PIRLS 2001 technical report

Chestnut Hill, MA: PIRLS International Study Center, Boston College, Lynch School of Education, 2003

PIRLS 2001 Technical Report

PIRLS 2001 Technical Report

Edited by

Michael O. Martin

Ina V.S. Mullis

Ann M. Kennedy

Contributors

Jay Campbell

Pierre Foy

Eugenio J. Gonzalez

Dirk Hastedt

Ursula Itzlinger

Marc Joncas

Dana L. Kelly

Ann M. Kennedy

Barbara Malak

Michael O. Martin

Ina V.S. Mullis

Marian Sainsbury

Knut Schwippert

IEA

International Association for
the Evaluation of
Educational Achievement

ISC
BOSTON
COLLEGE
LYNCH SCHOOL
OF EDUCATION

International
Study Center

April 2003

PIRLS 2001 Technical Report / Edited by Michael O.
Martin, Ina V.S. Mullis, and Ann M. Kennedy

Publisher: International Study Center,
Lynch School of Education, Boston College

Library of Congress
Catalog Card Number: 2003103847

ISBN: 1-889938-27-0

For more information about PIRLS contact:

PIRLS International Study Center
Lynch School of Education
Manresa House
Boston College
Chestnut Hill, MA 02467
United States

tel: +1-617-552-1600
fax: +1-617-552-1203

e-mail: pirls@bc.edu

http://pirls.bc.edu

Printed and bound in the United States.

Contents

1

Overview of PIRLS

Dana L. Kelly

1.1 Background

IEA has been conducting cross-national studies of educational achievement for more than 40 years – including periodic assessments of children's reading literacy. In 1973, reading was one of the subjects in IEA's six-subject study, which was conducted in 15 countries (Thorndike, 1973; Walker, 1976). In 1991, IEA's Reading Literacy Study was conducted in 32 educational systems (Elley, 1992; 1994). Most recently, PIRLS (Progress in International Reading Literacy Study) was established as the IEA's latest study to monitor progress in children's reading literacy into the future (Mullis, Martin, Gonzalez, & Kennedy, 2003).

In 1998, the IEA General Assembly formally agreed that PIRLS would be part of the IEA's regular cycle of assessments, which also includes mathematics and science. At that point, some basic principles were established.

- PIRLS would begin in 2001 with an assessment of children in fourth grade.

- PIRLS would focus on reading literacy achievement, as well as home and school contexts for learning to read.

- Reading literacy would be measured through a comprehensive assessment based on authentic reading materials requiring students to engage in a range of reading processes.

- The reading test would be designed so that future assessments could measure trends in achievement.

- Questionnaires would be administered to the tested students, their current reading teachers, and their school principals – to collect contextual data with which to interpret achievement.

In 1999, planning for the study began with a meeting among representatives from the IEA Secretariat, the International Study Center (ISC) at Boston College, Statistics Canada, and the National Foundation for Educational Research in England and Wales. At this meeting, it was established that – in addition to incorporating the General Assembly's basic principles – PIRLS would try to collect data from children's parents about literacy activities in the home, and also collect data about early reading instruction in schools to provide additional information on reading instruction (beyond what the current-year teachers would provide). These basic goals were supported by the Reading Development Group (RDG) and representatives from the participating countries (the National Research Coordinators).

The development of PIRLS spanned two years, beginning in 1999 and continuing until early 2001, when the final reading test and questionnaires were approved by the participating countries. As part of development, 30 countries conducted a field test of the test and questionnaires. Ultimately, 35 countries participated in the main data collection.

1.2 Participating Countries

Thirty-five countries joined together to conduct the first PIRLS assessment in 2001:

Argentina
Belize
Bulgaria
Canada (Ontario, Quebec)
Colombia
Cyprus
Czech Republic
England
France
Germany
Greece
Hong Kong
Hungary
Iceland
Iran
Israel
Italy
Kuwait
Latvia
Lithuania
Macedonia
Moldova
Morocco
Netherlands
New Zealand
Norway
Romania
Russian Federation
Scotland
Singapore
Slovak Republic
Slovenia
Sweden
Turkey
United States

1.3 Student Population Assessed

In 2001, PIRLS assessed the reading literacy of children in "the upper of the two grades with the most 9-year-olds at the time of testing" (PIRLS, 1999). This corresponds to the fourth grade in most countries. This population was chosen because it represents an important transition point in children's development as readers. In most countries, by the end of fourth grade, children are expected to have learned how to read, and are now reading to learn. This grade is also assessed in the IEA's Trends in International Mathematics and Science Study (TIMSS), to provide countries participating in both studies with achievement and background data for three subjects at the same grade level.

In each country, representative samples of students were selected using a two-stage sampling design. In the first stage, at least 150 schools were selected using probability-proportional-to-size sampling. Countries could incorporate in their sampling design important reporting variables (for example, urbanicity or school type) as stratification variables. At the second stage, one or two fourth-grade classes were randomly sampled in each school. This resulted in a sample size of at least 3,750 students in each country. Some countries opted to include more schools and classes, enabling additional analyses, which resulted in larger sample sizes.

1.4 Assessment Dates

PIRLS was administered near the end of the school year in each country. In countries in the Northern Hemisphere (where the school year typically ends in May or June) the assessment was conducted in April, May, or June 2001. In the Southern Hemisphere, the school year typically ends in November or December; so in these countries, the assessment was conducted in October or November 2001.

1.5 Study Management and Organization

PIRLS is directed in the United States by Ina V.S. Mullis and Michael O. Martin, at the International Study Center at Boston College; they also direct the IEA's TIMSS. The PIRLS International Study Center was responsible for the design, development, and implementation of the study – including developing the instruments and survey procedures; ensuring quality in data collection; and analyzing and reporting the study results. The International Study Center worked closely with the organizations responsible for particular aspects of the study, the PIRLS advisory committees, and representatives of the participating countries.

Each country appointed a National Research Coordinator (NRC) who, together with staff members at the PIRLS national center, was responsible for all aspects of the study within that country. The PIRLS ISC organized meetings of the NRCs several times a year to review study materials and procedures, and to receive training in scoring constructed-response items, and in entering the data using the prescribed software.

The IEA Secretariat provided guidance in all aspects of the study, and was responsible for managing the ambitious translation verification effort conducted for the field test and main assessment. Statistics Canada was responsible for all aspects of sampling – including working with countries to ensure that the international procedures are followed; adapting the international design to national conditions; documenting the national samples; and computing sampling weights.

The National Foundation for Educational Research in England and Wales had major responsibility for developing the reading test – including collecting reading passages; developing items and scoring guides; and conducting scoring training. The IEA Data Processing Center was responsible for processing and verifying the data from the 35 countries and for constructing the international database. Educational Testing Service provided software and support for scaling the achievement data.

The study directors and representatives from the International Study Center, IEA, Statistics Canada, the National Foundation for Educational Research in England and Wales, and Educational Testing Service met periodically to review the study's progress, procedures, and schedule.

The PIRLS Reading Development Group (see Appendix A) contributed their invaluable expertise to the framework and reading test. Committee members reviewed various drafts of the framework and assessment blocks, and reviewed and endorsed the final reading test. The PIRLS Questionnaire Development Group (see Appendix A) – comprising representatives from six countries – helped develop the PIRLS questionnaires (including writing items and reviewing drafts of all questionnaires).

1.6 Overview of Assessment Framework

At the heart of the PIRLS assessment is the definition of reading literacy established by the Reading Development Group, and refined by National Research Coordinators. The PIRLS definition of reading literacy builds on the definition used in the IEA 1991 study, but elaborates on that definition by making specific reference to reading by children. 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, and for enjoyment (Campbell, Kelly, Mullis, Martin, & Sainsbury, 2001).

Growing out of this definition are the three aspects of reading literacy assessed by PIRLS:

- Processes of comprehension
- Purposes for reading and
- Reading behaviors and attitudes.

Processes of comprehension and purposes for reading are the foundation of the written assessment of reading comprehension. The purposes for reading and processes of comprehension, as well as the percentages of the assessment devoted to each, are shown in Exhibit 1.1. Each process is assessed with each purpose for reading. Reading behaviors and attitudes are assessed through a questionnaire completed by the students.

Exhibit 1.1: Percentages of Reading Assessment Devoted to Reading Purposes and Processes[1]

		Purposes for Reading	
		Literary Experience	Acquire and Use Information
Processes of Comprehension	Focus on and Retrieve Explicitly Stated Information	9%	13%
	Make Straightforward Inferences	14%	9%
	Interpret and Integrate Ideas and Information	20%	20%
	Examine and Evaluate Content, Language, and Textual Elements	6%	8%

[1] Because numbers are rounded to the nearest whole number, some totals may appear inconsistent.

1.7 PIRLS Reading Assessment

PIRLS has ambitious goals for covering the domain of reading literacy. The Reading Development Group felt that at least eight passages and items (four for each reading purpose) were needed to provide a valid and reliable measure of reading achievement. Since it would not be possible to administer the entire test to any one child, PIRLS used a matrix sampling technique to distribute the assessment material among students, yet retain linkages necessary for scaling the achievement data.

1.7.1 Assessment Design

The material was divided into 40-minute "blocks," each comprising a passage (a story or article) and items representing at least 15 score points. There are eight such blocks, four for each reading purpose. Blocks containing literary passages are labeled L1 through L4, and those containing informational passages, I1 through I4. The eight assessment blocks are distributed across ten test booklets, and each student completed one booklet in an 80-minute testing session. Each booklet contains two blocks, and most blocks appeared in three booklets. One of the ten booklets is the PIRLS Reader, a color booklet containing two reading passages; the test items are located in a separate booklet. The two blocks comprising the Reader appear only in that booklet. The distribution of blocks across booklets "links" the booklets to enable the achievement data to be scaled using item response theory methods.

The design for the assessment booklets is presented in Exhibit 1.2, which shows that each booklet has two blocks – two literary, two informational, or one of each. It also shows that three of the literary and three of the informational blocks appear three times

Exhibit 1.2: PIRLS Assessment Design

	Booklet 1	Booklet 2	Booklet 3	Booklet 4	Booklet 5	Booklet 6	Booklet 7	Booklet 8	Booklet 9	Booklet R (Reader)
Assessment Block	L1	L2	L3	I1	I2	I3	L1	I2	I3	L4
	L2	L3	I1	I2	I3	L1	I1	L2	L3	I4

L = Reading for Literary Experience; I = Acquire and Use Information

Exhibit 1.3: Distribution of Items by Type and Reading Purpose

	Multiple-Choice Items	Constructed-Response Items			Total Number of Items	Total Score Points
		1-point	2-point	3-point		
Literary	25	14	9	3	51	66
Informational	21	10	12	4	47	67
Total	46	24	21	7	98	133

across test booklets 1 through 9, and that the fourth literacy and informational blocks appear only in the Reader.

1.7.2 Passages

The reading passages form the foundation of the reading literacy test. In accordance with the framework, four of the assessment blocks contain literary texts and four contain informational texts, and the passages are authentic texts drawn from children's storybooks and informational sources. Submitted and reviewed by the PIRLS countries, the passages represent a range of types of literary and informational texts. The literary passages include realistic stories and traditional tales; while the informational texts include chronological and non-chronological articles, a biographical article, and an informational leaflet.

1.7.3 Items and Scoring Guides

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

Each block of assessment material contained from 11 to 14 items that together represent at least 15 score points. Altogether, the PIRLS reading test includes 98 items representing 133 score points – enough to estimate achievement reliably. Exhibit 1.3 shows the distribution of items by type and reading purpose.

Scoring guides for constructed-response items were developed together with the items. Each scoring guide is unique to that item. It describes the essential features of appropriate and complete responses – including the kind of evidence of understanding required and example student responses to help scorers determine the score for a particular response. Actual student responses were used to develop the guides, and illustrate the kinds of responses garnering different points.

1.7.4 Releasing Assessment Material to the Public

The PIRLS test design provides for the release of half of the assessment material into the public domain after data collection, including the entire PIRLS Reader.

The remaining half will be kept secure and included in future PIRLS assessments so trends in achievement can be measured. After data collection in 2001, one literary and one informational block (as well as the blocks in the Reader) were released. Effectively, four blocks (or half of the assessment) were available to the public.

1.8 PIRLS Background Questionnaires

By gathering information about children's experiences together with reading achievement on the PIRLS test, it is possible to identify the factors or combinations of factors that relate to high reading literacy. An important part of the PIRLS design is a set of questionnaires targeting factors related to reading literacy. PIRLS administered four questionnaires: to the tested students, to their parents, to their reading teachers, and to their school principals.

1.8.2 Student Questionnaire

Each student taking the PIRLS reading assessment completes the student questionnaire. The questionnaire asks about aspects of students' home and school experiences – including instructional experiences and reading for homework, self-perceptions and attitudes towards reading, out-of-school reading habits, computer use, home literacy resources, and basic demographic information.

1.8.3 Learning to Read (Home) Survey

The learning to read survey is completed by the parents or primary caregivers of each student taking the PIRLS reading assessment. It addresses child/parent literacy interactions, home literacy resources, parents' reading habits and attitudes, home/school connections, and basic demographic and socioeconomic indicators.

1.8.4 Teacher Questionnaire

The reading teacher of each fourth-grade class sampled for PIRLS completes a questionnaire designed to gather information about classroom contexts for developing reading literacy. This questionnaire asks teachers about characteristics of the class tested (such as size, reading levels of the students, and the language abilities of the students). It also asks about instructional time, materials and activities for teaching reading and promoting the development of their students' reading literacy, and the grouping of students for reading instruction. Questions about classroom resources, assessment practices, and home/school connections also are included. The questionnaire also asks teachers for their views on opportunities for professional development and collaboration with other teachers, and for information about their education and training.

1.8.5 School Questionnaire

The principal of each school sampled for PIRLS responds to the school questionnaire. It asks school principals about enrollment and school characteristics (such as where the school is located, resources available in the surrounding area, and indicators of the socioeconomic background of the student body), characteristics of reading education in the school, instructional time, school resources (such as the availability of instructional materials and staff), home/school connections, and the school climate.

1.9 Translation and Verification of Instruments

The PIRLS reading tests and questionnaires were prepared in English, then translated into 31 other languages for use in the 35 participating countries. Countries were responsible for translating the instruments into their local language (or languages) following internationally prescribed procedures. To ensure standardization of instruments across countries, PIRLS undertook an extensive verification process, whereby each country's data collection instruments were independently reviewed and verified by an external translation company engaged by the IEA. The verifiers' reviews of the translated documents were used to improve the translations. Instruments were verified twice, once before the field test and once before the main data collection. In addition to the external review, the International Study Center also reviewed the countries' instruments against the verifiers' comments to ensure that all necessary corrections were made. Finally, statistical analyses of item data were conducted to check for evidence of differences in student performance across countries that could indicate a translation problem.

1.10 Data Collection

Each country was responsible for carrying out all aspects of the data collection, using standardized procedures developed for the study. Manuals provided explicit instructions to the NRCs and their staff members on all aspects of the data collection – from contacting sampled schools to packing and shipping materials to the IEA Data Processing Center for processing and verification. Manuals were also prepared for test administrators and for individuals in the sampled schools who work with the national centers to arrange for the data collection within the schools. These manuals addressed all aspects of the assessment administration within schools (including test security, distribution of booklets, timing and conduct of the testing session, and returning materials to the national center).

The PIRLS International Study Center placed great emphasis on monitoring the quality of the PIRLS data collection. In particular, the Study Center implemented an international program of site visits, whereby international quality control monitors visited a sample of 15 schools in each country and observed the test administration. In addition to the international program, NRCs were also expected to organize an independent national quality control program based upon the international model. The latter program required national Quality Control Observers to document data collection activities in their country. The national Quality Control Observers visited a random sample of 10 percent of the schools (in addition to those visited by the international Quality Control Monitors), and monitored the testing sessions – recording their observations for later analysis.

1.11 Scoring the Constructed-Response Items

Because almost two-thirds of the score points came from constructed-response items, PIRLS needed to develop procedures for reliably evaluating student responses within and across countries. The International Study Center prepared detailed guides containing the PIRLS scoring rubrics, and explanations of how to implement them – together with example student responses for the various rubric categories. These guides, along with training packets containing extensive examples of student responses for practice in applying the rubrics, were used as a basis for intensive training of national representatives in scoring the constructed-response items.

To gather and document empirical information about the within-country agreement among scorers, PIRLS arranged to have a sample of 200 students' responses to each item in each country scored independently by two readers. Scoring reliability within countries was high – the percentage of exact agreement, on average, across countries, was more than 90 percent. PIRLS also conducted a study of scoring reliability across countries, asking countries with scorers proficient in English to score a reference set of student responses chosen from students in English-speaking countries. This study revealed a high level of agreement between scorers also (85% on average).

1.12 Data Processing

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

Throughout the process, the data were checked and double-checked by the IEA Data Processing Center, the International Study Center, and the national centers. The national centers were contacted regularly and given multiple opportunities to review the data for their countries. In conjunction with the IEA Data Processing Center, the International Study Center reviewed item statistics for each cognitive item in each country to identify poorly performing items. In general, the items exhibited very good psychometric properties in all countries.

1.13 IRT Scaling

The general approach to reporting the PIRLS achievement data was based primarily on item response theory (IRT) scaling methods. Student reading achievement was summarized using a family of IRT models (2-parameter, 3-parameter, and generalized partial credit models). The IRT methodology was preferred for developing comparable estimates of performance for all students, since students responded to different passages and items depending upon which of the test booklets they received (Booklet 1 through 9 or the PIRLS Reader). This methodology produces a score by averaging the responses of each student to the items that he or she took in a way that takes into account the difficulty and discriminating power of each item. The approach followed in PIRLS uses information from the background questionnaires to provide improved estimates of student performance (a process known as conditioning) and multiple imputation to generate student scores (or "plausible values") for analysis and reporting.

The IRT analysis provides a common scale on which performance can be compared across countries. In addition to providing a basis for estimating mean achievement, scale scores permit estimates of how students within countries vary and provide information on percentiles of performance. Treating all participating countries equally, the PIRLS scale average across countries was set to 500 and the standard deviation to 100. Since the countries varied in size, each country was weighted to contribute equally to the mean and standard deviation of the scale. The average and standard deviation of the scale scores are arbitrary and do not affect scale interpretation.

In the PIRLS analysis, achievement scales were produced for each of the two reading purposes, reading for literary experience and reading for information, as well as for reading overall.

1.14 Data Analysis and Reporting

The PIRLS 2001 International Report (Mullis, Martin, Gonzalez & Kennedy, 2003) summarizes fourth-grade students' student reading achievement in each country. This report presents average student achievement in reading overall as well as in reading for literary experience and reading to acquire and use information, together with standard errors and tests of significance as appropriate. Average achievement is reported separately for girls and boys.

To provide additional information about reading achievement among high- and low-achieving students, PIRLS reported the percentage of students in each country performing at each of four international benchmarks of student achievement – corresponding to the 90th, 75th, 50th, and 25th percentiles of the international distribution of reading achievement. To enhance this reporting approach, PIRLS conducted a scale anchoring analysis to describe student performance at the international benchmarks in terms of the kind of reading students performing at each benchmark can do, and the level of comprehension they exhibit. Complementing this approach further, the PIRLS international report presents examples of questions from both literary and informational passages that anchor at each of the benchmarks (providing another perspective on the reading demands of the benchmarks), and also displays student performance in each country on the example questions.

PIRLS 2001 collected a wide array of information about the home and school context in which students learned to read (from parents, students, teachers, and school principals). The PIRLS international report summarized much of this information, combining data into composite indices showing an association with achievement where appropriate. In particular, student reading achievement is described in relation to literacy-related activities in the home, the school curriculum and organization for teaching reading, teachers and reading instruction, school contexts, and students' reading attitudes, self-concepts, and out-of-school activities.

Additional information about the countries participating in PIRLS 2001 may be found in the *PIRLS 2001 Encyclopedia* (Mullis, Martin, Kennedy, Flaherty, 2002), a volume providing general information on the cultural, societal, and economic situation in each country, as well as a more focused perspective on the structure and organization (of their respective educational systems as it pertains specifically to the promotion of reading literacy). Consisting of a chapter from each country, the *PIRLS 2001 Encyclopedia* describes primary/elementary schooling as it pertains to reading within each educational system – including teacher education and training, reading curricula, classroom organization and instruction, and assessment practices. As such, it is an extremely valuable companion publication to the international report providing insights and detailed information about the policies, practices, and resources within each country.

1.15 The Trends in IEA's Reading Literacy Study

While PIRLS 2001 is the first in a cycle of assessments designed to measure trends in reading achievement, some countries also measured achievement trends from 1991 to 2001. Countries that participated in the IEA's 1991 Reading Literacy Study were eligible to administer the 1991 reading test and student questionnaire to a sample of students in 2001 so that they could obtain information about how their children's reading literacy today compares with that of ten years ago. The following countries participated in the trend study:

Greece
Hungary
Iceland
Italy
New Zealand
Singapore
Slovenia
Sweden
United States

Countries sampled every other PIRLS school for the trend study, resulting in a sample size of at least 75 schools. In each school, one target-grade classroom was sampled and administered the 1991 test and student questionnaire. For some countries, the 1991 target grade and the PIRLS target grade were not the same. Statistics Canada worked with these countries to tailor the design so as to achieve a representative sample of students. The IRT scaling methodology used with PIRLS 2001 also was applied in the trends in reading literacy study. The results of the trend study are presented in Martin, Mullis, Gonzalez, and Kennedy (2003).

References

Campbell, J.R., Kelly, D.L., Mullis, I.V.S., Martin, M.O., & Sainsbury, M. (2001). *Framework and specifications for PIRLS assessment 2001. (2nd ed.)*. Chestnut Hill, MA: Boston College.

Elley, W.B. (1992). *How in the world do students read?* The Hague, Netherlands: IEA.

Elley, W.B. (Ed.). (1994). *The IEA study of reading literacy: Achievement and instruction in thirty-two school systems*. Oxford, England: Elsevier Science Ltd.

Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., & 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: Boston College.

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

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., & Flaherty, C.L. (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.

PIRLS (1999). *School Sampling Manual – Version 2* (PIRLS Ref. No. 99-0019). Prepared by Pierre Foy & Marc Joncas, Statistics Canada. Chestnut Hill, MA: Boston College.

Thorndike, R.L. (1973). Reading comprehension in fifteen countries: An empirical study. *International studies in evaluation: Vol. 3*. Stockholm: Almqvist & Wiksell.

Walker, D.A. (1976). *The IEA six subject survey: An empirical study of education in twenty-one countries*. New York: John Wiley & Sons Inc.

2

Developing the PIRLS Reading Assessment

Marian Sainsbury

Jay Campbell

2.1 Overview

The development of the PIRLS reading assessment took place over a two-year period, from 1999 to 2001. The work was undertaken by a team from the National Foundation for Educational Research in England and Wales (NFER[1]), with support and advice at all stages from the PIRLS Reading Coordinator,[2] the Reading Development Group (RDG), the National Research Coordinators (NRCs), the PIRLS Project Management Team, and staff of the PIRLS International Study Center at Boston College. Test development was based firmly on the *Framework and Specifications for the PIRLS Assessment 2001* (Campbell, Kelly, Mullis, Martin, & Sainsbury, 2001). The framework presents a view of reading literacy as a complex interactive process. It identifies two main purposes for reading relevant to the age group selected for the assessment: reading for literary experience, and reading to acquire and use information. The framework specifies four principal comprehension processes that readers use to construct meaning that are the same for both reading purposes. The assessment requires passages that offer students an authentic engagement with text, and items that draw upon the central qualities of that engagement.

The aim was to produce a set of reading passages and items (questions) related to those passages, arranged in a collection of blocks, or units – as described in the framework. Each block was to consist of one or more passages and accompanying items that would yield at

[1] The members of the NFER team were Chris Whetton, Marian Sainsbury, Jenny Bradshaw, Anne Kispal, Jenny Phillips and Jane Sowerby.

[2] Jay Campbell of Educational Testing Service served as the PIRLS Reading Coordinator.

least 15 score points. The initial development task was to develop 16 blocks, eight literary and eight informational, for field testing. Following the field test, four literary and four informational blocks were selected for use in the main survey from among the original 16 blocks.

The development of these reading literacy blocks involved, first, the selection of passages, and only then the generation, revision, and selection of items. This structure sets it apart from assessments in other curriculum areas such as mathematics or science, where items can be generated to an initial specification. For PIRLS, passages had to be selected before work could begin on the items.

Test development in an international context is an ambitious undertaking; a variety of cultural and linguistic factors must be considered in selecting passages and developing items. Moreover, the need to translate

Exhibit 2.1: Overview of the Test Development Process

Meeting Date	Group and Purpose of Meeting
May 1999	Reading Development Group: Initial drafting of the PIRLS assessment framework
July 1999	National Research Coordinators: Review of the draft PIRLS assessment framework Initial review of field-test passage pool, and feedback on the passage selection process
October 1999	Reading Development Group: Initial approval of the PIRLS assessment framework Initial review and selection of field-test passage pool and draft items
November 1999	National Research Coordinators: Final approval of the PIRLS assessment framework Review and final selection of field-test passage pool Review of draft items and scoring guides
January 2000	Reading Development Group: Review and initial selection of field-test item pool and scoring guides
March 2000	National Research Coordinators: Review and final selection of field-test item pool and scoring guides
July 2000	National Research Coordinators: Training on field-test scoring guides
December 2000	Reading Development Group: Review of field-test results, and initial selection of operational passages and items
January 2001	National Research Coordinators: Final review of field-test results, and selection of operational passages and items
May 2001	National Research Coordinators: Training on operational scoring guides

both passages and items into numerous languages required extreme sensitivity to the effects of sociolinguistic differences on assessing reading comprehension. As such, the development process required ongoing involvement of both the RDG (a seven-member multinational group of literacy experts) and the NRCs. Exhibit 2.1 provides a brief overview of the iterative process characterizing the development of this international assessment instrument. As suggested by this display, the process involved initial recommendations and guidance of the RDG, and final approval of the NRCs.

2.2 The PIRLS Assessment Framework

The PIRLS assessment development effort was guided by the description of reading literacy in the PIRLS assessment framework. The framework provided a theoretical understanding of reading literacy, and specified the types of reading materials and questions that were developed and selected for the assessment instrument. Central to the framework is its definition of reading literacy:

> *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, and for enjoyment.*

The view of literacy embodied in this definition – and described in more detail throughout the framework – is derived from and informed by numerous theories of reading. The framework was not intended to reflect any single theory of reading or approach to reading instruction. Rather, it was based on a multinational consensus about the nature of reading literacy, the goals of reading instruction, and the expectations for developing readers in a literate society.

Development of a thorough and theoretically cohesive framework was a necessary first step in the instrument development process. The framework provided explicit descriptions of the types of reading material that were to be represented in the assessment, and the types of comprehension questions that were to be developed to measure students' understandings of the reading material. In describing the types of reading materials to be used in the assessment, the focus was on purposes for reading. Because readers often approach different types of texts for different reading purposes, and because it is expected that students by age 9 should have developed the ability to read for a variety purposes, the characterization of test types by purposes for reading provided assurance of broad construct coverage in the assessment. While reading for different purposes, readers engage in a variety of processes to comprehend text. As such, a description of comprehension processes was included in the framework to guide the development of test questions.

The following sections provide a description of the text types (purposes for reading) and the item types (processes and strategies) that were included in the framework – and that guided the instrument development process.

2.2.1 Text Types

Readers interact with text in different ways to construct meaning. Their approach to constructing meaning varies by the purpose for reading and the type of text being read. Certain purposes for reading are associated with certain types of text. For nine- and ten-year-old students, the two most common purposes for reading are reading for enjoyment and reading to learn. As such, the PIRLS framework specifies the inclusion of two broad types of text in the assessment: literary texts read for *literary experience or enjoyment*, and informative texts read to *acquire and use information*.

In reading for literary experience, readers engage with the text in order to become immersed in the world portrayed by the author. Readers may vicariously experience a world unfamiliar to them, or make connections and find similarities between the text and their own experiences. Young readers by age 9 have already developed an awareness of narrative text structures and use of language, upon which they draw to construct meaning and to react to the text. The PIRLS framework called for the inclusion of literary texts that represent the types of narrative structures and language usages most common to 9-year-old readers. The main form of literary text used in the assessment was narrative fiction.

