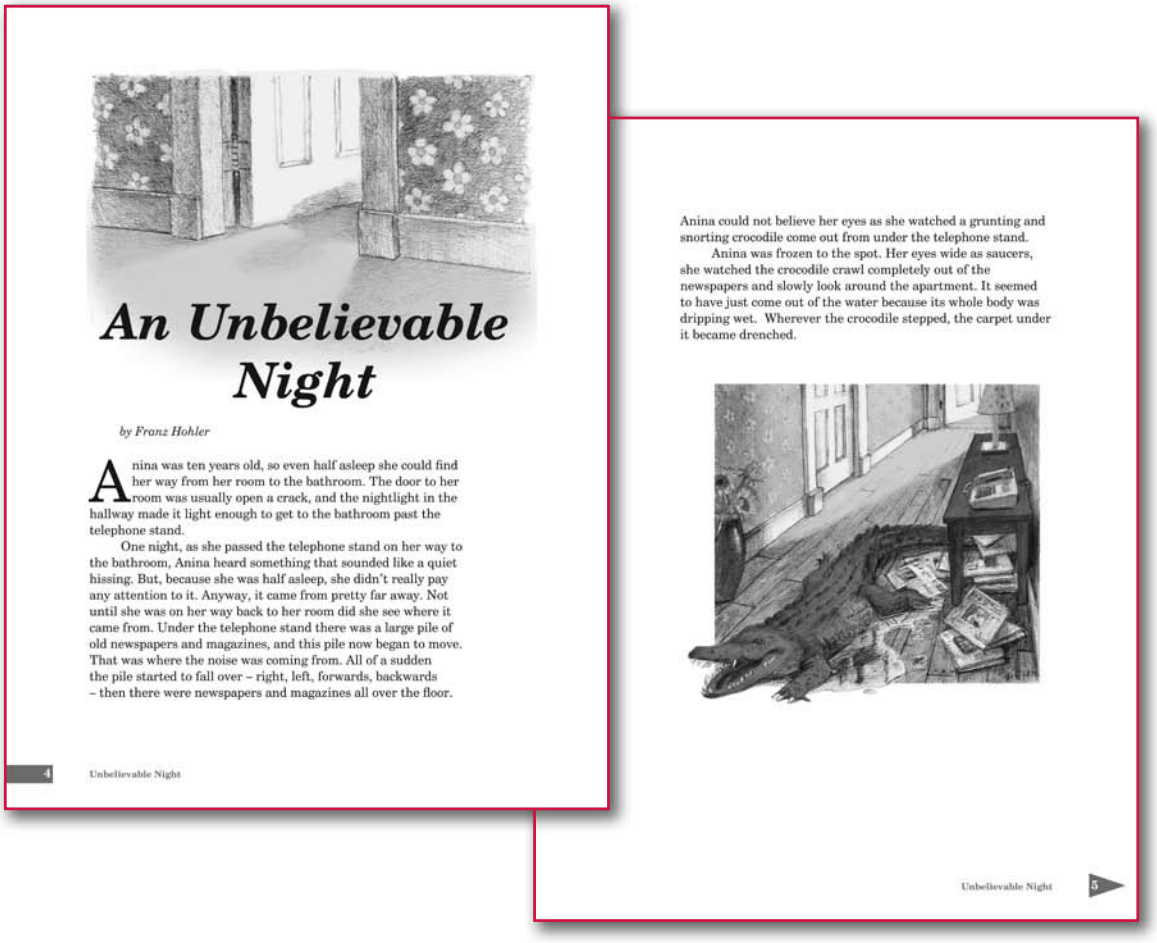
Appendix A: Sample Items From PIRLS 2006

This appendix contains a sample reading passage from PIRLS 2006 as well as several of the assessment items associated with the passage. The assessment items show actual student responses and also compare the performance of U.S. fourth-grade students on the item with the international average. The items demonstrate acceptable performance across the PIRLS international benchmarks (low, intermediate, high, and advanced). The reading passage and all associated items have been publicly released by the IEA.



An Unbelievable Night

by Franz Hohler

Anina was ten years old, so even half asleep she could find her way from her room to the bathroom. The door to her room was usually open a crack, and the nightlight in the hallway made it light enough to get to the bathroom past the telephone stand.

One night, as she passed the telephone stand on her way to the bathroom, Anina heard something that sounded like a quiet hissing. But, because she was half asleep, she didn't really pay any attention to it. Anyway, it came from pretty far away. Not until she was on her way back to her room did she see where it came from. Under the telephone stand there was a large pile of old newspapers and magazines, and this pile now began to move. That was where the noise was coming from. All of a sudden the pile started to fall over – right, left, forwards, backwards – then there were newspapers and magazines all over the floor.

4 Unbelievable Night

Anina could not believe her eyes as she watched a grunting and snorting crocodile come out from under the telephone stand.

Anina was frozen to the spot. Her eyes wide as saucers, she watched the crocodile crawl completely out of the newspapers and slowly look around the apartment. It seemed to have just come out of the water because its whole body was dripping wet. Wherever the crocodile stepped, the carpet under it became drenched.

Unbelievable Night 5

Reading passage continued on the next page.

The crocodile moved its head back and forth letting out a loud hissing sound. Anina swallowed hard, looking at the crocodile's snout with its terribly long row of teeth. It swung its tail slowly back and forth. Anina had read about that in "Animal Magazine"– how the crocodile whips the water with its tail to chase away or attack its enemies.

Her gaze fell on the last issue of "Animal Magazine," which had fallen from the pile and was lying at her feet. She got another shock. The cover of the magazine used to have a picture of a big crocodile on a river bank. The river bank was now empty!

Anina bent down and picked up the magazine. At that moment the crocodile whipped his tail so hard that he cracked the big vase of sunflowers on the floor and the sunflowers scattered everywhere. With a quick jump Anina was in her bedroom. She slammed the door shut, grabbed her bed and pushed it up against the door. She had built a barricade that would keep her safe from the crocodile. Relieved, she let her breath out.

But then she hesitated. What if the beast was simply hungry? Maybe to make the crocodile go away you had to give it something to eat?

