRLS is used to measure the reading knowledge and skills of 4th-graders over time.

PIRLS is designed to align broadly with reading curricula in the participating education systems. The results, therefore, suggest the degree to which students have learned reading concepts and skills likely to have been taught in school. PIRLS also collects background information on students, teachers, schools, curricula, and official education policies in order to allow cross-national comparison of educational contexts that may be related to student achievement. In 2011, there were 57 education systems that participated in PIRLS (table 1).[1] For the purposes of this report, "countries" are complete, independent political entities, whereas "other education systems" represent a portion of a country, nation, kingdom, or emirate or other non-national entities. Thus the category "other education systems" includes the U.S. state of Florida and Canadian provinces that participated as "benchmarking participants"[2] as well as French Belgium-BEL, Chinese Taipei-CHN, England-GBR, Northern Ireland-GBR, and Hong Kong-CHN.

This report presents the performance of U.S. 4th-grade students relative to their peers in other countries and educational systems and reports changes in reading achievement since 2001. Most of the findings in the report are based on the results presented in the report published by the IEA and available online at http://www.pirls.org.

It is important to note that comparisons in this report treat all participating education systems equally, as is done in the international report. Thus, the United States is compared with some education systems that participated in the absence of a complete national sample (e.g., Northern Ireland-GBR participated but there was no national United Kingdom sample) as well as with some education systems that participated as part of a complete national sample (e.g., Florida-USA participated as a separate state sample of public schools and as part of the United States national sample of all schools).

Countries or Education Systems?

The international bodies that coordinate international assessments vary in the labels they apply to participating entities. For example, the IEA, which coordinates PIRLS and the Trends in International Mathematics and Science Study (TIMSS), differentiates between IEA members, which the IEA refers to as "countries" in all cases, and "benchmarking participants." IEA members include countries such as the United States and Ireland, as well as subnational entities such as England and Scotland (which are both part of the United Kingdom), the Flemish community of Belgium, the French community of Belgium, and Hong Kong-CHN, which is a Special Administrative Region of China. IEA benchmarking participants are all subnational entities and include U.S. states, Dubai in the United Arab Emirates, and, in 2011, participating Canadian provinces (among others). The Organization for Economic Cooperation and Development (OECD), which coordinates the Program for International Student Assessment (PISA), differentiates between OECD member countries and all other participating entities (called "partner countries" or "partner economies"), which include countries and subnational entities. In PISA, the United Kingdom and Belgium are reported as whole countries. Hong Kong-CHN is a PISA partner country, as are countries like Singapore, which is not an OECD member but is an IEA member.

In an effort to increase the comparability of results across the international assessments in which the United States participates, this report uses a standard international classification of nation-states (see the U.S. State Department list of "independent states" at at http://www.state.gov/s/inr/rls/4250.htm) to report separately "countries" and "other education systems," systems," which include all other non-national entities that received a PIRLS score. This report's tables and figures, which are primarily adapted from the IEA's PIRLS 2011 report, follow the IEA PIRLS convention of placing members and nonmembers in separate parts of the tables and figures in order to facilitate readers' moving between the international and U.S. national report. However, the text of this report will refer to "countries" and "other education systems," following the standard classification of nation-states.

[1]This total count of education systems also includes those that only gave the 4th-grade assessment to 5th- and 6th-graders.
[2]Subnational entities that are not members of the IEA can participate in PIRLS as *benchmarking participants*, which affords them the opportunity to assess the comparative international standing of their students' achievement and to view their curriculum and instruction in an international context.

For a number of countries and education systems, changes in achievement can be documented over the last 10 years, from 2001 to 2011. For those that began participating in PIRLS data collections in 2006, changes can be documented over 5 years. Table 1 shows the education systems that participated in PIRLS 2011 as well as their participation status in the earlier PIRLS data collections. The PIRLS assessment was implemented in 2001, 2006, and 2011.

This report describes additional details about the achievement of U.S. students that are not available in the international report, such as the achievement of students of different racial and ethnic and socioeconomic backgrounds. Results are presented in tables and figures, and in text summaries of the tables and figures. In the interest of brevity, in most cases, the text reports only the number of countries and other education systems scoring higher than the United States (not the number scoring lower than or not measurably different from the United States). Because all education systems participating in PIRLS are treated equally, comparisons are made throughout this report between the United States (national sample) and the U.S. state of Florida that participated in PIRLS 2011 not only as part of the U.S. national sample of public and private schools, but also individually with a state-level public school sample. When scoring higher than the U.S. national average, Florida results are also listed. A state summary for Florida is included in the section "Performance within the United States."

Defining and measuring reading literacy

PIRLS defines reading literacy as

> *the ability to understand and use those written language forms required by society and/or valued by the individual. Young readers can construct meaning from a variety of texts. They read to learn, to participate in communities of readers in school and everyday life, and for enjoyment.* (Mullis et al. 2012)

Within this context, the study examines three dimensions of reading literacy:

- purposes of reading;
- processes of comprehension; and
- reading behaviors and attitudes.

The distribution of PIRLS items across the first two dimensions, purposes of reading and processes of comprehension, are shown in table 2. Both dimensions were measured through the PIRLS assessment items administered to each participating student. The third dimension, reading behaviors and attitudes, was measured through a separate background questionnaire administered to participating students.

The *purposes of reading* dimension describes the two main reasons why young students read printed materials: (1) for literary experience and (2) to acquire and use information. Fictional texts are used to measure the ability of students to read for literary experience, and nonfictional texts are used to measure their skills at acquiring and using information.

The *processes of comprehension* dimension describes how young readers interpret and make sense of text. PIRLS assesses students' abilities to (1) focus on and retrieve explicitly stated information, (2) make straightforward inferences, (3) interpret and integrate ideas and information, and (4) examine and evaluate content, language, and textual elements.

Results from the PIRLS assessment are reported on the content subscales that measure the two *purposes of reading*: reading for literary experience and reading to acquire and use information. Additionally, results are reported on a combined reading literacy scale, which captures students' overall literacy skills related to both the content dimension measuring the *purposes of reading* and the cognitive dimension measuring the *process of comprehension*. This report emphasizes results from the combined reading literacy scale because the scale summarizes student performance on the two *purposes of reading* dimensions in a single measure.[3]

The texts for the PIRLS assessment were submitted from the participating education systems and reflect the kinds of printed materials read by children in those education systems. All participating education systems used the same texts. The passages were reviewed by the PIRLS Reading Development Group, an international advisory panel that selected texts for the assessment that reflected the cultures of participating educational systems.

[3]See appendix B for more information about the items comprising the PIRLS scales.

Table 1. Participation in the PIRLS assessment, by education system: 2001, 2006, and 2011

Education system	2001	2006	2011
Total count	**36**	**45**	**57**
Total IEA members count	**34**	**40**	**48**
Argentina	✓		
Australia			✓
Austria		✓	✓
Azerbaijan			✓
Belgium (Flemish)-BEL		✓	
Belgium (French)-BEL		✓	✓
Belize	✓		
Botswana[1]			✓
Bulgaria	✓	✓	✓
Canada			✓
Chinese Taipei-CHN		✓	✓
Colombia	✓		✓
Croatia			✓
Cyprus	✓		
Czech Republic	✓		✓
Denmark		✓	✓
England-GBR	✓	✓	✓
Finland			✓
France	✓	✓	✓
Georgia		✓	✓
Germany	✓	✓	✓
Greece	✓		
Honduras[1]			✓
Hong Kong-CHN	✓	✓	✓
Hungary	✓	✓	✓
Iceland	✓	✓	
Indonesia		✓	✓
Iran, Islamic Rep. of	✓	✓	✓
Ireland			✓
Israel[2]	✓	✓	✓
Italy	✓	✓	✓
Kuwait[1]	✓	✓	✓
Latvia	✓	✓	
Lithuania	✓	✓	✓
Luxembourg		✓	
Macedonia	✓	✓	
Malta			✓
Moldova	✓	✓	
Morocco[3]	✓	✓	✓
Netherlands	✓	✓	✓
New Zealand	✓	✓	✓
Northern Ireland-GBR			✓
Norway	✓	✓	✓
Oman			✓
Poland		✓	✓
Portugal			✓
Qatar[2]		✓	✓
Romania	✓	✓	✓
Russian Federation	✓	✓	✓
Saudi Arabia			✓
Scotland-GBR	✓	✓	
Singapore	✓	✓	✓
Slovak Republic	✓	✓	✓
Slovenia	✓	✓	✓
South Africa		✓	
Spain		✓	✓
Sweden	✓	✓	✓
Turkey	✓		
Trinidad and Tobago		✓	✓
United Arab Emirates			✓
United States	✓	✓	✓

Benchmarking education systems

Total benchmarking	2	5	9
Abu Dhabi-UAE			✓
Alberta-CAN		✓	✓
Andalusia-ESP			✓
British Columbia-CAN		✓	
Dubai-UAE			✓
Eng/Afr(5)-RSA[4]			✓
Florida-USA			✓
Maltese-MLT			✓
Nova Scotia-CAN		✓	
Ontario-CAN	✓	✓	✓
Quebec-CAN	✓	✓	✓

[1]Administered the PIRLS 4th-grade assessment to 6th-grade students in 2011.
[2]Participated but data not comparable for measuring trends to 2011, primarily due to countries improving translations or increasing population coverage.
[3]Administered the PIRLS 4th-grade assessment to a national sample of 4th-grade students and a national sample of 6th-grade students in 2011.
[4]Republic of South Africa (RSA) tested 5th-grade students receiving instruction in English (ENG) or Afrikaans (AFR).
NOTE: Only education systems that completed the necessary steps for their data to appear in the reports from the International Study Center are listed. Included are eight benchmarking education systems that qualified for reporting participation in the Progress in International Reading Literacy Study (PIRLS) 2011: the provinces of Alberta, Ontario, and Quebec in Canada; Andalusia of Spain; Abu Dhabi and Dubai, UAE; Maltese Malta; and the U.S. state of Florida. Information on these education systems can be found in the international PIRLS 2011 report. In order to be reported on, education systems were required to sample students enrolled in the grade corresponding to the fourth year of schooling, beginning with International Standard Classification of Education (ISCED) Level 1, providing that the mean age at the time of testing was at least 9.5 years. In the United States and most education systems, this corresponds to grade 4. See table A-1 in appendix A for details.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Progress in International Reading Literacy Study (PIRLS), 2011.

Design and administration of PIRLS 2011

PIRLS 2011 is sponsored by the IEA and carried out under a contract with the TIMSS & PIRLS International Study Center at Boston College.[4] The National Center for Education Statistics (NCES), in the Institute of Education Sciences at the U.S. Department of Education, is responsible for the implementation of PIRLS in the United States. Data collection in the United States was carried out under contract to Westat and its subcontractor, Pearson Educational Measurement.

Participating countries and other education systems administered PIRLS to a probability sample of 4th-grade students and schools, based on a standardized definition. Participating countries and other education systems were required to draw samples of students who were nearing the end of their fourth year of formal schooling, counting from the first year of the International Standard Classification of Education (ISCED) Level 1.[5] In most countries, including the United States, these students were in the 4th grade. Details on the grades assessed in each education system are included in appendix A.

In the United States, one sample was drawn to represent the nation at grade 4. In addition to this national sample, a state public school sample was also drawn at grade 4 for Florida, which chose to participate in PIRLS separately from the nation in order to benchmark their student performance internationally.

In the United States, PIRLS was administered between April and June of 2011. The U.S. national sample included both public and private schools, randomly selected and weighted to be representative of the nation at grade 4.[6] In total, 370 schools and 12,726 students participated in PIRLS. The weighted school response rate in the United States was 80 percent before the use of substitute schools (schools substituted for originally sampled schools that refused to participate).[7] Student response rates are based on a combined total of students from both sampled and substitute schools.

Detailed information on sampling, administration, response rates, and other technical issues are included in appendix A.

The reading assessment

A total of 10 reading passages, two from PIRLS 2001 and 2006, four from 2006 only, and four new passages, were included in the assessment booklets used in all participating education systems. The use of common passages in the 2001 and 2011 assessments allows the analysis of changes in reading literacy over the 10-year period between administrations for education systems that participated in both cycles. The passages, as well as all other study materials, were translated into the primary language or languages of instruction in each education system.

The reading assessment items vary in terms of difficulty and the form of knowledge and skills addressed. *PIRLS 2011 Assessment Framework and Specifications* (Mullis et al. 2009) provides a more detailed description of the content and cognitive areas assessed in PIRLS.

The PIRLS reading assessment is focused on two dimensions: (1) a content dimension specifying the purpose for reading and (2) a cognitive dimension specifying the cognitive or thinking processes of comprehension. The two content domains assessed in PIRLS, called "purposes of reading," are *literary experience* and *acquire and use information*. PIRLS assesses students' reading literacy in four cognitive areas, called "processes of comprehension": *focus on and retrieve explicitly stated information*; *make straightforward inferences*; *interpret and integrate ideas and information*; and *examine and evaluate content, language, and textual elements*.[8] Example items from the PIRLS reading assessment are included in appendix B (see items 1 through 5).

The proportion of item score points devoted to purposes of reading and, therefore, the contribution of the purposes of reading domain to the overall reading scale is roughly 50 percent (as shown in table 2). For example, in 2011, literary experience made up 52 percent of the PIRLS reading assessment, while 48 percent of the PIRLS assessment focused on acquiring and using information. Table 2 also reports the percentage of items in the four processes of comprehension. This indicates the contribution of each process to the overall reading scale.

[4]The international study center takes its name from the two main IEA studies it coordinates; the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS).
[5]The ISCED was developed by the United Nations Educational, Scientific, and Cultural Organization (UNESCO) to assist countries in providing comparable, cross-national data. ISCED Level 1 is termed primary schooling, and in the United States is equivalent to the first through sixth grades (Matheson et al. 1996).
[6]The sample frame for public schools in the United States was based on the 2011 National Assessment of Educational Progress (NAEP) sampling frame. The 2011 NAEP sampling frame was based on the 2007–08 Common Core of Data (CCD). The data for private schools are from the 2007–08 Private School Universe Survey (PSS). Any school containing at least one grade 4 class was included in the school sampling frame. For more information about the NAEP sampling frame, see http://nces.ed.gov/nationsreportcard/tdw/sample_design/.
[7]Two kinds of response rates are reported here in the interest of comparability with the PIRLS international reports that report response rates before and after "replacement." However, NCES standards advise that substitute schools should not be included in the calculation of response rates (Statistical Standard 1-3-8; National Center for Education Statistics 2002). Thus, response rates calculated before the use of substitute schools ("before replacement") are consistent with this standard, while response rates calculated with the inclusion of substitute schools ("after replacement") are not consistent with NCES standards.
[8]In the interest of space, this report presents results only for the combined scale and the two purposes of reading content domains.

Table 2. Percentage of score points attributed to the purposes of reading and processes of comprehension assessed in PIRLS 2011

Purposes of reading	Percent of assessment
Literary experience	52
Acquire and use information	48
Processes of comprehension	
Focus on and retrieve explicitly stated information	22
Make straightforward inferences	28
Interpret and integrate ideas and information	38
Examine and evaluate content, language, and textual elements	12

NOTE: The percentages in this table are based on the number of score points and not the number of items. Some constructed-response items are worth more than one score point. For the corresponding percentages based on the number of items, see table A-3 in appendix A. The purposes of reading define the specific reading subject matter covered by the assessment, and the processes of comprehension define the sets of behaviors expected of students as they engage with the respective subject's content. The processes of comprehension are defined by the same four sets of expected processing behaviors—*focus on and retrieve explicitly stated information; make straightforward inferences; interpret and integrate ideas and information; and examine and evaluate content, language, and textual elements.* Detail may not sum to totals because of rounding.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Progress in International Reading Literacy Study (PIRLS), 2011.

Reporting student results on PIRLS

PIRLS achievement results are reported on a scale from 0 to 1,000, with an initial international PIRLS scale average of 500 and standard deviation of 100 (established at the first administration of PIRLS in 2001). PIRLS provides an overall reading scale score as well as scores on the two purposes of reading subscales. The scaling of data is conducted separately for the combined reading literacy scale and for the literary subscale and the informational subscale.

Although each scale was created to have a mean of 500 and a standard deviation of 100, the subject matter and the level of difficulty of items necessarily differ between areas. Therefore, direct comparisons between scores (e.g., between subscale scores for reading for literary experience and reading to acquire and use information) should not be made. For more details explaining why such comparisons are not warranted, see the "Weighting, scaling, and plausible values" section in appendix A.

However, scores are comparable over time. The PIRLS scale was established originally to have a mean of 500 based on the average of all of the education systems that participated in PIRLS 2001. Successive PIRLS assessments since then (PIRLS 2006 and 2011) have scaled the achievement data so that scores are in the same metric from assessment to assessment.[9] Thus, for example, a score of 500 in reading in 2011 is equivalent to a score of 500 in reading in 2006 and in 2001. More information on how the PIRLS scale was created can be found in the "Weighting, scaling, and plausible values" section in appendix A.

In addition to scale scores, PIRLS also has developed international benchmarks for reading. The PIRLS international benchmarks provide a way to interpret the scale scores and to understand how students' proficiency in reading varies along the PIRLS scale. The PIRLS benchmarks describe four levels of student achievement in reading (*Advanced*, *High*, *Intermediate* and *Low*), based on the kinds of skills and knowledge students at each score cutpoint would need to successfully answer the reading items.

In general, the score cutpoints for the PIRLS benchmarks were set based on the distribution of students along the PIRLS scale. More information on the development of the benchmarks and the procedures used to set the score cutpoints can be found in the *PIRLS 2011 Technical Report* (Martin, Mullis, and Foy forthcoming).

All differences described in this report are statistically significant at the .05 level. No statistical adjustments to account for multiple comparisons were used. Differences that are statistically significant are discussed using comparative terms such as "higher" and "lower." Differences that are not statistically significant are either not discussed or referred to as "not measurably different" or "not statistically significant." In the latter case, failure to find a difference as statistically significant does not necessarily mean that there was no difference. It could be that a real difference cannot be detected by the significance test because of small sample size or imprecise measurement in the sample. If the statistical test is significant, this means that there is convincing evidence (although no guarantee) of a real difference in the population. However, it is important to remember that statistically significant results, even if they are believed to reflect real population differences, do not necessarily identify those findings that have policy significance or practical importance. Supplemental tables providing all estimates and standard errors discussed in this report are available online at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2013010.

All data presented in this report are used to describe relationships between variables. These data are not intended, nor can they be used, to imply causality. Student performance can be affected by a complex mix of educational and other factors that are not examined here.