In reading to acquire and use information, the reader is mostly focused on understanding the aspects of the real world described in the text. In addition, depending on the nature of the text and the reader's orientation, the text may evoke an action or response – as in following a set of directions or reacting to a persuasive argument or appeal. The type of texts that fall into this category may be structured chronologically or logically. Examples of texts that may be structured in a chronological manner include biographical accounts of the lives of contemporary or historical figures and procedural documents that detail step-by-step directions to be followed in sequence. Examples of texts that are structured logically many include those that are written to provide information about a given topic and those that are intended to persuade or convince the reader to think and act in a certain manner. Often, these texts include adjunct aids (such as charts, pictures, and graphs to convey information). The PIRLS assessment included both chronologically and logically structured informational texts, some of which incorporated various types of adjunct aids.

2.2.2 Processes and Strategies

Across text types and purposes for reading, the reader engages in a variety of comprehension processes and strategies to gain and construct meaning from text. The PIRLS assessment framework described four specific processes of comprehension, which vary in terms of the degree of inference or interpretation required and in the focus on text content or structural features of the text. This description of comprehension processes in the framework served as a guide for developing the comprehension questions used to assess students' understandings of texts. Each question was written to engage students in one of four processes: 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. A brief description of each process is provided below.

In focusing on and retrieving explicitly stated information, the reader locates specific information or an idea in the text that is relevant to understanding the text's meaning. Little or no inference is required to understand the meaning of such information – it is explicit, and may be viewed as existing at the surface level of the text. Most often, the retrieved information resides locally in the text, within a specific sentence or phrase. A competent reader's understanding of the retrieved information is typically immediate or automatic.

In making straightforward inferences, the reader goes beyond what is stated explicitly in the text and infers some implied meaning or connection between textually-based ideas. Although not stated explicitly, the inference is very much constrained by the text. The text provides fairly obvious cues to guide the reader in making this type of inference. As such, skilled readers will often make such an inference automatically as they become engaged in constructing meaning within a specific part of the text, or as they develop a more global understanding of the text's overall meaning.

In order to construct a more complete and richer understanding of the text, readers must be able to interpret and integrate ideas and information. With this type of process, the reader moves beyond the phrase or sentence level of text to make connections between textual ideas, synthesize information, or consider the broader implications of textual meaning. In doing so, readers often draw on their background knowledge and experiences to develop interpretations, which may vary slightly – depending upon the reader's perspective.

Readers shift from constructing meaning to a critical consideration of the text as they examine and evaluate content, language, and textual elements. The reader recognizes that the text has been written to convey ideas, feelings, and information. The textual content may be evaluated for its overall value, believability, or relevance to the reader. Its structural and linguistic features may also be judged for its effectiveness, completeness, or impact. In examining and evaluating the text, readers may draw upon their understanding of the world, and on their past reading experiences.

2.2.3 Test Booklet Design

The test booklet design used in the operational PIRLS assessment was based on several considerations. First, in order to ensure broad coverage of reading comprehension (as described in the PIRLS framework) a total of eight reading blocks – each block consisting of a single passage or set of passages accompanied by comprehension questions – were developed. Each block was developed to assess a single purpose for reading; a total of four literary blocks and four informational blocks comprised the operational assessment. Secondly, it was acknowledged that the burden required for

Exhibit 2.2: Distribution of Literacy and Informational Blocks Across Booklets

	Booklet 1	Booklet 2	Booklet 3	Booklet 4	Booklet 5	Booklet 6	Booklet 7	Booklet 8	Booklet 9	Booklet R (Reader)
Assessment Block	L1	L2	L3	I1	I2	I3	L1	I2	I3	L4
	L2	L3	I1	I2	I3	L1	I1	L2	L3	I4

each student to take the entire assessment (a total of more than five hours testing time) would be too great. Consequently, a matrix sampling technique was employed so that each student would take only a portion of the assessment (two reading blocks), and that an appropriately representative sample of students would be administered to each portion. Finally, it was important to ensure adequate linking of results across blocks, since each student would not be administered the entire assessment.

With these considerations in mind, the four literary blocks and four informational blocks were distributed across 10 assessment booklets. Each student participating in the assessment was administered one of the 10 booklets. Because students were given 40 minutes to complete each block, the total assessment time was 80 minutes. (An additional 15 to 30 minutes was devoted to having students complete a background and instructional experience questionnaire.)

Exhibit 2.2 illustrates the distribution of literary and informational blocks across the 10 test booklets. The block designations L1, L2, L3, and L4 refer to the four literary blocks. The block designations I1, I2, I3, and I4 refer to the four informational blocks.

Although this booklet design does not provide for all possible combinations of literary and informational blocks (which would have resulted in twice the number of test booklets), it was determined that the block combinations represented here were more than adequate to provide for suitable linking between blocks. Each block appears in three booklets, and each block is combined with at least one block assessing the same purpose for reading, and at least one block assessing the other purpose for reading. Note that the nature of booklet 10 (the PIRLS Reader), which links one specific literary and informational block, and made it impossible to link these blocks to others in the design without substantially increasing student assessment time. Consequently, booklet 10 was distributed across sampled students at three times the rate of the other booklets.

2.3 Finding and Selecting the Passages

Finding a selection of passages that would suit the purposes of the PIRLS assessment was a major challenge. At each stage of the test development process, review by the RDG and the NRCs played a central part in ensuring the suitability of the materials. The passages had to be appropriate for valid assessment of reading literacy in all participating countries. The test materials, taken overall, had to be interesting and accessible for all the participating students – not favoring any particular national or cultural group.

2.3.1 The Initial Search for Passages

In order to achieve this, great import was placed on seeking passages that originated in the participating countries. Even before their first meeting, NRCs received a request to contribute to the pool of texts for consideration. This request incorporated the following criteria used throughout the test development process.

All passages:

- Must be suited in their content and reading level to 9- and 10-year-olds

- Should be well written in order to foster authentic engagement in the reader and to facilitate questioning across the PIRLS processes and strategies

- Could be either literary or informational, and should include as wide a range as possible within these two broad categories

- Should not exceed 1200 words in length

- Should avoid specific cultural references and material offensive to particular cultural or religious groups.

Representatives from participating countries were asked to contribute texts that met these criteria, and that would be typical of the reading matter available to students at the appropriate age and grade level in their countries.

The first meeting of the RDG, in May 1999, recommended an innovative approach to international literacy assessment, in the form of a "Reader." This was a reading booklet, produced in full color, including a number of different passages – both literary and informational – following a unifying theme. The questions on these passages appeared together in a separate question booklet. This approach found favor because of the attractive and authentic appearance of the Reader, and the possibility for thematic links between literary and informational reading. In searching for passages, therefore, ideas suitable for generating Readers were also sought.

At the first NRC meeting in July 1999, participants considered 68 passages that had been contributed by 11 different countries: Albania, Australia, Austria, Cyprus, France, Italy, Hungary, New Zealand, Russia, Singapore, and the United Kingdom. These comprised passages sent in advance or brought to the meeting by the NRCs themselves; texts suggested by members of the RDG; and passages found by the NFER research team. Although this collection already represented a wide range of material, it was agreed, at that meeting, that further texts should be sought and reviewed by the NRCs following the meeting.

2.3.2 Reviewing the Passages

The review materials presented at the July 1999 NRC meeting contained passages arranged for the first time as assessment blocks. Some of these blocks consisted of a single passage; others were combinations of shorter passages. There were 11 literary blocks, 12 informational blocks, and three possible Readers. Each Reader was the equivalent of two blocks, one literary and one informational. The texts ranged in length from 181 to 1,103 words. Passages for literary experience included contemporary realistic narrative, fantasy narrative, traditional tales, and myth and fable. The passages assessing the use and acquisition of information included instructions, explanatory texts, biographies, newspaper reports, information leaflets, tables, texts including diagrammatic information, and one that had originated as part of a website. The passages in these review books represented contributions from 14 countries: Australia, Austria, Canada, Cyprus, France, Iceland, Italy, The Netherlands, New Zealand, Russia, Singapore, the Slovak Republic, Sweden, and the United Kingdom.

NRCs responded to the review materials with a wide range of views. Their comments were summarized for discussion at the next meeting of the RDG, which took place in October 1999. Here, a shorter list of passages was agreed upon for consideration by the NRCs at their November meeting. At this stage, the passages were also illustrated and presented as they would be to students. In some cases, the illustrations were found in the original passage; in others, illustrations were specially commissioned. The illustrations were designed to support the reading of the text, without giving information that would distract or mislead the student. The passages proposed for the Readers had full-color illustrations.

The goal at the November 1999 NRC meeting was to arrive at final decisions about the 16 blocks to be used in the field test. Eight of these were to be literary blocks and eight informational. The two Readers were each to comprise one literary and one informational block, both taken from the 16. Exhibit 2.3 sets out the titles of the 16 passages finally chosen at the meeting, together with an indication of the textual features of each. The passages listed in the table were originally suggested by eight different countries: Canada, Iceland, Italy, New Zealand, Russia, the Slovak Republic, Sweden, and the United Kingdom. The involvement of participating countries from the earliest stage of development gave the resulting assessment its unique international flavor.

A comparison with the PIRLS framework shows that the passages selected at the end of the initial development process were a good reflection of the principles established there. All of the literary texts were narrative fiction, but within this overall category they represented a wide variety – in terms of story type, setting, characterization, plot

Exhibit 2.3: Passages Selected for Field Testing

Title	Content
Literary Blocks	
"The Upside-Down Mice"	Modern fable with a twist
"Flowers on the Roof"	Contemporary realistic story set in Iceland
"The Dressmaker"	Contemporary realistic story set in Africa
"Fathers and Sons"	Traditional fable (The Farmer and his Sons); Traditional moral tale (Equal Inheritance)
"The King with Dusty Feet"	Traditional tale from India
"Punch's Escape"	Fantasy tale from Italy about puppets
"Hare Heralds the Earthquake"	Traditional tale
"The Little Lump of Clay"	Contemporary moral tale
Informational Blocks	
"Leonardo da Vinci"	Biography
"Introducing Antarctica"	Nonchronological expository text including diagrams; letter
"Night of the Pufflings"	Mainly chronological informational text
"Puppy Walking"	Explanatory text
"River Trail"	Informative/persuasive leaflet
"Read Dinosaur Pox"	Book review information in a variety of forms, drawn from a website
"Finding Out About the Weather"	Information in chronological, nonchronological, and tabular forms
"All About Mobiles"	Information, biography, and instructions with diagrams

structure and length. The informational passages included both chronologically and nonchronologically organized texts with a variety of purposes and presentational features. Discussions with the RDG and the NRCs confirmed that this collection of passages adequately represented the range of purposes for engagement with texts envisaged for the PIRLS study.

2.4 Developing the Items

Item development started as soon as there began to be a consensus on the selected passages in August 1999. Once again, the writing, review, and revision of the items was closely guided by the principles established in the PIRLS framework. Repeated review by international reading experts, and by representatives of participating countries, provided valuable comments for improving the item pool.

There were two main types of items: multiple-choice questions, and constructed-response questions. The multiple-choice items offered students four plausible response options of which only one was correct or was clearly the best response to the question. Each of these carried one score point. Constructed-response items could yield one, two, or three score points. They were used in order to allow students to explain their interpretations and evaluations of the text, to show their reasoning, and to find for themselves the textual evidence that supported these views and reasons. In a typical block of 15 score points, the aim was to have seven multiple-choice items, two or three short-answer items of one or two points, and one extended-response item worth three points.

The items were written to address systematically the four PIRLS processes and strategies:

- Focus on and retrieve explicitly stated information (20%)
- Make straightforward inferences (30%)
- Interpret and integrate ideas and information (30%)
- Examine and evaluate content, language and textual elements (20%).

The varying nature of the texts, however, (in both the literary and informational categories) meant that the interpretation of these four comprehension processes also varied. For example, in a literary text with strong characterization, interpreting and integrating ideas and information would suggest some items addressing character and motive. In an informational piece, by contrast, items addressing this same process would be more likely to require the synthesis of information from different parts of the passage. The framework gives further details of these issues. The development of items was guided by the features of the text, on the one hand, and the PIRLS processes and strategies, on the other.

2.4.1 Item Piloting

Early drafts of items were reviewed by the RDG in October 1999, and by the NRCs in November 1999. At about the same time, these early drafts were subjected to some limited testing by NFER in schools in England, to check the suitability of the passages and to gauge student responses to the questions. The findings from these trials were mainly qualitative in nature.

On the basis of comments from the RDG and NRCs, and the findings from the small-scale trials, a major revision of the items was conducted in December 1999. This addressed a number of difficulties that had been identified by the reviews and trials of the early drafts. In some cases, styles of questioning were found to be inaccessible to some groups of students. In others, question wordings proved ambiguous. Some items were rejected because they were not central to important ideas in the texts, or were regarded as addressing peripheral aspects of the subject matter. At this stage, also, the proportion of multiple-choice items for each block was increased to about 50 percent from the previous target of 30-40 percent – because of feedback from participating countries.

2.4.2 Item Review and Revision

The revised assessment blocks were again reviewed in January 2000, at a meeting of the RDG. At the same time, the NRCs were consulted by means of a postal review, to which 22 countries provided responses.

Also in January, further trials were conducted by NFER in schools in England. Although these were again small in scale and conducted in only one country, they provided some valuable evidence as to how students responded to the passages and items – which were now approaching their final shape. A sample of 70-100 students completed each block. They were in Year 5, aged between 9.4 and 10.3 years.

The schools were not a representative sample; rather, they covered the full range of circumstances found in England, including students from socioeconomically deprived backgrounds, from ethnic minorities, and students for whom English was not their first language. A basic statistical analysis of the results showed that, in general, the draft blocks proved fairly easy for the sample, and that most of the blocks had a reasonable reliability index (Cronbach's alpha >0.70). Most students reached the end of the blocks in the time allowed.

2.4.3 Finalizing the Items

Once again, in February 2000, the items were revised to reflect the judgements of reviewers, paying attention (where appropriate) to the findings from the small-scale trials. In March 2000, the proposed blocks for the field test were submitted once more to the NRCs for a final review. After a final round of revisions (in response to these comments), the blocks were finalized and sent to the countries for translation in time for the field test.

2.5 Field Test

In order to ensure that the passages and items had good measurement properties in each country, PIRLS conducted a full-scale field test in September 2000. For the purposes of the field test, the 16 assessment blocks were divided among eight student booklets – six booklets containing passages and items, and two readers with accompanying answer booklets. Since a student was expected to complete only one booklet, countries were requested to draw probability samples of at least 1,600 students for the field test, so that at least 200 students would respond to each of the student booklets.

Approximately 48,000 students from almost 1,100 schools in 30 countries participated in the field test, providing about 6,000 student responses to each booklet. The field-test data showed that the passages and items generally had very good psychometric characteristics, with a wide range of difficulty levels and good discrimination indices, and would form a very good pool from which to select the passages and items for the main PIRLS assessment.

Exhibit 2.4: Blocks Selected for Main Survey

Title	Content
Literary Blocks	
"The Upside-Down Mice"	Modern fable with a twist
"Flowers on the Roof"	Contemporary realistic story set in Iceland
"The Little Lump of Clay"	Contemporary moral tale
"Hare Heralds the Earthquake" (Reader)	Traditional tale
Informational Blocks	
"Leonardo da Vinci"	Biography
"Introducing Antarctica"	Nonchronological expository text including diagrams, letter
"River Trail"	Informative/persuasive leaflet
"Night of the Pufflings" (Reader)	Mainly chronological informational text

2.6 Selection of Blocks for Main Survey

The results of the field test were reviewed at a meeting of the RDG in December 2000, and the assessment blocks for the main survey were selected. These were reviewed and approved (with minor modifications) at a meeting of the NRCs in January 2001. The blocks selected are listed in Exhibit 2.4.

2.7 Developing the Scoring Guides for Constructed-Response Items

For PIRLS, as with all tests of reading literacy with open response items, the development of the scoring guides was a major undertaking, and had to be informed by actual responses from students in test trials. The scoring guides needed to be explicit enough to credit all appropriate responses while ruling out all inappropriate responses.

However, students expressed these responses in a wide variety of ways. The scoring guides had to provide clear criteria against which the scorer could judge student responses, and these criteria needed to be supported by examples of actual student responses. At the item writing stage, it was impossible to envisage all the possible ways in which a student might express his or her understanding. Scrutiny and analysis of responses produced in test trials were essential in order to finalize the criteria and select the examples. In the PIRLS tests, the scoring guides were supported by scorer training materials consisting of anchor papers and practice papers.

2.7.1 Early Development of Scoring Guides

The initial development of scoring guides occurred while the corresponding constructed-response items were being developed. Items and scoring guides were developed concurrently so that item writers and reviewers would view the scoring criteria as an essential component of developing a reliable and valid constructed-response question. Drafting of scoring criteria must be part of constructed-response item development and review processes, so that thoughtful and ongoing considerations of how student responses will be scored can sharpen the focus and increase the measurement value of these open-ended item types.

The early drafts of the PIRLS items, in October-November 1999, had draft scoring guides describing the criteria to be applied in scoring the items, but without examples of student responses. The criteria were derived from a consideration of the process being assessed by means of one item in its relationship to the text, and specified the response (or a range of responses) expected to each open-ended question. These draft criteria were discussed alongside the items themselves during this review process, and were correspondingly revised afterwards.

2.7.2 Student Responses

The small-scale trials in January 2000 provided the first collection of student responses that could be used to develop and illustrate the criteria. The revision of the items in February 2000 included substantial attention to the scoring guides, aimed at clarifying the criteria and exemplifying a range of acceptable and unacceptable responses. Responses were listed and scrutinized against the draft criteria. At this stage, some appeared clearly acceptable and some clearly unacceptable. There were others that possessed some of the characteristics of an acceptable response, but not all, and so could be classified as borderline. For items carrying more than one score point, these classifications were made at each level of scoring. In the light of this collection of responses, the criteria were revisited and the fine distinctions that emerged were articulated. In many cases, it became clear that there were different ways of achieving the same score, for example, by choosing different but equally valid aspects of the text to support an answer. Examples of student responses were chosen to illustrate each level of scoring, demonstrating both frequent and unusual ways in which students expressed an acceptable response.

2.7.3 Finalization of Scoring Guides

Following the review meeting in March 2000, the scoring guides took on their final shape, giving fuller information and a wider range of examples. These examples were provided by further test trials that took place in May 2000, in four countries: the United States, Canada, Singapore, and England. They gave rise to at least 200 student responses to each item in its final form. Once again, responses were listed and classified, leading to a revision of the criteria and an increase in the number of listed examples.

The final scoring guide for each item was structured in the following way:

- Identification of the purpose for reading (literary or informational) being assessed

- Description of the comprehension process the item addressed

In addition, the following elements were included in each scoring guide in order to ensure that the scoring of students' responses was clearly related to the PIRLS framework, and to provide explicit guidance to scorers that would ensure reliability of scoring:

- The score to be awarded for each level of acceptable response
- The scoring criteria for each level of acceptable response
- The specific evidence to show that a response met the criteria; in many cases, this evidence could be in one of several forms, all of which were specified
- A series of example responses at each level of scoring, including examples for which no points were awarded.

To provide additional guidance and practice for scorers, further collections of student responses were assembled as anchor papers and practice papers. These were introduced to the NRCs at the scorer training meeting in July 2000. The anchor responses formed the basis for sometimes lengthy discussion and agreement by the NRCs, which served to clarify the distinctions between levels of scoring, and demonstrated the wide variety of ways in which acceptable responses might be framed. The practice papers gave opportunities for the NRCs to work through responses on their own, and to check their scoring against the agreed points.

In finalizing the scoring guides, the anchor papers were viewed as a critical extension of the scoring guides – providing further elaboration and more concrete examples of the levels of responses described in the scoring guide. The anchor sets were constructed to illustrate the expected range of responses and the most common approaches taken by students in answering the constructed-response questions. In addition, two sets of practice papers were compiled for each item. The first set represented the most common types of responses observed in the pilots and field test. The second set provided examples of student responses that might present some challenge in making scoring decisions. Taken together, the two practice sets were designed to prepare scorers for making appropriate and consistent decisions on the most common types of student responses, and on the types of responses that may fall close to the line separating the scoring guide levels. For further clarification, both the anchor and practice sets of sample responses included explicit annotations explaining the rationale for the assigned score.

2.7.4 Training Scorers

National Research Coordinators were responsible for training scoring staff and for conducting scoring in their countries. To prepare them for this task, the PIRLS International Study Center held a scoring training session in May of 2001. The primary purpose of this training session was to ensure that representatives of each participating country fully understood the scoring standards that were to be applied consistently and without variation after the collection of data was completed in each country. The representatives attending this training session were to train the group(s) of scorers in their respective countries, ensuring the comparability of scoring across countries.

At the May 2001 training session, NRCs were instructed on each constructed-response scoring guide. After an initial introduction to each scoring guide, the anchor papers were presented and a discussion of the annotated scoring rationales for each anchor paper ensued – to check that NRCs fully understood how the scoring standards were to be applied. For the majority of constructed-response items, the NRCs also practiced applying these standards with the sets of practice papers that had been compiled. During the practice scoring, NRCs were not shown the previously assigned scores or score rationales so that their ability to apply the scoring standards consistently could be verified.

During the same training session, NRCs were instructed on the specific procedures to be followed in training scorers, and to monitor intra- and inter-country reliability of scoring. NRCs were instructed to follow the same basic procedures in introducing and practicing scoring guides with their own scorers that were followed during their training session. The need for absolute standardization of scoring across countries was emphasized, and all NRCs acknowledged their responsibility for accomplishing this task. This, of course, meant that NRCs and their scorers could make no further changes to scoring guides or annotated sample papers after the May 2001 NRC meeting.

References

Campbell, J.R., Kelly, D. L, Mullis, I. V. S., Martin, M.O., & Sainsbury, M. (2001). *Framework and specifications for PIRLS assessment 2001*. (2nd ed.). Chestnut Hill, MA: Boston College.

3

Developing the PIRLS Background Questionnaires

Dana L. Kelly

3.1 Overview

Children are exposed to language and print at home and at school; receive formal reading instruction; and see others reading for recreation and to perform tasks. These and other experiences and activities at home and school combine to influence how well children read and how they feel about reading by the end of fourth grade. Beyond influences within the home and at school are those in the wider environments in which children live and learn. Community size and resources, organization of the educational system, and educational decision-making affect homes and schools, and thus children's literacy development. To be sure, not all children have the same experiences. Children have varying levels of home support for reading, and different levels of exposure to language and print throughout their lives. They also attend schools with different approaches to learning and resources with which to teach.

By gathering information about children's experiences in learning to read together with reading achievement on the PIRLS test, it is possible to identify the factors or combinations of factors associated with a high degree of reading literacy. The PIRLS design includes a set of questionnaires targeting important factors related to reading literacy. PIRLS administered four questionnaires to the tested students and their parents, reading teachers, and school principals. This chapter describes the conceptual framework underlying the questionnaires, the process used to develop them, and their content.[1]

[1] See *Framework and Specifications for the PIRLS Assessment 2001* (Campbell, Kelly, Mullis, Martin, & Sainsbury, 2001) for more information about the conceptual framework underlying the questionnaires.

Exhibit 3.1: Contexts for the Development of Reading Literacy

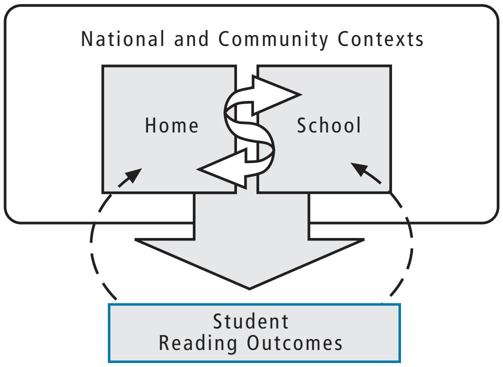


Exhibit 3.2: Factors within the Home, School, and National and Community Contexts Addressed by PIRLS

Factors
National and Community Contexts
Demographics and resources
Governance and organization of educational system
Curriculum characteristics and policies
Home Contexts
Activities fostering reading literacy
Language in the home
Home resources
Home/school connection
Students out-of-school literacy activities
School Contexts
School environment and resources
Teacher training and preparation
Classroom environment and structure
Instructional strategies and activities
Instructional materials and technology

3.2 Framework for the Questionnaires

The PIRLS questionnaires are grounded in a conceptual model relating reading outcomes – students' reading literacy achievement and attitudes – to home, school, and community and national contexts. Exhibit 3.1 illustrates how PIRLS conceptualizes the influences on children's reading by depicting the relationship between home and school, and how both are situated within the community and the country.

The PIRLS questionnaires address factors within each of the aspects that are deemed important for the development of reading literacy. In addition to reading achievement, reading outcomes include students' reading attitudes and behaviors. The factors within the home, school, and national and community contexts addressed by PIRLS are shown in Exhibit 3.2.

3.3 Process for Developing Questionnaire Items and Final Forms

The PIRLS questionnaires were developed through a collaborative process involving the PIRLS International Study Center, the National Research Coordinators (NRCs), the Questionnaire Development Group (QDG), the Reading Development Group (RDG), and the IEA Data Processing Center (DPC). The process included a series of reviews of draft instruments, a field test of five questionnaires in 30 countries, a review of field-test data, and a revision of the field-test questionnaires.

3.3.1 Plan for Questionnaires

In developing the PIRLS questionnaires, the aim was to create instruments that could be used to collect reliable information related to children's reading literacy achievement (as outlined in the framework) without unduly burdening students and schools. Altogether, the instruments were intended to provide a picture of children's experiences from early language and literacy development to the time of the PIRLS assessment. The plan initially called for five questionnaires:

- A student questionnaire to provide data on home and school factors related to reading

- A home questionnaire (to be completed by the students' parents or primary caregivers) to provide data on home support for literacy

- A school questionnaire to provide information on school policies and resources related to reading

- A teacher questionnaire to provide information on instructional approaches and resources at the fourth grade level

- An early-literacy instructional questionnaire to provide information on reading instruction in the grades below the grade tested (fourth grade).

This last questionnaire was originally intended to be administered to a sample of teachers in each of the grades prior to fourth grade. However, early on it became clear that this would be a burden for the schools, and for the national centers preparing and disseminating the instruments. Instead, a less burdensome approach was taken, whereby the school reading coordinator or a teacher familiar with the primary school reading program would complete a questionnaire about literacy instruction in the early grades. All five questionnaires were developed and field-tested.

3.3.2 Initial Item Development

Based on the home, school, and community factors addressed by the framework (shown in Exhibit 3.2), a detailed list of potential variables was developed and reviewed by the NRCs at their first meeting – in July 1999. The list of variables was refined and then used – together with the questionnaires from the Trends in IEA's Reading Literacy Study – as the basis for the initial questionnaire development.

Questionnaire scales used in other reading research studies also were consulted during this initial development phase.

3.3.3 Iterative Review

In September 1999, drafts of student, school, teacher, and home questionnaires were distributed to the NRCs for within-country review. At the same time, they were reviewed by the Reading Development Group. Comments from both reviews were used to revise the questionnaires for a second review by the NRCs in November 1999 – at their second meeting. The Questionnaire Development Group, comprising NRCs from six countries, then met with ISC staff in December 1999 to review the drafts, and to develop the early literacy instruction questionnaire.

In January 2000, the five draft field-test questionnaires were sent to the NRCs for within-country review, and also were reviewed by the RDG. The drafts were revised on the basis of these reviews. At the third NRC meeting – in March 2000 – NRCs reviewed the revised-draft field-test questionnaires, and suggested further revisions. In April 2000, the final field-test questionnaires, were sent to their respective countries for translation and production.

3.3.4 Field Test

The PIRLS field test was conducted in September 2000. Approximately 48,000 students from almost 1,100 schools in 28 countries participated, providing approximately: 1) 48,000 responses to the student questionnaire and the learning to read survey; 2) 2,000 responses to the teacher questionnaire; and 3) 1,000 responses to the school questionnaire and the early literacy instruction questionnaire.