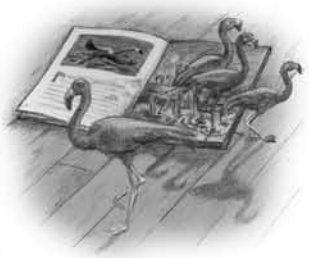
Anina looked again at the animal magazine. If the crocodile could crawl out of a picture then perhaps other animals could too. Anina hastily flipped through the magazine and stopped at a swarm of flamingos in a jungle swamp. Just right, she thought. They look like a birthday cake for crocodiles.

Suddenly there was a loud crack and the tip of the crocodile's tail pushed through the splintered door.

Quickly, Anina held the picture of the flamingos up to the hole in the door and called as loud as she could, "Get out of the swamp! Shoo! Shoo!" Then she threw the magazine through the hole into the hallway, clapped her hands and yelled and screamed.

She could hardly believe what happened next. The entire hallway was suddenly filled with screeching flamingos wildly flapping their wings and running around all over the place on their long, skinny legs. Anina saw one bird with a sunflower in its beak and another grabbing her mother's hat from its hook. She also saw a flamingo disappear into the crocodile's mouth. With two quick bites he swallowed the flamingo and quickly followed it with another, the one with the sunflower in its beak.

After two portions of flamingo the crocodile seemed to have had enough and lay down contentedly in the middle of the hallway. When he had closed his eyes and no longer moved, Anina quietly opened her door and slipped through it into the hallway. She placed the empty magazine cover in front of the crocodile's nose. "Please," she whispered, "please go back home." She crept back into the bedroom and looked through the hole in the door. She saw the crocodile back on the cover of the magazine.



She now went cautiously into the living room where the flamingos were crowded around the sofa and standing on the television. Anina opened the magazine to the page with the empty picture. "Thank you," she said, "thank you very much. You may now go back to your swamp."

In the morning, it was very difficult for her to explain the giant wet spot on the floor and the broken door to her parents. They weren't convinced about the crocodile even though her mother's hat was nowhere to be found.

Adapted from *Eine Wilde Nacht* in *Der Große Zwerg und Andere Geschichten* by Franz Hohler. Published in 2003 by Deutscher Taschenbuch Verlag, München, Germany.

Figure A-1. Example A of item at PIRLS low international benchmark: 2006

1 Point: Full-credit sample response

7. How did the bedroom door get broken?

- ● The crocodile's tail pushed through it.
- (B) The big vase cracked against it.
- (C) The flamingo's sharp beak crashed into it.
- (D) The bed smashed against it.

Percentage of students earning full-credit	
International average	77
United States	83*

*p < .05. Significantly different from international average at the .05 level of statistical significance.
SOURCE: International Association for the Evaluation of Educational Achievement, Progress in International Reading Literacy Study (PIRLS), 2006.

Figure A-2. Example B of item at PIRLS low international benchmark: 2006

1 Point: Full-credit sample response

9. At the end of the story, how did Anina feel toward the flamingos?

- (A) guilty
- (B) cautious
- ● grateful
- (D) annoyed

Percentage of students earning full-credit	
International average	69
United States	61*

*p < .05. Significantly different from international average at the .05 level of statistical significance.
SOURCE: International Association for the Evaluation of Educational Achievement, Progress in International Reading Literacy Study (PIRLS), 2006.

Figure A-3. Example of item at PIRLS intermediate international benchmark: 2006

1 Point: Full-credit sample response

5. Put the following sentences in the order in which they happened in the story.

The first one has been done for you.

2 Anina saw the crocodile.

4 The crocodile ate two flamingos.

5 Anina tried to explain the broken door to her parents.

1 Anina started to walk to the bathroom.

3 Anina ran to the bedroom and slammed the door.

Percentage of students earning full-credit	
International average	67
United States	79*

*$p < .05$. Significantly different from international average at the .05 level of statistical significance.
SOURCE: International Association for the Evaluation of Educational Achievement, Progress in International Reading Literacy Study (PIRLS), 2006.

Figure A-4. Example of item at PIRLS high international benchmark: 2006

2 out of 2 Points: Full-credit sample response

8. How did the magazine help Anina? Write **two** ways.

1. The magazine help Anina to tell where the crocodile came from.

2. It help Anina to know what the crocodile would do when it was going to attack.

Percentage of students earning full-credit	
International average	41
United States	54*

*$p < .05$. Significantly different from international average at the .05 level of statistical significance.
SOURCE: International Association for the Evaluation of Educational Achievement, Progress in International Reading Literacy Study (PIRLS), 2006.

Figure A-5. Example of item at PIRLS advanced international benchmark: 2006

3 out of 3 Points: Full-credit sample response

11. You learn what Anina was like from the things she did. Describe what she was like and give **two** examples of what she did that show this.

Anina was a girl who noticed things. In the story, she noticed the crocodile's tail swinging and she noticed he was gone from the magayene.

	Percentage of students earning full-credit
International average	16
United States	22*

*$p < .05$. Significantly different from international average at the .05 level of statistical significance.
SOURCE: International Association for the Evaluation of Educational Achievement, Progress in International Reading Literacy Study (PIRLS), 2006.

Page intentionally left blank

Appendix B: Technical Notes

Introduction

This appendix describes the sampling, data collection, test development and administration, weighting and variance estimation, scaling, and statistical testing procedures used to collect and analyze the data for the 2006 Progress in International Reading Literacy Study (PIRLS). Forty-five jurisdictions participated in PIRLS 2006, which collected data on the reading literacy of students in their fourth year of schooling (fourth-grade students in most participating jurisdictions, including the United States).

PIRLS 2006 is the second administration of the study, which was first administered in 2001. The study is conducted by the International Association for the Evaluation of Educational Achievement (IEA), with national sponsors in each participating jurisdiction. In the United States, PIRLS is sponsored by the National Center for Education Statistics (NCES), in the Institute of Education Sciences in the U.S. Department of Education. Further information about the technical aspects of the assessment are available in the international PIRLS 2006 technical report (Martin, Mullis, and Kennedy 2007).

Sampling, data collection, and response rate benchmarks

The PIRLS 2006 international project team instituted a series of sampling, data collection, and response rate benchmarks to ensure international comparability and to provide the ability to produce precise estimates of the main criterion variables for all jurisdictions.