[9]Even though the number and composition of education systems participating in PIRLS have changed between 2001 and 2011, comparisons between the 2011 results and prior results are still possible because the achievement scores in each of the PIRLS assessments are placed on a scale that is not dependent on the list of participating countries in any particular year. A brief description of the assessment equating and scaling is presented in appendix A to this volume. A more detailed presentation can be found in the *PIRLS 2011 Technical Report* (Martin, Mullis, and Foy forthcoming).

Nonresponse bias in the U.S. PIRLS sample

NCES standards require a nonresponse bias analysis if school-level response rates fall below 85 percent, as they did for the 4th-grade school sample in PIRLS 2011.[10] As a consequence, a nonresponse bias analysis was undertaken, similar to that used for TIMSS 2003 (Ferraro and Van de Kerckhove 2006).

Nonresponse bias analyses examined whether the participation status of schools (participant/nonparticipant) was related to seven school characteristics: the region of the education system in which the school was located (Northeast, Southeast, Central, West); the type of community served by the school (central city, urban fringe/large town, rural/small town); whether the school was public or private; percentage of students eligible for free or reduced-price lunch; number of students enrolled in grade 4; total number of students; and percentage of students from minority backgrounds. See appendix A for a detailed description of this analysis.

The findings indicate some potential for bias in the data arising from the fact that certain types of schools, including private and high-minority schools, were less likely to participate. The use of substitute schools, while not reducing the potential for bias, did not substantially increase the potential for bias. There are no significant group differences between participating and nonparticipating schools with respect to major demographic factors (e.g., gender, race/ethnicity), after substitute schools are included and nonresponse adjustments are applied.[11] This indicates that nonresponse adjustments eliminated almost all of the potential for nonresponse bias; however, this analysis only considered major demographic variables that are available on the school sampling frame. See appendix A for additional details on the findings.[12]

Further information

To assist the reader in understanding how PIRLS relates to the National Assessment of Educational Progress (NAEP), the primary source of national and state-level data on U.S. students' reading achievement, NCES compared the form and content of the PIRLS and NAEP reading assessments. A summary of the results of this comparison is included in appendix C. Appendix D includes a list of PIRLS publications and resources published by NCES and the IEA. Standard errors for the estimates discussed in the report are in appendix E, available online at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2013010. More information about U.S. participation in PIRLS is available at the NCES website at http://nces.ed.gov/surveys/pirls and the international PIRLS website at http://www.pirls.org

[10]See NCES Statistical Standard 2-2-2, available at: http://nces.ed.gov/statprog/2002/stdtoc.asp.

[11]The international weighting procedures created a nonresponse adjustment class for each explicit stratum; see the *TIMSS and PIRLS Methods and Procedures* (Martin and Mullis 2011) for details. In the case of the U.S. 4th-grade sample, 12 explicit strata were formed by poverty level, school control, and Census region. Beyond adjustments for explicit strata, the procedures could not be varied for individual countries to account for any specific needs. Therefore, the U.S. nonresponse bias analyses could have no influence on the weighting procedures and were undertaken after the weighting process was complete.

[12]Complete results of the nonresponse bias analysis conducted for PIRLS 2011 will be included in a technical report released with the U.S. national dataset.

Reading Literacy in the United States and Internationally

Average scores in 2011

In reading, the U.S. national average score was 556 (table 3). This score was higher than the PIRLS scale average, which is set at 500 for every administration of PIRLS. Five education systems had significantly higher average scores than the United States.[13]

Looking at all 53 education systems that participated in PIRLS at grade 4 (i.e., both countries and other education systems), the United States was among the top 13 education systems in average reading scores. The five education systems that had higher average scores were Hong Kong-CHN, Florida-USA, the Russian Federation, Finland, and Singapore. Seven education systems, Northern Ireland-GBR, Denmark, Croatia, Chinese Taipei-CHN, Ontario-CAN, Ireland, and England-GBR, had average scores not measurably different from the U.S. average score. The United States had higher average reading scores than 40 education systems.

Content scores in 2011

U.S. 4th-graders scored higher than the PIRLS scale average across the reading content domains in 2011 (table 3). U.S. 4th-graders' average scores were 563 in *literary experience* and 553 in *acquire and use information*, both above the PIRLS scale average of 500.

U.S. 4th-graders performed better on average in the *literary experience* domain than in the *acquire and use information* domain, at least in terms of comparisons with other education systems; that is, there were fewer education systems that outperformed the United States in *literacy experience* than in the *acquire and use information* domain. U.S. 4th-graders were outperformed by their peers in 1 country (Finland) and 1 education system (Florida-USA) in the *literary experience* domain, and 3 countries (Russian Federation, Singapore, and Finland,) and 2 education systems (Hong Kong-CHN, and Florida-USA) in the *acquire and use information* domain.

Change in scores

Several education systems participated in both the first administration of PIRLS in 2001 and the most recent administrations of PIRLS in 2006 and 2011. Comparing scores between previous administrations of PIRLS and the most recent administration provides perspective on change over time.

Changes between 2006 and 2011

Among those education systems that participated in both the 2006 and 2011 PIRLS assessments at grade 4 (24 countries and 7 other education systems), the average reading score increased in 10 countries (including the United States) and 3 other education systems and decreased in 7 countries and 1 other education system (figure 1). In the rest of the education systems that participated in PIRLS in both years, there was no measurable change in the average grade 4 reading scores between 2006 and 2011.

The U.S. average score for 4th-graders increased 16 score points (from 540 to 556). As a result of the U.S. increase and the changes in other education systems, the U.S. average went from below the averages of Alberta-CAN, Hungary, Italy, and Sweden in 2006 to above their averages in 2011, went from below the average score of Ontario-CAN in 2006 to not measurably different in 2011, and from not different from the average scores of Lithuania, Austria, Bulgaria, Germany, the Netherlands, and Quebec-CAN in 2006 to higher than their averages in 2011. No education systems with average scores lower than the United States in 2006 reached the U.S. average score in 2011; nor did any education system with average scores not measurably different from the U.S. average in 2006 surpass the U.S. average in 2011. Two countries (Iran and Trinidad and Tobago) had larger increases between 2006 and 2011 than did the United States.

Changes between 2001 and 2011

Among those that participated in both the 2001 and 2011 PIRLS assessments at grade 4 (19 countries and 4 education systems), the average reading score increased in 9 countries (including the United States) and 1 education system and decreased in 4 education systems (figure 1). In the rest of the education systems that participated in PIRLS in both years, there was no measurable change in the average grade 4 reading scores between 2001 and 2011.

The U.S. average score for 4th-graders increased 14 score points (from 542 to 556). In 4 countries and 1 other education system that participated in PIRLS in both 2001 and 2011, the average score of 4th-graders increased more than in the United States during this time: Iran, Hong Kong-CHN, the Russian Federation, Singapore, and Slovenia. As a result of changes in average scores among these and other education systems, the U.S. average went from below those of Sweden and the Netherlands in 2001 to above them in 2011, from below the average in England-GBR to not measurably different in 2011, from not different from the averages in Lithuania, Bulgaria, Hungary, Quebec-CAN, Italy, Germany, and the Czech Republic in 2001 to above their averages in 2011, and from above those of Singapore, Russian Federation, and Hong Kong-CHN in 2001 to below their averages in 2011 (and 2006).

[13]A score of 500 represents the international average of participants in the first administration of PIRLS in 2001. The PIRLS scale is the same in each administration such that a value of 500 in 2011 equals 500 in 2001.

Table 3. Overall reading average scale score and purposes of reading subscale scores of 4th-grade students, by education system: 2011

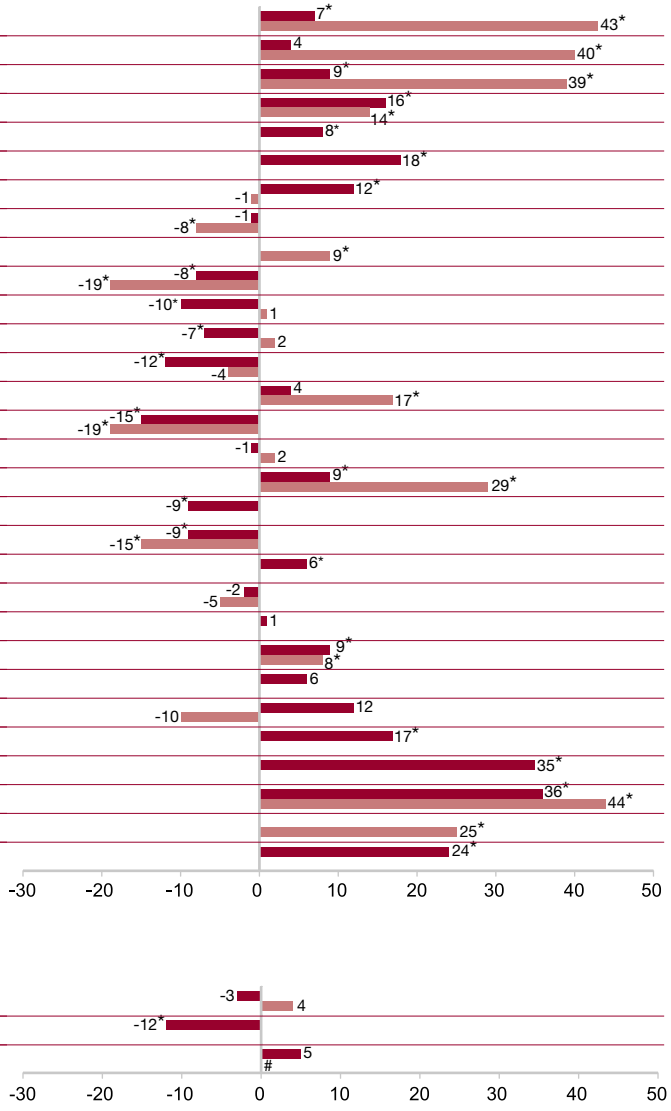
Education system	Overall reading average scale score	Purposes of reading: Literary experience	Purposes of reading: Acquire and use information
PIRLS scale average	500	500	500
Hong Kong-CHN[1]	571 ▲	565	578 ▲
Russian Federation	568 ▲	567	570 ▲
Finland	568 ▲	568 ▲	568 ▲
Singapore[2]	567 ▲	567	569 ▲
Northern Ireland-GBR[3]	558	564	555
United States[2]	**556**	**563**	**553**
Denmark[2]	554	555 ▼	553
Croatia[2]	553	555 ▼	552
Chinese Taipei-CHN	553	542 ▼	565 ▲
Ireland	552	557	549
England-GBR[3]	552	553 ▼	549
Canada[2]	548 ▼	553 ▼	545 ▼
Netherlands[3]	546 ▼	545 ▼	547 ▼
Czech Republic	545 ▼	545 ▼	545 ▼
Sweden	542 ▼	547 ▼	537 ▼
Italy	541 ▼	539 ▼	545 ▼
Germany	541 ▼	545 ▼	538 ▼
Israel[1]	541 ▼	542 ▼	541 ▼
Portugal	541 ▼	538 ▼	544 ▼
Hungary	539 ▼	542 ▼	536 ▼
Slovak Republic	535 ▼	540 ▼	530 ▼
Bulgaria	532 ▼	532 ▼	533 ▼
New Zealand	531 ▼	533 ▼	530 ▼
Slovenia	530 ▼	532 ▼	528 ▼
Austria	529 ▼	533 ▼	526 ▼
Lithuania[2,4]	528 ▼	529 ▼	527 ▼
Australia	527 ▼	527 ▼	528 ▼
Poland	526 ▼	531 ▼	519 ▼
France	520 ▼	521 ▼	519 ▼
Spain	513 ▼	516 ▼	512 ▼
Norway[5]	507 ▼	508 ▼	505 ▼
Belgium (French)-BEL[2,3]	506 ▼	508 ▼	504 ▼
Romania	502 ▼	504 ▼	500 ▼
Georgia[4,6]	488 ▼	491 ▼	482 ▼
Malta	477 ▼	470 ▼	485 ▼
Trinidad and Tobago	471 ▼	467 ▼	474 ▼
Azerbaijan[2,6]	462 ▼	461 ▼	460 ▼
Iran, Islamic Rep. of	457 ▼	459 ▼	455 ▼
Colombia	448 ▼	453 ▼	440 ▼
United Arab Emirates	439 ▼	427 ▼	452 ▼
Saudi Arabia	430 ▼	422 ▼	440 ▼
Indonesia	428 ▼	418 ▼	439 ▼
Qatar[2]	425 ▼	415 ▼	436 ▼
Oman[7]	391 ▼	379 ▼	404 ▼
Morocco[8]	310 ▼	299 ▼	321 ▼
Benchmarking education systems			
Florida-USA[1,4]	569 ▲	577 ▲	564 ▲
Ontario-CAN[2]	552	558	549
Alberta-CAN[2]	548 ▼	552 ▼	545 ▼
Quebec-CAN	538 ▼	539 ▼	536 ▼
Andalusia-ESP	515 ▼	518 ▼	512 ▼
Dubai-UAE	476 ▼	466 ▼	488 ▼
Maltese-MLT	457 ▼	458 ▼	455 ▼
Abu Dhabi-UAE	424 ▼	414 ▼	437 ▼

▲ Score is higher than U.S. average score.
▼ Score is lower than U.S. average score.
[1]National Defined Population covers less than 90 percent of National Target Population.
[2]National Defined Population covers 90 percent to 95 percent of National Target Population.
[3]Met guidelines for sample participation rates only after replacement schools were included.
[4]National Target Population does not include all of the International Target Population.
[5]Nearly satisfied guidelines for sample participation rates after replacement schools were included.
[6]Exclusion rates for Azerbaijan and Georgia are slightly underestimated as some conflict zones were not covered and no official statistics were available.
[7]The TIMSS & PIRLS International Study Center has reservations about the reliability of the average achievement score because the percentage of students with achievement too low for estimation exceeds 15 percent, though it is less than 25 percent.
[8]The TIMSS & PIRLS International Study Center has reservations about the reliability of the average achievement score because the percentage of students with achievement too low for estimation exceeds 25 percent.
NOTE: Education systems are ordered by 2011 average score. Italics indicate participants identified and counted in this report as an education system and not as a separate country. Participants that did not administer PIRLS at the target grade are not shown; see the international report for their results. All Florida-USA data are based on public school students only. All average scores reported as higher or lower than the U.S. average score are different at the .05 level of statistical significance. The tests for significance take into account the standard error for the reported difference. Thus, a small difference between the United States and one education system may be significant while a large apparent difference between the United States and another education system may not be significant. The standard errors of the estimates are shown in table E-1 available at http://nces.ed.gov/pubsearch/pubsinfor.asp?pubid=2013010.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Progress in International Reading Literacy Study (PIRLS), 2011.

Figure 1. Change in average reading scale scores of 4th-grade students, by education system: 2006 to 2011 and 2001 to 2011

Average score			
2001	2006	2011	Education system
528 ▼	564 ▲	571 ▲	*Hong Kong-CHN*[2]
528 ▼	565 ▲	568 ▲	Russian Federation
528 ▼	558 ▲	567 ▲	Singapore[3]
542	**540**	**556**	**United States[3]**
	546	554	Denmark[3]
	535	553	*Chinese Taipei-CHN*
553 ▲	539	552	*England-GBR*[4]
554 ▲	547	546 ▼	Netherlands[4]
537		545 ▼	Czech Republic
561 ▲	549 ▲	542 ▼	Sweden
541	551 ▲	541 ▼	Italy
539	548	541 ▼	Germany
543	551 ▲	539 ▼	Hungary
518 ▼	531 ▼	535 ▼	Slovak Republic
550	547	532 ▼	Bulgaria
529 ▼	532 ▼	531 ▼	New Zealand
502 ▼	522 ▼	530 ▼	Slovenia
	538	529 ▼	Austria
543	537	528 ▼	Lithuania[3,5]
	519 ▼	526 ▼	Poland
525 ▼	522 ▼	520 ▼	France
	513 ▼	513 ▼	Spain
499 ▼	498 ▼	507 ▼	Norway[6]
	500 ▼	506 ▼	*Belgium (French)-BEL*[3,4]
512 ▼	489 ▼	502 ▼	Romania
	471 ▼	488 ▼	Georgia[5,7]
	436 ▼	471 ▼	Trinidad and Tobago
414 ▼	421 ▼	457 ▼	Iran, Islamic Rep. of
422 ▼		448 ▼	Colombia
	405 ▼	428 ▼	Indonesia

			Benchmarking education systems
548	555 ▲	552	*Ontario-CAN*[3]
	560 ▲	548 ▼	*Alberta-CAN*[3]
537	533	538 ▼	*Quebec-CAN*



See notes at end of figure.

▲ Score is higher than U.S. average score.
▼ Score is lower than U.S. average score.
■ Change from 2006 to 2011.
■ Change from 2001 to 2011.
Rounds to zero.
*p<.05. Change in average scores is statistically significant.
[1]Differences are calculated by subtracting the 2006 from the 2011 estimate and the 2001 from the 2011 estimate, using unrounded numbers.
[2]National Defined Population covers less than 90 percent of National Target Population for 2011 (see appendix A).
[3]National Defined Population covers 90 percent to 95 percent of National Target Population for 2011 (see appendix A).
[4]Met guidelines for sample participation rates only after replacement schools were included for 2011.
[5]National Target Population does not include all of the International Target Population for 2011 (see appendix A).
[6]Nearly satisfied guidelines for sample participation rates after replacement schools were included for 2011.
[7]Exclusion rates for Georgia are slightly underestimated as some conflict zones were not covered and no official statistics were available for 2011.
NOTE: Education systems are ordered by 2011 average scores. All education systems at all years of assessment met international sampling and other guidelines in 2011, except as noted. Data are not shown for some education systems because comparable data from previous cycles are not available. Participants that did not administer PIRLS at the target grade are not shown; see the international report for their results. All Florida-USA data are based on public school students only. For 2001, Lithuania had a National Target Population that did not include all of the International Target Population; England-GBR, the Russian Federation, and the United States had a National Defined Population that covered 90 percent to 95 percent of the National Target Population; England-GBR, the Netherlands, and the United States met guidelines for sample participation rates only after replacement schools were included. For 2006, Georgia and Lithuania had a National Target Population that did not include all of the International Target Population; Georgia, the Russian Federation, Alberta-CAN, and Ontario-CAN had a National Defined Population that covered 90 percent to 95 percent of the National Target Population; the Netherlands and the United States met guidelines for sample participation rates only after replacement schools were included; Norway nearly satisfied guidelines for sample participation rates after replacement schools were included. All average scores reported as higher or lower than the U.S. average score are different at the .05 level of statistical significance. The tests for significance take into account the standard error for the reported difference. Thus, a small difference between the United States and one education system may be significant while a large difference between the United States and another education system may not be significant. Detail may not sum to totals because of rounding. The standard errors of the estimates are shown in table E-2 available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2013010.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Progress in International Reading Literacy Study (PIRLS), 2001, 2006, and 2011.