3.3.5 Item Analysis and Finalization of Questionnaires

After the field-test data files had been prepared by each country, then checked and processed by the IEA Data Processing Center, the International Study Center (ISC) prepared five data almanacs – one for each questionnaire – to facilitate review of the data. For each country, each almanac displayed appropriate student-weighted distributions of responses to each question in the questionnaires. In the case of categorical variables, the weighted percentage of respondents choosing each option were shown together with the corresponding average student achievement in reading. For questions with numeric responses, the mean, mode, and selected percentiles were displayed.

The QDG met in December 2000 to review the field-test data for the five questionnaires, and to recommend revisions to the items. Committee members were provided – in addition to the data almanacs – with results from scale-reliability analyses conducted by the DPC, analyses conducted by the Swedish national center, and comments from teachers in New Zealand – to inform their review.

In general, the committee recommended few revisions of the field-test questionnaires; however, there were improvements in the wording of some items in each questionnaire; the removal of as many "filter" questions as possible; and the reordering of items in the student questionnaire. The most significant decision was not to include the early literacy instruction questionnaire in the main survey. Field-test data indicated that, in many countries, the respondent (the school reading coordinator or teacher familiar with early reading instruction in the school) was not able to provide the detailed information required about reading instruction at each grade level; and that the questionnaire (as field-tested) was too burdensome – resulting in unreliable data on many questions. For these reasons, it was not taken forward to the main survey (although six of the more important questions were simplified and included in the school questionnaire).

The ISC prepared drafts of the four main survey questionnaires for review by NRCs at their January 2001 meeting. NRCs recommended few additional changes. Following the meeting, the ISC produced the final documents and electronic files, then distributed them to participating countries for translation and production.

3.4 PIRLS Main Survey Questionnaires

The contents of the PIRLS main survey questionnaires used to collect information about home, school, and community contexts for learning to read are described below.

3.4.1 Student Questionnaire

Each student taking the PIRLS reading assessment completed the student questionnaire. The questionnaire asks about aspects of students' home and school experiences – including instructional experiences and reading for homework, self-perceptions and attitudes towards reading, out-of-school reading habits, computer use, home literacy resources, and basic demographic information. The questionnaire was designed to take 15-30 minutes to complete. Exhibit 3.3 presents details regarding the items in the questionnaire.

3.4.2 Learning-to-Read Survey (Home)

The learning-to-read survey was completed by the parents or primary caregivers of each student taking the PIRLS reading assessment. It deals with child-parent literacy interactions, home literacy resources, parents' reading habits and attitudes, home-school connections, and basic demographic and socioeconomic indicators. This questionnaire was designed to take 10-15 minutes to complete. Exhibit 3.4 presents details regarding the items in the questionnaire.

Exhibit 3.3: Content of the PIRLS Student Questionnaire

Student Questionnaire		
Item Number	**Item Content**	**Description**
1	Gender	Whether student is a boy or girl
2	Date of birth	Month and year of student's birth
3	Out-of-school activities	Frequency student does various reading-related activities and watches television
4	Reading outside of school	Frequency student reads different types of texts outside school
5	Use of library	Frequency student borrows books from library for fun
6	Television watching	Frequency student watches television on a normal school day
7-8	Instructional activities	Frequency student does certain reading instructional activities in school
9-10	Homework	Frequency reading for homework is assigned and amount of time student spends on reading homework
11	Computer use	Frequency computer is used for different literacy activities, and where computer is used
12	Attitude toward reading	Student's attitude towards reading
13	Reading self-concept	Student's self-concept regarding his/her reading ability
14	Feelings about school	Student's feelings about school – safety, perception of students and teachers
15	School environment	Student's reports of problematic behavior by other students at school
16-18	Language in the home	Student's use of the language of the test at home (used as indicator of home environment and home support for reading in language of the test)
19	Books in the home	Number of books in student's home (used as indicator of home environment and socio-economic status)
20	Home possessions	The presence of various socio-economic indicators (used as indicator of home environment and socio-economic status)
21-22	Persons living in home	Number of people and children living in the home (used as indicator of home environment and socio-economic status)
23-25	Student and parents born in country	Provides information on immigrant status (used as indicator of home environment and home support for reading in language of the test)

3.4.3 Teacher Questionnaire

The reading teacher of each fourth-grade class sampled for PIRLS completed a teacher questionnaire, which was designed to gather information about classroom contexts for developing reading literacy. This questionnaire asks teachers to describe the general characteristics of the class tested, such as class size, and the reading levels and language abilities of the students. Several questions in the questionnaire focus on factors related to reading instruction, such as instructional time, materials, grouping of students for instruction, and activities to teach reading and promote the development of the students' reading literacy. The ques-

Exhibit 3.4: Content of the PIRLS Learning-to-Read Survey (Home Questionnaire)

Learning-to-Read Survey		
Item Number	**Item Content**	**Description**
1	Respondent	Who completed the survey
2	Parent/child literacy interactions	Frequency parents engaged in different literacy activities with child during early childhood
3	Attend kindergarten	Whether, and for how long, child attended kindergarten (or equivalent)
4	Age began school	Age when child began formal schooling
5	Literacy skills when began school	Child's literacy skills when he/she began formal schooling
6	Home literacy activities	Frequency parent engages in different literacy activities with child (now)
7	Home/school connection	Parents' perception of school's connection with home
8	View of school	Parents' opinion of school
9-10	Parents' literacy activities	Time parent spends reading for enjoyment, and frequency reading for different purposes
11	Parents' attitude towards reading	Parents' attitude toward reading
12-13	Books in home	Number of books (total and children's) in the home (used as an indicator of home environment and support for literacy, and for socio-economic indicator)
14	Parents' education	Highest level of education completed by both parents (used as an indicator of home environment and socio-economic status)
15-16	Occupational status	Employment status and type of profession of each parent
17-18	Wealth	Perception of wealth relative to others and annual income
19	Time	Amount of time required to complete questionnaire

tionnaire also asks teachers about classroom resources, assessment practices, and efforts to maintain home-school connections. It also asks teachers for their views about opportunities provided for cooperation and collaboration with other teachers, for professional development, and for information about themselves and their education and training. This questionnaire requires about 30 minutes of the teacher's time. Exhibit 3.5 presents details regarding the items in the questionnaire.

3.4.4 School Questionnaire

The principal of each school sampled for PIRLS responded to the school questionnaire. It asks school principals about enrollment and school demographic characteristics, such as school location, resources available in the surrounding area, and indicators of the socioeconomic background of the study body. The school questionnaire also asks principals about reading curriculum policies and total instructional time for the school year. It also includes questions

Exhibit 3.5: Content of the PIRLS Teacher Questionnaire

Teacher Questionnaire		
Item Number	**Item Content**	**Description**
1	Class size	Number of students total, and in the grade tested in the class
2-5	Students in class	Describes the students in the class with respect to reading level, language ability, and reading/language services received
6-7	Language instruction	Whether language instruction is conducted as part of instruction in different curriculum areas or as a separate subject, how much time is spent on language instruction, and how frequently language homework is assigned
8	Reading instruction	Whether reading instruction is conducted as part of instruction in different curriculum areas or as a separate subject
9	Reading instructional time	Amount of time spent on reading instruction, and if that time is for formal reading instruction
10	Reading instruction frequency	Number of days per week reading instruction is provided
11	Instructional grouping	Whether, and how, students are grouped for reading instruction
12	Instructional material	Frequency teacher uses different materials in reading instruction (worksheets, textbooks, etc.)
13	Reading material	Frequency teacher uses different types of texts in reading instruction
14	Reading instructional materials and different abilities	How teacher uses reading instructional materials for students at different reading levels
15-17	Instructional activities	Frequency teacher has students do different reading instructional activities
18	Instructional media	Frequency, and how teacher uses media in reading instruction
19	Computer use	Availability and use of computers and Internet for literacy activities
20	Classroom library	Availability, size, and use of classroom library or reading corner
21	School library use	Frequency teacher takes or sends students to the school library
22-23	Homework	Frequency teacher assigns reading for homework and how much time is expected to be spent on reading homework
24-25	Reading difficulties	Resources available to the teacher to deal with students' reading difficulties, and teacher's approach to dealing with reading difficulties
26-28	Assessment	Teacher's use of different assessment methods to monitor students' progress and performance in reading
29	Professional development	Teacher's perception of opportunities for professional development in school
30	Cooperation and collaboration	Frequency teacher meets with other teachers to discuss and plan reading curriculum or teaching approaches
31	Home/school connection	Frequency teacher meets with parents or sends students' work home
32	Expectations for success	Teacher's expectations for students' future success as readers
33-34	Teaching experience	Number of years teacher has been teaching altogether and teaching the grade tested in particular
35-36	Age and gender	Teacher's age and gender

Exhibit 3.5: Content of the PIRLS Teacher Questionnaire (continued)

Teacher Questionnaire		
Item Number	**Item Content**	**Description**
37-39	Education/training	Teacher's highest level of education, teaching certification, and academic preparation for teaching reading
40	Professional development	Time teacher has spent in professional development in the last two years
41	Reading habits	Frequency teacher reads different material and reads for different purposes
42	Teach class	Whether the class is taught by the teacher only, or by a team of teachers teaching different subjects
43	Time	Amount of time required to complete questionnaire

about resources, the availability of materials and staff, and perceptions of the school climate, as well as the interaction between the schools and the students' parents and families. The school questionnaire was designed to be completed in about 30 minutes. Exhibit 3.6 presents detail on the items in the questionnaire.

Exhibit 3.6: Content of the PIRLS School Questionnaire

School Questionnaire		
Item Number	**Item Content**	**Description**
1	Grades	Grades below the grade test that are present in the school
2-3	Enrollment	Number of boys and girls in school, and in grade tested
4-5	Community	Size and type of community in which the school is located
6	Community resources	Resources available in the community in which the school is located
7-9	Student body	Describes students in the school with respect to stability of student body, family socioeconomic status, and academic abilities
10	Tracking	Whether classes are formed on the basis of ability/performance
11	Instructional time	Number of instructional days per year, amount of instructional time per week, and number of days per week school is open
12	Years with same teacher	Number of years students typically stay with the same teacher
13	Influence on curriculum	Influence on the school's curriculum at grade tested by national or regional curriculum and examinations/assessments, standardized tests, and parents' and students' wishes
14	Literacy skills of students beginning formal schooling	Literacy skills of the students in the school when they begin formal schooling
15-16	Reading literacy emphasis	Relative emphasis placed on reading, writing, and oral language skills by school, presence of own reading curriculum and programs in support of reading education
17	Coordination of reading instruction	Whether school has a policy to coordinate reading instruction across primary school grades
18	Instructional materials	Emphasis on different types of reading material in reading instruction for primary school students
19	Instructional emphasis	Emphasis on different literacy skills and activities at different grades in primary school
20	Reading instruction and different abilities	How reading instructional program is implemented for students at different reading levels
21	School library	Availability, size, and staffing of school library
22	Classroom libraries	Availability of classroom libraries in school
23	Computer availability and Internet access	Availability of computers and access to the Internet for instructional purposes
24	Instructional resources	Material factors affecting school's capacity to provide instruction
25-27	Home/school connection	Availability of programs offered by the school to families, frequency of activities involving parents, and percentage of parents involved in school activities
28-29	School climate	Principal's perception of teachers', parents', and students' attitudes and the severity of students' problem behavior
30-31	Teacher collaboration	Existence of a school policy to promote cooperation and collaboration among teachers and the frequency with which teachers meet to share or develop instructional materials and approaches
32	Principal's time	Percentage of time the principal spends on various roles and functions
33	Time	Amount of time required to complete questionnaire

References

Campbell, J.R., Kelly, D. L, Mullis, I. V. S., Martin, M.O., & Sainsbury, M. (2001). *Framework and specifications for PIRLS assessment 2001*. (2nd ed.). Chestnut Hill, MA: Boston College.

4

Translating the PIRLS Reading Assessment and Questionnaires

Dana L. Kelly

Barbara Malak

4.1 Overview

Since English is the working language of IEA studies, the PIRLS reading assessment and background questionnaires were developed in English, and then translated by the participating countries into their local languages of instruction. In all, the PIRLS data collection instruments were translated from English into 31 languages. Five countries administered the assessment in two languages, and seven countries administered one or more questionnaires in more than one language. The languages in which the test was administered most often were English (seven countries), and Arabic (three countries). In translating the instruments, each country followed procedures established by the PIRLS International Study Center (ISC), and described in the *Survey Operations Manual – Main Survey* (PIRLS, 2001).

Before the translated instruments were used in schools, they were put through an exhaustive process of review and verification. This process – managed by the IEA Secretariat in Amsterdam – was intended to ensure that the instruments had been translated accurately and in accordance with the PIRLS guidelines, and that the translated versions were comparable to the originals (in terms of reading difficulty level and accessibility). As an essential component of the verification process, IEA engaged Berlitz GlobalNet (an independent translation company) to review and verify the translation and layout of each country's instruments. Verifiers reviewed the translated instruments and documented any deviations from the international versions in their reports to IEA. National Research Coordinators received a Translation Verification Report that listed corrections or improvements considered necessary by the verifiers. When all corrections had been completed, the ISC reviewed the revised instruments and gave final approval to the countries to print and administer the materials.

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

4.2 PIRLS Instruments to Be Translated

The instruments to be translated included the PIRLS reading assessment (passages and accompanying questions); the student, teacher, school, and learning-to-read survey questionnaires; and the administration manuals. Countries testing in English did not have to translate the instruments, but did need to adapt the American English of the originals to the vernacular, and make whatever adaptations were necessary for cultural reasons. The reading assessment and questionnaires were put through the verification process, but not the administration manuals.

4.2.1 Reading Assessment

The PIRLS reading assessment comprises ten booklets. Each contains two "blocks" of assessment material. A block is composed of a story or an article (referred to as a "passage" in this chapter) and accompanying questions or items. Nine of the assessment booklets comprise two blocks, each with a passage followed by test items. The tenth booklet contains stories and articles in the PIRLS Reader, a magazine-style booklet, in color, designed to create a more authentic reading experience for students. The questions for the Reader are presented in a separate booklet. Each student completes one of the ten booklets.

While there are ten assessment booklets altogether, there are just eight different blocks of assessment material, four for each reading purpose.[1] The eight blocks are systematically distributed across the ten booklets. Most of the blocks appear in three booklets; two blocks appear only in the PIRLS Reader and accompanying question booklet. The ISC provided each country with electronic files containing all of the material to be translated.

Translation of the reading assessment was based on blocks rather than booklets. Countries translated each block once and entered the translated text into the electronic file for the appropriate test booklet(s). In addition to the assessment blocks, the directions included in each of

[1] PIRLS assesses students' reading literacy for two purposes – reading for literary experience and reading to acquire and use information. See Chapter 2 for more information about the PIRLS test.

the ten booklets had to be translated. The directions were the same in each booklet, and thus needed to be translated only once.

4.2.2 Questionnaires

PIRLS administered four questionnaires: to the tested students, their parents, their reading teachers, and their school principals, to gather information about home and school contexts for learning to read.[2] Each questionnaire contained directions to respondents followed by the questionnaire items. Countries were provided with the electronic files for the questionnaires and entered translated text into the files.

4.3 Translation and Adaptation Guidelines

The survey operations manual developed by the PIRLS ISC provided countries with guidelines for producing a high-quality translation of the instruments and making appropriate cultural adaptations where necessary. These guidelines are summarized in the following sections.

4.3.1 Translating Text

A good translation follows the conventions of the target language and the cultural context while conveying the same essential meaning as the source text. This also is true of good adaptations of the American English of the international version to the variant of English used in another country or cultural context. More specifically:

- Translated text should have the same register (language level, degree of formality) as the source text.
- Translated text should have correct grammar and usage: subject/verb agreement, prepositions, verb tenses, etc.
- Translated text should neither clarify nor omit text from the source text, nor add information not given in the source text.
- Translated text should contain equivalent qualifiers and modifiers, in the order appropriate for the target language.
- Idiomatic expressions should be translated appropriately, not necessarily word-for-word.
- Spelling, punctuation, and capitalization in the target text should be appropriate for the target language and country/cultural context.

[2] See Chapter 3 for more detail about the PIRLS questionnaires.

4.3.2 Adaptations in Passages and Items

In order to make valid comparisons, it is important to ensure equivalence of the passages and items across languages. At the same time, it is important to acknowledge that there are differences in expressions across countries, and to incorporate those differences in the translations. Countries were advised to keep modifications to a minimum, but to make changes where necessary and appropriate. In particular, vocabulary, expressions, and names of people and places could be changed.

Countries were allowed to change particular words in a passage or item so that students would not be faced with unduly unfamiliar vocabulary or expressions. At the same time, the new word could not change the meaning or difficulty of the text. The primary concern was to convey the same meaning and style as the source text. In addition, the guidelines called for national conventions (such as measurement units, date formats, and punctuation to be followed). For example, miles could be replaced by kilometers, and quotation marks could be replaced by dashes to indicate dialogue.

The passages in the PIRLS reading test were collected from the countries participating in the study, and represent a range of cultural contexts. They contain names of characters, real people, and places from around the world. Still, in some instances the names of people and places may have been so unfamiliar to students that they could interfere with reading of the text. Countries were provided with a list of acceptable changes to the names of people and places in the passages. As with changes to vocabulary and expressions, these were not to affect the text in terms of meaning, context, or level of difficulty.

The translation of the questionnaires involved another type of adaptation: there are items in the questionnaires where adaptations were *required*. In the international version of the questionnaires, some items appear with carets (< >) around the text. The text in carets had to be replaced with a country-appropriate term. For example, <country> in the international version was replaced with "Iceland" in the Icelandic version. Questions about the highest level of education parents and teachers had completed were based on the ISCED-1997[3] system. Countries were required to replace the generic ISCED terms shown in carets (for example, <ISCED 3>) with country-appropriate names. For example, in the United States, <ISCED 3> was replaced with "high school." *The Operational Manual for ISCED-1997* (UNESCO, 1999) was provided to

[3] ISCED (International Standard Classification of Education) was developed by UNESCO for cross-national comparisons. *The Operational Manual for ISCED-1997*, provided to each PIRLS country, describes the nine levels of education in that system. Each country identified the levels of education that corresponded to the ISCED levels.

countries to help them determine the correspondence between ISCED levels and their specific educational system.

Countries received detailed information about how to adapt each item requiring modification. This information also clarified what information the item was designed to collect – to help translators select the appropriate word or expression.

4.4 Translation Procedures

The *Survey Operations Manual – Main Survey* also detailed the procedure to be followed in each country in translating the PIRLS instruments. This involved identifying the test language, engaging qualified translators, arranging for two independent translations to arrive at one final translation, documenting all adaptations, producing the translated test booklets and questionnaires, and submitting all materials to the IEA for review and verification.

4.4.1 Identifying the Target Language

In most cases, identifying the language to which the instruments should be translated was quite straightforward. Many countries have one predominant language that is used throughout their educational system. In some countries, however, there is more than one major language of instruction, and instruments needed to be prepared in those languages. For example, in Canada, French-speaking and English-speaking schools participated in the assessment, and so both French and English versions of the booklets were prepared. In some countries, one language is taught in schools and other languages are spoken in homes. In Singapore, for example, students are taught in English, but Chinese, Tamil, and Malay are commonly spoken in their homes. Some countries administered the reading assessment in more than one language, and some countries provided more than one language version of the home questionnaire – for parents for whom the language of the school was not their primary language. For each country, exhibit 4.1 shows the languages used for each PIRLS instrument.

4.4.2 Engaging Translators

The quality of a translation rests primarily on the ability of the translator. Therefore, it is important to hire experienced translators who can accomplish the task. To ensure high-quality translations of the PIRLS assessment and questionnaires, countries were advised to engage translators with the following characteristics:

- An excellent knowledge of English
- An excellent knowledge of the target language
- Experience in the country and cultural context
- Experience with students in the target population
- Familiarity with test development.

To accomplish the task of producing two independent translations, countries were advised to engage at least two translators for each target language.

Exhibit 4.1: Languages in which PIRLS Instruments Were Administered

Country	Language	Test	Student Questionnaire	School Questionnaire	Teacher Questionnaire	Home Questionnaire
Argentina	Spanish	x	x	x	x	x
Belize	English	x	x	x	x	x
Bulgaria	Bulgarian	x	x	x	x	x
Canada (Ontario and Quebec)	English	x	x	x	x	x
	French	x	x	x	x	x
Colombia	Spanish	x	x	x	x	x
Cyprus	Greek	x	x	x	x	x
Czech Republic	Czech	x	x	x	x	x
England	English	x	x	x	x	x
France	French	x	x	x	x	x
Germany	German	x	x	x	x	x
Greece	Greek	x	x	x	x	x
Hong Kong	Modern Chinese	x	x	x	x	x
Hungary	Hungarian	x	x	x	x	x
Iceland	Icelandic	x	x	x	x	x
Iran	Farsi	x	x	x	x	x
Israel	Hebrew	x	x	x	x	x
	Arabic	x	x	--	--	x
Italy	Italian	x	x	x	x	x
	German	x	x	x	x	x
Kuwait	Arabic	x	x	x	x	x
Latvia	Latvian	x	x	x	x	x
	Russian	x	x	x	x	x
Lithuania	Lithuanian	x	x	x	x	x
Macedonia	Macedonian	x	x	x	x	x
	Albanian	x	x	x	x	x
Moldova	Romanian	x	x	x	x	x
	Russian	x	x	x	x	x
Morocco	Arabic	x	x	x	x	x
Netherlands	Dutch	x	x	x	x	x
New Zealand	English	x	x	x	x	x
	Maori	x	x	--	--	x
Norway	Bokmaal	x	x	x	x	x
	Nynorsk	x	x	x	x	x
Philippines[1]	Filipino	x	x	--	--	x
	English	--	--	x	x	--
Romania	Romanian	x	x	x	x	x
	Hungarian	x	x	x	x	x
Russian Federation	Russian	x	x	x	x	x
Scotland	English	x	x	x	x	x
Singapore	English	x	x	x	x	x
	Chinese	--	--	--	--	x
	Malay	--	--	--	--	x
	Tamil	--	--	--	--	x

x indicates that the instrument was administered in that language

-- indicates that the instrument was not administered in that language

[1] The Philippines translated the PIRLS instruments into Filipino, but did not complete data collection.

Exhibit 4.1: Languages in which PIRLS Instruments Were Administered (continued)

Country	Language	Test	Student Questionnaire	School Questionnaire	Teacher Questionnaire	Home Questionnaire
Slovak Republic	Slovak	x	x	x	x	x
	Hungarian	x	x	x	x	x
Slovenia	Slovene	x	x	x	x	x
Sweden	Swedish	x	x	x	x	x
Turkey	Turkish	x	x	x	x	x
United States	English	x	x	x	x	x

x indicates that the instrument was administered in that language
-- indicates that the instrument was not administered in that language

4.4.3 Producing Independent Translations

Countries provided their translators with the international versions of the instruments, the PIRLS translation guidelines, and a blank set of Cultural Adaptation Records to document all adaptations. The two translators were each to translate the same document independently, and then come together to reconcile any differences into a single, finalized version. One set of Cultural Adaptation Records recorded adaptations, and was used during the translation verification to evaluate the quality of the translations.

Countries were allowed to add extra questions to the questionnaires – to collect information relevant to their country or educational system, provided the extra questions were included at the end of the questionnaire. These questions, often referred to as "national options," were to be documented on the Cultural Adaptation Records.

4.4.4 Submitting Materials for External Verification

After translating the test and questionnaires, and producing the booklets in the final layout, countries sent one set of translated and assembled booklets to IEA Headquarters to be verified by Berlitz GlobalNet.

With the exception of six countries (Cyprus, France, Greece, Hong Kong, Norway, and the United States), all countries submitted their instruments for translation verification twice: before the field test, and before main data collection. As those six countries did not participate in the field test, their instruments were verified before the main survey only. For all field-test participants, verification was completed beforehand, and necessary corrections were made before printing instruments. For the main study, due to a tight time schedule, some countries had to print and begin administering the assessment before the verification was completed. For these countries, the results of the international translation verification were used *a posteriori*.

4.5 International Translation Verification

Each country's translated documents went through a rigorous verification process that included verification by Berlitz translators; review by the ISC; and a final check by quality control monitors engaged by the ISC. In addition, item analyses were used to search for any items that had unusual psychometric properties for any country, which could indicate mistranslation.

4.5.1 Process of Translation Verification

IEA Headquarters managed the external verification of the PIRLS instruments. Translators from Berlitz GlobalNet were engaged to review the translated instruments, document all omissions and deviations from the international versions of the instruments, and make suggestions for improvements. Generally, a single verifier reviewed the instruments for each country. However, if a country was administering PIRLS in more than one language, a verifier was engaged for each language. The documentation prepared by the verifiers went to the National Research Coordinators (NRCs) and was used by them to revise their instruments.

The international translation verifiers for PIRLS were required to have the target language as their first language, to have formal credentials as translators working in English, to be educated at the university level, and to live and work in the country for which the verification was carried out. Verifiers were given general information about PIRLS, together with a description of the translation procedures used by the national centers. They also received detailed instructions for reviewing the instruments and registering deviations from the original versions. The standard package of materials for each verifier consisted of the following:

- The international version of each survey instrument (10 test booklets, 1 Reader, 4 questionnaires)
- One set of translated instruments to be verified
- Cultural Adaptation Records completed by the team that prepared the national version of the instruments
- Instructions for verifying translation and layout of the national version
- Guidelines for translation and cultural adaptation of the instruments (as provided to national translators)
- Blank Translation Verification Records to be used to document verification.

The main task of the translation verifiers was to evaluate the accuracy of the translation, the justification for and adequacy of any cultural adaptations, and the comparability of layout of the survey instruments. The instructions emphasized the importance of maintaining the meaning, difficulty level, and format of the text passages and related questions in the student assessment, as well as related questions included in each of the four questionnaires. Verifiers were also warned to pay attention to correspondence between the reading passages and the accompanying questions. Specifically, verifiers had to ensure that:

- The translation had not affected the meaning or difficulty of the text
- The questions had not become easier or more difficult when translated/adapted
- No information had been omitted or added in the translated text
- The assessment booklets contained the correct passages and all items
- The questionnaires contained all items
- The order of items placement on the page, and order of response options to items, were the same as in the international version
- Font, font size, paragraph spacing, and margins were the same
- Page order and numbering were the same
- Text placement on the pages was the same
- Graphics looked the same and were placed correctly
- Uses of boldface, shading, italics, etc., was the same.