The target population for PIRLS was defined by IEA using the International Standard Classification of Education (ISCED), developed by the United Nations Educational, Scientific, and Cultural Organization (UNESCO 1999). The target population of interest was all students enrolled in the grade corresponding to the fourth year of schooling, beginning with ISCED Level 1. For most jurisdictions, this was the fourth grade or its national equivalent. This definition is different from the one used in 2001, which targeted students in the upper of the two grades that include the most 9-year-olds, which in most jurisdictions was the fourth grade. Table B-1 provides information on ISCED levels for the United States.

Table B-1. International Standard Classification of Education (ISCED) levels, definitions, and U.S. equivalents in preprimary through 12th grade

ISCED level	Definition	U.S. equivalent
0	Preprimary	Kindergarten and below
1	Primary	1st–6th grades
2	Lower secondary	7th–9th grades
3	Upper secondary	10th–12th grades or first 3 years of vocational education

SOURCE: Matheson, N., Salganik, L., Phelps, R., Perie, M., Alsalam, N., and Smith, T., (1996). *Education Indicators: An International Perspective*. U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Standardized procedures for sampling were developed by IEA and disseminated in a school sampling manual. Statistics Canada was responsible for approving the designs and verifying the samples of all participating jurisdictions. The basic sample design called for a two-stage stratified cluster design, with schools selected at the first stage and classrooms at the second stage. Schools were sampled using a probability proportionate to size sampling method. Within each jurisdiction 150 schools were selected. Information on the number of classrooms containing fourth-grade students, and the size of the classes, were collected from participating schools and entered into the within school sampling software provided by IEA. In most jurisdictions, one or two classes per school were randomly selected using this software. All students in sampled classrooms were selected.

IEA also established sample size and response rate targets for all jurisdictions. As table B-2 shows, the response rate target for schools was set at 85 percent, with a minimum participation rate among "original sample schools" of 50 percent. When the original sample was drawn, the schools immediately before and immediately after each sampled school on the sampling frame were designated "replacement" schools and were contacted if the original sample school refused to participate. The response rate target for classrooms was 95 percent, and the target student response rate was set at 85 percent. In addition, classrooms with student participation below 50 percent were to be rejected from inclusion with the final data. Substitution of sampled classrooms was not permitted, and the school would be classified as a non-respondent if no other classrooms had been sampled. No U.S. schools were classified as non-respondents on the basis of these criteria.

The IEA's minimum acceptable rate for overall sample participation after replacement (the product of the school participation rate and the student participation rate) was 75 percent. In 2006, the overall sample participation rate for Norway was 71 percent. Consequently, all data reported for Norway in this report have the following footnote: "Did not meet guidelines for sample participation rates after replacement schools were included."

The goal of the study was to provide 95 percent coverage of the target population within each jurisdiction. Jurisdictions that excluded more than 5 percent of students for any reason are noted in the international report as having less than full coverage of the target population.

Sampling, data collection, and response rates in the United States

Sampling

The PIRLS sample in the United States was designed to be representative of all fourth-grade students in the 50 states and the District of Columbia. In addition to the base sample (designed to yield 150 participating schools), the United States sampled additional private schools and high-poverty schools, defined as those schools in which 50 percent or more of students were

Table B-2. IEA minimum sample size and unweighted response rate targets for participating PIRLS jurisdictions: 2006

Group	Minimum sample size (number)	Unweighted response rate (percent)
Schools	150	85[1]
Classrooms	1 per sampled school	95
Teachers	1 per sampled school	85
Students	4,500	85

[1]At least 50 percent must be original sample schools.
SOURCE: International Association for the Evaluation of Educational Achievement, Progress in International Reading Literacy Study (PIRLS), 2006.

eligible to receive free or reduced-price lunch in order to increase the precision of the estimates for these subgroups. The U.S. sample was designed to yield 180 participating schools.

The PIRLS school sample was drawn in March 2005. The sampling frame was constructed using data from the 2002–03 Common Core of Data (CCD) and Preliminary Data 2003–04 Private School Universe Survey (PSS).

To be consistent with the sampling design for PIRLS 2001, the frame was divided into two parts: (1) One stratum was created that included schools located in the 10 most populous Metropolitan Statistical Areas (MSAs); (2) all schools outside those MSAs were grouped into 451 Primary Sampling Units (PSUs) by sorting on MSA and then by the Federal Information Processing Standards (FIPS) code. PSUs were designed to fit within state boundaries and, where possible, within county and city boundaries. In the United States, schools were sorted by state, percentage of racial/ethnic minority students, control of school (public/private), percentage of students eligible for free or reduced-priced lunch, and locale before the selection process.

Locale was determined on the basis of a sampled school's address. School addresses were classified into one of three categories (central city, urban fringe/large town, or rural/small town) using the NCES locale code system in use at the time of sampling. The locale code system used the following designations:

- Large city: A central city of a Consolidated Metropolitan Statistical Area (CMSA) or MSA, with the city having a population greater than or equal to 250,000.
- Midsize city: A central city of a CMSA or MSA, with the city having a population less than 250,000.
- Urban fringe of a large city: Any territory within a CMSA or MSA of a large city and defined as urban by the Census Bureau.
- Urban fringe of a midsize city: Any territory within a CMSA or MSA of a midsize city and defined as urban by the Census Bureau.
- Large town: An incorporated place or Census-designated place with a population greater than or equal to 25,000 and located outside a CMSA or MSA.
- Small town: An incorporated place or Census-designated place with a population less than 25,000 and greater than or equal to 2,500 and located outside a CMSA or MSA.
- Rural, Outside MSA: Any territory designated as rural by the Census Bureau that is outside a CMSA or MSA of a large or midsize city.
- Rural, Inside MSA: Any territory designated as rural by the Census Bureau that is within a CMSA or MSA of a large or midsize city.

For this analysis, large city and midsize city were combined to form central city; urban fringe of a large city, urban fringe of a midsize city, and large town were combined to form urban fringe/large town; and small town, rural, outside MSA, and rural, inside MSA were combined to form rural/small town.