Performance on the PIRLS international benchmarks

The PIRLS international benchmarks provide a way to understand how students' proficiency in reading varies along the PIRLS scale (table 4). PIRLS defines four levels of student achievement: *Advanced*, *High*, *Intermediate,* and *Low*. The benchmarks can then be used to describe the kinds of skills and knowledge students at each score cutpoint needed to successfully answer the reading items included in the assessment.

In 2011, higher percentages of U.S. 4th-graders performed at or above each of the four PIRLS international benchmarks than the international medians[14] (figure 2). For example, 17 percent of U.S. 4th-graders performed at or above the *Advanced* benchmark (625) compared to the international median of 8 percent. These students demonstrated an ability to apply their understanding and knowledge to a variety of relatively complex reading situations (see description in table 4).

The percentage of 4th-graders performing at or above the *Advanced* international reading benchmark was higher than in the United States in 2 education systems; was not different in 7 education systems; and was lower than in the United States in 43 education systems.

Singapore and Florida-USA had a higher percentage of students performing at or above the *Advanced* international reading benchmark than the United States, and the Russian Federation, Northern Ireland-GBR, Finland, England-GBR, Hong Kong-CHN, Ireland, and Ontario-CAN had percentages not measurably different from the U.S. percentage.

At the other end of the scale, 98 percent of U.S. 4th-graders performed at or above the *Low* benchmark (400) compared to the international median of 95 percent. These students showed at least some basic reading skills by demonstrating an ability to retrieve explicitly stated details from literary or informational texts.

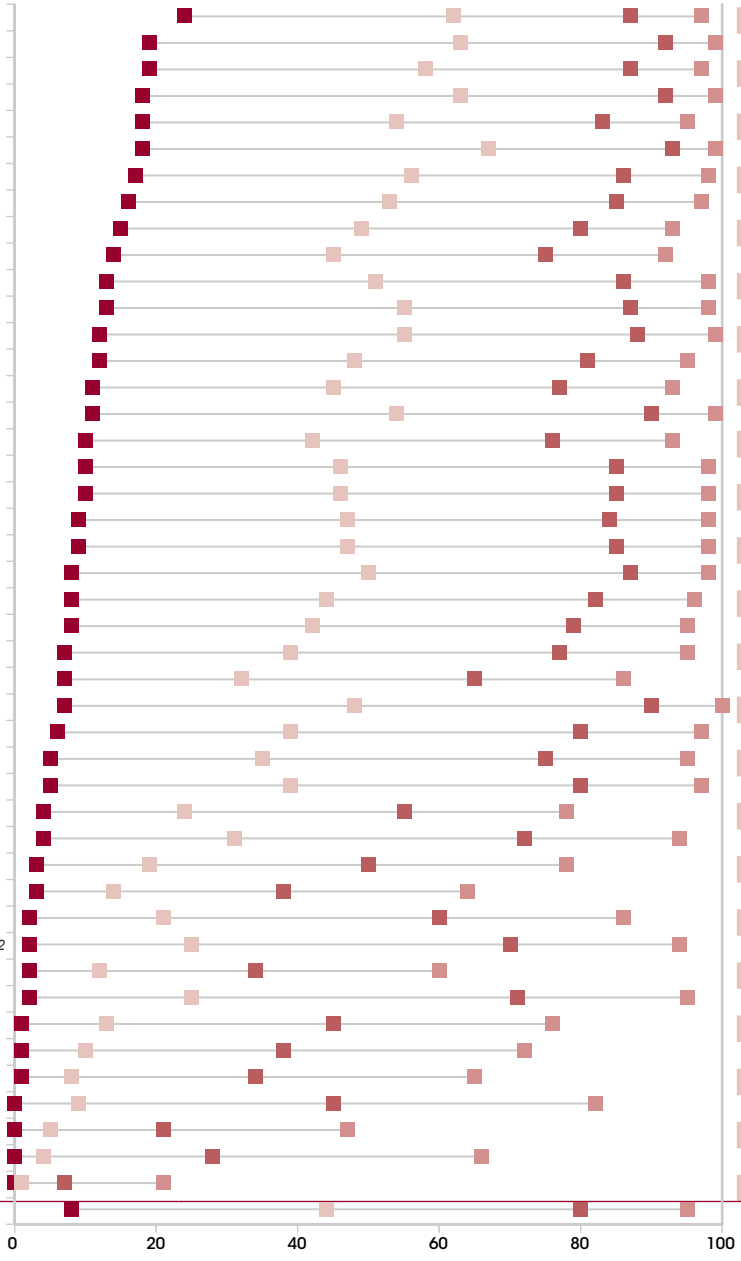
[14]The *international median* is the median percentage for all IEA member education systems (see the inset box on page 1 for IEA member education systems). Thus, the international median at each benchmark represents the percentage at which half of the participating IEA member education systems have that percentage of students at or above the median and half have that percentage of students below the median. For example, the *Low* international benchmark median of 95 percent at grade 4 indicates that half of the education systems have 95 percent or more of their students who met the *Low* benchmark, and half have less than 95 percent of their students who met the *Low* benchmark.

Table 4. Description of PIRLS international reading benchmarks: 2011

Benchmark (score cutpoint)	Description
Advanced (625)	Interpret figurative language Distinguish and interpret complex information from different parts of text Integrate ideas across text to provide interpretations about characters' feelings and behaviors
High (550)	Recognize some textual features, such as figurative language and abstract messages Make inferences on the basis of abstract or embedded information Integrate information to recognize main ideas and provide explanations
Intermediate (475)	Identify central events, plot sequences, and relevant story details Make straightforward inferences from the text Begin to make connections across parts of the text
Low (400)	Retrieve explicitly stated details from literary and informational texts

NOTE: Score cutpoints for the international benchmarks are determined through scale anchoring. Scale anchoring involves selecting benchmarks (scale points) on the achievement scales to be described in terms of student performance and then identifying items that students scoring at the anchor points can answer correctly. The score cutpoints are set at equal intervals along the achievement scales. The score cutpoints were selected to be as close as possible to the standard percentile cutpoints (i.e., 90th, 75th, 50th, and 25th percentiles). More information on the setting of the score cutpoints can be found in appendix A and Martin et al. (2012).
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Progress in International Reading Literacy Study (PIRLS), 2011.

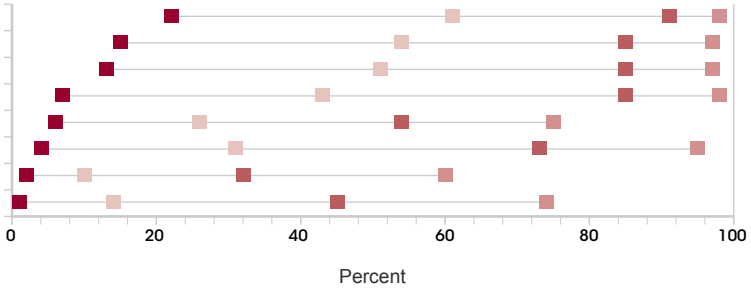
Figure 2. Percentage of 4th-grade students reaching the PIRLS international benchmarks in reading, by education system: 2011



Education system	Advanced (625)	High (550)	Intermediate (475)	Low (400)
	Percentage of students reaching each international benchmark			
Singapore[1]	24 *	62 *	87	97
Russian Federation	19	63 *	92 *	99 *
Northern Ireland-GBR[2]	19	58	87	97
Finland	18	63 *	92 *	99 *
England-GBR[2]	18	54	83 *	95 *
Hong Kong-CHN[3]	18	67 *	93 *	99 *
United States[1]	**17**	**56**	**86**	**98**
Ireland	16	53	85	97
Israel[3]	15 *	49 *	80 *	93 *
New Zealand	14 *	45 *	75 *	92 *
Canada[1]	13 *	51 *	86	98
Chinese Taipei-CHN	13 *	55	87	98
Denmark[1]	12 *	55	88 *	99 *
Hungary	12 *	48 *	81 *	95 *
Bulgaria	11 *	45 *	77 *	93 *
Croatia[1]	11 *	54	90 *	99 *
Australia	10 *	42 *	76 *	93 *
Italy	10 *	46 *	85	98
Germany	10 *	46 *	85	98
Portugal	9 *	47 *	84	98
Sweden	9 *	47 *	85	98
Czech Republic	8 *	50 *	87	98
Slovak Republic	8 *	44 *	82 *	96 *
Slovenia	8 *	42 *	79 *	95 *
Poland	7 *	39 *	77 *	95 *
Romania	7 *	32 *	65 *	86 *
Netherlands[2]	7 *	48 *	90 *	100 *
Lithuania[1,4]	6 *	39 *	80 *	97 *
France	5 *	35 *	75 *	95 *
Austria	5 *	39 *	80 *	97
Malta	4 *	24 *	55 *	78 *
Spain	4 *	31 *	72 *	94 *
Trinidad and Tobago	3 *	19 *	50 *	78 *
United Arab Emirates	3 *	14 *	38 *	64 *
Georgia[4,5]	2 *	21 *	60 *	86 *
Belgium (French)-BEL[1,2]	2 *	25 *	70 *	94 *
Qatar[1]	2 *	12 *	34 *	60 *
Norway[6]	2 *	25 *	71 *	95 *
Iran, Islamic Rep. of	1 *	13 *	45 *	76 *
Colombia	1 *	10 *	38 *	72 *
Saudi Arabia	1 *	8 *	34 *	65 *
Azerbaijan[1,5]	# *	9 *	45 *	82 *
Oman[7]	# *	5 *	21 *	47 *
Indonesia	# *	4 *	28 *	66 *
Morocco[8]	# *	1 *	7 *	21 *
International Median	**8 ***	**44 ***	**80 ***	**95 ***

See notes at end of figure.

Figure 2. Percentage of 4th-grade students reaching the PIRLS international benchmarks in reading, by education system: 2011—Continued



Benchmarking education systems	Percentage of students reaching each international benchmark			
	Advanced	High	Inter-mediate	Low
	(625)	(550)	(475)	(400)
Florida-USA[3,4]	22 *	61 *	91 *	98
Ontario-CAN[1]	15	54	85	97
Alberta-CAN[1]	13 *	51 *	85	97
Quebec-CAN	7 *	43 *	85	98
Dubai-UAE	6 *	26 *	54 *	75 *
Andalusia-ESP	4 *	31 *	73 *	95 *
Abu Dhabi-UAE	2 *	10 *	32 *	60 *
Maltese-MLT	1 *	14 *	45 *	74 *

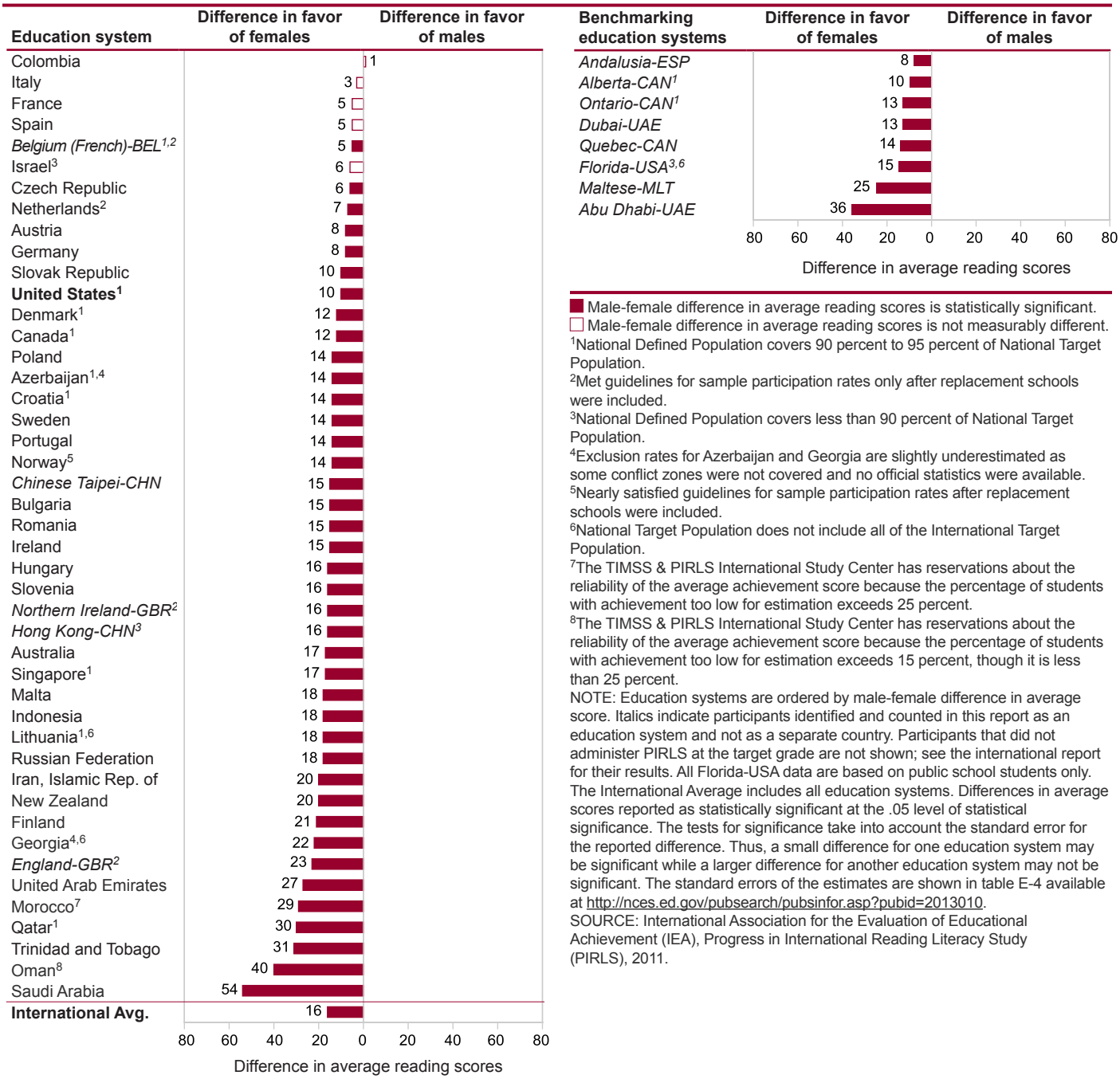
Advanced benchmark
High benchmark
Intermediate benchmark
Low benchmark

Rounds to zero.
**p*<.05. Percentage is significantly different from the U.S. percentage at the same benchmark.
[1]National Defined Population covers 90 percent to 95 percent of National Target Population.
[2]Met guidelines for sample participation rates only after replacement schools were included.
[3]National Defined Population covers less than 90 percent of National Target Population.
[4]National Target Population does not include all of the International Target Population.
[5]Exclusion rates for Azerbaijan and Georgia are slightly underestimated as some conflict zones were not covered and no official statistics were available.
[6]Nearly satisfied guidelines for sample participation rates after replacement schools were included.
[7]The TIMSS & PIRLS International Study Center has reservations about the reliability of the average achievement score because the percentage of students with achievement too low for estimation exceeds 15 percent, though it is less than 25 percent.
[8]The TIMSS & PIRLS International Study Center has reservations about the reliability of the average achievement score because the percentage of students with achievement too low for estimation exceeds 25 percent.
NOTE: Education systems are ordered by percentage at *Advanced* international benchmark. Italics indicate participants identified and counted in this report as an education system and not as a separate country. The PIRLS international median represents all participating PIRLS education systems, including the United States. The international median represents the percentage at which half of the participating education systems have that percentage of students at or above the median and half have that percentage of students below the median. Participants that did not administer PIRLS at the target grade are not shown; see the international report for their results. All Florida-USA data are based on public school students only. The tests for significance take into account the standard error for the reported difference. Thus, a small difference between the United States and one education system may be significant while a large difference between the United States and another education system may not be significant. The standard errors of the estimates are shown in table E-3 available at http://nces.ed.gov/pubsearch/pubsinfor.asp?pubid=2013010.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Progress in International Reading Literacy Study (PIRLS), 2011.

Average scores of male and female students

In 2011, U.S. 4th-grade females outperformed males by 10 score points on average (figure 3). Among all 53 education systems, 47 showed a significant difference in the average reading scores of males and females, all in favor of females. The difference in average scores between males and females ranged from 54 score points in Saudi Arabia to no measurable difference in 5 countries (Colombia, Italy, France, Spain, and Israel).

Figure 3. Difference in average reading scores of 4th-grade students, by sex and education system: 2011



Male-female difference in average reading scores is statistically significant.
Male-female difference in average reading scores is not measurably different.
[1]National Defined Population covers 90 percent to 95 percent of National Target Population.
[2]Met guidelines for sample participation rates only after replacement schools were included.
[3]National Defined Population covers less than 90 percent of National Target Population.
[4]Exclusion rates for Azerbaijan and Georgia are slightly underestimated as some conflict zones were not covered and no official statistics were available.
[5]Nearly satisfied guidelines for sample participation rates after replacement schools were included.
[6]National Target Population does not include all of the International Target Population.
[7]The TIMSS & PIRLS International Study Center has reservations about the reliability of the average achievement score because the percentage of students with achievement too low for estimation exceeds 25 percent.
[8]The TIMSS & PIRLS International Study Center has reservations about the reliability of the average achievement score because the percentage of students with achievement too low for estimation exceeds 15 percent, though it is less than 25 percent.
NOTE: Education systems are ordered by male-female difference in average score. Italics indicate participants identified and counted in this report as an education system and not as a separate country. Participants that did not administer PIRLS at the target grade are not shown; see the international report for their results. All Florida-USA data are based on public school students only. The International Average includes all education systems. Differences in average scores reported as statistically significant at the .05 level of statistical significance. The tests for significance take into account the standard error for the reported difference. Thus, a small difference for one education system may be significant while a larger difference for another education system may not be significant. The standard errors of the estimates are shown in table E-4 available at http://nces.ed.gov/pubsearch/pubsinfor.asp?pubid=2013010.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Progress in International Reading Literacy Study (PIRLS), 2011.

Performance within the United States

In 2011, PIRLS sampled enough schools and students in the United States to provide separate average reading scores for students by race/ethnicity and schools serving varying percentages of students from low-income families. In addition, PIRLS sampled enough schools and students in Florida to provide state results for these subpopulations among public school students. The reading results for Florida are reported at the end of this section.