4.5.2 Translation Verification Records

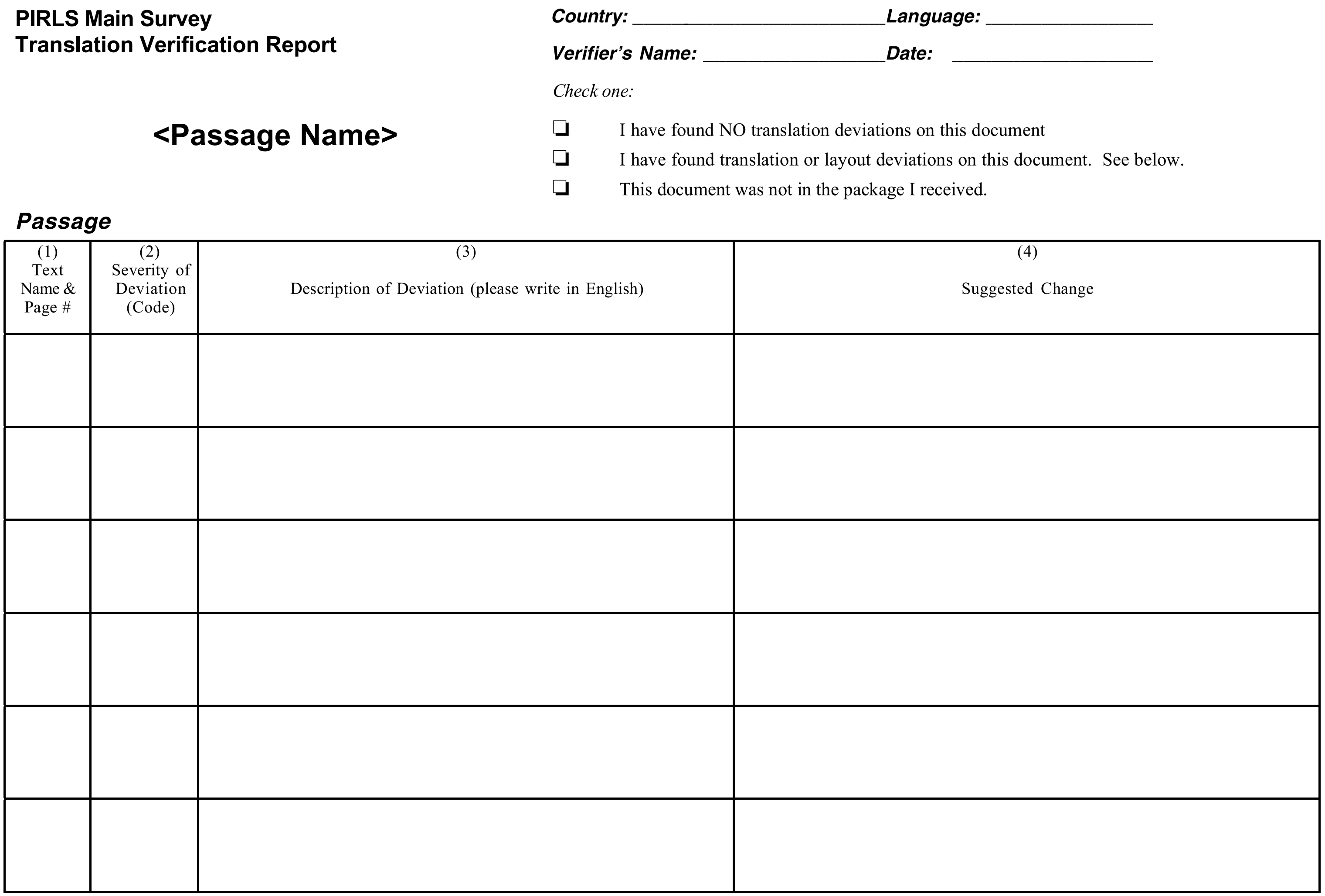
Translation Verification Records were used by verifiers to register all deviations in each participating country's translated or adapted instruments, including: additions, deletions, mistranslations, and changes in layout. There were separate forms for: assessment booklet directions, each of the eight blocks of assessment material, each of the four questionnaires, and assessment booklet layout and content.

For each form, the verifier completed the form header indicating whether or not deviations were found. If the verifier judged the translated or adapted version to be equivalent to the international version, no further entry was needed. If the verifier judged them to be different, an entry was made in the translation verification form – giving the location of the deviation (page #), the severity of the deviation (using the severity code below), a description of the deviation and a suggested change that would improve comparability. An example form for an assessment block is shown in Exhibit 4.2.

Severity codes were used to indicate the extent to which the translated text or format differed from the international version. The severity codes ranged from 1 (major change or error) to 4 (acceptable change).

- **Major Change or Error**: Examples include: incorrect order of choices in a multiple-choice question, omission of a graphic, omission of a question, incorrect translation resulting in the answer being revealed by the question, incorrect translation that changes the meaning or difficulty of the passage or question, and incorrect ordering of questions.

Exhibit 4.2: Example Translation Verification Record Form – Assessment Block



PIRLS Main Survey
Translation Verification Report

Country: ________________________***Language:*** ________________

Verifier's Name: __________________***Date:*** ____________________

Check one:

❑ I have found NO translation deviations on this document

❑ I have found translation or layout deviations on this document. See below.

❑ This document was not in the package I received.

<Passage Name>

Passage

(1) Text Name & Page #	(2) Severity of Deviation (Code)	(3) Description of Deviation (please write in English)	(4) Suggested Change

- **Minor Change or Error**: Examples include: spelling errors that do not affect comprehension, misalignment of margins or tabs, inappropriate changes in font or font sizes, discrepancies in the headers and footers of the document.

- **Suggestion for Alternative**: The translation may be adequate, but verifier suggests a different wording.

- **Acceptable Change**: Change is acceptable and appropriate. For example, a reference to winter is changed from January to July for the Southern Hemisphere.

The completed Translation Verification Records were sent to the NRCs and to the ISC at Boston College. The NRC was responsible for reviewing the report forms and revising the instruments based on the translation verifiers' suggestions. The NRC was not required to accept all recommendations made by the verifier; if a change did not seem warranted or appropriate, the NRC documented the disagreement along with a rationale for not changing the text.

4.5.3 Final Review by the International Study Center

After implementing the suggestions made by the verifier, the NRC submitted the translated assessment booklets and questionnaires to the International Study Center for final review. At the International Study Center, staff examined the translated versions of the instruments, using the Translation Verification Records and any documentation provided by the NRC. Any errors that were identified were reported to the NRC. When there were no remaining issues to resolve, the NRC could print the booklets and administer the assessment.

4.5.4 Quality Control Monitor Review

As part of the PIRLS quality control program, Quality Control Monitors were engaged by the ISC to visit each country and document the quality of the PIRLS assessment. One of the important tasks for the Quality Control Monitor during the visit was to check the translation of the PIRLS instruments by reviewing the Translation Verification Records alongside the translated instruments actually used in each country – to ensure that changes recommended by the verifier were indeed implemented in the final versions of the translated instruments.

References

Progress in International Reading Literacy Study (PIRLS). (2001). *Survey Operations Manual – Main Survey* (Doc. Ref.: PIRLS 01-0001). Prepared by the International Study Center. Chestnut Hill, MA: *PIRLS International Study Center*, Boston College.

UNESCO. (1999) *Operational Manual for ISCED-1997*. Paris: UNESCO Institute for Statistics.

5

PIRLS Sampling Design

Pierre Foy

Marc Joncas

5.1 Overview

This chapter describes the PIRLS 2001 procedures for sampling from the student population in each participating country. To be acceptable for PIRLS, national sample designs had to result in probability samples that gave accurate weighted estimates of population parameters such as means and percentages, and for which estimates of sampling variance could be computed. The PIRLS sample design is derived from the design of IEA's TIMSS (see Foy & Joncas, 2000), with minor refinements. Since sampling for PIRLS was to be implemented by the National Research Coordinator (NRC) in each participating country – often with limited resources – it was essential that the design be simple and easy to implement while yielding accurate and efficient samples of both schools and students. The design that was chosen for PIRLS strikes a good balance, providing accurate sample statistics while keeping the survey simple enough for all participants to implement.

The international project team provided manuals and expert advice to help NRCs adapt the PIRLS sample design to their national system, and to guide them through the phases of sampling. The *School Sampling Manual* (PIRLS, 1999) describes how to implement the international sample design to select the school sample; and offers advice on initial planning, adapting the design to national situations, establishing appropriate sample selection procedures, and conducting fieldwork. The *Survey Operations Manual – Main Survey* and *School Coordinator Manual – Main Survey* (PIRLS, 2001b, 2001a) provide information on sampling within schools, assigning assessment booklets and questionnaires to sampled students, and tracking respondents and non-respondents. To automate the rather complex within-school sampling procedures, NRCs were provided with sampling software jointly developed by the IEA Data processing Center and Statistics Canada (IEA, 2001).

As well as administering the PIRLS 2001 instruments, countries that had participated in IEA's 1991 Reading Literacy Study had the option of using their national 1991 Reading Literacy Study instruments to measure trends in reading achievement between 1991 and 2001. This component of PIRLS 2001 was known as the Trends in IEA's Reading Literacy Study. The *School Sampling Guide for the 10-Year Trend Study* (PIRLS, 2000) describes how to implement the international sample design for the trend study.

In addition to sampling manuals and software, expert support was made available to help NRCs with their sampling activities. Statistics Canada (in consultation with the PIRLS sampling referee) reviewed and approved the national sampling plans, sampling data, sampling frames, and sample implementation. Statistics Canada also provided advice and support to NRCs at all stages of the sampling process, drawing national school samples for more than half of the PIRLS participants.

Where the local situation required it, NRCs were permitted to adapt the sample design for their educational systems, using more sampling information, and more sophisticated designs and procedures than the base design required. However, these solutions had to be approved by the International Study Center (ISC) at Boston College, and by Statistics Canada.

5.2 PIRLS Target Population

In IEA studies, the target population for all countries is known as the *international desired target population*. This is the grade or age level that each country should address in its sampling activities. The international desired target population for PIRLS was the following:

> All students enrolled in the upper of the two adjacent grades that contain the largest proportion of 9-year-olds at the time of testing.

The PIRLS target grade was usually the fourth grade of primary school. Because fourth grade generally signals the completion of formal reading instruction, countries for which the target grade would have been the third grade (based on the international desired target population) were permitted to retain the fourth grade as their target grade. The PIRLS target population was derived from that used by TIMSS in 1995, and identical to that used by TIMSS 2003 at primary school level.

5.2.1 Sampling from the Target Population

PIRLS expected all participating countries to define their *national desired population* to correspond as closely as possible to its definition of the international desired population. For example, if fourth grade was the upper of the two adjacent grades containing the greatest proportion of 9-year-olds in a particular country, then fourth grade should be the national desired population for that country. Although countries were expected to include all students in the target grade in their definition of the population, sometimes they had to reduce their coverage. Lithuania, for example, planned to collect data only about students in Lithuanian-speaking schools, so their national desired population fell short of the international desired population. The international report documents such deviations from the international definition of the PIRLS target population.

Using its national desired population as a basis, each participating country had to define its population in operational terms for sampling purposes. This definition, known in IEA terminology as the *national defined population*, is essentially the sampling frame from which the first stage of sampling takes place. Ideally, the national defined population should coincide with the national desired population, although in reality there may be some school types or regions that cannot be included; consequently, the national defined population is usually a very large subset of the national desired population.All schools and students in the desired population not included in the defined population are referred to as the excluded population.

PIRLS participants were expected to ensure that the national defined population included at least 95 percent of the national desired population. Exclusions (which should be kept to a minimum) could occur at the school level, within the sampled schools, or both. Because the national desired population was restricted to schools that contained the required grade, schools not containing the target grade were considered to be outside the scope of the sample – not part of the target population.

Although countries were expected to do everything possible to maximize coverage of the population by the sampling plan, schools could be excluded, where necessary, from the sampling frame for the following reasons:

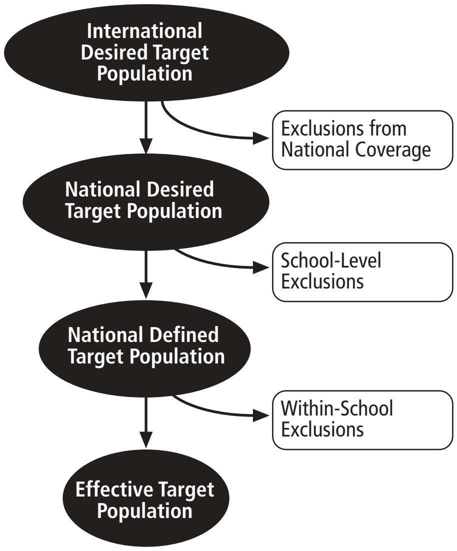
- They were in geographically remote regions.
- They were of extremely small size.
- They offered a curriculum or a school structure that was different from the mainstream educational system(s).
- They provided instruction only to students in the categories defined as "within-school exclusions."

Within-school exclusions were limited to students who, because of some disability, were unable to take the PIRLS tests. NRCs were asked to define anticipated within-school exclusions. Because these definitions can vary internationally, they were also asked to follow certain rules adapted to their jurisdictions. In addition, they were to estimate the size of the included population so that their compliance with the 95 percent rule could be projected.

The general PIRLS rules for defining within-school exclusions included the following three groups:

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

Exhibit 5.1: Relationship Between the Desired Populations and Exclusions


International Desired Target Population
Exclusions from National Coverage
National Desired Target Population
School-Level Exclusions
National Defined Target Population
Within-School Exclusions
Effective Target Population

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

- **Functionally disabled students**. These are students who were permanently physically disabled in such a way that they could not perform in the PIRLS tests. Functionally disabled students who could perform were included in the testing.
- **Non-native-language speakers**. These are students who could not read or speak the language of the test, and so could not overcome the language barrier of testing. Typically, a student who had received less than one year of instruction in the language of the test was excluded, but this definition was adapted in different countries.

A major objective of PIRLS was that the effective target population, the population actually sampled by PIRLS, be as close as possible to the international desired population. Exhibit 5.1 illustrates the relationship between the desired populations and the excluded populations. Each country had to account for any exclusion of eligible students from the international desired population. This applied to school-level exclusions as well as within-school exclusions.

5.3 Sample Design

The international sample design for PIRLS is generally referred to as a two-stage stratified cluster sample design. The first stage consists of a sample of schools,[1] which may be stratified; the second stage consists of a sample of one or more classrooms from the target grade in sampled schools.

[1] In some very large countries, it was necessary to include an extra preliminary stage, where school districts were sampled first, and then schools.

5.3.1 Units of Analysis and Sampling Units

The PIRLS analytical focus was on the cumulative learning of students, as well as on instructional characteristics affecting learning. The sample design, therefore, had to address the measurement both of characteristics thought to influence cumulative learning, and of those specific to the instructional settings. As a consequence, schools, classrooms, and students were all potential units of analysis; all had to be considered as sampling units in the sample design in order to meet specific requirements for data quality and sampling precision at all levels.

Although the second stage sampling units were intact classrooms, the ultimate sampling elements were students – making it important that each student from the target grade be a member of one (and only one) of the classes in a school from which the sampled classes would be selected.

5.3.2 Sampling Precision and Sample Size

Sampling sizes for the two stages of the PIRLS sampling had to be specified so as to meet the sampling precision requirements of the study. Since students were the principal units of analysis, the reliability of estimates of student characteristics was paramount. However, PIRLS planned to report extensively on school, teacher, and classroom characteristics, so it was necessary also to have sufficiently large samples of schools and classes. The PIRLS standard for sampling precision requires that all student samples have an effective sample size of at least 400 students for the main criterion variables. In other words, all student samples should yield sampling errors that are no greater than would be obtained from a simple random sample of 400 students.

An effective sample size of 400 students results in the following approximate 95 percent confidence limits for sample estimates of population means, percentages, and correlation coefficients.

- Means: m ± 0.1s (where *m* is the mean estimate, and *s* is the estimated standard deviation for students)
- Percentages: p ± 5% (where *p* is a percentage estimate)
- Correlations: r ± 0.1 (where *r* is a correlation estimate).

Furthermore, since PIRLS planned to conduct analyses at the school and classroom levels, at least 150 schools were to be selected from the target population. A sample of 150 schools yields 95 percent confidence limits for school-level and classroom-level mean estimates that are precise to within 16 percent of their standard deviations. To ensure sufficient sample precision for school-level analyses, some participants had to sample more schools than would have been selected otherwise.

The precision of multistage cluster sample designs is generally affected by the so-called clustering effect. Students are clustered in schools, and are also clustered in classrooms within the schools. A classroom – as a sampling unit – constitutes a cluster of students who tend to be more like each other than like other members of the population. The *intra-class correlation* is a measure of this within-class similarity. Sampling 30 students from a single classroom when the intra-class correlation is high will yield less information than a random sample of 30 students spread across all classrooms in a school. Such sample designs are less efficient, in terms of sampling precision, than a simple random sample of the same size. This clustering effect was considered in determining the overall sample size for PIRLS.

The size of the cluster (classroom) and the size of the intra-class correlation determine the magnitude of the clustering effect. For planning the sample size, therefore, each country had to identify a value for the intra-class correlation and a value for the expected cluster size (this was known as the minimum cluster size). For PIRLS, the intra-class correlation for each country was estimated from past studies (such as TIMSS) or from national assessments. In the absence of these sources, an intra-class correlation of 0.3 was assumed. Since participants were sampling intact classrooms, the minimum cluster size was in fact the average classroom size.

Exhibit 5.2: PIRLS Sample-Design Table

Minimum Cluster Size		Intraclass Correlations								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
5	a	150	169	201	233	265	297	329	361	393
	n	750	845	1 005	1 165	1 325	1 485	1 645	1 805	1 965
10	a	150	150	161	197	233	269	305	341	377
	n	1 500	1 500	1 610	1 970	2 330	2 690	3 050	3 410	3 770
15	a	150	150	150	184	222	259	296	334	371
	n	2 250	2 250	2 250	2 760	3 330	3 885	4 440	5 010	5 565
20	a	150	150	150	178	216	254	292	330	368
	n	3 000	3 000	3 000	3 560	4 320	5 080	5 840	6 660	7 360
25	a	150	150	150	175	213	251	291	328	367
	n	3 750	3 750	3 570	4 375	5 325	6 275	7 250	8 220	9 175
30	a	150	150	150	172	211	250	288	327	366
	n	4 500	4 500	4 500	5 160	6 330	7 500	8 640	9 810	10 980
35	a	150	150	150	170	209	248	287	326	365
	n	5 250	5 250	5 250	2 950	7 315	8 680	10 045	11 410	12 775
40	a	150	150	150	169	208	247	286	325	364
	n	6 000	6 000	6 000	6 760	8 320	9 880	11 440	13 000	14 960
45	a	150	150	150	168	207	246	285	325	364
	n	6 750	6 750	6 750	7 560	9 315	11 070	12 825	14 625	16 380
50	a	150	150	150	167	207	246	285	324	363
	n	7 500	7 500	7 500	8 350	10 350	12 300	14 250	16 200	18 150

a = Number of sampled schools

n = Number of sampled students in target grade

Note: Minimum Cluster Size is number of students selected in each sampled school (generally the average classroom size)

Sample-design tables, such as the one in Exhibit 5.2, were produced and included in the PIRLS *School Sampling Manual*. These tables illustrate the number of schools necessary to meet the PIRLS sampling precision requirements for a range of values of intra-class correlations and minimum cluster sizes. PIRLS participants could refer to the tables to determine how many schools they should sample. For example, on the basis of Exhibit 5.2, a participant whose intra-class correlation was expected to be 0.6, with an average classroom size of 30, would need to sample a minimum of 250 schools. Whenever the estimated number of schools to sample fell below 150, participants were asked to sample at least 150 schools.

The sample-design tables could be used also to determine sample sizes for more complex designs. For example, a number of strata could be constructed for which different minimum cluster sizes could be specified, thereby refining the national sample design in a way that might avoid special treatment of small schools (see section 5.4.1).

5.3.3 Stratification

Stratification is the grouping of sampling units (e.g., schools) in the sampling frame according to some attribute or variable prior to drawing the sample. It is generally used for the following reasons:

- To improve the efficiency of the sample design, thereby making survey estimates more reliable

- To apply different sample designs or disproportionate sample-size allocations to specific groups of schools (such as those within certain states or provinces)

- To ensure adequate representation in the sample of specific groups from the target population.

Examples of stratification variables for school samples are: geography (such as states or provinces), school type (such as public and private), and level of urbanization (such as rural and urban). Stratification variables in the PIRLS sample design could be used explicitly, implicitly, or both.

- **Explicit stratification** consists of building separate school lists, or sampling frames, according to the stratification variables under consideration. Where, for example, geographic regions are an explicit stratification variable, separate school-sampling frames would be constructed for each region. Different sample designs, or different sampling fractions, would then be applied to each school-sampling frame, to select the sample of schools. In PIRLS, the main reason for considering explicit stratification was to ensure disproportionate allocation of the school sample across strata. For example, a country stratifying by school size might require a specific number of schools from each stratum, regardless of the relative size of the stratum.

- **Implicit stratification** makes use of a single school-sampling frame, but sorts the schools in this frame by a set of stratification variables. This type of stratification is a simple way of ensuring proportional sample allocation without the complexity of explicit stratification. It can also improve the reliability of survey estimates – provided the variables are related to school mean student achievement in reading literacy.

5.3.4 Replacement Schools

Although PIRLS participants were expected to make great efforts to secure the participation of sampled schools, it was anticipated that a 100 percent participation rate would not be possible in all countries. To avoid sample-size losses, a mechanism was instituted to identify, *a priori*, replacement schools for each sampled school. For each sampled school, the next school on the ordered school sampling frame was identified as its replacement – and the one after that as a second replacement, should it be needed (see Exhibit 5.3 for an example).

The use of implicit stratification variables and the subsequent ordering of the school sampling frame by size ensured that any sampled school's replacement would have similar characteristics. Although this approach does not guarantee avoiding response bias, it tends to minimize the potential for bias, and was deemed more acceptable than over-sampling to accommodate a low response rate.

5.4 First Sampling Stage

The sample-selection method used for the first sampling stage in PIRLS made use of a systematic probability-proportional-to-size (PPS) technique. In order to use this method, it was necessary to have some measure of the size (MOS) of the sampling units. Ideally, this was the number of sampling elements within the unit (e.g., the number of students in the school in the target grade). If this was unavailable, some other highly correlated measure, such as total school enrollment, was used.

The schools in each explicit stratum were listed in order of the implicit stratification variables – together with the MOS for each school. Schools were further sorted by MOS within implicit stratification variables. The measures of sizes were accumulated from school to school, and the running total (the cumulative MOS) was listed next to each

Exhibit 5.3: Application of the PPS Systematic Sampling Method to PIRLS

Total MOS:	392 154	Sampling Interval:	2 614.3600
School Sample:	150	Random Start:	1 135.1551

School Code	School MOS	Cumulative MOS	Sample
939438	532	532	
26825	517	1049	
277618	487	1536	–
228882	461	1997	R1
833389	459	2456	R2
386017	437	2893	
986694	406	3299	
41733	385	3684	
56595	350	4034	–
945801	341	4375	R1
865982	328	4703	R2
700089	311	5014	
656616	299	5313	
647690	275	5588	
381836	266	5854	
510529	247	6101	
729813	215	6316	
294281	195	6511	–
16174	174	6685	R1
292526	152	6837	R2
541397	133	6970	
502014	121	7091	
662598	107	7198	
821732	103	7301	
436600	97	7398	

– = Sampled School

R1, R2 = Replacement Schools

school (see Exhibit 5.3). The cumulative MOS was a measure of the size of the population of sampling elements; dividing it by the number of schools to be sampled gave the sampling interval.

The first school was sampled by choosing a random number in the range between 1 and the sampling interval. The school whose cumulative MOS contained the random number was the sampled school. By adding the sampling interval to that first random number, a second school was identified. This process of consistently adding the sampling interval to the previous selection number resulted in a PPS sample of the required size.

Among the many benefits of this sample-selection method are that it was easy to implement, and that it was easy to verify that it was implemented properly. The latter is critical, since one of PIRLS's main objectives was to ensure that a sound sampling methodology had been used.

Exhibit 5.3 illustrates the PPS systematic sampling method applied to a fictitious sampling frame. The first three sampled schools are shown, as well as their preselected replacement schools – should the originally selected schools not participate.

5.4.1 Small Schools

Small schools tend to be problematic in PPS samples because students sampled from them get very large sampling weights, which can increase sampling variance. Also, when the school size falls below the minimum cluster size, it reduces the overall student sample size. In PIRLS, a school was deemed to be small if the number of students in the target grade was less than the minimum cluster size. For example, if the minimum cluster size was set at 20, then a school with fewer than 20 students in the target grade was considered a small school.

The PIRLS approach for dealing with small schools consisted of two steps:

- **Identifying extremely small schools**. Extremely small schools were defined as schools with fewer students than a quarter of the minimum cluster size. For example, if the minimum cluster size was set at 20, then schools with fewer than five students in the target grade were considered extremely small schools. If student enrollment in these schools was less than 2 percent of the eligible population, then these schools could be excluded – provided the overall inclusion rate met the 95 percent criterion (see section 5.2.1).

- **Creating an explicit stratum of small schools**. If fewer than 10 percent of eligible students were enrolled in small schools, then no additional actions were required. If, however, more than 10 percent of eligible students were enrolled in small schools, then it was necessary to create an explicit stratum for small schools. The number of schools to be sampled from this explicit stratum would remain proportional to the stratum size, but all schools would have an equal probability of selection. This action would ensure greater stability in the resulting sampling weights.

5.4.2 Optional Preliminary Sampling Stage

Very large countries had an opportunity to introduce a preliminary sampling stage before sampling schools. The Russian Federation and the United States availed themselves of this option. In these countries, the first step was to draw a sample of geographic regions using PPS sampling. Then a sample of schools was drawn from each sampled region. This design was used mostly as a cost-reduction measure, where the construction of a comprehensive list of schools would have been either impossible or prohibitively expensive. Also, the additional sampling stage reduced the dispersion of the school sample, thereby potentially reducing travel costs. Sampling guidelines were put in place to ensure than an adequate number of units would be sampled from this preliminary stage. The sampling frame had to consist of at least 80 primary sampling units, of which at least 40 had to be sampled at this stage.

5.5 Second Sampling Stage

The second sampling stage consisted of selecting classrooms within sampled schools. As a rule, one classroom per school was sampled, although some participants opted to sample two classrooms. All classrooms were selected with equal probabilities for all countries.

5.5.1 Small Classrooms

Generally, classrooms in an education system tend to be of roughly equal size. Occasionally, however, small classrooms are devoted to special situations, such as remedial or accelerated programs. These classrooms can become problematic – since they can lead to a shortfall in sample size – and thus introduce some instability in the resulting sampling weights, when classrooms are selected with PPS.

In order to avoid these problems, it was suggested that any classroom smaller than half the specified minimum cluster size be combined with another classroom from the same grade and school. For example, if the minimum cluster size was set at 30, then any classroom with fewer than 15 students should be combined with another. The resulting pseudo-classroom would then constitute a sampling unit.

5.6 Trends in IEA's Reading Literacy Study

PIRLS countries that had earlier participated in the 1991 IEA Reading Literacy Study had the option of undertaking the Trends in IEA's Reading Literacy Study, which measured trends in reading achievement using IEA's 1991 reading test and student questionnaire. Since the target population for the Trends in IEA's Reading Literacy Study was similar (but not identical to) the PIRLS target population, it was possible to use the PIRLS school sample as the basis for the trend study sample. Accordingly, the sampling plan for the Trends in IEA's Reading Literacy Study was simple: select every second school sampled for PIRLS, and from each of these, sample one additional classroom from the target grade. Since the sample of schools for the Trends in IEA's Reading Literacy Study is essentially a subsample of the PIRLS sample of schools, most of the required sampling tasks were carried out during the PIRLS school sampling.

Exhibit 5.4: Countries Participating in the Trends in IEA's Reading Literacy Study

Country	Primary School Target Grade
Greece	4
Hungary	3
Iceland	4
Italy	4
New Zealand	4
Singapore	3
Slovenia	3
Sweden	3
United States	4

The nine countries that took part in the Trends in IEA's Reading Literacy Study and their target grades are presented in Exhibit 5.4.

5.6.1 Trends in IEA's Reading Literacy Study Target Population

The target population in 1991 was the grade with the greatest number of nine-year-olds at the time of testing, and to maintain comparability, the same population was targeted by the trend data collection in 2001. However, the PIRLS 2001 target population differs somewhat from the 1991 population in that PIRLS targeted the upper of the *two* grades with most nine-year-olds, and so the target grade in each country was not always the same for the two studies. These definitions yield the same target grade in Greece, Iceland, Italy, New Zealand, Slovenia, and the United States, but different in Hungary, Singapore, and Sweden.

5.6.2 Sample Design

In general, the sample for the trend study consisted of half of the PIRLS school sample, with one classroom chosen at random from the target grade in each of the sampled schools. The procedure was as follows:

- Select every second school sampled for PIRLS starting randomly with the first or second school.

- Sample an extra classroom (in addition to the PIRLS classroom already sampled) within these selected schools.

- If a school sampled for both studies has only one classroom, assign that school and classroom to PIRLS and use the first replacement school for that school as the sampled school for the trend study.

5.6.3 Replacement Schools

Because schools sampled for the trend study were also sampled for PIRLS, first and second replacement schools for this study are the same ones identified for PIRLS.

There were, however, three exceptions to this rule:

- **A sampled school had only one classroom and agreed to participate in the study**. In this case, the only available classroom in the sampled school was assigned to PIRLS and the PIRLS first replacement school became the trend study sampled school. This left the PIRLS second replacement school as the only 10-year trend study replacement school.

- **A sampled school refused to participate, but the corresponding PIRLS first replacement school agreed to participate and had only one classroom**. In this case, the PIRLS first replacement was used for PIRLS and the PIRLS second replacement school became the trend study first replacement school. In this scenario, there is no second replacement school for the trend study.