Within each selected PSU or MSA stratum, schools were selected on the basis of the number of fourth-grade students in the school so that larger schools had a higher probability of selection than smaller schools. The final sample included 222 schools; 152 were chosen from PSUs and 70 were selected from the MSA stratum. The target number of students was designed to be similar across schools, both large and small, correcting for the greater likelihood of selection of large schools.

Data collection

School contacting began in April 2005, approximately 1 year prior to data collection. The suggested testing window for PIRLS in the southern hemisphere was October through December, 2005, and in the northern hemisphere it was March through June, 2006. The United States was allowed to begin early (on January 23) to accommodate schools that wished to participate before state-mandated tests occurred. Many U.S. schools also asked to participate after completing state tests, and so the United States was allowed to continue

testing through June 9, 2006, resulting in a 4½ month test window rather than the more typical 1 to 2 month test window. The mean score of students completing the exam in January through March was 539.5, which was not significantly different from the score (541.1) of the students completing the exam in April through June.

Response rates

Of the 222 sampled schools, 214 were eligible for inclusion in PIRLS. Eight schools had closed and were designated ineligible. Of the 214 eligible original sample schools, 120 participated (57 percent weighted). An additional 63 replacement schools were contacted and agreed to participate, for a total of 183 schools, or a weighted response rate, using final adjusted weights, of 86 percent of eligible schools.[1] Of the 120 participating schools from the original sample, 88 (73 percent) were from the PSU sample, while 40 of the 63 participating replacement schools (63 percent) were from the PSU sample. The United States met the international guidelines for school response rate, but only after using replacement schools.

Information on the number and size of classrooms containing fourth-grade students was collected from all participating schools. One or two classrooms were randomly selected from each school depending on the size of the school. Of the 256 classrooms sampled, 255 participated, or 99 percent. There were 5,601 fourth-grade students enrolled in the selected classrooms; 159 of these students were excluded from testing (see "Exclusions" for more information). Within these classrooms, 5,442 students were eligible, and 5,190 completed the assessment, for a weighted student response rate of 95 percent. The United States met the international guidelines for classroom and student response rates.

[1]All weighted response rates discussed in this report refer to final adjusted weights. Response rates were calculated using the formula developed by the IEA for PIRLS. The standard NCES formula for computing response rates would result in a lower school response rate of approximately 63 percent.

In addition to having students complete the assessment and a questionnaire, PIRLS asked teachers and school administrators to complete questionnaires. Of the 256 teachers sampled, 249 completed teacher questionnaires, or 97 percent. Among school administrators, 182 of the 183 questionnaires were completed, for a response rate of 99 percent.

Table B-3 presents information on the total number of participating schools, students assessed, and overall weighted response rates after replacement in all jurisdictions that participated in PIRLS.

Exclusions

Schools that were very small or that were classified as special education, vocational, or alternative schools (private and public) could be excluded from the sampling frame. In the United States these schools enrolled 3.2 percent of the expected number of fourth-grade students. Table B-4 shows the percentage of students excluded from the sample in 2001 and 2006.

International guidelines recognized that some students might not be eligible for inclusion in PIRLS because of limited exposure to the language of assessment (English in the case of the United States) or the need for special testing accommodations.

Within classrooms, students were excluded from participation in PIRLS if they met the criteria established by the IEA:

- *Functionally disabled students*. These are students who are permanently physically disabled in such a way that they cannot perform in the PIRLS testing situation. Functionally disabled students who could perform were included in the testing.
- *Intellectually disabled students*. These are students who are considered in the professional opinion of the school administrator or by other qualified staff members to be intellectually disabled or who have been psychologically tested as such. This includes students who are emotionally or mentally unable to

Table B-3. Total number of participating schools, students assessed, and overall weighted response rates, by participating PIRLS jurisdictions: 2006

Jurisdiction	Total number of participating schools	Total number of students assessed	Overall weighted response rate
Austria	158	5,067	97
Belgium (Flemish)	137	4,479	91
Belgium (French)	150	4,552	95
Bulgaria	143	3,863	94
Canada, Alberta	150	4,243	96
Canada, British Columbia	148	4,150	94
Canada, Nova Scotia	201	4,436	96
Canada, Ontario	180	3,988	87
Canada, Quebec	185	3,748	81
Chinese Taipei	150	4,589	99
Denmark	145	4,001	96
England	148	4,036	92
France	169	4,404	95
Georgia	149	4,402	98
Germany	405	7,899	92
Hong Kong, SAR	144	4,712	97
Hungary	149	4,068	97
Iceland	128	3,673	90
Indonesia	168	4,774	98
Iran	236	5,411	99
Israel	149	3,908	93
Italy	150	3,581	97
Kuwait	149	3,958	88
Latvia	147	4,162	92
Lithuania	146	4,701	92
Luxembourg	178	5,101	99
Macedonia	150	4,002	96
Moldova	150	4,036	95
Morocco	159	3,249	94
Netherlands	139	4,156	90
New Zealand	243	6,256	95
Norway	135	3,837	71
Poland	148	4,854	95
Qatar	119	6,680	94
Romania	146	4,273	97
Russian Federation	232	4,720	97
Scotland	130	3,775	81
Singapore	178	6,390	95
Slovak Republic	167	5,380	94
Slovenia	145	5,337	93
South Africa	397	14,657	88
Spain	152	4,094	97
Sweden	147	4,394	96
Trinidad and Tobago	147	3,951	94
United States	**183**	**5,190**	**82**

NOTE: The overall weighted response rate is the product of the school participation rate, after replacement, and the student participation rate, after replacement.
SOURCE: International Association for the Evaluation of Educational Achievement, Progress in International Reading Literacy Study (PIRLS), 2006.

follow the general instructions of the test. Students were not excluded solely because of poor academic performance or normal disciplinary problems.

- *Non-native language speakers.* These are students who are unable to read or speak the language of the test and would be unable to overcome the language barrier in the test situation. Typically, students who received less than 1 year of instruction in the language of the test were to be excluded, but this definition could be adapted in different jurisdictions. In the United States, students who had received less than 1 year of English instruction were defined as non-native language speakers.