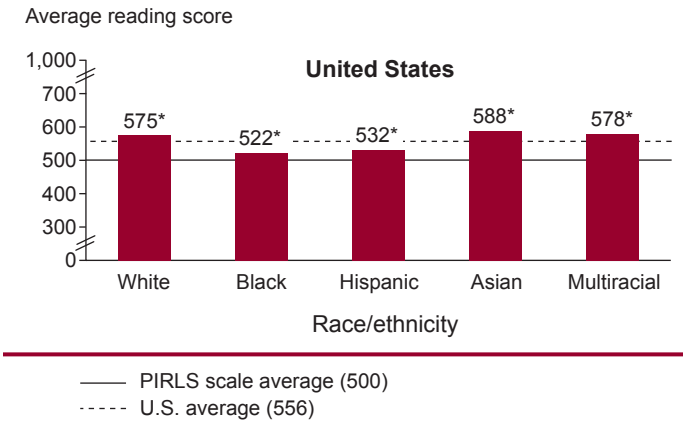
Average scores of students of different races and ethnicities

In 2011, all race/ethnicity groups (White, Black, Hispanic, Asian, and multiracial) of U.S. 4th-graders scored higher on average than the PIRLS scale average in reading (figure 4). In comparison to the U.S. national average, U.S. White, Asian, and multiracial 4th-graders scored higher, on average, while U.S. Black and Hispanic 4th-graders scored lower on average.

Average scores of students attending public schools of various poverty levels

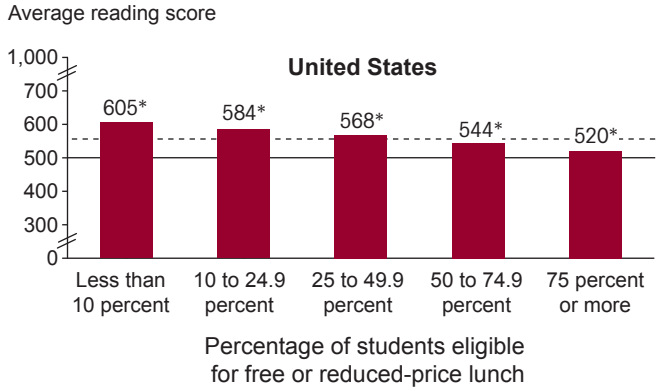
The U.S. results are also arrayed by the concentration of low-income enrollment in the public schools, as measured by eligibility for free or reduced-price lunch, and shown in relation to the PIRLS scale average and the U.S. national average. In comparison to the PIRLS scale average, the average reading score of U.S. 4th-graders in each of the categories of school poverty was higher than the PIRLS scale average (figure 5). In comparison to the U.S. national average score, 4th-graders in schools in very low to moderate poverty (from less than 10 percent to almost 50 percent of students eligible for free or reduced-price lunch) scored higher, on average, while those in schools with higher proportions of poverty (50 percent to 75 percent or more of students eligible for free or reduced-price lunch) scored lower, on average.

Figure 4. Average reading scores of U.S. 4th-grade students, by race/ethnicity: 2011



*p<.05. Significantly different from the U.S. average score.
NOTE: Reporting standards were not met for American Indian/Alaska Native and Native Hawaiian/Other Pacific Islander. Black includes African American, Hispanic includes Latino, and Asian includes Pacific Islander and Native Hawaiian. Racial categories exclude Hispanic origin. Students who identified themselves as being of Hispanic origin were classified as Hispanic, regardless of their race. Although data for some race/ethnicities are not shown separately because the reporting standards were not met, they are included in the U.S. and state totals shown throughout the report. See appendix A in this report for more information. The standard errors of the estimates are shown in table E-5 available at http://nces.ed.gov/pubsearch/pubsinfor.asp?pubid=2013010.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Progress in International Reading Literacy Study (PIRLS), 2011.

Figure 5. Average reading scores of U.S. 4th-grade students, by percentage of students in public school eligible for free or reduced-price lunch: 2011



*p<.05. Significantly different from the U.S. average score.
NOTE: Analyses are limited to public schools only, based on school reports of the percentage of students in public school eligible for the federal free or reduced-price lunch program. The standard errors of the estimates are shown in table E-6 available at http://nces.ed.gov/pubsearch/pubsinfor asp?pubid=2013010.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Progress in International Reading Literacy Study (PIRLS), 2011.

Page intentionally left blank

PIRLS 2011 Results for Florida Compared to Education Systems Outside the United States

Reading

- In Florida public schools, 4th-grade students' average score was 569 (see table 3). No education system scored higher than Florida, 4 were not measurably different, and 48 scored lower (table 5).
- Higher percentages of Florida 4th-graders performed at or above each of the four PIRLS international benchmarks than the international medians (figure 2). For example, 22 percent of 4th-graders in Florida performed at or above the *Advanced* benchmark (625) compared to the international median of 8 percent at grade 4.
- Females outperformed males by 15 score points on average in reading at grade 4 (figure 3).
- Male and female students in Florida scored higher in reading than the PIRLS international scale average (table 6).
- All racial and ethnic groups scored higher than the PIRLS international scale average.
- All categories of public school students eligible for free or reduced-price lunch scored higher than the PIRLS international scale average.

Table 5. Average reading scores of 4th-grade students in Florida public schools compared with other participating education systems: 2011

Education systems not measurably different from Florida	
Hong Kong-CHN	Finland
Russian Federation	Singapore
Education systems lower than Florida	
Northern Ireland-GBR	Lithuania
United States	Australia
Denmark	Poland
Croatia	France
Chinese Taipei-CHN	*Andalusia-ESP*
Ontario-CAN	Spain
Ireland	Norway
England-GBR	*Belgium (French)-BEL*
Canada	Romania
Alberta-CAN	Georgia
Netherlands	Malta
Czech Republic	*Dubai-UAE*
Sweden	Trinidad and Tobago
Italy	Azerbaijan
Germany	Iran, Islamic Rep. of
Israel	*Maltese-Malta*
Portugal	Colombia
Hungary	United Arab Emirates
Quebec-CAN	Saudi Arabia
Slovak Republic	Indonesia
Bulgaria	Qatar
New Zealand	*Abu Dhabi-UAE*
Slovenia	Oman
Austria	Morocco

NOTE: No education systems scored significantly higher than Florida-USA at $p<.05$.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Progress in International Reading Literacy Study (PIRLS), 2011.

Table 6. Average reading scores of 4th-grade students in Florida public schools, by sex, race/ethnicity, and percentage of students in public school eligible for free or reduced-price lunch: 2011

Reporting groups	Average score
PIRLS scale average	**500**
U.S. average	**556 ***
Florida average	**569 ***
Sex	
Female	576 *
Male	561 *
Race/ethnicity	
White	591 *
Black	537 *
Hispanic	564 *
Asian	604 *
Multiracial	591 *
Percentage of public school students eligible for free or reduced-price lunch	
Less than 10 percent	601 *
10 to 24.9 percent	610 *
25 to 49.9 percent	587 *
50 to 74.9 percent	566 *
75 percent or more	544 *

*$p<.05$. Difference between state score and PIRLS scale average is statistically significant.
NOTE: Reporting standards were not met for American Indian/Alaska Native and Native Hawaiian/Other Pacific Islander. Black includes African American and Hispanic includes Latino. Race categories exclude Hispanic origin. Students who identified themselves as being of Hispanic origin were classified as Hispanic, regardless of their race. Not all race/ethnicity categories are shown but they are included in the U.S. and state totals shown throughout the report. Multiracial students are those that identify themselves with more than one race. The standard errors of the estimates are shown in table E-7 available at http://nces.ed.gov/pubsearch/pubsinfor.asp?pubid=2013010.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Progress in International Reading Literacy Study (PIRLS), 2011.

Page intentionally left blank

References

Beaton, A.E., and González, E. (1995). *The NAEP Primer*. Chestnut Hill, MA: Boston College.

Chowdhury, S., Chu, A., and Kaufman, S. (2001). Minimizing Overlap in NCES Surveys. *Proceedings of the Survey Methods Research Section,* American Statistical Association, pp. 174-179.

Ferraro, D., and Van de Kerckhove, W. (2006). *Trends in International Mathematics and Science Study (TIMSS) 2003: Nonresponse Bias Analysis* (NCES 2007-044). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC

Foy, P., Joncas, M., and Zuhlke, O. (2009). *PIRLS 2011 School Sampling Manual*. Unpublished manuscript, Chestnut Hill, MA: Boston College.

IEA Data Processing Center. (2010). *PIRLS 2011 Data Entry Manager Manual*. Hamburg, Germany: Author.

Martin, M. O., & Mullis, I. V. S. (Eds.). (2011). *TIMSS and PIRLS Methods and Procedures*. Retrieved from http://timssandpirls.bc.edu/methods/index.html

Martin, M.O., Mullis, I.V.S., and Foy, P. (forthcoming). *PIRLS 2011 Technical Report*. Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Boston College.

Mullis, I.V.S., Kennedy, A.M., Martin, M.O., and Sainsbury, M. (2009). *PIRLS 2011: Assessment Framework and Specifications*. Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Boston College.

Mullis, I.V.S., Martin, M.O., Foy, P., and Arora, A. (2012). *PIRLS 2011 International Results in Reading*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

National Center for Education Statistics. (2002). *NCES Statistical Standards* (NCES 2003-601). Institute of Education Sciences, U.S. Department of Education. Washington, DC: Author.

United Nations Educational, Scientific, and Cultural Organization (UNESCO). (1999). *Operational Manual for ISCED 1997*. Paris: UNESCO Institute for Statistics.

Westat. (2007). *WesVar 5.0 User's Guide*. Rockville, MD: Author.

Page intentionally left blank

Appendix A: Technical Notes

Introduction

The Progress in International Reading Literacy Study (PIRLS) is a cross-national comparative study of the performance and schooling contexts of 4th-grade students in reading. In this third cycle of PIRLS, the reading literacy assessment and associated questionnaires were administered in 53 education systems at the 4th-grade level during April and June 2011. PIRLS is coordinated by the International Association for the Evaluation of Educational Achievement (IEA), with national sponsors in each participating education system. In the United States, PIRLS is sponsored by the National Center for Education Statistics (NCES), in the Institute of Education Sciences of the U.S. Department of Education.

This appendix provides an overview of technical aspects of PIRLS 2011, including

- International requirements for sampling design, data collection, and response rates;
- Sampling, data collection, and response rates in the United States and other countries;
- Test development;
- Recruitment, test administration, and quality assurance;
- Scoring and scoring reliability;
- Data entry and cleaning;
- Weighting, scaling, and plausible values;
- International benchmarks;
- Data limitations;
- Description of background variables;
- Confidentiality and disclosure limitations; and
- Statistical procedures.

More detailed information can be found in the *PIRLS 2011 Technical Report* (Martin, Mullis and Foy forthcoming).

International requirements for sampling, data collection, and response rates

In order to ensure comparability of the data across participating education systems, the IEA provided detailed international requirements on the various aspects of data collection described here, and implemented quality control procedures. Participating countries were obliged to follow these requirements. These requirements regarding the target populations, sampling design, sample size, exclusions, and defining participation rates are described below.

Target populations

In order to identify comparable populations of students to be sampled, the IEA defined the target populations as follows (Martin, Mullis and Foy forthcoming):

Fourth-grade student population. The international desired target population is all students enrolled in the grade that represents 4 years of schooling, counting from the first year of the International Standard Classification of Education (ISCED) Level 1,[1] providing that the mean age at the time of testing is at least 9.5 years. For most countries, the target grade should be grade 4, or its national equivalent. All students enrolled in the target grade, regardless of their age, belong to the international desired target population.

Teacher population. The target population is all teachers linked to the selected students. Note that these teachers are not a representative sample of teachers within the education system. Rather, they are the teachers who teach a representative sample of students in grade 4 within the education system.

School population. All eligible schools[2] containing one or more 4th-grade classrooms.

Sampling design

It was not feasible to assess every 4th-grade student in the United States. As is done in all participating countries, a representative sample of 4th-grade students was selected. The sample design employed by the PIRLS 2011 assessment is generally referred to as a two-stage stratified cluster sample. The sampling units at each stage were defined as follows.

First-stage sampling units. In the first stage of sampling, statisticians selected individual schools with a probability proportionate to size (PPS) approach, which means that the probability is proportional to the estimated number of students enrolled in the target grade. Prior to sampling, statisticians assigned schools in the sampling frame to a predetermined number of explicit or implicit strata. Then, sampling staff sampled schools using a PPS systematic sampling method. Statisticians also selected substitution schools, which were selected to replace those that we originally sampled but refused to participate. The original and substitution schools were selected simultaneously.

[1]The ISCED was developed by the United Nations Educational, Scientific, and Cultural Organization (UNESCO) to facilitate the comparability of educational levels across countries. ISCED Level 1 begins with the first year of formal, academic learning (UNESCO 1999). In the United States, ISCED Level 1 begins at grade 1.
[2]Some sampled schools may be considered ineligible for reasons noted in the section below titled "School exclusions."

Second-stage sampling units. In the second stage of sampling, statisticians selected classrooms within sampled schools using a sampling software provided by the International Study Center at Boston College. The software uses a sampling algorithm for selecting classes that standardized the class sampling across schools and assures that the class selection procedures are uniform across countries. Statistical staff followed the project rule that for a school to be selected there must be a minimum of one eligible classroom in that school. They chose classrooms from a list of eligible classrooms that sampling staff prepared. However, statistical staff were encouraged by PIRLS national research coordinators (NRCs) to select more than one eligible classroom per school. All students in sampled classrooms were selected for assessment.

Sample size for the main survey

PIRLS guidelines call for a minimum of 150 schools to be sampled, with a minimum of 4,000 students assessed. The basic sample design of one classroom per school was designed to yield a total sample of approximately 4,500 students per population. Countries with small class sizes or less than 30 students per school were directed to consider sampling more schools, more classrooms per school, or both, to meet the minimum target of 4,000 tested students.

In the United States, a sample of 450 schools was drawn at grade 4. These were larger sample sizes than used in previous administrations for PIRLS. The reason for a larger sample than in the past at grade 4 was that in 2011 both the Trends in International Mathematics and Science Study (administered every 4 years) and PIRLS (administered every 5 years) happened to coincide in the same year. Because the United States was participating in both studies and because both studies required a grade 4 sample of schools and students, the decision was made to draw a larger sample of schools and to request that both studies be administered in the same schools (where feasible), albeit to separate classroom samples of students.[3] Thus, TIMSS (grade 4) and PIRLS in the United States were administered in the same schools but to separately sampled classrooms of students.

Exclusions

The following discussion draws on the *PIRLS 2011 School Sampling Manual* (Foy, Joncas, and Zuhlke 2009). All schools and students excluded from the national defined target population are referred to as the excluded population. Exclusions could occur at the school level, with entire schools being excluded, or within schools, with specific students or entire classrooms excluded. PIRLS 2011 did not provide accommodations for students with disabilities or students who were unable to read or speak the language of the test. The IEA requirement with regard to exclusions is that they should not exceed more than 5 percent of the national desired target population (Foy, Joncas, and Zuhlke 2009). The specifications for school and student exclusions were applied equally to the U.S. national and the Florida-USA sample.

School exclusions. Countries could exclude schools that

- are geographically inaccessible;
- are of extremely small size;
- offer a curriculum or school structure radically different from the mainstream education system; or
- provide instruction only to students in the excluded categories defined under "within-school exclusions," such as schools for the blind.

Within-school exclusions. Countries were asked to adapt the following international within-school exclusion rules to define excluded students:

- **Students with intellectual disabilities—**Students who, in the professional opinion of the school principal or other qualified staff members, are considered to have intellectual disabilities or who have been tested psychologically as such. This includes students who are emotionally or mentally unable to follow even the general instructions of the test. Students were not to be excluded solely because of poor academic performance or normal disciplinary problems.
- **Students with functional disabilities—**Students who are permanently physically disabled in such a way that they cannot perform in the PIRLS testing situation. Students with functional disabilities who are able to respond were to be included in the testing.
- **Non-native-language speakers—**Students who are unable to read or speak the language(s) of the test and would be unable to overcome the language barrier of the test. Typically, a student who had received less than 1 year of instruction in the language(s) of the test was to be excluded.

[3]In some cases, sampled schools were unable to accommodate both studies due to small student enrollment in grade 4 or scheduling conflicts. Schools with at least two grade 4 classrooms were asked to participate in both studies, with one classroom being randomly assigned to TIMSS and the other to PIRLS. Up to two TIMSS classes and two PIRLS classes were selected in schools with sufficient student enrollment. In schools with only one grade 4 classroom, either the TIMSS or PIRLS assessment was randomly assigned, but not both. In no cases were the same students asked to complete both the TIMSS and PIRLS assessments at grade 4.

Defined participation rates

In order to minimize the potential for response biases, the IEA developed participation or response rate standards that apply to all participating education systems and govern whether or not an education system's data are included in the PIRLS 2011 international dataset and the way in which national statistics are presented in the international reports. These standards were set using composites of response rates at the school, classroom, and student and teacher levels, and response rates were calculated with and without the inclusion of substitute schools that were selected to replace schools refusing to participate.

The response rate standards determine how an education system's data will be reported in the international reports. These standards take the following two forms, distinguished primarily by whether or not meeting the school response rate of 85 percent requires the counting of substitute schools.

Category 1: Met requirements. Education systems that meet all of the following conditions are considered to have fulfilled the IEA requirements: (a) a minimum school participation rate of 85 percent, based on original sampled schools only; and (b) a minimum classroom participation rate of 95 percent, from both original and substitute schools; and (c) a minimum student participation rate of 85 percent, from both original and substitute schools.

Category 2: Met requirements after substitutes. In the case of education systems not meeting the category 1 requirements, provided that at least 50 percent of schools in the original sample participate, an education system's data are considered acceptable if the following requirements are met: a minimum combined school, classroom and student participation rate of 75 percent, based on the product of the participation rates described above. That is, the product of (a), (b), and (c), as defined in the category 1 standard, must be greater than or equal to 75 percent.

Education systems satisfying the category 1 standard are included in the international tabular presentations without annotation. Those able to satisfy only the category 2 standard are included as well but are annotated to indicate their response rate status. The data from education systems failing to meet either standard are presented separately in the international tabular presentations.