- **Both the sampled school and the PIRLS first replacement school refused to participate, but the PIRLS second replacement school agreed to participate and had only one classroom**. In this case, there was no trend study replacement school and the sampled school had a non-participation status.

5.7 Sampling Precision and Sample Size

With a single classroom sampled from only half of the 150 schools sampled for PIRLS, the number of students sampled for Trends in IEA's Reading Literacy Study should be roughly half the number of students sampled for PIRLS. This translates into a loss of sampling precision when compared with PIRLS. To get an idea of the resulting standard of sampling precision, the 95 percent confidence limits given earlier in Section 5.3.2 are simply multiplied by $\sqrt{2}$. This gives the following 95 percent confidence limits for sample estimates of population means, percentages and correlation coefficients:

- Means: m ± 0.14s (where *m* is a student mean estimate and *s* is its estimated standard deviation for students)

- Percentages: p ± 7 percent (where *p* is a student-level percentage estimate)

- Correlations: r ± 0.14 (where *r* is a student-level correlation estimate).

The Trends in IEA's Reading Literacy Study focuses mainly on student achievement, but can also report results from schools and classrooms. Based on a minimum sample size of 75 schools, such results should have 95 percent confidence limits for means and percentages in the range ± 23 percent of their standard deviations.

References

IEA. (2001). *W3S: Within-School Sampling Software*. Hamburg: IEA Data Processing Center.

Foy, P., & Joncas, M. (2000). TIMSS Sample Design. In M.O. Martin, K.D. Gregory, & S.E. Stemler. *TIMSS 1999 Technical Report*. Chestnut Hill, MA: Boston College.

Progress in International Reading Literacy Study (PIRLS). (1999). *School Sampling Manual – Version 2* (Doc. Ref.: PIRLS 99-0019). Prepared by Pierre Foy & Marc Joncas, Statistics Canada. Chestnut Hill, MA: Boston College.

Progress in International Reading Literacy Study (PIRLS). (2000). *School Sampling Guide for the 10-Year Trend Study – Version 2* (Doc. Ref.: PIRLS 00-0001). Prepared by Pierre Foy & Marc Joncas, Statistics Canada. Chestnut Hill, MA: Boston College.

Progress in International Reading Literacy Study (PIRLS). (2001a). *School Coordinator Manual – Main Survey* (Doc. Ref.: PIRLS 01-0003). Prepared by the International Study Center. Chestnut Hill, MA: Boston College.

Progress in International Reading Literacy Study (PIRLS). (2001b). *Survey Operations Manual – Main Survey* (Doc. Ref.: PIRLS 01-0001). Prepared by the International Study Center. Chestnut Hill, MA: Boston College.

6

PIRLS Survey Operations Procedures

Eugenio J. Gonzalez

Dirk Hastedt

Ann M. Kennedy

6.1 Overview

The PIRLS 2001 data collection was a very complex undertaking in each country, requiring close cooperation between the National Research Coordinator (NRC) and school personnel – including students and their parents, school principals, and teachers. Survey operations and procedures for administering the PIRLS 2001 assessment were developed by the international project team, and documented in a series of manuals provided to the national centers. Each country was responsible for implementing the procedures according to the international standards.

The PIRLS 2001 survey operations were designed collaboratively by the International Study Center (ISC) at Boston College, the IEA Data Processing Center, and Statistics Canada. They were based on procedures used successfully in TIMSS and other IEA studies, and refined on the basis of the PIRLS 2001 field-test experience. As well as providing data to inform the instrument development process, the field test, which was conducted in 30 countries in September 2000, allowed participating countries to gain practical experience with the procedures described in the manuals, and provided an opportunity to identify areas in need of improvement.

This chapter describes the survey operations used to collect the PIRLS data, including: the procedure for sampling classrooms within schools and tracking students and teachers, the steps involved in administering the achievement tests and background questionnaires, and the requirements for monitoring the quality of the data collection. It also describes the activities involved in preparing the data files at the national center, particularly those for scoring the constructed-response items, creating and checking data files for achievement test and questionnaire responses, and dispatching the completed files to the IEA Data Processing Center in Hamburg.

In addition to administering the PIRLS 2001 instruments, the Trends in IEA's Reading Literacy Study was an option for PIRLS countries that also participated in the 1991 IEA Reading Literacy Study. A separate section in this chapter describes the survey operations for the trend study.

6.2 Responsibilities of the National Research Coordinator

The NRC for each country had primary responsibility for carrying out the survey operations. The NRC was responsible for collecting and preparing the data for the PIRLS assessment according to the procedures specified internationally. Earlier chapters of this report describe the tasks of the NRC with regard to choosing a sample of schools, and translating the achievement tests and questionnaires. This chapter focuses on activities associated with the data collection itself.

An important responsibility of the NRC was to identify a School Coordinator for each of the sampled schools to act as a liaison between the national center and the school. The primary role of the School Coordinator was to assist the NRC in the assessment activities within the school, such as: the sampling and identification of classes, and ensuring that the administrative and testing materials were correctly distributed, completed, and collected. The School Coordinator also was responsible for identifying and training a Test Administrator to conduct the testing sessions. In some countries, the School Coordinator assumed the roles and responsibilities of the Test Administrator. Both the School Coordinator and Test Administrator received training materials that described their responsibilities in detail.

6.3 Documentation and Software

NRCs were provided with a comprehensive set of procedural manuals detailing all aspects of data collection:

- The *Survey Operations Manual* (PIRLS, 2001f) was the essential handbook of the NRC. It described in detail all of the operational activities and responsibilities of the NRC, including: translating and verifying the achievement tests and questionnaires, preparing the assessment materials for use in schools, securing school cooperation, conducting within-school sampling activities, distributing materials to schools, administering the tests and questionnaires and retrieving the completed instruments from schools, training scoring staff and scoring the constructed response achievement questions, and preparing the data files for dispatch to the IEA Data Processing Center.

- The *School Sampling Manual* (PIRLS, 1999) gave an operational definition of the PIRLS main survey target population and sampling goals, and detailed the procedures for sampling schools.

- The *School Coordinator Manual* (PIRLS, 2001d) described the steps to be taken by the School Coordinator from the selection of the school for testing through the receipt of the survey tracking forms and testing materials at the school. It also specified procedures for returning the completed testing materials to the national center.

- The *Test Administrator Manual* (PIRLS, 2001g) covered the procedures from the beginning of testing to the return of the testing materials and completed Student Tracking Forms to the School Coordinator, and contained the administration script for the testing sessions.

- The *Manual for Entering the PIRLS Data* (PIRLS, 2001a) defined the variables and file formats in the data files, and provided instructions for coding, entering, and verifying the data. Codebooks containing detailed specifications for each variable were part of the documentation provided along with this manual.

- The *Scoring Guides for the Constructed-Response Items* (PIRLS, 2001e) contained the guides developed for scoring each of the constructed-response items.

- The *Manual for National Quality Control Observers* (PIRLS, 2001c) provided instructions for conducting classroom observations in a sample of 10 percent of the participating schools. The observers who conducted the observations were hired by the national centers.

- The *Manual for International Quality Control Monitors* (PIRLS, 2001b) described the procedures employed by International Quality Control Monitors, who were hired by the International Study Center to visit each PIRLS country, interview the NRC, and observe the administration of the PIRLS achievement test in a sample of 15 of the PIRLS schools.

Additionally, three software packages and their corresponding manuals were supplied by the IEA Data Processing Center to assist NRCs in the main survey. These were:

1. The Within-School Sampling Software (W3S), a computer program designed to help NRCs select the within-school sample, prepare the survey tracking forms, assign test booklets to students, and print labels for test booklets and questionnaires (IEA, 2001a).

2. The DataEntryManager (WinDEM), a computer program for data entry and verification (IEA, 2001b).

3. The Linkcheck program (LINKPIRL), a computer program for testing and verifying the PIRLS data files.

The staff of the IEA Data Processing Center conducted hands-on training sessions in the installation and use of these software packages.

6.4 Within-School Sampling Procedures

The goal of the PIRLS sampling procedures was to select a nationally representative sample of students in each country. Sampling intact classrooms within randomly selected schools offered the simplest solution from an operational perspective, while optimizing the information gathered about the students and their teachers. Although this was the standard procedure, it could only be implemented where classes in a school constituted an exhaustive and mutually exclusive partition of the students in the grade. In order for a random sample of classes to result in a representative sample of students, every student in the target grade in each country had to belong to one (and only one) of the classes in the school.

A key step in the PIRLS within-school sample selection was the correct identification of classes and teachers. Before classes could be sampled, all eligible classes within the sampled school had to be identified. To that end, the NRC asked the School Coordinator to list (on a Class Listing Form) all classes in the target grade, along with the names of the teacher or teachers responsible for teaching reading to the students in those classes. From the list, the NRC then prepared the Class Sampling Form and applied a prescribed sampling algorithm to select at least one class at random. Within each school, a class identification number was assigned to each class in the target grades listed on the Class Sampling Form. The six-digit class ID consisted of the four-digit school ID plus the two-digit identification number for the class within the school. All students in this class were then selected to participate in the testing.

6.4.1 Survey Tracking Forms

PIRLS 2001 relied on a set of tracking forms to implement and record the sampling of classes, teachers, and students. It was essential that these were used and completed accurately, since they indicated the particular test booklet that each student was to receive. They were also used to record participation in each school. In addition to facilitating the data collection, the tracking forms provided essential information for computing sampling weights, and in evaluating the quality of the sampling procedures. Although the tracking forms could be produced manually, NRCs were urged to use the W3S sampling software, which automated many of the sampling and tracking procedures. Once completed, all tracking forms were retained for review by staff at the ISC.

For each sampled class, the NRC asked the School Coordinator to provide a list of students, giving their names, dates of birth, and genders. The School Coordinator also confirmed the name(s) of the teacher(s) of the sampled class. Once the NRC received the list and verified that the requested information had been supplied, he or she transcribed it onto a Student Tracking Form. Each student listed was assigned an eight-digit student identification number consisting of the six-digit class ID plus a two-digit number corresponding to the student's sequential position on the Student Tracking Form. Three extra student records were created at the end of the list to provide for additional students, or for assigning replacement materials in the event that a student received a damaged test booklet or questionnaire. These three extra records were also assigned a student ID.

The NRC produced a Teacher Tracking Form listing each teacher of the students in the sampled class. Each teacher of the selected class was assigned a teacher ID that consisted of the four-digit school ID, followed by a two-digit number for the teacher within the school, and a two-digit running number that sequentially numbered the entries on the Teacher Tracking Form. The two-digit running number was referred to as the Teacher Link Number, which identified each unique occurrence of a teacher in the Teacher Tracking Form.

During the test administration, the Test Administrator and School Coordinator used the tracking forms to record student, parent, and teacher participation, then returned them to the NRC after the test – together with the completed test booklets and questionnaires.

6.4.2 Excluding Students from Testing

Although all students enrolled in the target grade were part of the target population and were eligible to be selected for testing, PIRLS recognized that some students in some schools would be unable to take part in the 2001 assessment because of a physical or mental disability. Accordingly, the sampling procedures provided for the exclusion of students with specified disabilities (see Chapter 5). Countries were required to track and account for all excluded students, and were cautioned that excluding an excessive proportion would lead to their results being annotated in international reports. The conditions under which students could be excluded were carefully delineated, because the definition of "disabled" students varied considerably from country to country.

6.4.3 Assigning Instruments to Students and Teachers

The PIRLS reading assessment was packaged into nine student booklets and one magazine-like booklet known as the PIRLS Reader, which came with its own answer booklet (see Chapter 2). Each student was asked to complete just one of the 10 assessment booklets. Students recorded their answers directly in the booklet, except in the case of the Reader, where the accompanying answer booklet was used. All students completed the same student questionnaire. The assessment booklets were numbered sequentially 1 through 9, with R for the Reader answer booklet. Booklets were assigned to students following a systematic procedure that ensured an

Exhibit 6.1: Booklet Rotation Positions

Rotation Position	Booklet Number
1	Booklet 1
2	Booklet 2
3	Booklet 3
4	Booklet R and Reader
5	Booklet 4
6	Booklet 5
7	Booklet 6
8	Booklet R and Reader
9	Booklet 7
10	Booklet 8
11	Booklet 9
12	Booklet R and Reader

even distribution of booklets throughout each class. Each student's booklet assignment was recorded in advance on the Student Tracking Form, and the Test Administrator was expected to ensure that the correct booklet was given to each student. To facilitate proper booklet distribution, each booklet was individually labeled. Exhibit 6.1 provides the rotation scheme whereby booklets were assigned to students. Within each class, a number from one to twelve was randomly selected and used as a starting point to cycle through the list. For example, if 10 was chosen as the starting number in a particular class, the first student in the tracking form was assigned Booklet 8, the second Booklet 9, the third Booklet R, the fourth Booklet 1, and so on until all students in the class were assigned a booklet.[1] It was critical that the test booklets be assigned to the students before the testing day to ensure their correct distribution.

A student questionnaire was prepared for each entry in the Student Tracking Form, including the three extra entries. The student IDs and names were identified on the student questionnaires so that students' responses could be linked with their assessment booklets.

The teachers who taught the selected students were each given a teacher questionnaire. This questionnaire focused on the teacher's instructional practices as they applied to the target classes in the grade tested for PIRLS. The Teacher Tracking Form indicated the target grade and class with respect to which the teacher should have responded.

[1] To link the assessment passages together in the assessment, each passage appeared in three of Booklets 1 through 9. However, the passages in the Reader appear only in the Reader. To ensure that the same number of students respond to the Reader passages as to the others, the reader was assigned at three times the rate of the other booklets, as shown in Exhibit 6.1. On average, each assessment passage was seen by one quarter of the students in the assessment.

6.5 Packaging and Sending Materials to Schools

The NRC prepared three packages for each sampled class. One package contained the test booklets for all students listed on the Student Tracking Form. The second and third packages contained the Student Questionnaires and Learning to Read Surveys, respectively. For each participating school, the packages for each sampled class were bundled together with the Teacher Tracking Form, the Teacher Questionnaire, the School Questionnaire, and any materials prepared for briefing school coordinators and test administrators, and were sent to the School Coordinator. Labels and prepaid envelopes addressed to the NRC were included – to facilitate the return of the testing materials.

6.6 Within-School Assessment Activities

The School Coordinator in each school was responsible for organizing the administration of the PIRLS 2001 test. The coordinator could be the principal, the principal's designee, or an outsider appointed by the NRC with the approval of the principal. The NRC was responsible for ensuring that the School Coordinator was familiar with his or her responsibilities.

The tasks of the School Coordinator were detailed in the *School Coordinator Manual*. Before the test administration, the School Coordinator, working with the NRC, had to arrange a testing date, select a Test Administrator to conduct the testing sessions, and ensure that the Test Administrator was fully acquainted with the assessment procedures. In some countries, the School Coordinator assumed the roles and responsibilities of the Test Administrator as well as those of the School Coordinator.

6.6.1 Arranging the Testing Sessions

In preparation for the testing day, the School Coordinator worked with the school principal, Test Administrator, and the teacher, to plan the testing sessions by arranging rooms, classes, and materials. In countries where obtaining parental permission for testing was required, the School Coordinator ensured that permission forms were signed and returned in time. Once the testing materials arrived from the national center, the School Coordinator checked that they were for the appropriate students and teachers, that there were enough copies, and that the materials would be kept in a secure place until the testing day.

6.6.2 Distributing Materials

The School Coordinator distributed Teacher Questionnaires to the teachers listed in the Teacher Tracking Form, and a School Questionnaire to the school principal, and ensured that they were completed and returned. Teacher participation was recorded on the Teacher Tracking Form when the questionnaires were returned. In some countries, it was also the responsibility of the School Coordinator to collect the completed Learning to Read Surveys from the schools, record parent participation on the Student Tracking Forms, and return the questionnaires to the NRC.

6.6.3 Test Administration

The Test Administrator was responsible for administering the PIRLS test and student questionnaire. Specific responsibilities of the Test Administrator were described in the *Test Administrator Manual*. The Test Administrator distributed the test booklets and questionnaires according to the assignment documented on the Student Tracking Form, ensuring that each student received the correct testing materials. The Test Administrator conducted the testing sessions in accordance with a script provided in the *Test Administrator Manual*, and recorded the timing of the testing sessions on the Test Administration Form. After the testing session, the Test Administrator recorded student participation on the Student Tracking Form, and returned the testing materials to the School Coordinator.

6.6.4 Timing of the Testing Sessions

Testing was conducted in two consecutive sessions: one for administering the PIRLS achievement test booklets, and one for administering the student questionnaire. The first session was conducted in two parts, one for each part of the test booklet. During the field test, 30 minutes were provided for students to answer each part of the booklet. At the end of the 30 minutes, up to 10 extra minutes were allowed if less than 90 percent of the students had not completed answering the questions. The Test Administrator was required to document the timing on the Test Administration Form. As part of the analysis of the field test results, the time requirements were analyzed to determine what amount of time was sufficient across all countries, and, as a consequence, the timing for the main survey was adjusted. For the main survey, the allotted time was increased to 40 minutes – and, of course, no additional time was permitted beyond what was specified. The timing of the session was as follows:

- Approximately 10 minutes for preparation (i.e., reading instructions, distributing test booklets)
- 40 minutes for answering Part 1 of the test booklet
- Approximately 15 minutes for a break
- 5 minutes for preparing students for Part 2
- 40 minutes for answering Part 2 of the test booklet
- At least 20 minutes for the completion of the Student Questionnaire
- 5 minutes for distributing the Learning to Read Surveys.

The Student Questionnaire was to be administered on the same day as the achievement test, following the testing session, or, if this was not possible, on the following day.

6.6.5 Activities Following the Test Administration

After the test administration, the School Coordinator was responsible for calculating the student response rate in each class and, if the participation rate in any class in the school was below 90 percent, for arranging for a makeup session. The School Coordinator then returned to the NRC all testing materials, including the completed Student Tracking Form, the Test Administration Form, and any unused materials. Any relevant information about the test administration in the school was communicated to the NRC by means of a Test Administration Form that was completed by the Test Administrator and School Coordinator.

6.7 Monitoring Data Collection

The ISC implemented an international quality control program whereby international quality control monitors visited a sample of 15 schools in each country and observed the test administration. In addition, NRCs were also expected to organize an independent national quality control program based upon the international model. The latter program required Quality Control Observers to document data collection activities in their country. The Quality Control Observers visited a random sample of 10 percent of the schools (additional to those visited by the international Quality Control Monitors) and monitored the testing sessions, recording their observations for later analysis.

To assist NRCs in conducting their national quality control program, the International Study Center prepared the *Manual for National Quality Control Observers*, which contained information about PIRLS 2001, detailing the roles and responsibilities of the National Quality Control Observers.

6.8 Data Preparation

In the period immediately following the administration of the PIRLS tests, the main tasks for the NRC included retrieving the materials from the schools and preparing the constructed-response items for scoring. This involved recruiting and training scorers; scoring the constructed-response items, including independent double-scoring of the reliability sample; entering the data from the achievement tests and background questionnaires; submitting the data files and materials to the IEA Data Processing Center; and preparing a report on survey activities.

When the testing materials were received back from the schools, NRCs were to do the following:

- Check that all survey tracking forms were returned from the schools

- Check that the appropriate testing materials were received for every student listed on the Student Tracking Form

- Verify that all identification codes were correctly recorded on all of the test booklets

- Check that the participation status recorded on the tracking forms matched the information on the test instruments

- Contact schools that did not return the testing materials, or for which forms were missing, incomplete, or inconsistent.

NRCs then organized the test booklets and questionnaires for scoring and data entry. Procedures were provided to minimize the burden of sorting and handling the booklets, ensure reliability in the constructed-response coding, and document the reliability of the coding.

6.9 Scoring the Constructed-Response Items

Reliable application of the scoring guides to the constructed-response questions and empirical documentation of the reliability of that process were essential to the success of PIRLS 2001. The *PIRLS Survey Operations Manual* contained information about arranging for staff and facilities for the constructed-response scoring effort required for the main survey. The manual outlined how to select and train the scorers, and specified the procedures for scoring the constructed-response items and double-scoring a random sample of at least 200 responses per item, to document scoring reliability.

In selecting those who were to do the scoring, NRCs took care to arrange for persons who were conscientious and attentive to detail, knowledgeable in reading, and willing to apply the scoring guides as stated – even if they disagreed with a particular definition or category. Good candidates for scoring included teachers, retired teachers, college or graduate students, and staff members from educational agencies, ministries, or research centers.

6.9.1 Preparing Materials to Train the Scorers

The success of assessments containing constructed-response questions depends upon the reliability of scoring student responses. In PIRLS 2001, reliability was assured through the provision of scoring guides (manuals), extensive training in their use, and monitoring of the quality of the work. In addition, PIRLS 2001 provided training packets for training in selected questions, along with practice papers to help scorers achieve a consistent level of scoring.

Each scorer received a copy of the *PIRLS Scoring Guides for Constructed-Response Items*. This document explained the PIRLS scoring system, which was designed to produce a rich and varied profile of the range of students' competencies in reading literacy. A description of the development and content of the scoring guides is provided in Chapter 2.

At international scoring training meetings for both the field test and the main survey, NRCs received training packets containing "anchor papers" (example student papers) and practice papers to help them achieve accuracy and consistency in scoring. About 10 to 15 responses were sufficient for most items, but the complexity of some scoring guides made additional practice papers necessary.

6.9.2 Documenting the Reliability of the Constructed-Response Scoring

In order to demonstrate the quality of the PIRLS 2001 scoring, it was important to document the agreement between scorers. To establish the scoring reliability, NRCs were required to have a random sample of at least 200 responses for each constructed-response item independently scored by two scorers. This number is equal to approximately 25 percent of the responses based on the typical sample size. With the exception of the test booklet that accompanied the Reader, each item appeared in three booklets, meaning that the scorers needed to score a sample of 67 of each booklet. The items in the booklet accompanying the Reader do not appear in any other booklet; the scorers needed to score a sample of 200 of these booklets.

The *Survey Operations Manual* provides a procedure for interleaving the double scoring of the sample of reliability booklets with the regular booklet scoring, so that the reliability sample is scored in the same way and at the same time as the other booklets.

The activity described above provides evidence of the extent of agreement among scorers within each country, but does not address the question of scoring consistency across countries. Since PIRLS is administered in each country's own language, and since scorers generally do not know other countries' languages, cross-country scoring reliability is much more difficult to establish. As a partial solution to this problem, PIRLS took samples of student responses to a selection of constructed-response questions from each of the PIRLS countries that tested in English, and had this common set of English responses scored by two scorers in each country where scorers could operate through English. The common set consisted of 200 student responses to 25 questions from four of the PIRLS reading passages, two of which were literary and two informational. The cross-country reliability scoring was conducted in each country after the scoring of the national PIRLS data had been completed.

6.10 Trends in IEA's Reading Literacy Study

The Trends in IEA's Reading Literacy Study was an option for countries that participated in the 1991 IEA Reading Literacy Study, allowing them to administer the 1991 test booklets and questionnaires again in 2001 – to compare reading achievement over time. A list of countries participating in the Trends in IEA's Reading Literacy Study is provided in Chapter 5.

Documentation for implementing the trend study was incorporated in the manuals for PIRLS 2001. The *Survey Operations Manual* provided instructions for preparing tracking forms, packaging materials for the schools, preparing, administering, and returning the test booklets and questionnaires, and for data entry. The *School Coordinator Manual* and *Test Administrator Manual* included sections specific to the 10-Year Trend Study wherever necessary.

The *School Sampling Guide for the 10-Year Trend Study* (PIRLS, 2000) documented procedures for sampling schools for the Trends in IEA's Reading Literacy Study. Essentially, the procedure was to sample an additional class from the target grade in half of the schools in the PIRLS sample. The survey tracking forms used in PIRLS were also used for tracking trend study schools, classes, and students.

Countries were expected to use test booklets identical to those used in 1991. Schools in which classes were selected for the trend study received two packages from the NRC. One contained the test booklets with the Student Tracking Forms, and the other the Student Questionnaires. Teachers of the classes selected for the trend study were not given a Teacher Questionnaire, nor was the School Questionnaire administered.

The data collection was conducted under the same conditions as in 1991. There were three data-collection sessions, with the same time limitations as in 1991. There was a short break between sessions:

- 1 minute and 30 seconds for answering questions in the Word Test (word recognition), followed by 35 minutes for answering questions in the first part of the Reading Test

- 40 minutes for answering the questions in the second part of the Reading Test

- At least 25 minutes for the completion of the Student Questionnaire.

The Test Administrator and School Coordinator followed the PIRLS procedures for collecting the test instruments, checking for proper documentation on the survey tracking forms, and for calculating student response rates. Once any necessary makeup sessions had been held, the School Coordinator returned the materials to the NRC for data preparation. Since the trend study test booklets did not include constructed-response items for scoring, they were sorted separately for data entry – along with the trend study Student Questionnaires.

6.11 Data Entry

The IEA Data Processing Center provided each NRC with a copy of WinDEM (an integrated computer program for data entry and data verification) designed specifically for use with IEA studies. This program allowed data entry directly from the tracking forms and test instruments, and provided a convenient checking and editing mechanism. WinDEM also offered interactive error detection, error reporting, and quality control procedures. Detailed information and operational instructions were provided in the manual for the WinDEM software. WinDEM for PIRLS incorporates – for each PIRLS instrument – the international codebook, which describes the format and data characteristics of each variable in the instrument. Correct use of the WinDEM software ensured that the data files were produced according to the PIRLS 2001 standards for data entry.

Although WinDEM was strongly recommended for all data entry tasks, NRCs sometimes chose to use their own procedures and computer programs, which was acceptable – provided all data files conformed to the specifications of the international codebooks. NRCs who did not to use WinDEM were responsible for ensuring that all data files were delivered to the Data Processing Center in the international format.

During the PIRLS 2001 main survey operations, data were gathered from several instruments, including: student assessment booklets, questionnaires from students, teachers, parents, and principals, as well as from a range of tracking forms. Before beginning data entry, the NRC needed to ensure that the instruments and corresponding tracking forms had been completed and sorted correctly. Data entry involved the following files:

- The school background file – information from the School Questionnaire and the School Tracking Form

- The teacher background file – information from the Teacher Questionnaire and the Teacher Tracking Form

- The student background file – data from the Student Questionnaire, the Test Administration Form, and Student and Teacher Tracking Forms

- The student achievement file – data from the student assessment booklets and the Data Entry Batch Header

- The home background file – data from the Learning to Read Survey

- The constructed-response scoring reliability file – data from the scoring sheets for the constructed-response items that were double-scored

- The Trends in IEA's Reading Literacy Study file consisted of the data from the Student Questionnaire, the achievement test booklet, the Test Administration Form, and the Student Tracking Form.

Quality control throughout the data entry process is essential in maintaining accurate data. Countries were responsible for performing periodic reliability checks on the data entry, and for applying a series of data verification options provided as part of WinDEM. NRCs not using WinDEM for data entry still had to apply the WinDEM data verification checks to their data before sending their files to the IEA Data Processing Center. The WinDEM data-checking facility could identify a range of problems that could then be fixed before submission to the Data Processing Center. Specifically, WinDEM checks for the following:

- Duplicate identification codes

- Inconsistencies in the hierarchical identification system

- Out-of-range values

- Mismatches between different student-level files.

In addition to the checks performed by WinDEM, the Linkcheck program detected mismatches between different files, as well as inconsistencies in important identification variables. Students whose IDs were not numbered sequentially are reported, and classrooms with unusually small numbers of students were detected. The booklet assignment was checked, and information about the scorers of the achievement test booklets and the reliability booklets were compared.

Data files were regarded as having been satisfactorily checked only if the reports generated by the WinDEM program and the Linkcheck program indicated no errors. When all data files had passed these quality control checks, they were dispatched to the Data Processing Center for further checking and processing.

6.12 Survey Activities Report

NRCs were requested to maintain a record of their experiences during the PIRLS 2001 data collection, and to send a report to the ISC when data-collection activities were completed. This should describe any problems or unusual occurrences in selecting the sample or securing school participation, translating or preparing the data-collection instruments, administering the test and questionnaires in the schools, scoring the constructed-response items, or creating and checking the data files.

References

IEA. (2001a). *W3S: Within-School Sampling Software*. Hamburg: IEA Data Processing Center.