In the United States, 2.8 percent of students were excluded from PIRLS on the basis of these criteria. In keeping with international protocol, no testing accommodations were offered to students.

The overall exclusion rate was 5.9 percent in the United States, which means that the overall U.S. coverage rate is .09 percent below the recommended 95 percent. Other jurisdictions that had exclusion rates above 5.0 percent included Bulgaria (6.4); the province of Ontario, Canada (8.3); Israel (22.5); Italy (5.3); Lithuania (5.1); New Zealand (5.3); and the Russian Federation (7.7).

Table B-4. Percentage of U.S. students excluded from PIRLS at the school-listing level and student-listing level: 2001 and 2006

Level	2001	2006
Total	5.3	5.9
Excluded at the school listing level	0.6	3.2
Excluded at the student listing level	4.7	2.8

NOTE: Detail may not sum to totals because of rounding.
SOURCE: International Association for the Evaluation of Educational Achievement, Progress in International Reading Literacy Study (PIRLS), 2001 and 2006.

Nonresponse bias analysis

The analysis of school nonresponse was conducted in two parts. The basis for both analyses was the original sample of 214 eligible schools. First, the distribution of the 120 responding original sample schools was compared with that of the total sample of eligible original schools. All original schools in the sample that declined to participate in the study were treated as nonparticipants regardless of whether they were substituted by a replacement school. In the second part, replacement schools were included in the analysis, reflecting the final sample of schools that participated in PIRLS 2006.

Seven variables were examined using the original sample, the participating schools from the original sample, and the participating schools in the final sample: (1) public/private school control, (2) locale, (3) region, (4) percentage of students eligible for free or reduced-price lunch, (5) total school enrollment, (6) fourth-grade enrollment, and (7) relative enrollment of racial and ethnic groups (White, non-Hispanic; Black, non-Hispanic; Hispanic; Asian or Pacific Islander; American Indian or Alaska Native; and other).

Measures of bias and relative bias were computed, and the hypothesis of independence between the characteristic and participation status was tested using chi-square statistics. In addition, logistic regression models were used to evaluate whether any of these characteristics were significant in predicting response status. A comparison of the participating schools from the original sample with the total eligible sample of schools found that school composition was significantly different across the two groups: the mean percentage of Asian students in schools in the eligible sample was 3.5 percent, while among participating original sample schools it was 2.4 percent; the measure of bias is 1.07. No other variables were found to differ significantly between these two groups.

In the second analysis, the final sample of all participating schools (both original and replacement) was compared to the total eligible sample. In this analysis, the percentage of Asian students in the school was not significantly different between the two groups. However, the number of fourth-grade students enrolled in the school was related to nonresponse. Schools with fewer students enrolled in fourth grade (schools with an average of 67 students in the fourth grade) were

less likely to participate than larger schools (schools with an average of 71.2 students in fourth grade); the measure of bias is 4.17. It is unclear whether this bias has any impact on student achievement scores. More detailed information on nonresponse bias analysis, including item nonresponse analysis, can be found in Krotki and Bland (2007).

Test development

The International Study Center (ISC), which organized and managed the international components of PIRLS, developed an assessment framework used to guide the test development process (Mullis et al. 2006). PIRLS was designed to assess two purposes of reading: reading for literary experience and reading to acquire and use information. In addition, the PIRLS assessment evaluates four processes of comprehension: (1) to focus on and retrieve explicitly stated information; (2) to make straightforward inferences; (3) to interpret and integrate ideas and information; and (4) to examine and evaluate content, language, and textual elements.

Jurisdictions participating in PIRLS 2006 were invited to submit reading passages to be used in the test. Two types of passages were sought: literary texts, which were typically narrative fiction, and informational texts, which could include biographies, step-by-step directions, informational leaflets, and scientific or other nonfictional material. All passages were to be authentic texts typical of the reading material in their jurisdictions, well suited to fourth-grade students, and no longer than 1,000 words. The national research coordinators from participating jurisdictions were asked to review the texts and work together to agree on a shortened list of passages to be illustrated and formatted. Questions for each passage were refined by PIRLS project staff and reviewed by a group of reading experts. Each reading passage, including text and questions, was designed to be completed in 40 minutes.

Twelve new passages were created and tested during a field trial in spring 2005. Item statistics, including item difficulties, point biserial correlations, and item discrimination statistics, were calculated for each item for each jurisdiction. After a careful review of the quality of all items across jurisdictions, 6 of these passages, 3 literary and 3 informational, were selected for the main study.

These passages, along with 4 passages from PIRLS 2001, were used to create the test booklets for the main study. The same 10 passages were used in all participating PIRLS jurisdictions. Each test booklet contained 2 reading passages. Students were given 40 minutes to complete each passage, or 80 minutes in all. The passages were distributed across 13 booklet types. Students were asked to answer a number of items related to each passage, including both multiple-choice and constructed-response items. The distribution of the items by type of passage and type of item is shown in table B-5.

In addition to the assessment, students were asked to complete a 20-minute questionnaire. The questionnaire included items about students' reading experiences in

Table B-5. Distribution of items on the PIRLS 2006 assessment

Reading purpose	Multiple-choice items	Constructed-response items			Total number of items	Total score points
		1 point	2 points	3 points		
Total	64	28	27	7	126	167
Literary	34	13	13	4	64	85
Informational	30	15	14	3	62	82

SOURCE: International Association for the Evaluation of Educational Achievement, Progress in International Reading Literacy Study (PIRLS), 2006.

school, self-perception and attitudes toward reading, out-of-school reading habits and computer use, home literacy resources, and basic demographic information.

Translation

Source versions of all instruments (assessment booklets, questionnaires, and manuals) were prepared in English and then translated into the primary language or languages of instruction in each jurisdiction. In addition, it was sometimes necessary to adapt the instrument for cultural purposes, even in jurisdictions such as the United States that use English as the primary language of instruction. For example, words such as "lift" might be adapted to "elevator" for the United States. The IEA and ISC verified the translations and adaptations used by all participating jurisdictions. Certified translators were retained by the IEA to compare national versions with the source versions of all documents.