Sampling, data collection, and response rates in the United States and other education systems

The U.S. PIRLS sample design

In the United States and most other education systems, the target populations of students corresponded to grade 4. In sampling these populations, PIRLS used a two-stage stratified cluster sampling design.[4] The U.S. sampling frame was explicitly stratified by three categorical stratification variables: percentage of students eligible for free or reduced-price lunch, type of school (public or private), and region of the country (Northeast, Central, West, Southeast).[5] The U.S. sample was implicitly stratified (that is, sorted for sampling) by two categorical stratification variables: locality (four levels),[6] and minority status (above or below 15 percent of the student population).

The first stage selected schools for the original sample using Probability Proportional to Size (PPS). Using a sampling frame based on the 2011 National Assessment of Educational Progress (NAEP) school sampling frame,[7] schools were selected with a probability proportionate to the school's estimated enrollment of grade 4 students. Data for public schools were taken from the Common Core of Data (CCD), and data for private schools were taken from the Private School Universe Survey (PSS). In addition, for each original school selected, the two neighboring schools in the sampling frame were designated as substitute schools. The first school following the original sample school was the first substitute and the first school preceding it was the second substitute. If an original school refused to participate, the first substitute was contacted. If that school also refused to participate, the second substitute was contacted. There were several constraints on the assignment of substitutes. One sampled

[4]The primary purpose of stratification is to improve the precision of the survey estimates. If explicit stratification of the population is used, the units of interest (schools, for example) are sorted into mutually exclusive subgroups–strata. Units in the same stratum are as homogeneous as possible, and units in different strata are as heterogeneous as possible, with respect to the characteristics of interest to the survey. Separate samples are then selected from each stratum. In the case of implicit stratification, the units of interest are simply sorted with respect to one or more variables known to have a high correlation with the variable of interest. In this way, implicit stratification guarantees that the sample of units selected will be spread across the categories of the stratification variables.

[5]The Northeast region consists of Connecticut, Delaware, the District of Columbia, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, and Vermont. The Central region consists of Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, Wisconsin, and South Dakota. The West region consists of Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oklahoma, Oregon, Texas, Utah, Washington, and Wyoming. The Southeast region consists of Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Virginia, and West Virginia.

[6]Four locality levels are distinguished: city of 250,000 or more; suburb less than 250,000; town of 25,000 or more; rural metropolitan statistical area (MSA).

[7]To maximize response rates from both districts and schools, it was necessary to begin the recruitment of both prior to the end of the 2009-10 school year. Since the 2011 NAEP sampling frame was not available until March 2010, it was necessary to base the PIRLS sample on the 2010 NAEP sampling frame.

school was not allowed to substitute for another, and a given school could not be assigned to substitute for more than one sampled school. Furthermore, substitutes were required to be in the same implicit stratum as the sampled school.

The second stage consisted of selecting intact 4th-grade classes within each participating school. Schools provided lists of 4th-grade classrooms. Within schools, classrooms with fewer than 15 students were collapsed into pseudo-classrooms so that each classroom in the school's classroom sampling frame had at least 20 students.[8] An equal probability sample of two classrooms[9] was identified from the classroom frame for the school. In schools where there was only one classroom, this classroom was selected with certainty. For PIRLS, 16 pseudo-classrooms were created prior to classroom sampling, with 7 of these being selected in the final classroom sample.

All students in sampled classrooms and pseudo-classrooms were selected for assessment. In this way, the overall sample design for the United States results in an approximately self-weighting sample of students, with each 4th-grade student having a roughly equal probability of selection. While in small schools we select a higher proportion of the classes, and therefore of the students, this is counterbalanced by the selection of schools with probability proportional to size.

Selecting a school sample for the U.S. benchmarking state

The PIRLS state benchmarking sample was selected from Florida-USA and consisted of public schools only. The school frame was identical to the national frame of public schools in that state. The state sample included the schools in the state that were previously selected as part of the TIMSS-PIRLS national sample at grade 4 plus a supplement of schools with the target of 100 assessed classrooms. The target reference is classrooms due to the national design where in each school up to four classes are selected and randomly assigned to PIRLS or TIMSS. The additional number of schools needed in the state is then ([100 – number of national public schools] / 2) plus an additional five schools to account for ineligible schools (schools with no grade 4 students).

The benchmarking sample was selected using a version of the Keyfitz procedure. Chowdhury, Chu, and Kaufman (2001) have described the implementation of the procedure. The method is generally used to minimize overlap but it can also be used to maximize overlap by ordering the rows in descending order of the response load indicator. By following the process outlined in table 2 of the paper, the rows in the table can be thought of as a hierarchy of selection preference, where the top row maximizes the probability and the bottom row minimizes it. This property allowed for maximization of the overlap with the TIMSS-PIRLS national sample (in fact select all national schools) and minimization the overlap with the NAEP validation public school sample. This minimization was undertaken to reduce the burden for schools selected in the NAEP sample and to improve response rates. This was accomplished by partitioning the frame into the following three groups shown in order as in table 2 of the paper. The three groups were:

1. schools selected for the TIMSS-PIRLS national sample (including schools also selected for the NAEP validation public school sample);
2. schools not selected for either the TIMSS-PIRLS national or NAEP validation public school samples; and
3. schools selected for the NAEP validation public school sample and not TIMSS-PIRLS national sample.

The method guarantees all schools in group 1 will be selected with certainty since the probability of being selected for the state sample is always larger than being selected for the national sample, since more schools were selected in the state sample (the national schools plus a state supplement) than in the national sample, with the frames being identical. The method minimized the overlap with schools in group 3 (NAEP validation public school sample) and selected the majority of the state supplement from schools in group 2.

U.S. PIRLS sample

School sample. The 4th-grade school sample consisted of 450 public and private schools. As described previously, the joint administration of TIMSS and PIRLS at grade 4 required a larger sample of schools to ensure an adequate number of participating classes and students in both studies. Twelve ineligible schools and one excluded school were identified on the basis that they served special student populations or had closed or altered their grade makeup since the sampling frame was developed. This left 437 schools eligible to participate, and 349 agreed to do so. The school response rate before substitution then was 80 percent unweighted. The analogous weighted school response rate was also 80 percent (see table A-1) and is given by the following formula:

$$\text{weighted school response rate before replacement} = \frac{\sum_{i \in Y} W_i E_i}{\sum_{i \in (Y \cup N)} W_i E_i}$$

[8]Since classrooms are sampled with equal probability within schools, small classrooms would have the same probability of selection as large classrooms. Selecting classrooms under these conditions would likely mean that student sample size would be reduced, and some instability in the sampling weights created. To avoid these problems, pseudo-classes are created for the purposes of classroom sampling, in which small classrooms are joined to reach a larger student count. These pseudo-classrooms are treated as single classes in the class sampling process. Following sampling, the pseudo-class combinations are dissolved and the small classes involved retain their own identity. In this way, data on students, teachers, and classroom practices are linked in small classes in the same way as with larger classes.

[9]The classrooms selected could be "pseudo-classrooms," previously defined as classrooms within a school with fewer than 15 students that were merged with other classes within the school for sampling purposes.

where *Y* denotes the set of responding original-sample schools; *N* denotes the set of eligible non-responding original sample schools; *Wi* denotes the base weight for school *i;* *Wi* = 1*/Pi*, where *Pi* denotes the school selection probability for school *i;* and *Ei* denotes the enrollment size of age-eligible students, as indicated on the sampling frame.

In addition to the 349 participating schools from the original sample, 21 substitute schools participated for a total of 370 participating schools at grade 4 in the United States (see table A-2). This gives a weighted (and unweighted) school participation rate after substitution of 85 percent (see table A-1).[10]

Classroom sample. Schools agreeing to participate in PIRLS were asked to list their 4th-grade classes as the basis for sampling at the classroom level. At this time, schools were given the opportunity to identify special classes—classes in which all or most of the students had intellectual or functional disabilities or were non-native-language speakers. While these classes were regarded as eligible, the students as a group were treated as "excluded" since, in the opinion of the school, their disabilities or language capabilities would render meaningless their performance on the assessment. Fifty 4th-grade schools excluded classes and 669 students were excluded from participation in PIRLS as a result.

Prior to sampling, classes with fewer than 15 students were collapsed with other classes into what are called pseudo-classrooms. Creating pseudo-classrooms in this way ensured that all eligible classrooms in a school had at least 20 students. Up to four eligible classrooms were selected, with classes being randomly assigned to TIMSS or PIRLS. In schools with only one classroom, this classroom was selected with certainty and randomly assigned to TIMSS or PIRLS. Some 1,257 classrooms were selected as a result of this process. All selected classrooms participated in PIRLS yielding a classroom response rate of 100 percent (Mullis, et al. 2012, exhibit C.8).

Student sample. Schools were asked to list the students in each of the classrooms. A total of 14,253 students were listed as a result, and 12,726 4th-grade students participated in PIRLS 2011. These students are identified by IEA as "sampled students in participating schools" (see table A-2).

This pool of students is reduced by within-school exclusions and withdrawals. At the time schools listed the students in the sampled classrooms, they had the opportunity to identify particular students who were not suited to take the test because of physical or intellectual disabilities (i.e., students with disabilities who had been mainstreamed) or because they were non-English-language speakers. Schools identified a total of 830 students they wished to have excluded from the assessment; also by the time of the assessment a further 169 of the listed students had withdrawn from the school or classroom. In total, the pool of 14,253 sampled students was reduced by 999 students (830 excluded and 169 withdrawn) to yield 13,254 "eligible" students. The number of eligible students is used as the base for calculating student response rates (Mullis, et al. 2012, exhibit C.6).

The number of eligible students was further reduced on assessment day by 528 student absences, leaving 12,726 "assessed students" identified as having completed a PIRLS 2011 assessment booklet (see table A-2). IEA defines the student response rate as the number of students assessed as a percentage of the number of eligible students which, in this case, yields a weighted (and unweighted) student response rate of 96 percent (see table A-1).

Note that the 669 students excluded because whole classes were excluded do not figure in the calculation of student response rates. They do, however, figure in the calculation of the coverage of the International Target Population. Together, these 669 students excluded prior to classroom sampling, plus the 830 within-class exclusions, resulted in an overall student exclusion rate of 7 percent (see table A-1 and Mullis, et.al. 2012, exhibit C.3). The reported coverage of the International Target Population, then, is 93 percent (see Mullis, et. al. 2012, exhibit C.3).

Combined participation rates. For the results for an education system to be included in the PIRLS international report without a response rate annotation, the IEA requires a "combined" or overall response rate—expressed as the product of (a) the (unrounded) weighted school response rate without substitute schools and (b) the (unrounded) weighted student response rate—of at least 75 percent (after rounding to the nearest whole percent). The overall response rate for the United States, 76.6 percent without substitute schools, meets this requirement. However, the United States did include substitute schools because its school-level response rate was less than 85 percent, and, absent advance knowledge of the student-level response rate, introducing substitute schools was a prudent approach to take. For the results of an education system to be included in the PIRLS international report without a student inclusion annotation, the IEA requires a student inclusion rate of at least 95 percent. Because 7 percent of the 4th-grade student population was excluded in the United States, the overall U.S. student inclusion rate was 93 percent. For this reason, the U.S. 4th-grade results in the PIRLS international report carry a coverage annotation indicating that coverage of the defined student population was less than the IEA standard of 95 percent.

Tables A-1 and A-2 are extracts from the international report exhibits noted above and are designed to summarize information on school and student responses rates and coverage of the target populations in each nation.

[10]Substitute schools are matched pairs and do not have an independent probability of selection. NCES standards (Standard 1-3-8) indicate that, in these circumstances, response rates should be calculated without including substitute schools (National Center for Education Statistics 2002). PIRLS response rates denoted as "before replacement" conform to this standard. PIRLS response rates denoted as "after replacement" are not consistent with NCES standards since, in the calculation of these rates, substitute schools are treated as the equivalent of sampled schools.

Table A-1. Coverage of target populations, school participation rates, and student response rates, by education system: 2011

Education system	Percentage of international desired population coverage	National desired population overall exclusion rate	Weighted school participation rate before substitution	Weighted school participation rate after substitution	Weighted student response rate	Combined weighted school participation and student response rate
Australia	100	4	96	98	95	93
Austria	100	5	100	100	98	98
Azerbaijan	100	7	84	100	100	100
Belgium (French)-BEL	100	6	77	85	97	82
Bulgaria	100	3	97	100	95	95
Canada	100	10	98	98	96	94
Chinese Taipei-CHN	100	1	100	100	99	99
Colombia	100	2	89	99	97	95
Croatia	100	8	99	100	95	95
Czech Republic	100	5	90	99	94	94
Denmark	100	7	87	98	97	95
England-GBR	100	2	73	87	94	82
Finland	100	3	97	99	96	95
France	100	5	98	100	98	97
Georgia	92	5	97	98	98	96
Germany	100	2	96	99	96	95
Hong Kong-CHN	100	12	86	88	94	83
Hungary	100	4	98	99	97	96
Indonesia	100	3	100	100	97	97
Iran, Islamic Rep. Of	100	5	100	100	99	99
Ireland	100	3	98	100	95	95
Israel	100	25	98	99	94	93
Italy	100	4	81	98	96	95
Lithuania	93	6	94	100	94	94
Malta	100	4	100	100	95	95
Morocco	100	2	99	99	96	95
Netherlands	100	4	68	92	97	89
New Zealand	100	3	93	99	94	93
Northern Ireland-GBR	100	4	62	85	93	79
Norway	100	4	57	83	86	71
Oman	100	2	98	98	98	96
Poland	100	4	100	100	96	96
Portugal	100	3	87	99	95	93
Qatar	100	6	100	100	99	99
Romania	100	4	99	100	97	97
Russian Federation	100	5	100	100	98	98
Saudi Arabia	100	2	95	100	98	98
Singapore	100	6	100	100	96	96
Slovak Republic	100	5	95	99	97	96
Slovenia	100	3	96	97	97	95
Spain	100	5	96	99	97	96
Sweden	100	4	97	99	92	91
Trinidad and Tobago	100	1	99	99	96	95
United Arab Emirates	100	3	100	100	97	97
United States	100	7	80	85	96	81

See notes at end of table.

Table A-1. Coverage of target populations, school participation rates, and student response rates, by education system: 2011—Continued

Benchmarking education systems	Percentage of international desired population coverage	National desired population overall exclusion rate	Weighted school participation rate before substitution	Weighted school participation rate after substitution	Weighted student response rate	Combined weighted school participation and student response rate
Alberta-CAN	100	7	97	99	95	94
Ontario-CAN	100	8	99	99	96	95
Quebec-CAN	100	4	95	96	96	92
Maltese-MLT	100	4	100	100	94	94
Andalusia-ESP	100	5	99	99	97	96
Abu Dhabi-UAE	100	3	99	99	97	96
Dubai-UAE	100	5	99	99	96	94
Florida-USA	89	13	96	96	95	91

NOTE: Education systems in the Southern hemisphere administered PIRLS 2011 in the fall of 2010 while those in the Northern hemisphere administered the assessment in the spring of 2011. Italics indicate participants identified and counted in this report as an education system and not as a separate country. The international desired population refers to the sample and not the responding schools, classes, and students.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Progress in International Reading Literacy Study (PIRLS), 2011.

Table A-2. Total number of schools and students, by education system: 2011

Education system	Schools in original sample	Eligible schools in original sample	Schools in original sample that participated	Substitute schools	Total schools that participated	Sampled students in participating schools	Students assessed
Australia	290	284	275	5	280	6,709	6,126
Austria	160	158	158	0	158	4,976	4,670
Azerbaijan	170	169	142	27	169	5,098	4,881
Belgium (French)-BEL	150	150	115	12	127	3,910	3,727
Bulgaria	150	147	142	5	147	5,725	5,261
Canada	1,142	1,125	1,106	5	1,111	25,707	23,206
Chinese Taipei-CHN	150	150	150	0	150	4,376	4,293
Colombia	157	152	131	19	150	4,309	3,966
Croatia	152	152	150	2	152	5,097	4,587
Czech Republic	180	178	161	16	177	4,895	4,556
Denmark	240	236	207	25	232	4,994	4,594
England-GBR	150	148	109	20	129	4,243	3,927
Finland	150	146	141	4	145	4,914	4,640
France	175	175	170	4	174	4,638	4,438
Georgia	180	177	172	1	173	4,958	4,796
Germany	200	199	190	7	197	4,229	4,000
Hong Kong-CHN	154	150	130	2	132	4,189	3,875
Hungary	150	150	146	3	149	5,488	5,204
Indonesia	158	158	158	0	158	5,049	4,791
Iran, Islamic Rep. Of	250	244	244	0	244	5,932	5,758
Ireland	152	151	148	3	151	4,849	4,524
Israel	153	153	150	2	152	4,579	4,186
Italy	205	205	166	36	202	4,529	4,189
Lithuania	160	154	145	9	154	5,140	4,661
Malta	99	96	96	0	96	3,958	3,598
Morocco	289	287	284	0	284	8,381	7,805
Netherlands	151	151	97	41	138	4,179	3,995
New Zealand	201	199	180	12	192	6,192	5,644
Northern Ireland-GBR	160	160	100	36	136	3,942	3,586
Norway	150	145	85	35	120	3,921	3,190
Oman	338	333	327	0	327	10,840	10,394
Poland	150	150	150	0	150	5,316	5,005
Portugal	150	150	133	15	148	4,428	4,085
Qatar	175	167	166	0	166	4,394	4,120
Romania	150	148	147	1	148	4,879	4,665
Russian Federation	202	202	202	0	202	4,693	4,461
Saudi Arabia	175	171	163	8	171	4,625	4,507
Singapore	176	176	176	0	176	6,687	6,367
Slovak Republic	200	198	187	10	197	5,933	5,630
Slovenia	202	201	193	2	195	4,674	4,512
Spain	314	314	308	4	312	9,223	8,580
Sweden	161	153	148	4	152	5,209	4,622
Trinidad and Tobago	150	150	149	0	149	4,190	3,948
United Arab Emirates	478	460	458	0	458	15,372	14,618
United States	450	437	349	21	370	14,253	12,726

See notes at end of table.

Table A-2. Total number of schools and students, by education system: 2011—Continued

Benchmarking education systems	Schools in original sample	Eligible schools in original sample	Schools in original sample that participated	Substitute schools	Total schools that participated	Sampled students in participating schools	Students assessed
Alberta-CAN	150	147	143	2	145	4,292	3,789
Ontario-CAN	200	191	188	1	189	4,932	4,561
Quebec-CAN	200	197	189	1	190	4,529	4,244
Maltese-MLT	99	95	95	0	95	3,942	3,548
Andalusia-ESP	150	150	149	0	149	4,652	4,333
Abu Dhabi-UAE	168	165	164	0	164	4,308	4,146
Dubai-UAE	152	139	138	0	138	6,497	6,061
Florida-USA	81	80	77	0	77	3,052	2,598

NOTE: Education systems in the Southern hemisphere administered PIRLS 2011 in the fall of 2010, while those in the Northern hemisphere administered the assessment in the spring of 2011. Italics indicate participants identified and counted in this report as an education system and not as a separate country.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Progress in International Reading Literacy Study (PIRLS), 2011.