IEA. (2001b). *WinDem: Software for Data Entry and Verification*. Hamburg: IEA Data Processing Center.

Progress in International Reading Literacy Study (PIRLS). (1999). *School Sampling Manual* (PIRLS Ref. No. 99-0019). Prepared by Pierre Foy and Marc Joncas, Statistics Canada. Chestnut Hill, MA: Boston College.

Progress in International Reading Literacy Study (PIRLS). (2000). *School Sampling Guide for the 10-Year Trend Study – Version 2* (PIRLS Ref. No. 00-0001). Prepared by Pierre Foy & Marc Joncas, Statistics Canada. Chestnut Hill, MA: Boston College.

Progress in International Reading Literacy Study (PIRLS). (2001a). *Manual for Entering the PIRLS Data* (PIRLS Ref. No. 01-0004). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

Progress in International Reading Literacy Study (PIRLS). (2001b). *Manual for International Quality Control Monitors* (PIRLS Ref. No. 01-0005). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

Progress in International Reading Literacy Study (PIRLS). (2001c). *Manual for National Quality Control Observers* (PIRLS Ref. No. 01-0006). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

Progress in International Reading Literacy Study (PIRLS). (2001d). *School Coordinator Manual* (PIRLS Ref. No. 01-0003). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

Progress in International Reading Literacy Study (PIRLS). (2001e). *Scoring Guides for the Constructed-Response Items* (PIRLS Ref. No. 01-0007). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

Progress in International Reading Literacy Study (PIRLS). (2001f). *Survey Operations Manual* (PIRLS Ref. No. 01-0001). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

Progress in International Reading Literacy Study (PIRLS). (2001g). *Test Administrator Manual* (PIRLS Ref. No. 01-0002). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

7

Quality Control in the PIRLS Data Collection

Eugenio J. Gonzalez

Ann M. Kennedy

7.1 Overview

The International Study Center (ISC) conducted an ambitious program of site visits to document the quality of the PIRLS 2001 data collection. Together with the IEA Secretariat and the national centers, the ISC identified and appointed one international Quality Control Monitor (QCM) in each country to observe data collection procedures at both national and school levels.

Quality Control Monitors had two major responsibilities: to interview the National Research Coordinator (NRC) about the survey operations and activities, and to arrange visits to a random sample of 15 schools in their country during the test administration. An Interview with the NRC Form was used to record the NRC's responses during the interview. For each testing session observed, QCMs completed a Classroom Observation Record.

More than 30 monitors attended a two-day training session conducted by the staff of the ISC, where they were introduced to the PIRLS 2001 survey operations procedures and instructed on how to conduct their site visit observations and interviews. At the training session, QCMs received a copy of the *Manual for International Quality Control Monitors* (PIRLS, 2000), which explained their duties in detail, and copies of the PIRLS survey operations manual and manuals for school coordinators and test administrators.

The QCMs who attended the training session were asked to recruit other QCMs within their country when necessary, in order to allow for efficiency in the coverage of the territory and testing timetable. A total of 71 QCMs were trained across the 33 countries where the

Exhibit 7.1: Preliminary Activities of the Test Administrator

Question	Yes	No	Not Answered
Had the Test Administrator verified adequate supplies of the test booklets?	454*	21**	0
Had the Test Administrator familiarized himself or herself with the test administration script prior to the testing?	449*	23**	3
Did the student identification information on the test booklets and student questionnaires correspond with the Student Tracking Form?	465	8	2
Was there adequate seating space for the students to work without distractions?	462	12	1
Was there adequate room for the Test Administrator to move about during the testing to ensure that student were following directions correctly?	470	4	1
Did the Test Administrator have a stop watch or timer for accurately timing the testing session?	451	21	3

* Represents the number of respondents answering either Definitely Yes or Probably Yes

** Represents the number of respondents answering either Definitely No or Probably No

international quality control was conducted.[1] All together, these monitors observed 475 testing sessions and conducted interviews with the national research coordinator in each of the 33 PIRLS countries.

7.2 Observing the PIRLS Test Administration

When visiting the school, the QCM was to complete a Classroom Observation Record Form. This form was organized into four sections to facilitate the accurate recording of the test administration's major activities. The four sections are:

- Preliminary activities of the Test Administrator
- Test session activities
- General impressions
- Interview with the School Coordinator.

7.2.1 Preliminary Activities of the Test Administrator

Section A of the Classroom Observation Record addressed the extent to which the Test Administrator had prepared for the testing session. Monitors were asked to note the following activities of the Test Administrator: checking the testing materials, reading the administration script, organizing space for the session, and arranging for the necessary equipment (e.g., pencils, a watch for timing the testing session).

Exhibit 7.1 summarizes the results for Section A. In almost all testing sessions, test administrators observed the proper preparatory procedures. When deviations occurred,

[1] Operational constraints did not permit QCM visits to be conducted in Argentina or Iceland.

the QCMs provided reasonable explanations for the discrepancies. For example, QCMs noted that the main reason for students receiving booklets with student identifications that did not correspond to the Student Tracking Form was because new students did not appear on the list because the tracking forms had been created before they were enrolled. In the few cases where there reportedly was not enough room for students, QCMs reported unavoidable circumstances (e.g., the test was administered in a small classroom, the desks were too narrow, students had to sit three to a table).

The absence of a stopwatch was considered a negligible limitation. Test Administrators who did not have a stopwatch had a wristwatch available to monitor the time remaining on the test sessions. In general, QCMs observed no procedural deviations in test preparations severe enough to jeopardize the integrity of the test administration.

7.2.2 Test Session Activities

Section B of the Classroom Observation Record addressed the activities that took place during the actual testing session. These activities included following the Test Administrator script, distributing and collecting test booklets, and making announcements during the testing sessions.

The achievement test was administered in two parts with a short break in-between. Activities during the first part of the testing session are presented in Exhibit 7.2. In at least 80 percent of the schools visited, the Test Administrators followed their script exactly when preparing the students, distributing the test materials, and reading the directions and examples. Of the changes that were made, the majority were considered minor. Changes made to the script were most frequently acceptable additions – rather than revisions or deletions.

In about 15 percent of the sessions visited, the total testing time for Part 1 was not equal to the time allowed. However, in most of these sessions, this was because all students had completed Part 1 before the allotted time had elapsed, and so the test administrator reasonably went on with the next part of the session according to the prescribed procedures. The average testing session for Part 1 was approximately 36 minutes in duration instead of the 40 minutes allocated. Students were instructed to close their test booklets and leave them on their desk during the break. In most sessions, the room was then either secured or supervised during the break. In no instance did a QCM report a breach of security during the break.

In more than 80 percent of the testing sessions visited, the total time for the break between parts was equal to or less than 15 minutes. Of those sessions with breaks longer than 15 minutes, most reportedly took up to 20 minutes for the break. The total break time across all countries ranged between 1 and 40 minutes.

Exhibit 7.2: Testing Session Part 1

Question	Yes	No	Not Answered
Did the test administrator follow the test administrator's script exactly in each of the following tasks?			
Preparing the students	404	63 (Minor changes) 6 (Major)	2
Distributing the materials	449	23 (Minor) 1 (Major)	2
Reading the directions	381	88 (Minor) 5 (Major)	1
Reading the examples	410	59 (Minor) 5 (Major)	1
If the Test Administrator made changes to the script, how would you describe them?			
Additions	107	136	232
Revisions	57	161	257
Deletions	30	177	268
Did the Test Administrator distribute test booklets one at a time to each student?	468	7	0
Did the Test Administrator distribute the test booklets according to the booklet assignments on the Student Tracking Form?	463	12	0
Did the Test Administrator record attendance correctly on the Student Tracking Form?	458	11	6
Did the total testing time for Part 1 equal the time allowed?	402	71	2
Did the Test Administrator announce "you have 5 minutes left" prior to the end of Part 1?	419	55	1
Were there any other time remaining announcements made during Part 1?	57	413	5
At the end of Part 1, did the Test Administrator make sure all students had closed their booklets?	460	10	5
Was the total time for the break equal to or less than 15 minutes?	391	71	13
Were the booklets left unattended or unsecured during the break?	21	443	11

Exhibit 7.3: Testing Session Part 2

Question	Yes	No	Not Answered
Was the time spent to restart the testing for Part 2 equal to or less than 5 minutes?	445	18	12
Was the total time for testing in Part 2 correct as indicated in the script?	355	107	13
Did the Test Administrator announce "you have 5 minutes left" prior to the end of Part 2?	359	100	16
Were there any other time remaining announcements made during Part 2?	35	420	20
At the end of Part 2, did the Test Administrator collect the test books one at a time from each student?	425	41	9
When the Test Administrator read the script to end the testing for Part 2, did he/she announce a break to be followed by the Student Questionnaire?	374	76	25
Did the Test Administrator accurately read the script to end the testing and signal a break?	321 (No changes)	91 (Minor) 23 (Major)	40
If there were changes, how would you describe them?			
Additions	46	131	298
Some minor changes	57	130	288
Omissions	38	137	300
Did the Test Administrator distribute the Student Questionnaires and give directions as specified in the script?	407	17	51
Did the students ask for additional time to complete the questionnaire?	150	252	73
Did the Test Administrator distribute a Learning to Read Survey to each student who participated in the testing?	321	115	39
At the end of the session, prior to dismissing the students, did the Test Administrator thank the students for participating in the study?	391	50	34

Exhibit 7.3 summarizes the QCMs' observations during the second part of the testing session. In over 90 percent of the sessions, the Test Administrator adhered to the prescribed time limits in the directions; the time spent to restart the testing session was 5 minutes or less. The rest of the sessions took up to 10 minutes to restart the testing session. Similar to the timing of Part 1, the average testing session in Part 2 was shorter than the 40 minutes allotted because students had finished the achievement test early.

Exhibit 7.4: Testing Session Activities

Question	Very Well	Well	Fairly Well	Not well at all
When the Test Administrator ended Part 1, how well did the student comply with the instruction to stop work (close their booklets and put their pencils down)?	418	50	6	1
When the Test Administrator ended Part 2, how well did the student comply with the instruction to stop work (close their booklets and put their pencils down)?	414	46	5	10

Question	Exactly the same	Longer	Shorter	Not Answered
How does the total time allocated for the administration of the Student Questionnaire compare to the time specified in the script?	158	225	25	67

Question	Very orderly	Somewhat orderly	Not orderly at all	Not Answered
How orderly was the dismissal of the student?	350	88	8	29

About 65 percent of the Test Administrators kept to the testing script for signaling a break before administering the student questionnaire. Of those who did make changes, most made acceptable additions or other minor changes, such as paraphrasing the directions. More than 80 percent of the students requested additional time to complete the student questionnaire, which, in most cases, was granted.

Results of the remaining questions that focused on the test session activities are provided in Exhibit 7.4. These questions dealt with topics such as student compliance with instructions, and the alignment of the scripted instructions with their implementation.

Exhibit 7.4 shows that in almost all of the sessions, the students complied well or very well with the instructions to stop testing. In more than half the sessions, however, the amount of time needed to complete the student questionnaire was longer than the time specified in the script. Usually this was because the Test Administrators read each question aloud to the students, a practice that was encouraged to help students complete the questionnaire accurately.

Exhibit 7.5: Summary Observations of the QCM

Question	Yes	No	Not Answered
During the testing sessions did the Test Administrator walk around the room to be sure students were working on the correct section of the test and/or behaving properly?	462	11	2
Did the Test Administrator address students' questions appropriately?	473	1	1
Did you see any evidence of students attempting to cheat on the tests (e.g., by copying from a neighbor)?	21	454	0
Were any defective test books detected and replaced *before* the testing began?	27	445	3
Were any defective test books detected and replaced *after* the testing began?	14	452	9
If any defective test books were replaced, did the Test Administrator replace them appropriately?	32	11	432
Did any students refuse to take the test either prior to the testing or during the testing?	11	462	2
If a student refused, did the Test Administrator accurately follow the instructions for excusing the student (collect the test book and record the incident on the Student Tracking Form)?	23	4	448
Did any students leave the room for an "emergency" during the testing?	58	411	6
If a student left the room for an emergency during the testing, did the Test Administrator address the situation appropriately (collect the test booklet, and if re-admitted, return the test booklet)?	61	11	403

7.2.3 General Impressions

Section C of the Classroom Observation Record asked QCMs to reflect on their observations. The QCMs reported overall impressions of the test administration – including how well the Test Administrator monitored students' conduct, and any unusual circumstances that arose during the testing session (e.g., student refusal to participate, defective instrumentation, emergency situations, cheating).

The results presented in Exhibit 7.5 show that in almost all sessions, the testing took place without any problems. In the few sessions where problems arose due to defective instrumentation, the Test Administrator replaced the instruments appropriately.

In less than 5 percent of sessions, QCMs reported evidence of students attempting to cheat on the exam. However, when asked to explain the situation, QCMs generally indicated that students were merely looking around at their neighbors to see whether their test booklets were indeed different. Because the PIRLS test design involves 10 different booklets, students were unlikely to have the same booklet as their neighbors. Anyone who may have

Exhibit 7.6: Summary Observations of Student Behavior

Question	Extremely	Moderately	Somewhat	Hardly	Not answered
To what extent would you describe the students as orderly and cooperative?	333	131	10	1	0

	Definitely	Some effort	Hardly any effort	Not answered
If the students were not cooperative and orderly, did the Test Administrator make an effort to control the students and the situation?	129	24	0	322

	No, there were no late students	No, they were not admitted	Yes, but before testing began	Yes, after testing began	Not answered
Were any late students admitted to the testing room?	439	3	15	13	5

	Excellent	Very good	Good	Fair	Poor	Not answered
In general, how would you describe the overall quality of the testing session?	224	181	55	8	3	4

tried to copy a neighbor’s answers would have had to find a student with the same booklet around them, and this is very unlikely – given the test design and booklet rotation. The QCMs reported that on the rare occasions when they observed serious efforts to cheat, the Test Administrator intervened to prevent cheating.

Most of the 58 students who reportedly left the room for an “emergency” during the testing session had already completed the test. When students left the room for an emergency, Test Administrators handled the situation appropriately by ensuring the security of the test booklets until the students returned. Students were permitted to complete the test when they returned to the classroom.

Finally, Exhibit 7.6 indicates that in almost all of the testing sessions, QCMs found the behavior of students to be orderly and cooperative. The problem cited most often by QCMs as the reason for disorderly behavior was the noise level of those students who had completed the test well before the prescribed 40 minutes had passed. In the few cases where it was less than perfect, the Test Administrator was able to control the students and the situation. For the great majority of sessions, QCMs reported that the overall quality of the sessions was either excellent or very good.

7.2.4 Interview with the School Coordinator

The QCM recorded details of the interview with the School Coordinator in Section D of the Classroom Observation Record. The interview addressed the shipment of assessment materials, arrangements for the test administration, the responsiveness of the NRC to queries, the necessity for make-up sessions, and, as a validation of within-school sampling procedures, the organization of classes in the school.

PIRLS' administrative success, according to the school coordinators, is exemplified by the results presented in Exhibit 7.7. School Coordinators received the correct shipment of the test materials in at least 80 percent of all the testing sessions. School Coordinators reportedly not having received materials provided legitimate reasons (such as materials were brought by the Test Administrators as planned, etc.). In those cases where shipment errors occurred, they tended to be minor and were remedied prior to testing. More than 85 percent of School Coordinators reported that the NRCs were responsive to their questions or concerns.

More than half of the School Coordinators reported that they were able to collect the completed teacher questionnaires prior to student testing. Of those who did not, most reported that teachers completed their questionnaires during the testing sessions. Almost half of the School Coordinators indicated that the estimate of 30 minutes to complete the questionnaire was accurate; while about 35 percent reported that the questionnaires took longer, and about 15 percent that they took less time to complete.

In about 35 percent of the observed classes, School Coordinators indicated that students were given special instructions, motivational talks, or incentives prior to testing. The majority of students received motivational talks either by a school official, classroom teacher, or the PIRLS Test Administrator. Only a few classes received special instructions or practice, such as reading competitions or extra reading assignments prior to the testing session.

A tribute to the planning and implementation of PIRLS 2001 was the fact that about 90 percent of respondents said they would be willing to serve as a School Coordinator in future international assessments. Furthermore, the results shown in Exhibit 7.8 suggest that the majority of School Coordinators believed the testing session went very well, and that the school staff members had positive attitudes towards the PIRLS testing.

Exhibit 7.7: Results of the QCM Interviews with the School Coordinator

Question	Yes	No	Not Answered
Prior to the test day did you have time to check your shipment of materials from your PIRLS National Coordinator?	393	50	32
Did you receive the correct shipment of the following items?			
School Coordinator Manual	373	70	32
Test Administrator Manual	423	6	46
Student Tracking Forms	440	4	31
Test booklets	411	18	46
Student Questionnaires	417	12	46
Learning to Read Surveys	396	33	46
Teacher Questionnaires	442	2	31
School Questionnaire	444	1	30
Test Administration Form	424	4	47
Teacher Tracking Form	322	102	51
Envelopes or boxes addressed to the National Center for the purpose of returning the materials after the assessment	313	113	49
Was the National Coordinator responsive to your questions or concerns?	426	19	30
Were you able to collect completed Teacher Questionnaire(s) prior to the test administration?	282	174	19
Was the estimated time of 30 minutes to complete the Teacher Questionnaires a correct estimate?	230	166 (Took longer) 34 (Took less time)	45
Were you able to collect the completed School Questionnaire prior to the test administration?	275	181	19
Were you satisfied with the accommodations (testing room) you were able to arrange for the testing?	462	10	3

Exhibit 7.7: Results of the QCM Interviews with the School Coordinator (continued)

Question	Yes	No	Not Answered
Do you anticipate that makeup session will be required at your school?	56	411	8
If you anticipate makeup sessions, do you intend to conduct one?	75	71	329
Did the students receive any special instructions, a motivational talk, or incentives to prepare them for the assessment?	178	278	19
Is this a complete list of the classes in this grade in this school?	390	35	50
To the best of your knowledge, are there any students in this grade level who are not in any of these classes?	17	401	57
To the best of your knowledge, are there any students in this grade level in more than one of these classes?	6	409	60
If there were another international assessment, would you be willing to serve as a School Coordinator?	434	29	12

Exhibit 7.8: Overall Impressions from the QCM Interviews with the School Coordinator

Question	Very well, no problems	Satisfactorily, few problems	Unsatisfactorily, many problems	Not Answered
Overall, how would you say the session went?	385	81	6	3

	Positive	Neutral	Negative	Not Answered
Overall, how would you rate the attitude of the other school staff members towards the PIRLS testing?	345	112	16	2

	Worked well	Needs improvement	N/A
Overall, do you feel the PIRLS School Coordinator Manual worked well or does it need improvement?	342	24	79

7.3 Interview with the National Research Coordinator

In addition to observing testing sessions, QCMs conducted face-to-face interviews with the National Research Coordinator for their country. The QCM who attended the training session was responsible for conducting this interview, and for completing an Interview with the NRC Form.

The interview questions were designed to examine NRCs' experiences in preparing for, and conducting, the PIRLS data collection – with a focus on identifying and selecting samples, working with school coordinators, translating the instruments, assembling and printing the test materials, packing and shipping the test materials, scoring constructed-response questions, entering and verifying data, choosing quality assurance samples, and suggesting improvement in the process.

7.3.1 Sampling

Section A of the NRC interview form involved questions about the sampling process. Topics covered in this section included the extent to which the NRCs used the manuals and sampling software provided by the International Study Center, and the extent to which the process was difficult in terms of the complexity of the tasks.

Exhibit 7.9 shows that only one country did not use the sampling manuals provided, mainly because Statistics Canada performed the sampling for the country. Just over two-thirds of the NRCs used the within-school sampling software provided by the IEA DPC to select classes. In the cases where the sampling software was not used, the within-school sampling was done manually, or using other sampling software not provided by the ISC.

Some NRCs reported deviations from the sample design due to organizational constraints in their systems. A sampling expert was consulted in each case, to verify that the adopted design remained compatible with the PIRLS standards. Of those who found the sampling process very difficult, some NRCs cited the lack of personnel as a major obstacle. Despite any problems, all NRCs provided high-quality school and student samples for the data collection.

7.3.2 Working with School Coordinators

Questions in Section B of the NRC interview asked about cooperation with the School Coordinators, specifically about communication, shipment of materials, and training.

A summary of the responses to the questions in Section B is presented in Exhibit 7.10. At the time the interviews were conducted, nearly all NRCs had contacted the School Coordinators for their sample, and sent the appropriate materials on the testing procedures. Where this was not the case, it was often because a meeting had been scheduled but not yet held. About half of the NRCs planned to conduct formal training sessions for school coordinators prior to the test administration.

Exhibit 7.9: Results of the QCM Interviews with Their NRC – Sampling

Question	Yes	No	Not Answered
Were you able to select a sample of schools and students within schools using the manuals provided by the International Study Center?	31	1	1
Did you use the Within-School Sampling Software provided by the International Study Center to select classes or students?	22	11	0
Were there any conditions or organizational constraints that necessitated deviations from the basic PIRLS sampling design?	9	24	0

	Very difficult	Somewhat difficult	Not difficult at all	Not Answered
In terms of the complexity of the procedures and number of personnel needed, how would you describe the process of sample selection?	5	10	17	1

Exhibit 7.10: Results of the QCM Interviews with Their NRC – School Coordinator

Question	Yes	No	Not Answered
Have all the School Coordinators for your sample been contacted?	24	9	0
If all School Coordinators have been contacted, have you sent them materials about the testing procedures?	20	9	4
Did you or do you plan to have formal training sessions for the School Coordinators?	15	18	0

7.3.3 Translating the Instruments

Section C of the NRC interview dealt with the difficulty of translating and adapting the assessment instruments and manuals.

Exhibit 7.11 shows that most NRCs reported little difficulty in translating and adapting the test booklets and questionnaires, and even less difficulty in translating the Test Administrator and School Coordinator manuals.

NRCs generally used their own staff (or a combination of staff and outside experts) to translate the test booklets. The majority of NRCs reported that they already had submitted the achievement test booklets to the translation verification program at the ISC. Of those that did not, one country did not make adaptations to the international version, and the other two had submitted their test booklets and questionnaires for verification – but did not receive verifier's comments in time to make all recommended changes.

Exhibit 7.11: Results of the QCM Interviews with Their NRC – Translation

Question	Own Staff	Outside Experts	Combination	Not Answered
Did you use your own staff or outside experts to translate the test booklets for verification?	8	6	17	2

	Very difficult	Somewhat difficult	Not difficult at all	Not Answered
How difficult was it to translate and/or adapt the test booklets?	1	15	15	2
How difficult was it to adapt the questionnaires?	0	18	14	1
How difficult was it to adapt the Test Administrator Manual?	0	10	22	1
How difficult was it to adapt the School Coordinator Manual?	0	10	19	4

	Yes	No	Not Answered
Did you go through the process of submitting test booklets and receiving a translation verification report from the IEA?	29	3	1
Did you translate, or do you plan to translate, the Scoring Guides for Constructed-Response Items?	20	12	1

7.3.4 Assembling and Printing the Test Materials

Section D of the NRC survey addressed assembling and printing the test materials. Also, it included instructions for quality control issues related to checking the materials and securely storing them.

The results in Exhibit 7.12 show that NRCs were able to assemble the test booklets according to the instructions provided, and that almost all NRCs conducted the recommended quality control checks during the process. In the cases where the NRCs did not conduct quality assurance procedures during the printing process, it was because of a shortage of time.

Most countries elected to send their test booklets and questionnaires to an external printer, but printed their manuals in-house. All NRCs reported having followed procedures to protect the security of the tests during assembly and printing. In no instance was there a breach of security reported.

Exhibit 7.12: Interview with the NRC – Assembling and Printing Test Materials

Question	Yes	No	Not Answered
Were you able to assemble the test booklets according to the instructions provided by the International Study Center?	29	4	0
Did you conduct the quality assurance procedures for checking the test booklets during the printing process?	28	5	0
Were any errors detected during the printing process?	11	19	3
If errors were detected, what was the nature of the errors?			
Poor print quality	6	5	22
Pages missing	1	9	23
Page order	2	8	23
Upside down pages	1	9	23
Did you follow procedures to protect the security of the tests during the assembly and printing process?	31	1	1
Did you discover any potential breaches of security?	0	32	1

Question	In-House	External	Combination	Not Answered
Where did you print the test booklets?	6	21	6	0
Where did you print the questionnaires?	8	18	7	0
Where did you print the manuals?	22	7	3	1

7.3.5 Packing and Shipping the Testing Materials

Section E of the NRC interview addressed the extent to which NRCs detected errors in the testing materials as they were packed for shipping to School Coordinators. As shown in Exhibit 7.13, very few errors were found in any of the materials. Errors that were discovered before distribution were remedied.

In addition, almost half of the NRCs reported having established a procedure to confirm the schools' receipt of the testing materials, and for verification of their contents. In most countries, NRCs reported that the deadline for return of materials from the schools was within a day or two of testing. All NRCs reported that the deadline was within two weeks of testing.

Exhibit 7.13: Interview with the NRC – Packaging Test Materials

Question	No Errors, or not used	Errors found before distribution	Errors found after distribution	Not Answered
In packing the assessment materials for shipment to schools, did you detect any errors in any of the following items?				
Supply of test booklets	18	2	1	12
Supply of Student Questionnaires	18	2	1	12
Supply of Learning to Read Surveys	17	1	1	14
Student tracking Forms	21	0	0	12
Teacher tracking Forms	21	0	0	12
Test administrator Manual	21	0	0	12
School coordinator Manual	19	0	0	14
Supply of Teacher Questionnaires	20	1	0	12
School Questionnaire	21	0	0	12
Test book ID labels	19	1	1	12
Sequencing of books or questionnaires	19	2	0	12
Return labels	19	0	0	14
Self-addressed post-cards for test dates	19	0	0	14

Exhibit 7.14: Interview with the NRC – Scoring

Question	Yes	No	Not Answered
Have you selected your scorers for the constructed-response questions?	23	8	2
If you have selected them, have you trained the scorers?	10	16	7
Have you scheduled the scoring sessions for the constructed-response questions?	21	11	1
Do you understand the procedure for scoring the 25 percent reliability sample as explained in the survey operations manual?	30	3	0

7.3.6 Scoring Constructed-Response Questions

Section F of the NRC interview form focused on the NRC's preparation for scoring the constructed-response items. The scoring process was an ambitious effort, requiring the recruitment and training of scoring staff to score student responses – including double scoring 25 percent of the responses to verify reliability.

Exhibit 7.14 indicates that, at the time of the NRC interview, at least two-thirds of the NRCs had selected their scoring staff, and about half of these had already begun the training process. Each country planned to use about 15 scorers, on average. Almost all NRCs reported that they understood the procedures for scoring the 25 percent reliability sample as explained in the *Survey Operations Manual*.

7.3.7 Data Entry and Verification

Section G of the NRC interview addressed preparations for data entry and verification. As shown in Exhibit 7.15, at the time of the interviews about two-thirds of the NRCs had selected their data entry staff and more than half of those selected had taken part in training sessions.

Exhibit 7.15: Interview with the NRC – Data Entry and Verification

Question	Yes	No	Not Answered
Have you selected the data entry staff?	23	9	1
If yes, have you conducted training sessions for the data entry staff?	15	9	9
Do you plan to key enter a percentage of test booklets twice as a verification procedure?	22	10	1
Have you established a secure storage area for the returned tests after coding and until the original documents can be discarded?	33	0	0

About two-thirds of the NRCs reported that they planned to enter the data from a percentage of booklets twice – as a verification procedure. The estimated proportion of booklets to be entered twice ranged from 5 percent to 25 percent, with one country reporting that it planned to re-enter 100 percent of the data.

7.3.8 Quality Assurance Sample

As part of their national quality assurance activities, NRCs were required to send National Quality Control Observers to a 10 percent sample of the PIRLS schools to observe the test administration and document compliance with prescribed procedures. These site visits were over and above those visits to 15 schools conducted by the International Quality Control Monitors.

At the time of the NRC interviews, two-thirds of the NRCs had selected their 10 percent quality assurance sample for site visits. Three NRCs reported that an external agency would conduct the observations, eleven reported that a member of their staff would do so, and eight reported that a combination of staff and external agency people would conduct the observations. Five NRCs reported that teachers would be recruited to conduct the on-site observations.