Test administration and quality assurance

PIRLS 2006 emphasized the use of standardized procedures in all jurisdictions. Each jurisdiction was responsible for its own data collection; however, the IEA insisted that all jurisdictions use the procedures and materials developed by the international project team. The ISC developed standardized survey operations manuals that were used in all jurisdictions, as well as manuals for participating schools and test administrators, to ensure that data collection processes were consistent across jurisdictions. In addition, jurisdictions used standardized listing forms for student participation and standardized session report forms.

Test administration in the United States was carried out by a professional staff trained according to the international guidelines. School personnel were asked to assist with listing classrooms and students, selecting a test day, and selecting the parental consent procedures to be used at that school. Test administrators were responsible for all other aspects of the administration.

The ISC conducted quality monitoring visits at approximately 15 sampled schools in each jurisdiction. The international quality monitors were trained by the staff of the ISC and the IEA Secretariat. After each visit, the quality monitor completed a standard form describing the test session and any deviations from international protocols.

In addition, each jurisdiction was encouraged to conduct its own national quality monitoring operation. In the United States, a sample of 10 percent of schools was selected for monitoring. Project staff and field supervisory staff visited selected schools during the assessment administration and completed a classroom observation record immediately after the visit.

Both international and national quality monitors were asked to verify that student and class lists were prepared correctly by the school personnel; verify the completeness and security of the test booklets; check when possible that the international guidelines concerning the exclusion of students had been properly followed; keep an independent record of session timing; verify adherence to the script and instructions outlined in the test administrator manual; check that materials were distributed correctly; indicate whether the students were cooperative during the test session; and note whether the test administrator monitored that students were working in the correct section of the test booklet.

Scoring

PIRLS contained a large number of constructed-response items, as discussed in the test development section. The process of scoring these items was an important step in ensuring the quality and cross-jurisdiction comparability of the PIRLS data. Detailed guidelines were developed for the scoring guides themselves, and training materials were prepared including an extensive set of anchor and practice papers. These materials were prepared by the ISC with the advice and guidance of an international group of experts.

In spring 2006, the ISC organized an international training session to present the material and train the scoring coordinators from participating jurisdictions, who in turn trained the national scorers. For each test item, the scoring guide described the intent of the question and how to code students' responses to each item. This description included guidelines for assigning full credit, partial credit, or no credit for each item. During the training session, PIRLS staff discussed the scoring guidelines for each item and reviewed the anchor papers (selected examples of real student answers) for each item. Trainees were asked to complete the practice papers, and the answers were then discussed.

The criteria described in the scoring guides related only to evidence of reading comprehension. Students' writing abilities were not evaluated. A student could receive a high score for an item if the ideas expressed in the response exhibited a high level of understanding, even if the response contained misspellings or grammatical errors. Given that PIRLS was a timed test, responses were considered "first-draft writing."

The reliability of coding was assessed in three ways. First, to establish within-jurisdiction scoring reliability, it was necessary for two different scorers to independently score a random sample of 200 responses for each constructed-response item. The degree of agreement between the scores assigned by the two scorers was a measure of the reliability of the scoring process. The average of each percent agreement across items was 93 percent, both for the United States and the international average. Second, international scoring reliability was assessed by having each jurisdiction use the IEA's Cross-Country Scoring Reliability software to score a common set of answers selected from field test and PIRLS 2001 responses. Finally, in jurisdictions that participated in both 2001 and 2006, the staff scoring the 2006 responses were also asked to score a sample of 2001 responses. The scores assigned in 2006 were then compared with the actual scores assigned to those responses in 2001. Information on trend and cross-jurisdiction reliability is available in the international technical report (Martin et al. 2007).

Data entry and cleaning

The national research coordinator for each jurisdiction assumed responsibility for data entry. All data were entered into a data entry system developed by the IEA Data Processing Center (IEA-DPC) with a number of built-in data quality checks. In addition, each jurisdiction was required to run a number of validity checks (e.g., checking the links among teachers, schools, and students) before delivering the data to the IEA-DPC. The IEA-DPC conducted a number of additional cleaning steps before providing each jurisdiction with a version of the cleaned data to be reviewed and accepted by the jurisdiction. The U.S. data were cleared through this process and no major issues were found.

Weighting and variance estimation

Using sampling weights is necessary for computing statistically sound, nationally representative estimates. Survey weights help adjust for the intentional over- or undersampling of certain sectors of the population, school or student nonresponse, or errors in estimating the size of a school at the time of sampling. Survey weighting for the entire international PIRLS 2006 sample was carried out by Statistics Canada.

The internationally defined weighting specifications for PIRLS required that each assessed student's sampling weight be the product of six weighting factors: the inverse of the school's probability of selection, an adjustment for school-level nonresponse, the inverse of the classroom's probability of selection, an adjustment for classroom-level nonresponse, the inverse of the student's probability of selection (always equal to 1 because whole classrooms were selected), and an adjustment of student-level nonresponse.

The statistics presented in this report are estimates of group and subgroup performance based on a sample of fourth-graders, rather than the values that could be calculated if every fourth-grader answered every question on the instrument. It is therefore important to have

measures of the degree of uncertainty of the estimates. Accordingly, in addition to providing estimates of percentages of respondents and their average scale score, this report provides information about the uncertainty of each statistic.

Because PIRLS used clustered sampling, conventional formulas for estimating sampling variability that assume simple random sampling and hence independence of observations are inappropriate. For this reason, PIRLS used a jackknife repeated replication method (Johnson and Rust 1992) to estimate standard errors that capture the sampling variance.

Scaling and plausible values

Each student who completed the PIRLS assessment read 2 passages, rather than all 10 passages developed for the study, to keep individual response burden to a minimum. PIRLS used a matrix-sampling design to assign passages to booklets. Item Response Theory (IRT) was then used to combine these responses to provide accurate estimates of reading achievement in the student population in each jurisdiction.

As was done in 2001, PIRLS used three distinct scaling models: a three-parameter model for multiple-choice items, a two-parameter model for constructed-response items that were scored as correct or incorrect, and a partial credit model for constructed-response items with more than two score points.