Nonresponse bias in the U.S. PIRLS sample

NCES standards require a nonresponse bias analysis when the response rate of any sampled unit falls below 85 percent (Standard 2-2-2, NCES Statistical Standards, 2002). Because the response rate for U.S. schools was below 85 percent, a nonresponse bias analysis was conducted. It took a form similar to that adopted for TIMSS 2003 (Ferraro and Van de Kerckhove 2006). A full report of this study will be included in a technical report to be released with the U.S. national PIRLS dataset. The response rate in Florida was sufficiently high so that a nonresponse bias analysis was not required.

Three methods were chosen to perform this analysis. The first method focused exclusively on the *sampled schools* and ignored substitute schools. The schools were weighted by their school base weights, excluding any nonresponse adjustment factor. The second method focused on *sampled schools plus substitute schools*, treating as nonrespondents those schools from which a final response was not received from the original or substitute school. Again, schools were weighted by their base weights, with the base weight for each substitute school set to the base weight of the original school that it replaced. The third method repeated the analyses from the second method using nonresponse adjusted weights.[11]

In order to compare PIRLS respondents and nonrespondents, it was necessary to match the sample of schools back to the sample frame to identify as many characteristics as possible that might provide information about the presence of nonresponse bias.[12] The characteristics available for analysis in the sampling frame were taken from the CCD for public schools, and from the PSS for private schools. For categorical variables, the distribution of the characteristics for respondents was compared with the distribution for all schools. The hypothesis of independence between a given school characteristic and the response status (whether or not the school participated) was tested using a Rao-Scott modified chi-square statistic. For continuous variables, summary means were calculated and the difference between means was tested using a *t* test. Note that this procedure took account of the fact that the two samples in question were not independent samples, but in fact the responding sample was a subsample of the full sample. This effect was accounted for in calculating the standard error of the difference. Note also that in those cases where both samples were weighted using just the base weights, the test is exactly equivalent to testing that the mean of the respondents was equal to the mean of the nonrespondents.

In addition, multivariate logistic regression models were set up to identify whether any of the school characteristics were significant in predicting response status when the effects of all potential influences were considered simultaneously.

Public and private schools were modeled together using the following variables:[13] community level (central city, urban fringe/large town, rural/small town); control of school (public or private); census region (Northeast, Southeast, Central, West); poverty level (percentage of students in school eligible for free or reduced-price lunch);[14] number of students enrolled in grade 4; total number of students; and percentage minority students.[15]

Results for the original sample of schools. In the analyses for the original sample of schools, all substituted schools were treated as nonresponding schools. The results of these analyses follow.

In the investigation into nonresponse bias at the school level for PIRLS 4th-grade schools, comparisons between schools in the eligible sample and participating schools showed that there was no relationship between response status for eight of the twelve school characteristics available for analysis. In the original sample, a separate variable-by-variable bivariate analyses identified four variables that were found to be statistically significantly predictors of response status related to: school control, community level, 4th-grade enrollment, and students eligible for free or reduced-price lunch. When all school-level factors were considered simultaneously in a regression analysis, four variables were found to be statistically significant predictors of response status: private schools, high poverty, total school enrollment, and 4th-grade enrollment. The second method focused on sampled schools plus substitute schools, treating as nonrespondents those schools from which a final response was not received from the original or substitute school. This model (using as a predictor percent minority rather than percent in various race/ethnicity

[11]A detailed treatment of the meaning and calculation of sampling weights, including the nonresponse adjustment factors, is provided in the *TIMSS and PIRLS Methods and Procedures* (Martin and Mullis 2011).

[12]Comparing characteristics for respondents and nonrespondents is not always a good measure of nonresponse bias if the characteristics are either unrelated or weakly related to more substantive items in the survey. Nevertheless, this is often the only approach available.

[13]NAEP region and community level were dummy coded for the purposes of these analyses. In the case of NAEP region, "West" was used as the reference group. For community level, "urban fringe/large town" was chosen as the reference group.

[14]The measure of school poverty is based on the proportion of students in a school eligible for the free or reduced-price lunch (FRPL) program, a federally assisted meal program that provides nutritionally balanced, low-cost or free lunches to eligible children each school day. For the purposes of the nonresponse bias analyses, schools were classified as "low poverty" if less than 50 percent of the students were eligible for FRPL, and "high poverty" if 50 percent or more of students were eligible. Since the nonresponse bias analyses involve both participating and nonparticipating schools, they are based, out of necessity, on data from the sampling frame. PIRLS data are not available for nonparticipating schools. The school frame data are derived from the CCD and PSS. The CCD data provide information on the percentage of students in each school who are eligible for free or reduced-price lunch, but are limited to public schools. The PSS data do not provide the same information for private schools. In the interest of retaining all of the schools and students in these analyses, private schools were assumed to be low-poverty schools–that is, they were assumed to be schools in which less than 50 percent of students were eligible for FRPL.

[15]Two forms of this school attribute were used in the analyses. In the bivariate analyses the percentage of each race/ethnic group was related separately to participation status. In the logistic regression analyses a single measure was used to characterize each school, namely, "percentage of minority students."

categories) showed that private schools, high poverty, and 4th-grade enrollment were significant predictors of participation.

Results for the final sample of schools. In the analyses for the final sample of schools, all substitute schools were included with the original schools as responding schools, leaving nonresponding schools as those for which no assessment data were available.

The bivariate results for the final sample of 4th-grade schools indicated that three variables were statistically significant: school control, 4th-grade enrollment, and the percentage of Hispanic students. When all of these factors were considered simultaneously in a regression analysis, two variables remained significant predictors of participation: private schools and 4th-grade enrollment.

For the final sample of schools in grade 4 with school nonresponse adjustments applied to the weights,[16] there were no statistically significant variables in the bivariate analysis. Note that the multivariate regression analysis cannot be conducted after the school nonresponse adjustments are applied to the weights.

These results suggest that there is some potential for nonresponse bias in the U.S. 4th-grade original sample based on the characteristics studied. It also suggests that, while there is little evidence that the use of substitute schools reduced the potential for bias, it has not added to it substantially. The application of school nonresponse adjustments substantially reduced the measurable potential for bias as no variables remained statistically significant.

Test development

PIRLS is a cooperative effort involving representatives from every education system participating in the study. For PIRLS 2011, the test development effort began with a review and revision of the frameworks that are used to guide the construction of the assessment (Mullis et al. 2009). The frameworks were updated to reflect changes in the curriculum and instruction of participating education systems. Extensive input from experts in reading education, assessment, and curriculum, and representatives from national educational centers around the world contributed to the final shape of the frameworks. Maintaining the ability to measure change over time was an important factor in revising the frameworks.

As part of the PIRLS dissemination strategy, approximately one-half of the 2006 assessment items were released for public use. To replace assessment items that had been released, education systems submitted items for review by subject-matter specialists, and additional items were written by the IEA Reading Review Committee in consultation with item-writing specialists in various countries to ensure that the content, as explicated in the frameworks, was covered adequately. Items were reviewed by an international Reading Item Review Committee and field-tested in most of the participating countries. Results from the field test were used to evaluate item difficulty, how well items discriminated between high- and low-performing students, the effectiveness of distracters in multiple-choice items, scoring suitability and reliability for constructed-response items, and evidence of bias toward or against individual countries or in favor of boys or girls. As a result of this review, 60 new items were selected for inclusion in the international assessment. In total, 135 reading items were included in the 2011 PIRLS assessment booklets. More detail on the distribution of new and trend items is included in table A-3.

[16]The international weighting procedures created a nonresponse adjustment class for each explicit stratum; see the *TIMSS and PIRLS Methods and Procedures* (Martin and Mullis 2011) for details. In the case of the U.S. 4th-grade sample, 12 explicit strata were formed by poverty level, school control, and Census region. The procedures could not be varied for individual education systems to account for any specific needs. Therefore, the U.S. nonresponse bias analyses could have no influence on the weighting procedures and were undertaken after the weighting process was complete.

Table A-3. Number and percentage distribution of reading items in the PIRLS assessment, by content domain and process: 2011

Content domain and process	All items		New items		Trend items	
	Number	Percent	Number	Percent	Number	Percent
Total items	**135**	**100**	**60**	**100**	**75**	**100**
Purposes of reading						
Literary experience	72	53	33	55	39	52
Acquire and use information	63	47	27	45	36	48
Processes of comprehension						
Focus on and retrieve explicitly stated information	33	24	14	23	19	25
Make straightforward inferences	46	34	20	33	26	35
Interpret and integrate ideas and information	38	28	18	30	20	27
Examine and evaluate content, language, and textual elements	18	13	8	13	10	13

NOTE: Detail may not sum to 100 percent due to rounding.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Progress in International Reading Literacy Study (PIRLS), 2011.

Design of instruments

PIRLS 2011 included booklets containing assessment items as well as self-administered background questionnaires for principals, teachers, and students.

Assessment booklets

The assessment booklets were constructed such that not all of the students responded to all of the items. This is consistent with other large-scale assessments, such as NAEP.

The 2011 assessment consisted of 12 booklets and one reader (presented in a magazine-type format with the questions in a separate booklet). The assessment is given in 40-minute parts with a 5- to 10-minute break in between. The student questionnaire given after the second part of the assessment, while untimed, is allotted approximately 30 minutes of response time.

The booklets were rotated among students, with each participating student completing one booklet only. The reading items were each assembled separately into 10 blocks, or clusters, of items. Each of the 13 PIRLS 2011 booklets contained two blocks in total. Each booklet contained one block of literary experience items and one block of informational items only and each block occurred twice across the 13 booklets. Six of the ten blocks were included in previous PIRLS assessments. The remaining four blocks were new for PIRLS 2011.

The PIRLS booklets administered in the state sample were exactly the same as those administered in the national sample.

As part of the design process, it was necessary to ensure that the booklets showed an item distribution across the reading content domains as specified in the framework as well as a relatively equal distribution of items by item type. The number of reading items in the PIRLS 2011 assessment is shown in table A-4.

Table A-4. Number and percentage of reading items in the PIRLS assessment, by item format: 2011

Item Format	Number of items	Percent of items
Total	**135**	**100**
Multiple choice	74	55
Constructed response	61	45

SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Progress in International Reading Literacy Study (PIRLS), 2011.

Background questionnaires

As in prior administrations, PIRLS 2011 included self-administered questionnaires for principals, teachers, and students. To create the questionnaires for 2011, the 2006 versions were reviewed extensively by the NRCs from the participating countries as well as a Questionnaire Item Review Committee (QIRC). The QIRC comprises 10–12 experienced NRCs from different participating countries who have analyzed PIRLS data and use it in their countries. The QIRC review resulted in items being deleted or revised, and the addition of several new ones. Like the assessment items, all questionnaire items were field tested, and the results reviewed carefully. As a result, some of the questionnaire items needed to be revised prior to their inclusion in the final questionnaires. The questionnaires requested information to help provide a context for the performance scores, focusing on such topics as students' attitudes and beliefs about learning, their habits and homework, and their lives both in and outside of school; teachers' attitudes and beliefs about teaching and learning, teaching assignments, class size and organization, instructional practices, and participation in professional development activities; and principals' viewpoints on policy and budget responsibilities, curriculum and instruction issues and student behavior, as well as descriptions of the organization of schools and courses. For 2011, online versions of the school and teacher questionnaires were offered to respondents as the primary mode of data collection. Detailed results from the student, teacher, and school surveys are not discussed in this report but are available in the international report, the *PIRLS 2011 International Report in Reading* (Mullis et al. 2012).

Translation

Source versions of all instruments (assessment booklets, questionnaires, and manuals) were prepared in English and translated into the primary language or languages of instruction in each education system. In addition, it was sometimes necessary to adapt the instrument for cultural purposes, even in countries that use English as the primary language of instruction. All adaptations were reviewed and approved by the International Study Center to ensure they did not change the substance or intent of the question or answer choices. For example, proper names were sometimes changed to names that would be more familiar to students (e.g., Marja-leena to Maria).

Each education system prepared translations of the instruments according to translation guidelines established by the International Study Center. Adaptations to the instruments were documented by each education system and submitted for review. The goal of the translation guidelines was to produce translated instruments of the highest quality that would provide comparable data across countries.

Translated instruments were verified by an independent, professional translation agency prior to final approval and printing of the instruments. Countries were required to submit copies of the final printed instruments to the International Study Center. Further details on the translation process can be found in the *PIRLS 2011 Technical Report* (Martin, Mullis, and Foy forthcoming).

Recruitment, test administration, and quality assurance

PIRLS 2011 emphasized the use of standardized procedures in all participating education systems, so that each collected its own data, based on comprehensive manuals and training materials provided by the international project team. These materials explained the survey's implementation, including precise instructions for the work of school coordinators and scripts for test administrators to use in testing sessions.

Recruitment of schools and students

With the exception of private schools, the recruitment of schools required several steps. Beginning with the sampled schools, the first step entailed obtaining permission from the school district to approach the sampled school(s) in that district. If a district refused permission, then the district of the first substitute school was approached and the procedure was repeated. With permission from the district, the school(s) was contacted in a second step. If a sampled school refused to participate, the district of the first substitute was approached and the permission procedure repeated. During most of the recruitment period sampled schools and substitute schools were being recruited concurrently. Each participating school was asked to nominate a school coordinator as the main point of contact for the study. The school coordinator worked with project staff to arrange logistics and liaise with staff, students, and parents as necessary.

On the advice of the school, parental permission for students to participate was sought with one of three approaches to parents: a simple notification; a notification with a refusal form; and a notification with a consent form for parents to sign. In each approach, parents were informed that their students could opt out of participating in the assessment.

Gifts to schools, school coordinators, and students

Schools, school coordinators, and students were provided with small gifts in appreciation for their willingness to participate. Schools were offered $200, school coordinators received $100, and students were given a clock-compass carabiner.

Test administration

Test administration in the United States was carried out by professional staff trained according to the international guidelines. School personnel were asked only to assist with listings of students, identify space for testing in the school, and specify any parental consent procedures needed for sampled students.

Quality assurance

The International Study Center monitored compliance with the standardized procedures. NRCs were asked to nominate one or more persons unconnected with their national center, such as retired school teachers, to serve as quality control monitors for their education systems. The International Study Center developed manuals for the monitors and briefed them in 2-day training sessions about PIRLS, the responsibilities of the national centers in conducting the study, and their own roles and responsibilities. Some 30 of the 370 schools in the PIRLS sample were visited by monitors. These schools included schools in Florida and were scattered geographically across the nation.

Scoring and scoring reliability

The PIRLS assessment items included both multiple-choice and constructed-response items. A scoring rubric (guide) was created for every constructed-response item included in the PIRLS assessments. The rubrics were carefully written and reviewed by NRCs and other experts as part of the field test of items, and revised accordingly (Martin, Mullis, and Foy forthcoming).

The NRC in each education system was responsible for the scoring and coding of data in that education system, following established guidelines. The NRC and, sometimes, additional staff attended scoring training sessions held by the International Study Center. The training sessions focused on the scoring rubrics and coding system employed in PIRLS. Participants in these training sessions were provided extensive practice in scoring example items over several days. Information on within-education-system agreement among coders was collected and documented by the International Study Center. Information on scoring and coding reliability was also used to calculate cross-education-system agreement among coders.

Data entry and cleaning

The NRC from each education system was responsible for data entry. In the United States, Westat was contracted to collect data for PIRLS 2011 and entered the data into data files with a common international format. This format was specified in the *PIRLS Data Entry Manager Manual* (IEA Data Processing Center 2010), which accompanied the IEA-supplied data-entry software (WinDEM) given to all participating countries to create data files. This software facilitated the checking and correction of data by providing various data consistency checks. The data were then sent to the IEA Data Processing Center (DPC) in Hamburg, Germany, for cleaning. The DPC checked that the international data structure was followed; checked the identification system within and between files; corrected single case problems manually; and applied standard cleaning procedures to questionnaire files. Results of the data cleaning process were documented by the DPC. This documentation was then sent to the NRC along with any remaining questions about the data. The NRC then provided the DPC with revisions to coding or solutions for anomalies. The DPC subsequently compiled background univariate statistics and preliminary test scores based on classical item analysis and item response theory (IRT). Detailed information on the entire data entry and cleaning process can be found in the *PIRLS 2011 Technical Report* (Martin, Mullis, and Foy forthcoming).

Weighting, scaling, and plausible values

Before the data were analyzed, the assessed students were assigned sampling weights to ensure that their representation in the PIRLS 2011 analysis more closely matched the prevalence of groups in the student population of the grade assessed. With these sampling weights in place, the analyses of PIRLS 2011 data proceeded in two phases: scaling and estimation. During the scaling phase, IRT procedures were used to estimate the measurement characteristics of each assessment question. During the estimation phase, the results of the scaling were used to produce estimates of student achievement. Subsequent analyses related these achievement results to the background variables collected by PIRLS 2011.

Weighting

Students were assigned sampling weights to adjust for over- or under-representation of particular groups in the final sample. The weight assigned to a student is the inverse of the probability that the student is selected for the sample. When students are weighted, none are discarded, and each contributes to the results for the total number of students represented by the individual student assessed. The use of sampling weights is necessary for the computation of sound, nationally representative estimates. Weighting also adjusts for various situations (such as school and student nonresponse) because data cannot be assumed to be randomly missing. The internationally defined weighting specifications for PIRLS require that each assessed student's sampling weight should be the product of (1) the inverse of the school's probability of selection, (2) an adjustment for school-level nonresponse, (3) the inverse of the classroom's probability of selection, and (4) an adjustment for student-

level nonresponse.[17] All PIRLS 2001, 2006, and 2011 analyses are conducted using sampling weights. A detailed description of this process is provided in the *PIRLS 2011 Technical Report* (Martin, Mullis, and Foy forthcoming). For 2011, though the national and state samples share schools, the samples are not identical and, thus, weights are estimated separately for the national and state samples.