7.3.9 The Survey Activities Report

The final section of the NRC interview asked the NRC for comments on any aspects of the study they felt might improve the assessment process. A major concern expressed by many NRCs was a time constraint for accomplishing all that was required to keep up with the demanding PIRLS schedule – particularly the translation and instrument preparation aspects. Some NRCs indicated they did not have ample staff.

References

Progress in International Reading Literacy Study (PIRLS). (2001) *Manual for International Quality Control Monitors* (PIRLS Ref. No. 01-0005). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

8

Creating and Checking the PIRLS Database

Ursula Itzlinger

Knut Schwippert

8.1 Overview

Creating the PIRLS 2001 database, and ensuring its integrity, was a complex endeavor – requiring close coordination and cooperation among the staff at the IEA Data Processing Center (DPC), the PIRLS International Study Center at Boston College (ISC), Statistics Canada, and the national research centers of the participating countries. The overriding concerns were: to ensure that all information in the database conformed to the internationally defined data structure; that national adaptations to questionnaires were reflected appropriately in the codebooks and documentation; and that all variables used for international comparisons were indeed comparable across countries. Quality control measures were applied throughout the process to assure the quality and accuracy of the PIRLS data.

This chapter describes the data entry and verification tasks undertaken by the National Research Coordinators and data entry managers of participating countries, the data checking and database creation procedures implemented by the IEA Data Processing Center, and the steps taken at all institutions to confirm the integrity of the international database.

Database construction began with each national research center entering the data collected in the PIRLS 2001 survey into data files following the standard international format. Before sending the files to the IEA DPC, national center staff applied a system of checks to verify the structure of the data files. Checking and editing the national data sets was a matter of cooperation between the national centers, the ISC, Statistics Canada, and the DPC team.

The IEA DPC was responsible for checking the data files and applying standard cleaning rules to verify the accuracy and consistency of the data. Any queries were addressed to the national research centers, and modifications were made to the data files as necessary. The IEA DPC produced summary statistics for all variables in the background and achievement data for the national research centers, which were then reviewed by the ISC for any apparent oversights in recoding or valid range issues.

After all modifications had been applied, all data were processed and checked again. This process of editing the data, checking the reports, and implementing corrections was repeated as many times as necessary until all data were consistent and comparable within and between countries.

In preparation for creating the international database, the IEA DPC provided data almanacs containing international univariate statistics and item statistics to the national centers so that they could examine their data from an international perspective. This was one of the most important checks (in terms of international comparability of the data). While in a national context some statistics may seem plausible, it may become apparent in comparing data across countries that such interpretations lead to dubious results in an international context, despite accurate translation of the questionnaires. Any such instances were addressed, and the corresponding variables were either recoded or subject to removal from the international database.

The final tasks of database construction included achievement scores and sampling weights, distributing national data files and documentation to each of the participating countries, and creating the international database. National research centers received their processed national databases approximately six months after arrival at the DPC. At the same time, processed data files also were sent to Statistics Canada for the calculation of sampling weights (see Chapter 9) and to the ISC, where the achievement scores were computed (see Chapter 12).

8.2 Data Entry at the National Research Centers

To assist with data entry, the IEA DPC supplied the DataEntryManager (WinDEM) software and manual (IEA, 2001b), and held a training session on the use of the software. The International Study Center provided each national research center with a *Manual for Entering the PIRLS Data* (PIRLS, 2001a), which details prescribed procedures for data entry and verification. In addition, the *Survey Operations Manual* (PIRLS, 2001b) includes directions for submitting the data files to the IEA DPC.

The data manager at each PIRLS national research center gathered data from tracking forms used to record information on students selected to participate in the study, as well as about their schools, teachers, and parents. Together with the responses from the student achievement booklets and student, teacher, school, and parent questionnaires, the information from the tracking forms were entered into computer data files. Codebooks specifying the standardized format and layout of the data were provided as a supplement to the WinDEM software and the *Manual for Entering the PIRLS Data* (PIRLS, 2001a). While strongly encouraged to use the recommended WinDEM software, a few participating countries elected to use a different data entry system. However, they were required to conform to all specifications established in the international codebooks.

In order to facilitate data entry, the codebooks and data files were structured to match the tests and questionnaires. This meant that for each survey instrument there was a corresponding data file and codebook. Furthermore, countries administering the test booklets or questionnaires in more than one language had to carefully prepare for data entry. They needed to determine whether the different versions of the test booklets or questionnaire could be entered into one database, or if they required one database for each version.

8.3 Data Checking and Editing at the National Centers

Before sending the data to the DPC for further data processing, countries were responsible for checking data files with programs specifically prepared for PIRLS and for making corrections as necessary. The first step was the application of the checking programs that are a feature of the WinDEM program. These tools are intended mainly to identify invalid data, but also can check the consistency between some basic variables. An important feature of WinDEM is the ability to check for unique identification codes. These checks were obligatory for all countries.

In the application of the LinkPIRL program (IEA, 2001c), the identification variables (student, teacher, class, or school ID) were checked against one another both within and between all files. Examples of linkage errors include: schools that were reported as non-participating, but for which there was a questionnaire in the teacher file; or students listed in the achievement files for whom there was no corresponding identification number in the background files. NRCs were asked to recheck their records, and resolve the problems identified in the within-country cleaning process.

8.4 Submitting Data Files to the IEA Data Processing Center

Each country was responsible for submitting six data files to the IEA Data Processing Center: the student background questionnaire file, student achievement file, home background file, teacher background file, school background file, and the constructed-response scoring reliability file. Countries administering the 1991 Reading Literacy Study test booklets and questionnaires submitted a seventh file: the 10-year trend study file. (For details of these files, see section 6.11.)

In addition to the data files, countries were required to submit copies of all tracking forms, copies of their national versions of translated test booklets and questionnaires, Data Management Forms documenting all national adaptations to the background questionnaires, and those booklets selected for the double scoring of constructed-response items.

8.5 DPC Quality Assurance Program

The IEA DPC has established a Quality Assurance Program to ensure that data is of high-quality, and that it is internationally comparable. Quality assurance was initiated before the first data arrived at the DPC through the provision of software to countries participating in PIRLS.

- The W3S software (IEA, 2001a) performs within school sampling and creates the required tracking forms.
- The WinDEM (IEA, 2001b) program performs data entry and data quality checks.
- The LinkPIRL program allows the NRCs to perform consistency checks between files.

A study as complex as PIRLS required a complex data cleaning design. To ensure that programs ran in the correct sequence, that no special requirements were overlooked, and that the cleaning process ran independently of the persons in charge, the following steps were undertaken:

- All incoming data and documents were read into a specific database. The date of arrival was stored, along with any specific issues, with the person in charge of monitoring the characteristics of the data and documents.
- Thorough testing of all cleaning programs took place prior to their implementation by means of simulated data sets containing all possible problems and inconsistencies.
- The cleaning was organized following strict rules. Deviations in the cleaning sequence were not possible, and the scope for involuntary changes to the cleaning procedures was minimal.
- Regular reviews of the country-specific data processing were done by a quality-assurance work group.
- A validity check was implemented for all cleaning steps, once the cleaning for a specific country was done. A country's data were virtually treated as new incoming data, and was again subjected to the entire cleaning process. There could be no new findings; all findings at this stage had already been justified.

A comparison was made between the original data set and the final, clean data set. Any changes in the data set had to be documented in the country's cleaning documentation.

8.6 Data Checking and Editing at the IEA Data Processing Center

Once the data were entered into data files at the national research center, the data files were submitted to the IEA Data Processing Center for checking and input into the international database. This process is generally referred to as data cleaning. The program-based data cleaning consisted of the following steps:

- Documentation and structure check

- Identification number cleaning and linkage check

- Valid range check and cleaning of inconsistencies within and between background files

- Quality control cleaning.

Special issues addressed by the IEA DPC during the cleaning process included the handling of missing data, and cleaning of Trends in IEA's Reading Literacy Study data.

8.6.1 Documentation and Structure Check

For each country, data cleaning began with an exploration of its data file structures and a review of its data documentation: Data Management Forms, Student Tracking Forms, Class Sampling Forms, Teacher Tracking Forms, and Test Administration Forms. Most countries sent all required documentation along with their data, which greatly facilitated the data checking. The IEA DPC contacted those countries for which documentation was incomplete, and obtained all forms necessary to complete the documentation.

The first checks implemented at the DPC looked for differences between the international file structure and national file structures. Some adaptations (such as adding national variables, or omitting or modifying international variables) were made to the background questionnaires in some countries. The extent and nature of such changes differed across the countries: some countries administered the questionnaires without any changes (apart from the translations), whereas other countries inserted items or options within existing international variables or added entirely new national variables. To keep track of any adaptations, NRCs were asked to complete Data Management Forms as they adapted the codebooks. Where necessary, the DPC modified the structure of the countries' data to ensure that the resulting data remained comparable between countries.

8.6.2 ID Cleaning and Linkage Check

Each record in a data file should have a unique identification number. Duplicate ID numbers imply an error of some kind. If two records shared the same ID, and contained exactly the same data, one of the records was deleted and the other remained in the database. If the records contained different data apart from the ID, and it was impossible to detect which record contained the "true data," both records were removed from the database. The DPC tried to keep losses at a minimum, and, in only in a few cases, were data actually deleted.

The ID cleaning focused on the student background questionnaire file, because most of the critical variables were present in this file. Apart from the unique student ID, there were variables pertaining to the students' participation and exclusion status – as well as dates of birth and dates of testing used to calculate age at the time of testing. The Student Tracking Forms[1] were

[1] Tracking Forms are used to record the sampling of schools, classes, teachers, and students. (see also Chapter 6).

essential in resolving any anomalies, as was close cooperation with NRCs (in most cases, the Student Tracking Forms were completed in the country's official language). The information about participation and exclusion was sent to Statistics Canada, where it was used to calculate students' participation rates, exclusion rates, and student sampling weights.

In PIRLS, data about students and their homes, schools, and teachers appear in several files. It is crucial that the records from these files were linked to each other correctly, to obtain meaningful results. Therefore, the second important check run at the DPC was the check for linkage between the files. The students' entries in the achievement file and in the student background file must match one another; the home background file must match the student file; the reliability scoring file must represent a specific part of the achievement file; the teachers must be linked to the correct students; and the schools must be linked to the correct teachers and students. The linkage is implemented through a hierarchical ID numbering system incorporating a school, class, and student component,[2] and is cross-checked against the tracking forms.

8.6.3 Valid Range Check, Filter-Dependent Check, and Consistency Check

"Valid range" indicates the range of values considered to be correct and meaningful for a specific variable. For example, the student gender variable had two valid values: "1" for a girl, and "2" for a boy. All other values are invalid. There were also questions in the school and teacher questionnaires for the respondent to write in a number – for example, the principal was asked to supply the school enrollment. For such variables, valid ranges may vary from country to country, and the acceptable ranges were set very wide to accommodate variations. It was possible for countries to adapt these ranges according to their needs, although countries were advised that a smaller range would decrease the possibility of mispunches. Cleaning at the DPC did not take smaller national ranges into account; only if values were found outside the international accepted range were the cases mentioned in the list of inquiries sent to countries. In cases where out-of-range values were found in the achievement file, the data were set to "Omitted" if the true value could not be retrieved.

Filter questions, which appear in some questionnaires, were used to direct the respondent to a particular section of the questionnaire. Depending on the response to a filter question, responses to subsequent questions are either expected or not expected. During data entry, these dependent

[2] The ID of a higher level is repeated in the ID of a lower sampling level: the class ID holds the school ID, and the student ID contains the class ID (e.g., student 1220523 can be described as student 23 of class 5 in school 122).

variables are not treated differently from any others. However, a special missing code is applied to dependent variables during data processing (for details on the handling of missing data, see section 8.6.5).

The number of inconsistent and implausible responses in background files varied from country to country, but no country's data was completely free of inconsistent responses. Treatment of these responses was determined on a question-by-question basis, using available documentation to make an informed decision. One example of inconsistencies between files is when a school principal states that his or her school has no library, but the teacher in the same school indicates that students are taken to the school library regularly. These cases were not changed in either file, provided mispunches were ruled out as cause.

8.6.4 Quality Control Cleaning

Quality control cleaning ensures that all necessary recoding of variables was performed correctly, and that consistency within and between files could be verified. The variables in the database have complex interrelationships. To avoid changes that make the relationship between two variables consistent but breaks the relationship with a third variable, a final cleaning step was established to take care of such multiple relationships within the database. This quality control cleaning can be interpreted as a check of the results of all earlier checks. After this variable-level cleaning, the consistency check between files was performed.

8.6.5 Handling of Missing Data

When the PIRLS data were entered using WinDEM, two types of entries were possible: valid data values or missing data values. Missing data can be assigned a value of omitted, not administered, or invalid during data entry.

At the IEA DPC, additional missing codes were applied to the data to be used for further analyses. In the international database, five missing codes are used:

- Not administered – the respondent was not administered the actual item. He or she had no chance to read and answer the question (assigned both during data entry and data processing).

- Omitted – the respondent had a chance to answer the question, but did not do so (assigned both during data entry and data processing).

- Logically not applicable – the respondent answered a preceding filter question in a way that made the following dependent questions not applicable to him or her (assigned during data processing only).

- Not reached (only used in the achievement files) – this code indicates those items not reached by the students, due to a lack of time (assigned during data processing only).

- Not interpretable (only used in the achievement files) – this code was used for multiple-choice items that were answered, but the chosen answer options were not clear – as well as for constructed-response items where the scorer assigned two or more scores (assigned during data entry and data processing).

8.6.6 Specific Cleaning Issues of the Trends in IEA's Reading Literacy Study

The Trends in IEA's Reading Literacy Study is a repetition of the IEA's 1991 Reading Literacy Study. Nine of the countries that participated in the 1991 study elected to readminister the test in 2001 (for a list of these countries, see Exhibit 5.4). The requirements for the Trends in IEA's Reading Literacy Study were that the achievement test and the student background questionnaires must be administered in exactly the same way, and that the cleaning procedures be applied in the same way as in 1991.

As a result, data cleaning for the Trends in IEA's Reading Literacy Study data is somewhat different in comparison to the cleaning rules for PIRLS (International Association for the Evaluation of Educational Achievement, 1995):

- All items following the last item containing a valid value were recoded to "Not reached."
- An additional missing value, "Invalid," indicates that the data were recorded in an invalid or inconsistent way. This value was used only in the student background file. A more detailed description of the Trends in IEA's Reading Literacy Study data cleaning can be found in the cleaning documentation of PIRLS 2001 (Barth, Itzlinger, Niemeyer, & Schwippert, 2001).

8.7 Returning Data to National Centers

As soon as the ID cleaning was complete, and the file structures had been standardized, participating countries received their national data files back from the DPC, in order to conduct preliminary national analyses. These preliminary data sets did not include national variables, derived variables, scaled scores, or sampling weights. Due to the timelines in PIRLS, several versions of the data were sent to the national research centers, with each subsequent version containing more features.

When data processing was complete, final national data sets were sent to countries along with final sampling weights, international scores, derived variables, and all international and national variables. National variables were placed in extra files that could be merged with the files containing the international variables.

8.8 Creating the International Database

The international database incorporates all national data files. After data processing by the DPC, it can be ensured that:

- Information coded in each variable is internationally comparable.

- National adaptations are reflected appropriately in all variables.

- Questions that are not internationally comparable have been removed from the database.

- All entries in the database can be linked to the appropriate respondent – student, teacher, parent, or principal.

- Sampling weights and student achievement scores are available for international comparisons.

In a joint effort between the IEA DPC and the ISC at Boston College, a National Adaptations Database containing all adaptations to questionnaires made by individual countries (documenting how they were handled) was constructed. The meaning of country-specific items can also be found in this database, as well as recoding requirements of the ISC. Information contained in this database is provided in the user guide for the international database upon release of the PIRLS 2001 data.

The PIRLS 2001 international database is a unique resource for policy makers and analysts, containing student reading achievement and background data from representative samples of fourth grade students from 35 countries. In all, the database contains more than 713 variables, with data from 5,777 schools, 7,041 teachers, 153,340 students, and 131,047 parents.

References

Barth, J., Itzlinger, U., Niemeyer, A. & Schwippert, K. (2001). *Cleaning documentation for PIRLS 2001*. Hamburg: Unpublished document, IEA Data Processing Center.

IEA. (1995). *The IEA Reading Literacy Study: technical report*. The Hague: IEA.

IEA. (2001a). *W3S: Within-school sampling software*. Hamburg: IEA Data Processing Center.

IEA. (2001b). *WinDEM: Software for data entry and verification*. Hamburg: IEA Data Processing Center.

IEA. (2001c). *LinkPIRL guide*. Hamburg: IEA Data Processing Center.

Progress in International Reading Literacy Study (PIRLS). (2001a). *Manual for entering the PIRLS data* (PIRLS Ref. No. 01-0004) Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

Progress in International Reading Literacy Study (PIRLS). (2001b). *Survey operations manual* (PIRLS Ref. No. 01-0001). Prepared by the International Study Center at Boston College. Chestnut Hill, MA: Boston College.

9

PIRLS Sampling Weights and Participation Rates

Marc Joncas

9.1 Overview

Selecting valid and efficient samples is critical to the quality and success of an international comparative study, such as PIRLS. The accuracy of the survey results depends on the quality of the sampling information available when planning the sample, and on the care with which the sampling activities themselves are conducted. For PIRLS 2001, National Research Coordinators (NRCs) worked on all phases of sampling, in conjunction with staff from Statistics Canada. NRCs were trained in how to select the school and student samples, and in how to use the sampling software provided by the IEA Data Processing Center. This chapter summarizes major characteristics of the national samples, and describes the procedure for computing sampling weights and participation rates for each country. In consultation with the PIRLS 2001 sampling referee,[1] staff from Statistics Canada reviewed the national sampling plans, sampling data, sampling frames, and sample selection. The PIRLS International Study Center (ISC) at Boston College, jointly with Statistics Canada and the sampling referee, used this information to evaluate the quality of the samples. Summaries of the sample design for each country, including details of population coverage and exclusions, stratification variables, and participation rates, are provided in Appendix B.

[1] Keith Rust, Westat.

9.2 Sampling implementation

9.2.1 PIRLS 2001 Target Population

In IEA studies, the target population for all countries is known as the *international desired population*. The international desired population for PIRLS 2001 was defined as:[2]

- All students enrolled in the upper of the two adjacent grades that contain the largest proportion of 9-year-olds at the time of testing.

Beyond the age criterion embedded in the above definition, the target grade should represent that point in the curriculum where students have essentially finished learning the basic reading skills, and will then focus more on "reading to learn" in the subsequent grades. Thus, the PIRLS 2001 target grade was expected to be the fourth grade in most countries (some countries, therefore, have students significantly older than nine years of age).[3]

Exhibit 9.1 summarizes the grades identified as the target grade in all participating countries. For most countries, the target grade did indeed turn out to be the fourth grade. Average student ages ranged from 9.7 (in Cyprus and Iceland) to 11.2 (in Morocco).

Exhibit 9.1: National Grade Definitions

Country	Country's Name for Grade Tested	Years of Formal Schooling	Mean Age of Students Tested
Argentina	4	4	10.2
Belize	Standard II	4	9.8
Bulgaria	4	4	10.9
Canada (O, Q)[1]	4	4	10.0
Colombia	4	4	10.5
Cyprus	4	4	9.7
Czech Republic	4	4	10.5
England	Year 5	5	10.2
France	Cours Moyen 1	4	10.1
Germany	4	4	10.5
Greece	4	4	9.9
Hong Kong, SAR	Primary 4	4	10.2
Hungary	4	4	10.7
Iceland	4	4	9.7
Iran, Islamic Rep. of	4	4	10.4
Israel	4	4	10.0
Italy	4	4	9.8
Kuwait	4	4	9.9
Latvia	4	4	11.0
Lithuania	4	4	10.9
Macedonia, Rep. of	4	4	10.7
Moldova	4	4	10.8
Morocco	4	4	11.2
Netherlands	6th group	4	10.3
New Zealand	Year 5[2]	4	10.1
Norway	4	4	10.0
Romania	4	4	11.1
Russian Federation	3 in stream I and 4 in stream II	3 or 4	10.3
Scotland	Primary 5	5	9.8
Singapore	Primary 4	4	10.1
Slovak Republic	4	4	10.3
Slovenia	3	3	9.8
Sweden	4	4	10.8
Turkey	4	4	10.2
United States	4	4	10.2

[1] Canada is represented by the provinces of Ontario and Quebec only

[2] The official nomenclature used in New Zealand since 1996 refers to students' years of schooling rather than a class/grade level. Year 5 students were at a class level equivalent to Grade 4.

[2] This is also the population definition used by TIMSS for primary-school students.

[3] The target population for each participating country is described in Appendix B.

9.2.2 Population Coverage and Exclusions

Exhibit 9.2 summarizes population coverage and exclusions for the PIRLS 2001 target populations. National coverage of the international desired target population was generally comprehensive. Only Canada and Lithuania chose a national desired population less than the international desired

Exhibit 9.2: Population Coverage and Exclusions

Country	International Desired Population		National Desired Population		
	Coverage	Notes on Coverage	School-Level Exclusions	Within-Sample Exclusions	Overall Exclusions
Argentina	100%		3.7%	0.4%	4.1%
Belize	100%		0.8%	0.0%	0.8%
Bulgaria	100%		2.7%	0.0%	2.7%
Canada (O, Q)[1]	60%	Provinces of Ontario and Quebec only	3.1%	2.2%	5.4%
Colombia	100%		3.2%	0.1%	3.3%
Cyprus	100%		0.0%	2.0%	2.0%
Czech Republic	100%		5.0%	0.0%	5.0%
England	100%		1.8%	3.9%	5.7%
France	100%		5.1%	0.3%	5.3%
Germany	100%		0.8%	1.0%	1.8%
Greece	100%		2.0%	5.3%	7.3%
Hong Kong, SAR	100%		2.8%	0.0%	2.8%
Hungary	100%		2.1%	0.0%	2.1%
Iceland	100%		1.8%	1.3%	3.1%
Iran, Islamic Rep. of	100%		0.5%	0.0%	0.5%
Israel	100%		16.5%	5.9%	22.4%
Italy	100%		0.0%	2.9%	2.9%
Kuwait	100%		0.0%	0.0%	0.0%
Latvia	100%		4.3%	0.3%	4.6%
Lithuania	90%	Lithuanian speaking students only	1.3%	2.5%	3.8%
Macedonia, Rep. of	100%		3.8%	0.4%	4.2%
Moldova	100%		0.5%	0.0%	0.5%
Morocco	100%		1.0%	0.0%	1.0%
Netherlands	100%		3.4%	0.3%	3.7%
New Zealand	100%		1.6%	1.7%	3.2%
Norway	100%		1.9%	0.8%	2.8%
Romania	100%		2.6%	1.9%	4.5%
Russian Federation	100%		2.8%	3.8%	6.6%
Scotland	100%		3.8%	0.8%	4.7%
Singapore	100%		1.3%	0.1%	1.4%
Slovak Republic	100%		1.4%	0.6%	2.0%
Slovenia	100%		0.0%	0.3%	0.3%
Sweden	100%		2.5%	2.5%	5.0%
Turkey	100%		3.9%	0.0%	3.9%
United States	100%		0.6%	4.7%	5.3%

[1] Canada is represented by the provinces of Ontario and Quebec only

population.[4] Because coverage of the international desired population fell below 65 percent for Canada, the Canadian results have been labeled "Canada (O,Q)" in the international report. Coverage was more inclusive in Lithuania, but since it was less than 100 percent, the Lithuanian results were footnoted to reflect this.

For the most part, school-level exclusions consisted of schools for the disabled and very small schools; however, there were some exceptions that are documented in Appendix B. Within-school exclusions generally consisted of disabled students and students who could not be assessed in the language of the test. Only in Israel did the level of excluded students exceed 10 percent. Three other countries (England, Greece, and the Russian Federation) have an exclusion rate above 5 percent (but below 7%). This was reflected in footnotes in the international reports. A few countries had no within-school exclusions.

9.2.3 General Sample Design

The basic design of the sample used in PIRLS 2001 was a two-stage stratified cluster design.[5] The first stage consisted of a sampling of schools, and the second stage of a sampling of intact classrooms from the target grade in the sampled schools.

The PIRLS 2001 design allowed countries to stratify the school sampling frame in order to improve the precision of survey results. Countries could use an explicit stratification procedure, by which schools were categorized according to some criterion (e.g., regions of the country), ensuring a predetermined number of schools would be selected from each stratum. Countries also could use an implicit stratification procedure, by which schools were sorted according to a set of stratification variables prior to sampling. This approach provided an efficient method of allocating the school sample in proportion to the size of the implicit stratum, when used in conjunction with a systematic PPS method. Stratification variables and procedures for each country are described in Appendix B.

Most countries sampled 150 schools and one intact classroom (with all of its students) from each school. Countries that selected larger school samples included large countries such as the United States and the Russian Federation, and countries such as Canada, Germany, and Hungary that required accurate survey estimates for regions or provinces. Schools were selected with probability proportional to size, and classrooms with equal probabilities. Upon recommendation from Statistics Canada, some countries chose to sample more than one classroom per selected school. Details of the sampling of schools and students for each country are provided in Appendix B.

[4] The Lithuanian population was restricted to schools catering to Lithuanian-speaking students only, the Canadian population to schools from the provinces of Ontario and Quebec only.

[5] The PIRLS sample design is described in Chapter 5.

9.2.4 Target Population Sizes

Exhibit 9.3 summarizes the number of schools and students in each country's target population, as well as the number of schools and students that participated in the study. Most of the target population sizes are derived from the sampling frames from which the PIRLS samples were drawn. The school and student population sizes for the United States and the Russian Federation, however, were not computed from the sampling frame, but were instead provided by their respective NRC. Using the sampling weights computed for each

Exhibit 9.3: Population and Sample Sizes

Country	Population		Sample			Mean Age
	Schools	Students	Schools	Students	Estimated Population	
Argentina	14 055	709 772	138	3 300	709 193	10.2
Belize	237	9 261	120	2 909	7 408	9.8
Bulgaria	2 424	98 270	170	3 460	95 702	10.9
Canada (O, Q)[1]	5 357	241 805	372	8 253	222 012	10.0
Colombia	46 805	867 583	147	5 131	975 170	10.5
Cyprus	242	10 209	150	3 001	10 206	9.7
Czech Republic	3 830	121 330	141	3 022	123 831	10.5
England	15 191	629 524	131	3 156	592 787	10.2
France	31 056	748 424	145	3 538	717 378	10.1
Germany	19 207	941 200	211	7 726	899 014	10.5
Greece	4 999	102 927	145	2 494	97 288	9.9
Hong Kong, SAR	760	81 207	147	5 050	88 645	10.2
Hungary	2 700	113 594	216	4 666	117 238	10.7
Iceland	140	4 566	133	3 676	4 456	9.7
Iran, Islamic Rep. of	61 110	1 741 673	184	7 430	1 812 810	10.4
Israel	1 462	90 905	147	3 973	85 802	10.0
Italy	7 162	573 571	184	3 502	573 318	9.8
Kuwait	184	21 414	135	7 133	22 318	9.9
Latvia	940	34 216	141	3 019	34 213	11.0
Lithuania	1 146	44 188	146	2 567	43 094	10.9
Macedonia, Rep. of	351	27 726	146	3 711	27,365	10.7
Moldova	1 395	64 467	150	3 533	60 634	10.8
Morocco	14 828	529 105	117	3 153	554 573	11.2
Netherlands	7 185	183 599	134	4 112	181 387	10.3
New Zealand	1 984	59 705	156	2 488	58 122	10.1
Norway	2 468	60 503	136	3 459	58 174	10.0
Romania	10 582	306 891	144	3 625	283 340	11.1
Russian Federation	63 641	2 009 900	206	4 093	1 823 855	10.3
Scotland	2 045	62 783	118	2 717	64 375	9.8
Singapore	196	50 772	196	7 002	49 301	10.1
Slovak Republic	2 165	76 182	150	3 807	71 409	10.3
Slovenia	443	21 906	148	2 952	21 066	9.8
Sweden	3 727	117 767	146	6 044	118 134	10.8
Turkey	13 941	1 111 470	154	5 125	977 316	10.2
United States	71 498	3 871 487	174	3 763	3 802 557	10.2

[1] Canada is represented by the provinces of Ontario and Quebec only

country (see section 9.3), PIRLS derived an estimate of the student population size, which matched closely the student population size from the sampling frame.