Because each student completed only a limited set of items, plausible values were calculated to estimate student-level scores. PIRLS generated five possible scale scores for each student, which represented selections from the distribution of scale scores of students with similar backgrounds who answered the assessment items the same way. The plausible values methodology is one way to ensure that the estimates of the mean performance of student subpopulations and the estimates of variability in those means are more accurate than those determined through traditional procedures, which estimate a single score for each student. During the construction of plausible values, careful quality control steps ensure that the subpopulation estimates based on these plausible values are accurate.

It is important to recognize that plausible values are not test scores for individuals, and they should not be treated as such. Plausible values are randomly drawn from the distribution of scores that could be reasonably assigned to each individual. As such, the plausible values contain random error variance components and are not optimal as scores for individuals. The PIRLS student file contains 15 plausible values per student, 5 for each of the three scales (the combined reading literacy scale, the literary subscale, and the informational subscale). If an analysis is to be undertaken with one of these scales, then (ideally) the analysis should be undertaken five times, once with each of the 5 relevant plausible value variables. The results of these five analyses are averaged, and then significance tests that adjust for variation between the five sets of results are computed.

Descriptions of background variables

In the United States, background questionnaires were administered to students, school administrators, and teachers. The information collected from the background questionnaires provides a context for interpreting the results from the assessment. The following background variables are presented in this report:

Sex

Students were asked to indicate whether they were a boy or a girl.

Race/ethnicity

School administrators were asked to classify the race/ethnicity of each sampled student into one or more of the following categories:

White
Black
Hispanic
Asian
American Indian/Alaska Native
Pacific Islander

For reporting, all students who were identified as Hispanic by their school's administrator were classified as Hispanic, regardless of their race. The remaining categories include only students who were identified as non-Hispanic. The other, non-Hispanic category includes non-Hispanic students identified as Pacific Islander as well as those non-Hispanic students identified as belonging to multiple racial groups. Because the number of Pacific Islander and multiple-race students was each too small to report separately (fewer than 30 students in each group), the two groups were combined into the other, non-Hispanic category.

In 2001, data about the race and ethnicity of students were collected directly from student responses. The 2001 student background questionnaire also defined White and Black as White (not Hispanic) and Black (not Hispanic), respectively. Because the classification of racial/ethnic categories and procedures for collecting data on race/ethnicity changed between 2001 and 2006, no comparisons between racial/ethnic groups in 2001 and 2006 are presented in this report.

School poverty level

In this report, the percentage of students in schools eligible for the National School Lunch Program (NSLP) is used as a measure of a school's poverty level. The guidelines for the NSLP stipulate that children from families with incomes at or below 130 percent of the federal poverty level are eligible for free meals, while those between 130 percent and 185 percent of the federal poverty level are qualified for reduced-price meals. (For the period July 1, 2005, through June 30, 2006, for a family of four, 130 percent of the poverty level was $25,155 per year, and 185 percent was $35,798. See http://www.fns.usda.gov/cnd/lunch/ for more information.)

Information about the poverty level of a school was collected from school administrators. Administrators were asked to indicate the percentage of students in their schools eligible for free or reduced-price lunch using the following categories: All, Some, or None.

Data limitations

As with any study, there are limitations to PIRLS that researchers should take into consideration. Estimates produced using data from PIRLS are subject to two types of error: nonsampling errors and sampling errors. Nonsampling errors can be due to errors made in the collection and processing of data. Sampling errors can occur because the data were collected from a sample rather than a complete census of the population. In addition to sampling errors, researchers should also be aware of missing data issues and how these issues were addressed.

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. For example, the sampling frame was limited to regular public and private schools in the 50 states and the District of Columbia and did not include Puerto Rico or the U.S. Trust Territories. The sources of nonsampling errors are typically problems such as unit and item nonresponse, the differences 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. Some of these issues (particularly unit nonresponse) are discussed above in the section entitled "Sampling, data collection, and response rates in the United States." Note that this is a school-based sample; home-schooled children are not included.

It is difficult to identify and estimate either the amount of nonsampling error or the bias caused by this error. In PIRLS, efforts were made to prevent such errors from occurring and to compensate for them when possible. For example, the design phase entailed a field test that evaluated items as well as the implementation procedures for the survey. It should also be recognized that background information was obtained from students' self-reports, which are subject to several different forms of response bias.

Sampling errors

Sampling errors occur when a discrepancy between a population characteristic and the sample estimate arises because not all members of the target population are sampled for the survey. Both the size of the sample relative to the population and the variability of the population characteristics influence the magnitude of sampling error. The particular sample of students drawn in March 2005 was just one of many possible samples that could have been selected. Therefore, estimates produced from the PIRLS 2006 sample may differ from estimates that would have been produced had another sample of fourth-grade students been selected. This type of variability is called sampling error because it arises from using a sample of fourth-grade students in 2006 rather than all fourth-grade students that year.

Missing data

Items with missing data were designated with one of four missing data codes: (1) omitted response or uninterpretable, (2) not administered, (3) not reached, and (4) not applicable. An "omitted response" occurred when a respondent was expected to answer an item but gave no response. An item was coded as "uninterpretable" if some type of response was given but it was either invalid or unreadable. Items that were not administered, either by design or by error (e.g., a printing problem), were coded as "not administered." For assessment questions, the missing data code "not reached" was assigned for consecutive missing values starting from the end of the assessment passage. In the questionnaire data files, a code of "not applicable" was assigned to items that respondents were instructed to skip. All five kinds of missing data were coded distinctly in the PIRLS database.

Background data were not imputed for cases with missing data. Item response rates for variables discussed in this report were over the NCES standard of 85 percent (weighted) to report without notation.

Confidentiality and limitations disclosure

The PIRLS data are hierarchical and include school, teacher, and student data from the participating schools. Confidentiality analyses for the United States were designed to provide reasonable assurance that public-use data files issued by the IEA would not allow the identification of individual U.S. schools, students, or teachers when compared against public-use data collections. Disclosure limitation included identifying and masking potential disclosure risk to PIRLS schools and adding an additional measure of uncertainty to school and student identification through random swapping of data elements within the student, teacher, and school files.

Statistical procedures

Tests of significance

All comparisons discussed in this report have been tested for statistical significance using the t statistic. Statistical significance was determined by calculating a t value for the difference between a pair of means, or proportions, and comparing this value with published tables of values at a certain level of significance, called the alpha level. The alpha level is an a priori statement of the probability of inferring that a difference exists when, in fact, it does not. The alpha level used in this report is .05, based on a two-tailed test.