Scaling

In PIRLS, the propensity of students to answer questions correctly was estimated with a two-parameter IRT model for dichotomous constructed response items, a three-parameter IRT model for multiple choice response items, and a generalized partial credit IRT model for polytomous constructed-response items. The scale scores assigned to each student were estimated using a procedure described below in the "Plausible values" section, with input from the IRT results. With IRT, the difficulty of each item, or item category, is deduced using information about how likely it is for students to get some items correct (or to get a higher rating on a constructed response item) versus other items. Once the parameters of each item are determined, the ability of each student can be estimated even when different students have been administered different items. At this point in the estimation process achievement scores are expressed in a standardized logit scale, which ranges from -4 to +4. In order to make the scores more meaningful and to facilitate their interpretation, the scores for the first year (2001) are transformed to a scale with a mean of 500 and a standard deviation of 100. Subsequent waves of assessment are linked to this metric (see below).

To make scores from the second (2006) wave of data comparable to the first (2001) wave of data, two steps had to be taken. First, the 2001 and 2006 data for countries that participated in both years were scaled together to estimate item parameters. Ability estimates for all students (those assessed in 2001 and those assessed in 2006) based on the new item parameters were then estimated. To put these jointly calibrated 2001 and 2006 scores on the 2001 metric, a linear transformation was applied such that the jointly calibrated 2001 scores have the same mean and standard deviation as the original 2001 scores. Such a transformation also preserves any differences in average scores between the 2001 and 2006 waves of assessment.

In order for scores resulting from subsequent waves of assessment (2011) to be made comparable to 2001 scores (and to each other), the two steps above are applied sequentially for each pair of 2006 and 2011 data: two adjacent years of data are jointly scaled, then resulting ability estimates are linearly transformed so that the mean and standard deviation of the prior year is preserved. As a result, the transformed 2011 scores are comparable to all previous waves of assessment and longitudinal comparisons between all waves of data are meaningful.

To facilitate the joint calibration of scores from adjacent years of assessment, common test items are included in successive administrations. This also enables the comparison of item parameters (difficulty and discrimination) across administrations. If item parameters change dramatically across administrations, they are treated as unique items across administration so that scales can be more accurately linked across years. In this way even if the average ability levels of students in education systems participating in PIRLS changes over time, the scales still can be linked across administrations.

Plausible values

To keep student burden to a minimum, PIRLS administered a limited number of assessment items to each student—too few to produce accurate content-related scale scores for each student. To accommodate this situation, during the scaling process plausible values were estimated to characterize students participating in the assessment, given their background characteristics. Plausible values are imputed values and not test scores for individuals in the usual sense. In fact, they are biased estimates of the proficiencies of individual students. Plausible values do, however, provide unbiased estimates of population characteristics (e.g., means and variances of demographic subgroups).

Plausible values represent what the performance of an individual on the entire assessment might have been, had it been observed. They are estimated as random draws (usually five) from an empirically derived distribution of score values based on the student's observed responses to assessment items and on background variables. Each random draw from the distribution is considered a representative value from the distribution of potential scale scores for all students in the sample who have similar characteristics and identical patterns of item responses. Differences between plausible values drawn for a single individual quantify the degree of error (the width of the spread) in the underlying distribution of possible scale scores that could have caused the observed performances.

An accessible treatment of the derivation and use of plausible values can be found in Beaton and González (1995). A more technical treatment can be found in the *PIRLS 2011 Technical Report* (Martin, Mullis, and Foy forthcoming).

[17] These adjustments are for overall response rates and did not include any of the characteristics associated with differential nonresponse as identified in the nonresponse bias analyses reported above.

International benchmarks

International benchmarks for achievement were developed in an attempt to provide a concrete interpretation of what the scores on the PIRLS reading achievement scale mean (for example, what does it imply about what a student knows and can do if he or she has an achievement score of 625). To describe student performance at various points along the PIRLS reading scale, PIRLS uses scale anchoring to summarize and describe student achievement at four points on the reading scale—*Advanced* (625), *High* (550), *Intermediate* (475), and *Low* (400) international benchmarks. Scale anchoring involves selecting benchmarks (scale points) on the PIRLS achievement scales to be described in terms of student performance. Once benchmark scores have been chosen, items are identified that students are likely to score highly on. The content of these items describe what students know and can do who are at the benchmark level of achievement. To interpret the content of anchored items, these items are grouped by content area within benchmarks and reviewed by reading experts. These experts focus on the content of each item and describe the kind of reading domain or process demonstrated by students answering the item correctly. The experts then provide a summary description of performance at each anchor point leading to a content-referenced interpretation of the achievement results. (Detailed information on the creation of the benchmarks is provided in Mullis et al. 2012, and Martin, Mullis, and Foy forthcoming.)

Data limitations

As with any study, there are limitations to PIRLS 2011 that researchers should take into consideration. Estimates produced using data from PIRLS 2011 are subject to two types of error—nonsampling and sampling errors. Nonsampling errors can be due to errors made in collecting and processing data. Sampling errors can occur because the data were collected from a sample rather than a complete census of the population.

Nonsampling errors

Nonsampling error is a term used to describe variations in the estimates that may be caused by population coverage limitations, nonresponse bias, and measurement error, as well as data collection, processing, and reporting procedures. The sources of nonsampling errors are typically problems like unit and item nonresponse, the difference in respondents' interpretations of the meaning of the survey questions, response differences related to the particular time the survey was conducted, and mistakes in data preparation.

Missing data. Five kinds of missing data were identified by separate missing data codes: omitted, uninterpretable, not administered, not applicable, and not reached. An item was considered *omitted* if the respondent was expected to answer the item but no response was given (e.g., no box was checked in the item which asked "Are you a girl or a boy?"). Items with invalid responses (e.g., multiple responses to a question calling for a single response) were coded as *uninterpretable*. The *not administered* code was used to identify items not administered to the student, teacher, or principal (e.g., those items excluded from the student's test booklet because of the BIB-spiraling of the items). An item was coded as *not applicable* when it is not logical that the respondent answer the question (e.g., when the opportunity to make the response is dependent on a filter question). Finally, items that are *not reached* were identified by a string of consecutive items without responses continuing through to the end of the assessment or questionnaire.

Missing background data on other than key variables[18] are not included in the analyses for this report and are not imputed, thus only unimputed variables are used in this report. Item response rates for variables discussed in this report exceeded the NCES standard of 85 percent and so can be reported without notation.

Of the three key variables identified in the PIRLS 2011 data for the United States—sex, race/ethnicity, and the percentage of students eligible for free or reduced-price lunch (FRPL)—as table A-5 indicates, sex has no missing responses and race/ethnicity missing responses are minimal at some 2 percent. The FRPL variable has some 5 percent missing responses at grade 4 among the public schools in the sample and these were imputed by substituting values taken from the CCD for the schools in question. Note, however, that the CCD provides this information only for public schools. The comparable database for private schools (PSS) does not include data on participation in the FRPL program.

Sampling errors

Sampling errors arise when a sample of the population, rather than the whole population, is used to estimate some statistic. Different samples from the same population would likely produce somewhat different estimates of the statistic in question. This fact means that there is a degree of uncertainty associated with statistics estimated from a sample. This uncertainty is referred to as sampling variance and is usually expressed as the standard error of a statistic estimated from sample data. The approach used for calculating standard errors in PIRLS was jackknife repeated replication (JRR). Standard errors can be used as a measure for the precision expected from a particular sample. Standard errors for all of the reported estimates are included in appendix E (online only).

[18]Key variables include survey-specific items for which aggregate estimates are commonly published by NCES. They include, but are not restricted to, variables most commonly used in table row stubs. Key variables also include important analytic composites and other policy-relevant variables that are essential elements of the data collection. For example, the National Assessment of Educational Progress (NAEP) consistently uses gender, race/ethnicity, urbanicity, region, and school type (public/private) as key reporting variables.

Table A-5. Weighted response rates for unimputed variables for PIRLS: 2011

Variable	Source of Information	U.S. response rate	Range of response rates in other countries
Sex	Classroom tracking form	100	99.5-100
Race/ethnicity	Student questionnaire	98	†
Free or reduced-price lunch	School questionnaire	95	†

† Not applicable (U.S.-only variables).
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Progress in International Reading Literacy Study (PIRLS), 2011.

Confidence intervals provide a way to make inferences about population statistics in a manner that reflects the sampling error associated with the statistic. Assuming a normal distribution, the population value of this statistic can be inferred to lie within the confidence interval in 95 out of 100 replications of the measurement on different samples drawn from the same population.

For example, the average reading score for the U.S. 4th-grade students was 556 in 2011, and this statistic had a standard error of 1.5. Therefore, it can be stated with 95 percent confidence that the actual average of U.S. 4th-grade students in 2011 was between 553 and 559 (1.96 x 1.5 = 2.94; confidence interval = 556 +/- 2.94).

Description of background variables

The international versions of the PIRLS 2011 student, teacher, and school questionnaires are available at http://PIRLS.bc.edu. The U.S. versions of these questionnaires are available at http://nces.ed.gov/surveys/pirls/.

Race/ethnicity

Students' race/ethnicity was obtained through student responses to a two-part question. Students were asked first whether they were Hispanic or Latino, and then whether they were members of the following racial groups: American Indian or Alaska Native; Asian; Black or African American; Native Hawaiian or other Pacific Islander; or White. Multiple responses to the race classification question were allowed. Students who responded that they are Hispanic or Latino were categorized as Hispanic, regardless of their reported race. Results are shown separately for Blacks, Hispanics, Whites, Asians, and multiracial as distinct groups. The small numbers of students indicating that they were American Indian or Alaska Native or Native Hawaiian or other Pacific Islander are included in the total but not reported separately.

Poverty level in public schools (percentage of students eligible for free or reduced-price lunch)

The poverty level in public schools was obtained from principals' responses to the school questionnaire. The question asked the principal to report, as of approximately the first of October 2010, the percentage of students at the school eligible to receive free or reduced-price lunch through the National School Lunch Program. The answers were grouped into five categories: less than 10 percent; 10 to 24.9 percent; 25 to 49.9 percent; 50 to 74.9 percent; and 75 percent or more. Analysis was limited to public schools only. Missing data on this variable were replaced with measures taken from the CCD. The effect of this replacement on the confidentiality of the data was examined as part of the confidentiality analyses described in the following section.

Confidentiality and disclosure limitations

In accord with NCES standard 4-2-6 (National Center for Education Statistics 2002), confidentiality analyses for the United States and Florida data were implemented to provide reasonable assurance that public-use data files issued by the IEA and NCES would not allow identification of individual U.S. schools or students when compared against publicly available data collections. Disclosure limitations included the identification and masking of potential disclosure risks for PIRLS schools and adding an additional measure of uncertainty of school, teacher, and student identification through random swapping of a small number of data elements within the student, teacher, and school files.[19] These procedures were applied to the national and state samples.

[19]The NCES standards describe such techniques as follows: perturbation disclosure limitation techniques directly alter the individual respondent's data for some variables, but preserve the level of detail in all variables included in the microdata file. Blanking and imputing for randomly selected records; blurring (e.g., combining multiple records through some averaging process into a single record); adding random noise; and data swapping or switching (e.g., switching the sex variable from a predetermined pair of individuals) are all examples of perturbation techniques (National Center for Education Statistics 2002).

Statistical procedures

Tests of significance

Comparisons made in the text of this report were tested for statistical significance. For example, in the commonly made comparison of education systems' 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. The estimation of the standard errors that are required in order to undertake the tests of significance is complicated by the complex sample and assessment designs, both of which generate error variance. Together they mandate a set of statistically complex procedures in order to estimate the correct standard errors. As a consequence, the estimated standard errors contain a sampling variance component estimated by the jackknife repeated replication (JRR) procedure; and, where the assessments are concerned, an additional imputation variance component arising from the assessment design. Details on the procedures used can be found in the *WesVar 5.0 User's Guide* (Westat 2007).

In almost all instances, the tests for significance used were standard *t* tests.[20] These fell into two categories according to the nature of the comparison being made: comparisons of independent and nonindependent samples. Before describing the *t* tests used, some background on the two types of comparisons is provided below.

The variance of a difference is equal to the sum of the variances of the two initial variables minus two times the covariance between the two initial variables. A sampling distribution has the same characteristics as any distribution, except that units consist of sample estimates and not observations. Therefore,

$$\sigma^2_{(\hat{\mu}_X - \hat{\mu}_Y)} = \sigma^2_{(\hat{\mu}_X)} + \sigma^2_{(\hat{\mu}_Y)} - 2\mathrm{cov}(\hat{\mu}_X, \hat{\mu}_Y)$$

The sampling variance of a difference is equal to the sum of the two initial sampling variances minus two times the covariance between the two sampling distributions on the estimates.

If one wants to determine whether girls' performance differs from boys' performance, for example, then, as for all statistical analyses, a null hypothesis has to be tested. In this particular example, it consists of computing the difference between the boys' performance mean and the girls' performance mean (or the inverse). The null hypothesis is

$$H_0 : \hat{\mu}_{(boys)} - \hat{\mu}_{(girls)} = 0$$

To test this null hypothesis, the standard error on this difference is computed and then compared to the observed difference. The respective standard errors on the mean estimate for boys and girls can be easily computed.

The expected value of the covariance will be equal to 0 if the two sampled groups are independent. If the two groups are not independent, as is the case with girls and boys attending the same schools within an education system, or comparing an education system mean with the international mean that includes that particular education system, the expected value of the covariance might differ from 0.

In PIRLS, education system samples are independent. Therefore, for any comparison between two education systems, the expected value of the covariance will be equal to 0, and thus the standard error on the estimate is

$$\sigma_{(\hat{\theta}_i - \hat{\theta}_j)} = \sqrt{\sigma^2_{(\hat{\theta}_i)} + \sigma^2_{(\hat{\theta}_j)}}$$

with θ being a tested statistic.

Within a particular education system, any subsamples will be considered as independent only if the categorical variable used to define the subsamples was used as an explicit stratification variable.

Therefore, as for any computation of a standard error in PIRLS, replication methods using the supplied replicate weights are used to estimate the standard error on a difference. Use of the replicate weights implicitly incorporates the covariance between the two estimates into the estimate of the standard error on the difference.

Thus, in simple comparisons of independent averages, such as the U.S. average with other education system averages, the following formula was used to compute the *t* statistic:

$$t = \frac{(est_1 - est_2)}{\sqrt{(se_1)^2 + (se_2)^2}}$$

Est_1 and est_2 are the estimates being compared (e.g., average of education system A and the U.S. average), and se_1 and se_2 are the corresponding standard errors of these averages.

The second type of comparison used in this report occurred when comparing differences of nonsubset, nonindependent groups (e.g., when comparing the average scores of

[20] Adjustments for multiple comparisons were not applied in any of the t-tests undertaken.

males versus females within the United States). In such comparisons, the following formula was used to compute the *t* statistic:

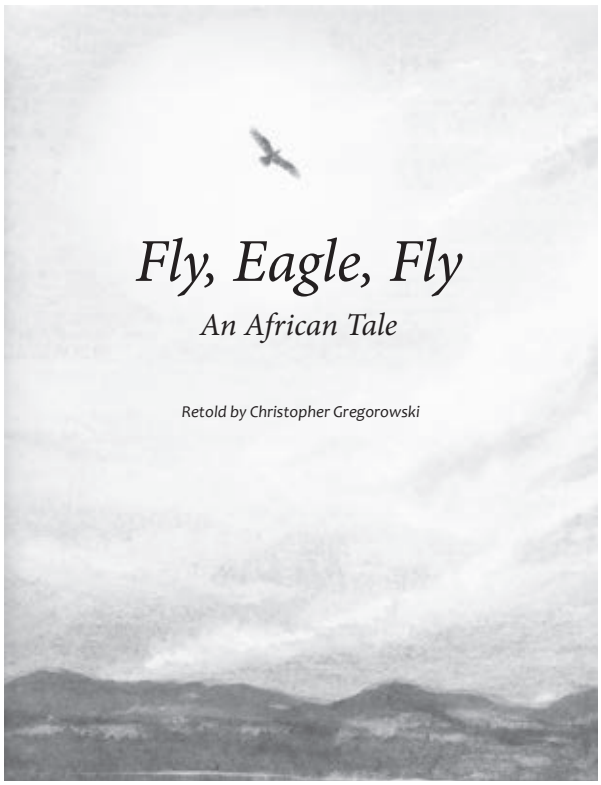
$$t = \frac{(est_{grp1} - est_{grp2})}{se(est_{grp1} - est_{grp2})}$$

est_{grp1} and est_{grp2} are the nonindependent group estimates being compared. $Se(est_{grp1} - est_{grp2})$ is the standard error of the difference calculated using a JRR procedure, which accounts for any covariance between the estimates for the two nonindependent groups.

$$sd_{pooled} = \sqrt{\frac{sd_1^2 + sd_2^2}{2}}$$

Page intentionally left blank

Appendix B: Reading Passages and Items



A farmer went out one day to search for a lost calf. The herders had returned without it the evening before. And that night there had been a terrible storm.

He went to the valley and searched by the riverbed, among the reeds, behind the rocks and in the rushing water.

He climbed the slopes of the high mountain with its rocky cliffs. He looked behind a large rock in case the calf had huddled there to escape the storm. And that was where he stopped. There, on a ledge of rock, was a most unusual sight. An eagle chick had hatched from its egg a day or two earlier, and had been blown from its nest by the terrible storm.

He reached out and cradled the chick in both hands. He would take it home and care for it.

He was almost home when the children ran out to meet him.

"The calf came back by itself!" they shouted.

Reading passage continued on the next page.

The farmer was very pleased. He showed the eagle chick to his family, then placed it carefully in the chicken house among the hens and chicks.

"The eagle is the king of the birds," he said, "but we shall train it to be a chicken."

So, the eagle lived among the chickens, learning their ways. As it grew, it began to look quite different from any chicken they had ever seen.

One day a friend dropped in for a visit. The friend saw the bird among the chickens.

"Hey! That is not a chicken. It's an eagle!"

The farmer smiled at him and said, "Of course it's a chicken. Look—it walks like a chicken, it eats like a chicken. It thinks like a chicken. Of course it's a chicken."

But the friend was not convinced. "I will show you that it is an eagle," he said.

The farmer's children helped his friend catch the bird. It was fairly heavy, but the farmer's friend lifted it above his head and said, "You are not a chicken but an eagle. You belong not to the earth but to the sky. Fly, Eagle, fly!"

The bird stretched out its wings, looked about, saw the chickens feeding, and jumped down to scratch with them for food.

"I told you it was a chicken," the farmer said, and he roared with laughter.

Reading passage continued on the next page.

Very early the next morning the farmer's dogs began to bark. A voice was calling outside in the darkness. The farmer ran to the door. It was his friend again. "Give me another chance with the bird," he begged.

"Do you know the time? It is long before dawn."

"Come with me. Fetch the bird."

Reluctantly, the farmer picked up the bird, which was fast asleep among the chickens. The two men set off, disappearing into the darkness.

"Where are we going?" asked the farmer sleepily.

"To the mountains where you found the bird."

"And why at this ridiculous time of the night?"

"So that our eagle may see the sun rise over the mountain and follow it into the sky where it belongs."

They went into the valley and crossed the river, the friend leading the way. "Hurry," he said, "for the dawn will arrive before we do."

The first light crept into the sky as they began to climb the mountain. The wispy clouds in the sky were pink at first, and then began to shimmer with a golden brilliance. Sometimes their path was dangerous as it clung to the side of the mountain, crossing narrow shelves of rock and taking them into dark crevices and out again. At last he said, "This will do." He looked down the cliff and saw the ground thousands of feet below. They were very near the top.

Carefully, the friend carried the bird onto a ledge. He set it down so that it looked toward the east, and began talking to it. The farmer chuckled. "It talks only chicken-talk."

But the friend talked on, telling the bird about the sun, how it gives life to the world, and how it reigns in the heavens, giving light to each new day. "Look at the sun, Eagle. And when it rises, rise with it. You belong to the sky, not to the earth." At that moment the sun's first rays shot out over the mountain, and suddenly the world was ablaze with light.

The sun rose majestically. The great bird stretched out its wings to greet the sun and feel the warmth on its feathers. The farmer was quiet. The friend said, "You belong not to the earth, but to the sky. Fly, Eagle, fly!" He scrambled back to the farmer. All was silent. The eagle's head stretched up, its wings stretched outwards, and its legs leaned forward as its claws clutched the rock.

Then, without really moving, feeling the updraft of a wind more powerful than any man or bird, the great eagle leaned forward and was swept upward higher and higher, lost to sight in the brightness of the rising sun, never again to live among the chickens.

Passage	FLY, EAGLE, FLY
Reading Purpose	Literary Experience

Item 1: What farmer set out to look for (Example of item at PIRLS Low International Benchmark)

Comprehension Process: Focus on and Retrieve Explicitly Stated Information and Ideas

1. What did the farmer set out to look for at the beginning of the story?

   A a calf

   B herders

   C rocky cliffs

   D an eagle chick

Percentage of students earning full-credit	
International Avg.	89
United States	90

Item 2: Where farmer found eagle chick (Example of item at PIRLS High International Benchmark)

Comprehension Process: Focus on and Retrieve Explicitly Stated Information and Ideas

2. Where did the farmer find the eagle chick?

   A in its nest

   B by the riverbed

   C on a ledge of rock

   D among the reeds

Percentage of students earning full-credit	
International Avg.	73
United States	75

Item 3: What shows farmer was careful (Example of item at PIRLS Intermediate International Benchmark)

Comprehension Process: Make Straightforward Inferences

3. What in the story shows that the farmer was careful with the eagle chick?

   A He carried the eagle chick in both hands.

   B He brought the eagle chick to his family.

   C He put the eagle chick back in its nest.

   D He searched the riverbed for the eagle chick.

Percentage of students earning full-credit	
International Avg.	64
United States	78

Item 4: What farmer did with the chick(Example of item at PIRLS Intermediate International Benchmark)

Comprehension Process: Focus on and Retrieve Explicitly Stated Information and Ideas

4. What did the farmer do with the eagle chick when he brought it home?

   (A) He taught it to fly.

   (B) He set it free.

   (C) He trained it to be a chicken.

   (D) He made a new nest for it.

Percentage of students earning full-credit	
International Avg.	88
United States	89

Item 5: Eagle chick behaved like a chicken (Example of item at PIRLS High International Benchmark)

Comprehension Process: Focus on and Retrieve Explicitly Stated Information and Ideas

5. During the friend's first visit, the eagle chick behaved like a chicken. Give two examples that show this.

   1. _____

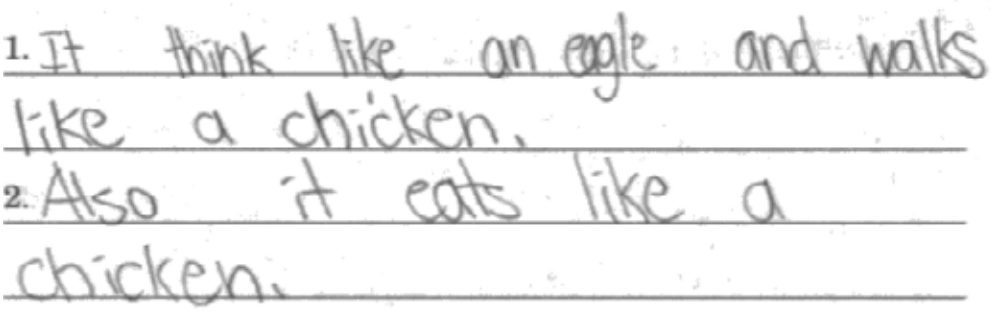
   _____

   2. _____

   _____

Percentage of students earning full-credit	
International Avg.	56
United States	69

Correct Response:



Item 6: How friend tried making eagle fly (Example of item at PIRLS High International Benchmark)

Comprehension Process: Make Straightforward Inferences

6. When the farmer's friend first met the eagle, how did he try to make the eagle fly?

   (A) He lifted it above his head.

   (B) He set it on the ground.

   (C) He threw it in the air.

   (D) He brought it to the mountain.

Percentage of students earning full-credit	
International Avg.	70
United States	74

Item 7: Explanation of friend's words (Example of item at PIRLS High International Benchmark)

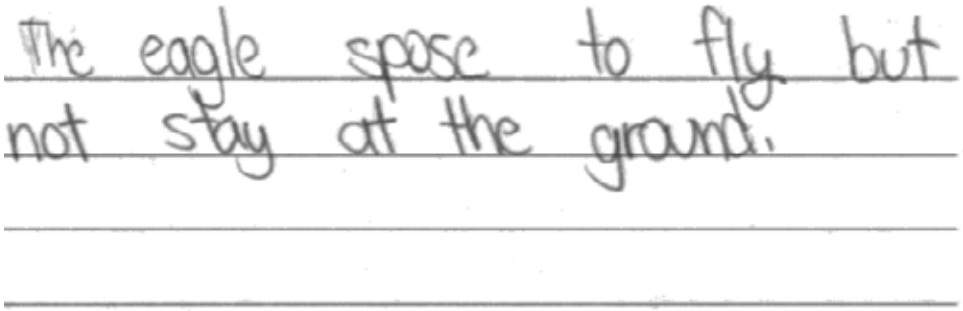
Comprehension Process: Interpret and Integrate Ideas and Information

7. Explain what the farmer's friend meant when he told the eagle, "You belong not to the earth but to the sky."

_____
_____
_____
_____

Percentage of students earning full-credit	
International Avg.	42
United States	62

Correct Response:



Item 8: Why farmer roared with laughter (Example of item at PIRLS Advanced International Benchmark)

Comprehension Process: Interpret and Integrate Ideas and Information

8. Why did the farmer roar with laughter during his friend's first visit?

   (A) The eagle was too heavy to fly.

   (B) The eagle was difficult to catch.

   (C) The eagle looked different from the chickens.

   (D) The eagle proved him right.

Percentage of students earning full-credit	
International Avg.	46
United States	59

Item 9: Eagle taken to the high mountains (Example of item at PIRLS Advanced International Benchmark)

Comprehension Process: Interpret and Integrate Ideas and Information

9. Why did the farmer's friend take the eagle to the high mountains to make it fly? Give two reasons.

1. _____

_____

2. _____

_____

Percentage of students earning full-credit	
International Avg.	17
United States	20

Correct Response:

1. He did it because the sun would make him fly.
2. He also did this because the eagle will try to fill the warmth.

Item 10: Beautiful sky at dawn (Example of item at PIRLS High International Benchmark)

Comprehension Process: Examine and Evaluate Content, Language, and Textual Elements

10. Find and copy words that tell you how beautiful the sky was at dawn.

_____

_____

Percentage of students earning full-credit	
International Avg.	56
United States	67

Correct Response:

majestic

Item 11: Why sun rising was important (Example of item at PIRLS High International Benchmark)

Comprehension Process: Examine and Evaluate Content, Language, and Textual Elements

11. Why was the rising sun important to the story?

- **A** It awakened the eagle's instinct to fly.
- B It reigned in the heavens.
- C It warmed the eagle's feathers.
- D It provided light on the mountain paths.

Percentage of students earning full-credit	
International Avg.	57
United States	73

Item 12: What farmer's friend was like (Example of item at PIRLS Advanced International Benchmark)

Comprehension Process: Interpret and Integrate Ideas and Information

12. You learn what the farmer's friend was like from the things he did.

Describe what the friend was like and give an example of what he did that shows this.

_____

_____

_____

_____

Percentage of students earning full-credit	
International Avg.	29
United States	42

Correct Response:

The friend tried to convince the farmer that the eagle isn't a chicken. So the friend proved by letting the eggle fly with the sun rising.

Appendix C: PIRLS-NAEP Comparison

How Does the Content of PIRLS 2011 Compare With That of the NAEP 2011 Reading Assessment?

In reporting results on how U.S. students perform, the National Center for Education Statistics (NCES) draws on multiple sources of data in order to capitalize on the information presented in national and international assessments. In the United States, data on 4th-grade students' reading achievement come primarily from two sources: the National Assessment of Educational Progress (NAEP) and the Progress in International Reading Literacy Study (PIRLS). PIRLS provides internationally comparable data on student performance, while NAEP tracks performance nationally as well as in state and national subpopulations. A comparative study of PIRLS 2011 and NAEP 2009/2011 revealed important similarities and differences between the two assessments. The purpose of this current study was to examine how the PIRLS 2011 international assessment relates to the NAEP 2011 national reading assessment. The study examined how reading is defined by each assessment broadly, and in terms of content and cognitive dimensions. Also, NCES compared the form and content of the PIRLS and NAEP reading assessments. Answers to these questions provided background information that is useful in interpreting the 2011 results from PIRLS by comparing its design, features, and framework with that of NAEP.

Both the PIRLS and NAEP assessments have a similar definition of reading. Both define reading literacy as an active and constructive process between readers and texts, and both emphasize how readers draw connections across sentences and interpret meanings in the text. The two assessments also employ literary texts and informational texts as the main text types of the reading passages used in the assessments. In addition, both assessments involve two types of items: multiple-choice and constructed response. However, there are important differences between the PIRLS 2011 and NAEP 2011 reading assessments.

Passages text type analyses reveal that PIRLS 2011 and NAEP 2011 have relatively equal proportions of literary texts and informational texts and both assessments have more literary texts than informational texts. However, NAEP 2011 included poetry in its assessment, whereas PIRLS 2011 did not. NAEP 2011 also included paired texts in its assessment in which readers compare two different texts on a similar topic simultaneously, while PIRLS 2011 readers only read one text at a time.

In examining passage length and difficulty, PIRLS 2011 passages were shorter on average than the NAEP 2011 passages. Readability analyses indicate that, on average, the PIRLS 2011 passages were about one grade level lower than the NAEP 2011 passages. However, it should be noted that NAEP included items in its 4th-grade assessment intended for students in 8th grade, where PIRLS did not.

Item-by-item content showed some differences between the assessments. About half of the PIRLS 2011 items were mapped to the NAEP "locate and recall" cognitive target. Most of the remaining PIRLS 2011 items were mapped to the NAEP "integrate and interpret" cognitive target. Very few items were mapped to the NAEP "critique and evaluate" cognitive target. By contrast, NAEP 2011 had more items to assess the "integrate and interpret" as well as the "critique and evaluate" cognitive targets than did PIRLS 2011. The comparison on the cognitive dimensions measured in each assessment indicates that PIRLS 2011 focused more on assessing readers' skills in analyzing information within the text and drawing text-based inferences, whereas NAEP 2011 placed more emphasis on how readers develop inferences and personal interpretations by utilizing personal knowledge or perspectives to examine and evaluate the text in relation to that knowledge or perspectives.

Although both assessments used both multiple-choice and constructed-response items, the constructed-response items in PIRLS 2011 listed separately on the answer sheet the number for each written response needed as a way to scaffold the answering process for readers. PIRLS 2011 also used pictures or symbols within the text to cue test-takers to a specific part of the text where information for answers could be found. These features were absent in NAEP; NAEP 2011 did not provide a scaffolding structure, nor did it offer cues in the form of visual aids to help test-takers.

In summary, there are distinctive differences between PIRLS 2011 and NAEP 2011. Overall, these differences suggest that the NAEP 2011 reading assessment may be more cognitively challenging than PIRLS 2011 for U.S. 4th-grade students. Taken together, these findings suggest that caution should be exercised when attempting to compare 4th-grade students' performance on PIRLS 2011 with 4th-grade students' performance on the NAEP 2011 reading assessment.

Page intentionally left blank

Appendix D: Online Resources and Publications

Online Resources

The NCES website (http://nces.ed.gov/surveys/pirls/) provides background information on the PIRLS surveys, copies of NCES publications that relate to PIRLS, information for educators about ways to use PIRLS in the classroom, and data files. The international PIRLS website (http://www.pirls.org) includes extensive information on the study, including the international reports and databases.

NCES Publications

The following publications are intended to serve as examples of some of the numerous reports published on the Progress in International Reading Literacy Study (PIRLS) by NCES. All of the publications listed here are available at http://nces.ed.gov/pubsearch/getpubcats.asp?sid=099.

PIRLS 2006 Achievement Report

Baer, J., Baldi, S., Ayotte, K., and Green, P. (2007). *The Reading Literacy of U.S. Fourth-Grade Students in an International Context: Results From the 2001 and 2006 Progress in International Reading Literacy Study (PIRLS)* (NCES 2008-017). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2008017

PIRLS 2001 Achievement Report

Ogle, L., Sen, A., Pahlke, E., Jocelyn, L., Kastberg, D., Roey, S., and Williams, T. (2003). *International Comparisons in Fourth-Grade Reading Literacy: Findings from the Progress in International Reading Literacy Study (PIRLS) of 2001* (NCES 2003-073). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2003073

IEA Publications

The following publications are intended to serve as examples of some of the numerous reports that have been published on PIRLS by the International Association for the Evaluation of Educational Achievement (IEA). All of the publications listed here are available at http://www.pirls.org.

PIRLS 2011 Achievement Report

Martin, M.O., Mullis, I.V.S., and Foy, P. (2012). *PIRLS 2011 International Report: IEA's Progress in International Reading Literacy Study in Primary School*. Chestnut Hill, MA: Boston College. http://timssandpirls.bc.edu/isc/publications.html

PIRLS 2006 Achievement Report

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., and Foy, P. (2007). *PIRLS 2006 International Report: IEA's Progress in International Reading Literacy Study in Primary School in 40 Countries*. Chestnut Hill, MA: Boston College. http://timssandpirls.bc.edu/pirls2006/intl_rpt.html

PIRLS 2001 Achievement Reports

Mullis, I.V.S., Martin, M.O., Gonzalez, E., and Kennedy, A.M, (2003). *PIRLS 2001 International Report: IEA's Study of Reading Literacy Achievement in Primary Schools*. Chestnut Hill, MA: Boston College. http://timssandpirls.bc.edu/pirls2001i/PIRLS2001_Pubs_IR.html

Mullis, I.V.S., Martin, M.O., and Gonzalez, E. (2004). *International Achievement in the Processes of Reading Comprehension: Results From PIRLS 2001 in 35 Countries*. Chestnut Hill, MA: Boston College. http://timssandpirls.bc.edu/pirls2001i/PIRLS2001_Pubs_PR.html

Mullis, I.V.S., Martin, M.O., Gonzalez, E., and Kennedy, A.M. (Eds.) (2003). *Trends in Children's Reading Literacy Achievement 1991-2001: IEA's Repeat in Nine Countries of the 1991 Reading Literacy Study*. Chestnut Hill, MA: Boston College. http://timssandpirls.bc.edu/pirls2001i/PIRLS2001_Pubs_TrR.html

PIRLS Encyclopedia

Kennedy, A.M., Mullis, I.V.S., Martin, M. O., and Trong, K. (Eds.) (2007). *PIRLS 2006 Encyclopedia: A Guide to Reading Education in the Forty PIRLS 2006 Countries*. Chestnut Hill, MA: Boston College. http://timssandpirls.bc.edu/pirls2006/encyclopedia.html

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., and Flaherty, C.L. (Eds.) (2002). *PIRLS 2001 Encyclopedia: A Reference Guide to Reading Education in the Countries Participating in IEA's Progress in International Reading Literacy Study (PIRLS)* Chestnut Hill, MA: Boston College. http://timssandpirls.bc.edu/pirls2001i/PIRLS2001_Pubs_ER.html

Mullis, I.V.S., Martin, M.O., Minnich, C.A., Drucker, K.T., and Ragan, M.A. (Eds.) (2012). *PIRLS 2011 Encyclopedia: Education Policy and Curriculum in Reading, Volumes 1 and 2*. Chestnut Hill, MA: Boston College. http://timssandpirls.bc.edu/pirls2011/encyclopedia-pirls.html

PIRLS Technical Reports and Frameworks

Campbell, J.R., Kelly, D., Mullis, I.V.S., Martin, M.O., and Sainsbury, M. (2001). *Framework and Specifications for PIRLS Assessment 2001, 2nd ed*. Chestnut Hill, MA: Boston College. http://timssandpirls.bc.edu/pirls2001i/PIRLS2001_Pubs_AF.html

Martin, M.O., Mullis, I.V.S., and Kennedy, A.M. (Eds.) (2003). *PIRLS 2001 Technical Report*. Chestnut Hill, MA: Boston College. http://timssandpirls.bc.edu/pirls2001i/PIRLS2001_Pubs_TR.html

Martin, M.O., Mullis, I.V.S., and Kennedy, A.M. (Eds.) (2007). *PIRLS 2006 Technical Report*. Chestnut Hill, MA: Boston College. http://timssandpirls.bc.edu/pirls2006/tech_rpt.html

Mullis, I.V.S., Kennedy, A.M., Martin, M.O., and Sainsbury, M. (2006). *PIRLS 2006 Assessment Framework and Specifications, 2nd ed*. Chestnut Hill, MA: Boston College College http://timssandpirls.bc.edu/pirls2006/framework.html

Mullis, I.V.S., Martin, M. O., Kennedy, A.M., Trong, K., and Sainsbury, M. (2009). *PIRLS 2011 Assessment Framework*. Chestnut Hill, MA: Boston College. http://timssandpirls.bc.edu/pirls2011/framework.html

Page intentionally left blank

www.ed.gov

ies.ed.gov