The calculation of the t statistic varied depending on the type of analysis. For comparisons between independent samples (e.g., an average score for U.S. students compared with an average score for students in another jurisdiction) or between the U.S. average and the international average,[2] the t statistic was calculated as follows:

$$t = \frac{(p_1 - p_2)}{\sqrt{(se_1^2 + se_2^2)}}$$

where p_1 and p_2 are the estimates to be compared and se_1 and se_2 are their corresponding standard errors.

For all other comparisons, the t statistic was calculated by running the jackknife regression procedure available in the International Database (IDB) Analyzer software provided by the IEA. Because of the clustered nature of the PIRLS sample (students within classrooms within schools), seemingly independent samples (e.g., boys and girls) may in fact be correlated. To estimate the standard error of the difference between groups in correlated samples, the jackknife regression calculated the standard error of the difference between the groups being compared for each of the replicate PIRLS samples.[3] The t statistic was calculated by dividing the difference between the two estimates being compared by the average standard error of the difference between the two comparison groups.

Effect Size

Tests of statistical significance are, in part, influenced by sample sizes. To provide the reader with an increased understanding of the size of the significant difference between student populations in the United States, effect sizes for selected results are included in the report. Effect sizes use standard deviations, rather than standard errors, and are therefore not influenced by the size of the student population samples.[4] Following Cohen (1988) and Rosnow and Rosenthal (1996), effect size is calculated by finding the difference between the means of two groups and dividing that result by the pooled standard deviation of the two groups. The formula used to compute effect size (d) is as follows:

$$d = \frac{est_{grp1} - est_{grp2}}{sd_{pooled}}.$$

est_{grp1} and est_{grp2} are the student group estimates being compared. sd_{pooled} is the pooled standard deviation of the groups being compared. The formula for the pooled standard deviation is as follows (Rosnow and Rosenthal 1996):

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

sd_1 and sd_2 are the standard deviations of the groups being compared.

[2]Because U.S. students contribute to the international average, the two samples are not entirely independent. When dependent samples are compared, it is most appropriate to use a different t-test formula that takes account of the overlap between the two samples. Tests of differences between the U.S. average and the international average could not be performed using dependent samples t-tests because the international data were unavailable during the time in which the U.S. data were analyzed. Consequently, the independent samples t statistic was used when comparing a jurisdiction average to the international average.

[3]See Martin et al. (2007) for details on the tests of statistical significance used for correlated samples.

[4]The IDB Analyzer software provided by IEA does not provide the variance or standard deviations of estimates. To calculate these statistics for effect sizes, the estimates for sex, race/ethnicity, school control, and school poverty level were re-run using the *AM* statistical software package.

Page intentionally left blank

Appendix C: PIRLS Technical Review Panel, United States

Linda Baker
Professor, Department of Psychology
University of Maryland, Baltimore County
Baltimore, MD

Richard Durán
Professor, Gevirtz Graduate School of Education
University of California, Santa Barbara
Santa Barbara, CA

Douglas Fuchs
Professor, Department of Special Education
Vanderbilt University
Nashville, TN

Ann Kennedy
Coordinator of Project Development and Operations
PIRLS International Study Center
Boston College
Chestnut Hill, MA

John Sabatini
Research Scientist
Educational Testing Service
Princeton, NJ

Timothy Shanahan
Professor, College of Education
University of Illinois at Chicago
Chicago, IL

Joanna Williams
Professor, Teachers College
Columbia University
New York, NY

Page intentionally left blank

Appendix D: PIRLS Publications

NCES Publications

PIRLS 2006

Krotki, K., and Bland, J. (2007). *PIRLS 2006 Nonresponse Bias Analysis Report* (NCES 2008-019). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

PIRLS 2001

Binkley, M., and Kelly, D.L. (2003). *A Comparison of the NAEP and PIRLS Fourth-Grade Reading Assessments* (NCES 2003–10). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Kapinus, B. (2003). *PIRLS-IEA Reading Literacy Framework: Comparative Analysis of the 1991 IEA Reading Study and the Progress in International Reading Literacy Study* (NCES 2003-05). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Ogle, L.T., 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.

Piesse, A., and Rust, K. (2003). *2001 U.S. PIRLS Nonresponse Bias Analysis* (NCES 2003–21). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

International Publications

PIRLS 2006

Martin, M.O., Mullis, I.V.S., and Kennedy, A.M. (2007). *PIRLS 2006 Technical Report*. Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Lynch School of Education, Boston College.

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: TIMSS and PIRLS International Study Center, Lynch School of Education, Boston College.

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 Schools in 40 Countries*. Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Lynch School of Education, Boston College.

PIRLS 2001

Campbell, J.R., Kelly, D.L., Mullis, I.V.S., Martin, M.O., and Sainsbury, M. (2001). *Framework and Specifications for PIRLS Assessment 2001* (2nd ed.). Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Lynch School of Education, Boston College.

Gonzalez, U.J., and Kennedy, A.M. (2003). *PIRLS 2001 User Guide for the International Database*. Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Lynch School of Education, Boston College.

Martin, M.O., Mullis, I.V.S., Gonzalez, U.J., and Kennedy, A.M. (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: TIMSS and PIRLS International Study Center, Lynch School of Education, Boston College.

Martin, M.O., Mullis, I.V.S., and Kennedy, A.M. (2003). *PIRLS 2001 Technical Report*. Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Lynch School of Education, Boston College.

Mullis, I.V.S., Martin, M.O., and Gonzalez, U.J. (2004). *International Achievement in the Processes of Reading Comprehension*. Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Lynch School of Education, Boston College.

Mullis, I.V.S., Martin, M.O., Gonzalez, U.J., and Kennedy, A.M. (2003). *PIRLS 2001 International Report: IEA's Study of Reading Literacy Achievement in Primary Schools in 35 Countries*. Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Lynch School of Education, Boston College.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., and Flaherty, C. (2002). *PIRLS 2001 Encyclopedia*. Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Lynch School of Education, Boston College.

www.ed.gov

ies.ed.gov