9.3 Calculating Sampling Weights

The PIRLS 2001 sampling design required schools to be sampled with a probability proportional to size (PPS), and for classrooms to be sampled with equal probabilities.[6] PIRLS 2001 participants adapted the basic design to the requirements of their educational systems, with guidance from the PIRLS sampling consultants at Statistics Canada and the sampling referee. Very large countries could add an extra preliminary stage, where districts or regions were sampled first, and then schools within districts.[7] Participants used stratification in order to improve the precision of their samples where appropriate. Individual country designs could be quite complex, as may be seen from the information in Appendix B – showing how the design was implemented in each country.

While the PIRLS 2001 multistage stratified cluster design provided very economical and effective data collection in a school environment, it resulted in differential probabilities of selection of the students. To adjust for these differential selection probabilities and ensure proper survey estimates, PIRLS 2001 computed a sampling weight for each participating student. Because appropriate sampling weights were essential for the computation of accurate survey results, the ability to provide proper sampling weights was an essential requirement of an acceptable sample design. This section describes the procedures for calculating sampling weights for the PIRLS 2001 data.

Sampling weights were calculated according to a three-step procedure involving selection probabilities for schools, classrooms, and students. The first step consisted of calculating a school weight, which also incorporated weighting factors from any additional front-end sampling stages such as districts or regions. A school-level participation adjustment was then made to the school weight to compensate for any sampled schools that did not participate. This adjustment was calculated independently for each explicit stratum.

In the second step, a classroom weight reflecting the probability of the sampled classroom(s) being selected from among all the classrooms in the school at the target grade level was calculated. No classroom-level participation adjustment was necessary, since in most cases a single classroom was sampled in each school. If a school agreed to take part in the study, but the classroom refused to participate, adjustment for non-participation was made at the school level. If one of two selected class-

[6] The PIRLS 2001 sampling design is presented in Chapter 5.

[7] For example, the United States sampled school districts as primary sampling units and then schools within the school districts.

rooms in a school did not participate, then the classroom weight was calculated as though a single classroom had been selected in the first place. The classroom weight was calculated independently for each school.

Because intact classrooms were sampled in PIRLS, each student in the sampled classrooms was certain of selection, and so the student weight was 1.0. However, as a third and final step, a non-participation adjustment was made to compensate for students who did not take part in the testing. This was calculated independently for each sampled classroom. The basic sampling weight attached to each student record was the product of the three intermediate weights: the first stage (school) weight, the second stage (classroom) weight, and the third stage (student) weight. The overall student sampling weight was the product of the three weights including the non-participation adjustments.

9.3.1 The First Stage (School) Weight

Essentially, the first stage weight represented the inverse of the probability of a school being sampled on the first stage. The PIRLS 2001 sample design required that school selection probabilities be proportional to the school size, defined as enrollment in the target grade. The basic first stage weight for the ith sampled school was thus defined as:

$$BW_{sc}^{i} = \frac{M}{n \cdot m_i}$$

where n was the number of sampled schools, m_i was the measure of size for the ith school, and

$$M = \sum_{i=1}^{N} m_i$$

where N was the total number of schools in the explicit stratum containing the school.

For countries with a preliminary sampling stage (such as the United States and the Russian Federation), the basic first stage weight also incorporated the probability of selection in this preliminary stage. The first stage weight in such cases was simply the product of the "region" weight and the first stage weight, as described earlier.

In some countries, schools were selected with equal probabilities. This generally occurred when a large sampling ratio was used. In some countries also, explicit or implicit strata were defined to deal with very large schools or small schools. Equal probability sampling was necessary in these strata.

Under equal probability sampling, the basic first stage weight for the ith sampled school was defined as:

$$BW_{sc}^{i} = \frac{N}{n}$$

where n was the number of sampled schools and N was the total number of schools in the explicit stratum. The basic weight for all sampled schools in a stratum was identical in this context.

9.3.2 School Non-Participation Adjustment

First stage weights were calculated for all sampled and replacement schools that participated. A school-level participation adjustment was required to compensate for those schools that were sampled but did not participate, and hence were not replaced. Sampled schools that were found to be ineligible were removed from the calculation of this adjustment.[8] The school-level participation adjustment was calculated separately for each explicit stratum.

The adjustment was calculated as follows:

$$A_{sc} = \frac{n_s + n_{r1} + n_{r2} + n_{nr}}{n_s + n_{r1} + n_{r2}}$$

where n_s was the number of originally sampled schools that participated, n_{r1} and n_{r2} the number of first and second replacement schools, respectively, that participated, and n_{nr} the number of schools that did not participate.

The final first stage weight for the ith School, corrected for non-participating schools, thus became:

$$FW_{sc}^{i} = A_{sc} \cdot BW_{sc}^{i}$$

9.3.3 The Second Stage (Classroom) Weight

The second stage weight represented the inverse of the second stage selection probability assigned to a sampled classroom. All classrooms were sampled with equal probability. For the ith school, let C^i be the total number of classrooms and c^i the number of sampled classrooms that participated in the study. Using equal probability sampling, the final second stage weight assigned to all sampled classrooms in the ith school was:

$$FW_{cl}^{i} = \frac{C^i}{c^i}$$

For most countries, c^i took the values 1 or 2, and remained fixed for all sampled schools. Some countries sampled all classrooms in a selected school.

9.3.4 The Third Stage (Student) Weight

The third stage weight represented the inverse of the third stage selection probability attached to a sampled student. Because intact classrooms were sampled, and all students in the classroom were expected to participate, the basic third stage weight for the jth classroom in the ith school was simply:

$$BW_{st}^{i,j} = 1.0$$

9.3.5 Adjustment for Student Non-Participation

The student non-participation adjustment was calculated for each participating classroom as follows:

$$A_{st}^{i,j} = \frac{s_{rs}^{i,j} + s_{nr}^{i,j}}{s_{rs}^{i,j}}$$

[8] A sampled school was ineligible if it was found to contain no eligible (i.e., fourth-grade) students. Such schools usually were in the sampling frame by mistake, or schools that had recently closed.

where $s_{rs}^{i,j}$ was the number of eligible students that participated in the jth classroom of the ith school and $s_{nr}^{i,j}$ was the number of eligible students that did not participate in the jth classroom of the ith school.

The third, and final, stage weight for students in the jth classroom in the ith school thus became:

$$FW_{st}^{i,j} = A_{st}^{i,j} \cdot BW_{st}^{i,j}$$

9.3.6 Overall Sampling Weight

The overall sampling weight was simply the product of the final first stage weight, the final second stage weight, and the final third stage weight and is given by:

$$W^{i,j} = FW_{sc}^{i} \cdot FW_{cl}^{i,j} \cdot FW_{st}^{i,j}$$

or

$$W^{i,j} = A_{sc} \cdot BW_{sc}^{i} \cdot FW_{cl}^{i,j} \cdot A_{st}^{i,j} \cdot BW_{st}^{i,j}$$

It is important to note that sampling weights vary by school and classroom, but that students within the same classroom have the same sampling weights. It is also important to note that sampling weights were calculated separately by explicit strata.

9.4 Calculating School and Student Participation Rates

Since non-participation by sampled schools or students can lead to bias in the study results, a variety of participation rates were computed to reveal the level of success each country achieved in securing participation from their sampled schools and students. To monitor school participation, three school participation rates were computed: one using originally sampled schools only; one using sampled and first replacement schools; and one using sampled and both first and second replacement schools. Student participation rates were also computed, as were overall participation rates.

9.4.1 Unweighted School Participation Rates

The three unweighted school participation rates that were computed were the following:

R_{unw}^{sc-s} = unweighted school participation rate for originally sampled schools only,

R_{unw}^{sc-r1} = unweighted school participation rate, including sampled and first replacement schools,

R_{unw}^{sc-r2} = unweighted school participation rate, including sampled, first, and second replacement schools.

Each unweighted school participation rate was defined as the ratio of the number of participating schools to the number of originally sampled schools, excluding any ineligible schools. The rates were calculated as follows:

$$R_{unw}^{sc-s} = \frac{n_s}{n_s + n_{r1} + n_{r2} + n_{nr}}$$

$$R_{unw}^{sc-r1} = \frac{n_s + n_{r1}}{n_s + n_{r1} + n_{r2} + n_{nr}}$$

$$R_{unw}^{sc-r2} = \frac{n_s + n_{r1} + n_{r2}}{n_s + n_{r1} + n_{r2} + n_{nr}}$$

9.4.2 Unweighted Student Participation Rates

The unweighted student participation rate was computed as follows:

$$R_{unw}^{st} = \frac{\sum_{i,j} s_{rs}^{i,j}}{\sum_{i,j} s_{rs}^{i,j} + \sum_{i,j} s_{nr}^{i,j}}$$

9.4.3 Unweighted Overall Participation Rates

Three unweighted overall participation rates were computed for each country. They were as follows:

$R_{unw}^{ov-s} =$ unweighted overall participation rate for originally sampled schools only,

$R_{unw}^{ov-r1} =$ unweighted overall participation rate, including sampled and first replacement schools,

$R_{unw}^{ov-r2} =$ unweighted overall participation rate, including sampled, first, and second replacement schools.

For each country, the overall participation rate was defined as the product of the unweighted school participation rate and the unweighted student participation rate. They were calculated as follows:

$$R_{unw}^{ov-s} = R_{unw}^{sc-s} \cdot R_{unw}^{st}$$

$$R_{unw}^{ov-r1} = R_{unw}^{sc-r1} \cdot R_{unw}^{st}$$

$$R_{unw}^{ov-r2} = R_{unw}^{sc-r2} \cdot R_{unw}^{st}$$

9.4.4 Weighted School Participation Rates

Three weighted school-level participation rates were computed for each country. They were as follows:

$R_{wtd}^{sc-s} =$ weighted school participation rate for originally sampled schools only,

$R_{wtd}^{sc-r1} =$ weighted school participation rate, including sampled and first replacement schools,

$R_{wtd}^{sc-r2} =$ weighted school participation rate, including sampled, first, and second replacement schools.

The weighted school participation rates were calculated as follows:

$$R_{wtd}^{sc-s} = \frac{\sum_{i,j}^{s} BW_{sc}^{i} \cdot FW_{cl}^{i,j} \cdot FW_{st}^{i,j}}{\sum_{i,j}^{s+r1+r2} FW_{sc}^{i} \cdot FW_{cl}^{i,j} \cdot FW_{st}^{i,j}}$$

$$R_{wtd}^{sc-r1} = \frac{\sum_{i,j}^{s+r1} BW_{sc}^{i} \cdot FW_{cl}^{i,j} \cdot FW_{st}^{i,j}}{\sum_{i,j}^{s+r1+r2} FW_{sc}^{i} \cdot FW_{cl}^{i,j} \cdot FW_{st}^{i,j}}$$

$$R_{wtd}^{sc-r2} = \frac{\sum_{i,j}^{s+r1+r2} BW_{sc}^{i} \cdot FW_{cl}^{i,j} \cdot FW_{st}^{i,j}}{\sum_{i,j}^{s+r1+r2} FW_{sc}^{i} \cdot FW_{cl}^{i,j} \cdot FW_{st}^{i,j}}$$

where both the numerator and denominator were summations over all responding students and the appropriate classroom-level and student-level sampling weights were used. Note that the basic school-level weight appears in the numerator, whereas the final school-level weight appears in the denominator.

The denominator remains unchanged in all three equations and is the weighted estimate of the total enrollment in the target population. The numerator, however, changes from one equation to the next. Only students from originally-sampled schools were included in the first equation. Students from first replacement schools were added in the second equation, and students from first and second replacement schools were added in the third equation.

9.4.5 Weighted Student Participation Rates

The weighted student participation rate was computed as follows:

$$R_{wtd}^{st} = \frac{\sum_{i,j}^{s+r1+r2} BW_{sc}^{i} \cdot FW_{cl}^{i,j} \cdot BW_{st}^{i,j}}{\sum_{i,j}^{s+r1+r2} BW_{sc}^{i} \cdot FW_{cl}^{i,j} \cdot FW_{st}^{i,j}}$$

where both the numerator and denominator were summations over all responding students, and the appropriate classroom-level and student-level sampling weights were used. Note that the basic student-level weight appears in the numerator, whereas the final student-level weight appears in the denominator. Furthermore, the denominator in this formula was the same quantity that appears in the numerator of the weighted school-level participation rate for all participating schools, sampled and replacement.

9.4.6 Weighted Overall Participation Rates

Three weighted overall participation rates were computed. They were as follows:

R_{wtd}^{ov-s} = weighted overall participation rate for originally sampled schools only,

R_{wtd}^{ov-r1} = weighted overall participation rate, including sampled and first replacement schools,

R_{wtd}^{ov-r2} = weighted overall participation rate, including sampled, first, and second replacement schools.

Each weighted overall participation rate was defined as the product of the appropriate weighted school participation rate and the weighted student participation rate. They were computed as follows:

$$R_{wtd}^{ov-s} = R_{wtd}^{sc-s} \cdot R_{wtd}^{st}$$

$$R_{wtd}^{ov-r1} = R_{wtd}^{sc-r1} \cdot R_{wtd}^{st}$$

$$R_{wtd}^{ov-r2} = R_{wtd}^{sc-r2} \cdot R_{wtd}^{st}$$

Weighted school, student, and overall participation rates were computed for each participating country using these procedures. Countries understood that the goal for sampling participation was 100 percent for all sampled schools and students. Guidelines for reporting achievement data for countries securing less than full participation were modeled after IEA's TIMSS study. Countries were assigned to one of three categories on the basis of their sampling participation (Exhibit 9.4). Countries in Category 1 were considered to have met the PIRLS sampling requirements, and to have an acceptable participation rate. Countries in Category 2 met the sampling requirements only after including replacement schools. Countries that failed to meet the participation requirements even with the use of replacement schools were assigned to Category 3. One of the main goals for quality data in PIRLS 2001 was to have as many countries as possible achieve Category 1 status, and to have no countries in Category 3.

Exhibits 9.5 through 9.8 present the school, student, and overall participation rates and achieved sample sizes for each participating country. As can be seen from these exhibits, almost all countries met the PIRLS sampling requirements, and belong in Category 1. Because they met the sampling requirements only after including replacement schools – England, The Netherlands, and the United States belong in Category 2, and their results were annotated with an obelisk in the achievement exhibits in the international report. Although Morocco and Scotland had overall weighted participation rates of 69 and 74 percent, respectively (even after including replacement schools), it was decided during the sampling adjudication that these rates did not warrant the placement of the countries in Category 3. Instead, results for Morocco and Scotland were annotated with a double-obelisk indicating that they nearly satisfied the guidelines for sample participation rates after including replacement schools.

Exhibit 9.4: Categories of Sampling Participation

Category 1	Acceptable sampling participation rate **without** the use of replacement school. In order to be placed in this category, a country had to have: • An **unweighted** school response rate **without** replacement of at least 85% (after rounding to the nearest whole percent) AND an unweighted student response rate (after rounding) of at least 85%. OR • A **weighted** school response rate **without** replacement of at least 85% (after rounding to the nearest whole percent) AND a **weighted** student response rate (after rounding) of at least 85%. OR • The product of the (unrounded) **weighted** school response rate **without** replacement and the (unrounded) **weighted** student response rate of at least 75% (after rounding to the nearest whole percent). Countries in this category appeared in the tables and figures in international reports without annotation ordered by achievement as appropriate.
Category 2	Acceptable sampling participation rate **only when replacement schools were included**. A country was placed in category 2 if: • It failed to meet the requirements for Category 1 but had either an unweighted or weighted school response rate **without** replacement of at least 50% (after rounding to the nearest percent). AND HAD EITHER • An **unweighted** school response rate **with** replacement of at least 85% (after rounding to the nearest whole percent) AND an **unweighted** student response rate (after rounding) of at least 85%. OR • A **weighted** school response rate **with** replacement of at least 85% (after rounding to nearest whole percent) AND a **weighted** student response rate (after rounding) of at least 85%. OR • The product of the (unrounded) **weighted** school response rate **with** replacement and the (unrounded) **weighted** student response rate of at least 75% (after rounding to the neasest whole percent). Countries in this category were annotated in the tables and figures in international reports and ordered by achievement as appropriate.
Category 3	Unacceptable sampling response rate even when replacement schools are included. Countries that could provide documentation to show that they complied with PIRLS sampling procedures and requirements but did not meet the requirements for Category 1 or Category 2 were placed in Category 3. Countries in this category would appear in a separate section of the achievement tables, below the other countries, in international reports. These countries were presented in alphabetical order.

Exhibit 9.5: School Participation Rates and Sample Sizes

Country	School Participation Before Replacement (Weighted Percentage)	School Participation After Replacement (Weighted Percentage)	Number of Schools in Original Sample	Number of Eligible Schools in Original Sample	Number of Schools in Original Sample That Participated	Number of Replacement Schools That Participated	Total Number of Schools That Participated
Argentina	89%	92%	150	150	133	5	138
Belize	80%	80%	150	150	119	1	120
Bulgaria	97%	97%	177	176	170	0	170
Canada (O, Q)[1]	90%	97%	387	387	359	13	372
Colombia	80%	98%	150	150	119	28	147
Cyprus	98%	100%	150	150	148	2	150
Czech Republic	90%	95%	150	148	135	6	141
England	57%	87%	150	150	88	43	131
France	93%	97%	150	150	140	5	145
Germany	98%	98%	216	215	209	2	211
Greece	78%	85%	170	170	133	12	145
Hong Kong, SAR	73%	98%	150	150	115	32	147
Hungary	98%	98%	220	220	216	0	216
Iceland	95%	95%	140	140	133	0	133
Iran, Islamic Rep. of	97%	100%	184	184	180	4	184
Israel	96%	98%	150	150	144	3	147
Italy	90%	100%	184	184	164	20	184
Kuwait	87%	89%	150	150	133	2	135
Latvia	89%	96%	148	147	133	8	141
Lithuania	56%	97%	150	150	84	62	146
Macedonia, Rep. of	97%	97%	150	150	145	1	146
Moldova	84%	100%	150	150	133	17	150
Morocco	74%	74%	158	158	117	0	117
Netherlands	53%	89%	150	150	80	54	134
New Zealand	94%	100%	156	156	144	12	156
Norway	82%	89%	162	160	119	17	136
Romania	96%	96%	150	150	144	0	144
Russian Federation	100%	100%	206	206	205	1	206
Scotland	76%	79%	150	150	113	5	118
Singapore	100%	100%	196	196	196	0	196
Slovak Republic	88%	100%	150	150	130	20	150
Slovenia	98%	99%	150	150	147	1	148
Sweden	97%	99%	150	149	142	4	146
Turkey	100%	100%	154	154	154	0	154
United States	61%	86%	200	200	125	49	174

[1] Canada is represented by the provinces of Ontario and Quebec only

Exhibit 9.6: Student Participation Rates and Sample Sizes

Country	Within School Student Participation (Weighted Percentage)	Number of Sampled Students in Participating Schools	Number of Students Withdrawn from Class/School	Number of Students Excluded	Number of Students Eligible	Number of Students Absent	Number of Students Assessed
Argentina	91%	3 769	132	13	3 624	324	3 300
Belize	94%	3 137	32	0	3 105	196	2 909
Bulgaria	97%	3 633	53	0	3 580	120	3 460
Canada (O, Q)[1]	94%	9 151	99	228	8 824	571	8 253
Colombia	96%	5 582	225	5	5 352	221	5 131
Cyprus	97%	3 149	2	63	3 084	83	3 001
Czech Republic	94%	3 220	10	0	3 210	188	3 022
England	94%	3 528	46	122	3 360	204	3 156
France	97%	3 673	20	11	3 642	104	3 538
Germany	88%	8 997	27	58	8 912	1186	7 726
Greece	97%	2 718	0	151	2 567	73	2 494
Hong Kong, SAR	99%	5 192	69	0	5 123	73	5 050
Hungary	97%	4 819	14	0	4 805	139	4 666
Iceland	87%	4 320	29	58	4 233	557	3 676
Iran, Islamic Rep. of	98%	7 703	104	0	7 599	169	7 430
Israel	96%	4 400	33	214	4 153	180	3 973
Italy	98%	3 703	15	103	3 585	83	3 502
Kuwait	91%	7 874	0	0	7 874	741	7 133
Latvia	93%	3 266	8	11	3 247	228	3 019
Lithuania	85%	3 114	7	72	3 035	468	2 567
Macedonia, Rep. of	97%	3 904	42	14	3 848	137	3 711
Moldova	96%	3 679	9	0	3 670	137	3 533
Morocco	93%	3 452	35	0	3 417	264	3 153
Netherlands	98%	4 256	11	14	4 231	119	4 112
New Zealand	96%	2 720	68	53	2 599	111	2 488
Norway	92%	3 784	25	26	3 733	274	3 459
Romania	97%	3 744	23	2	3 719	94	3 625
Russian Federation	97%	4 281	24	42	4 215	122	4 093
Scotland	95%	2 912	20	26	2 866	149	2 717
Singapore	98%	7 162	46	4	7 112	110	7 002
Slovak Republic	96%	4 034	33	18	3 983	176	3 807
Slovenia	95%	3 112	10	8	3 094	142	2 952
Sweden	93%	6 678	38	145	6 495	451	6 044
Turkey	97%	5 390	123	0	5 267	142	5 125
United States	96%	4 091	55	121	3 915	152	3 763

[1] Canada is represented by the provinces of Ontario and Quebec only

Exhibit 9.7: School and Student Participation Rates

Country	School Participation Before Replacement	School Participation After Replacement	Student Participation	Overall Participation Before Replacement	Overall Participation After Replacement
Argentina	89%	92%	91%	81%	84%
Belize	79%	80%	94%	74%	75%
Bulgaria	97%	97%	97%	93%	93%
Canada (O, Q)[1]	93%	96%	94%	87%	90%
Colombia	79%	98%	96%	76%	94%
Cyprus	99%	100%	97%	96%	97%
Czech Republic	91%	95%	94%	86%	90%
England	59%	87%	94%	55%	82%
France	93%	97%	97%	91%	94%
Germany	97%	98%	87%	84%	85%
Greece	78%	85%	97%	76%	83%
Hong Kong, SAR	77%	98%	99%	76%	97%
Hungary	98%	98%	97%	95%	95%
Iceland	95%	95%	87%	82%	82%
Iran, Islamic Rep. of	98%	100%	98%	96%	98%
Israel	96%	98%	96%	92%	94%
Italy	89%	100%	98%	87%	98%
Kuwait	89%	90%	91%	80%	82%
Latvia	90%	96%	93%	84%	89%
Lithuania	56%	97%	85%	47%	82%
Macedonia, Rep. of	97%	97%	96%	93%	94%
Moldova	89%	100%	96%	85%	96%
Morocco	74%	74%	92%	68%	68%
Netherlands	53%	89%	97%	52%	87%
New Zealand	92%	100%	96%	88%	96%
Norway	74%	85%	93%	69%	79%
Romania	96%	96%	97%	94%	94%
Russian Federation	100%	100%	97%	97%	97%
Scotland	75%	79%	95%	71%	75%
Singapore	100%	100%	98%	98%	98%
Slovak Republic	87%	100%	96%	83%	96%
Slovenia	98%	99%	95%	94%	94%
Sweden	95%	98%	93%	89%	91%
Turkey	100%	100%	97%	97%	97%
United States	63%	87%	96%	60%	84%

[1] Canada is represented by the provinces of Ontario and Quebec only

Exhibit 9.8: School and Students Participation Rates (Weighted)

Country	School Participation Before Replacement	School Participation After Replacement	Student Participation	Overall Participation Before Replacement	Overall Participation After Replacement
Argentina	89%	92%	91%	81%	84%
Belize	80%	80%	94%	75%	75%
Bulgaria	97%	97%	97%	93%	93%
Canada (O, Q)[1]	90%	97%	94%	85%	
Colombia	80%	98%	96%	76%	94%
Cyprus	98%	100%	97%	95%	97%
Czech Republic	90%	95%	94%	85%	90%
England	57%	87%	94%	54%	82%
France	93%	97%	97%	90%	94%
Germany	98%	98%	88%	86%	86%
Greece	78%	85%	97%	76%	82%
Hong Kong, SAR	73%	98%	99%	72%	97%
Hungary	98%	98%	97%	95%	95%
Iceland	95%	95%	87%	82%	82%
Iran, Islamic Rep. of	97%	100%	98%	95%	98%
Israel	96%	98%	96%	92%	94%
Italy	90%	100%	98%	88%	98%
Kuwait	87%	89%	91%	80%	81%
Latvia	89%	96%	93%	83%	89%
Lithuania	56%	97%	85%	47%	83%
Macedonia, Rep. of	97%	97%	97%	94%	94%
Moldova	84%	100%	96%	81%	96%
Morocco	74%	74%	93%	69%	69%
Netherlands	53%	89%	98%	52%	87%
New Zealand	94%	100%	96%	90%	96%
Norway	82%	89%	92%	76%	82%
Romania	96%	96%	97%	93%	93%
Russian Federation	100%	100%	97%	97%	97%
Scotland	76%	79%	95%	72%	74%
Singapore	100%	100%	98%	98%	98%
Slovak Republic	88%	100%	96%	84%	96%
Slovenia	98%	99%	95%	94%	94%
Sweden	97%	99%	93%	90%	92%
Turkey	100%	100%	97%	97%	97%
United States	61%	86%	96%	59%	83%

[1] Canada is represented by the provinces of Ontario and Quebec only

9.5 Trends in IEA's Reading Literacy Study

9.5.1 Overview

Because the data collection for PIRLS 2001 was scheduled 10 years after IEA's 1991 Reading Literacy Study, PIRLS 2001 provided an option for countries that participated in the earlier study to measure trends in their children's reading literacy since 1991 by readministering the 1991 Reading Literacy Test at the same time as the PIRLS assessment.

9.5.2 Target Population

The target population in 1991 was the grade with the greatest number of nine-year-olds at the time of testing, and, to maintain comparability, the same population was targeted by the Trends in IEA's Reading Literacy Study data collection in 2001. However, the PIRLS 2001 target population differs somewhat from the 1991 population in that PIRLS targeted the upper of the *two* grades with most nine-year-olds, and so the target grade in each country was not always the same for the two studies. These definitions yield the same target grade in Greece, Iceland, Italy, New Zealand, Slovenia, and the United States – but different ones in Hungary, Singapore, and Sweden. Average student ages ranged from 9.1 in Singapore to 10.2 in the United States. All definitions and quality criteria regarding the national desired and defined target populations (described in Chapter 5 and section 9.2), applied also to the Trends in IEA's Reading Literacy Study. Exhibit 9.9 provides the country's name for the grade tested, the corresponding number of years of formal schooling, and the average age of the students tested in each of the nine participating countries.

Exhibit 9.9: Countries Participating in the Trends in IEA's Reading Literacy Study

Country	Country's Name for Grade Tested	Years of Formal Schooling	Mean Age of Students Tested
Greece	4	4	9.9
Hungary	3	3	9.7
Iceland	4	4	9.8
Italy	4	4	9.9
New Zealand	Year 5[1]	4	10.0
Singapore	Primary 3	3	9.1
Slovenia	3	3	9.8
Sweden	3	3	9.8
United States	4	4	10.2

[1] The official nomenclature used in New Zealand since 1996 refers to students' years of schooling rather than a class/grade level. Year 5 students were at a class level equivalent to Grade 4.

9.5.3 Population Coverage and Exclusions

Exhibit 9.10 summarizes population coverage and exclusions for the Trends in IEA's Reading Literacy Study target populations. The national desired target population corresponded to 100 percent of the international desired target population in each country. The percentage of students excluded from testing because of disabilities was below the maximum permitted (10%) in all countries, and below 5 percent in all countries except Greece.

9.5.4 General Sampling Design

The basic idea behind the sampling approach for the Trends in IEA's Reading Literacy Study is rather simple: to select every second school sampled for PIRLS. From each of these selected schools, an additional classroom was sampled for the Trends in IEA's Reading Literacy Study. When there weren't enough classrooms in the sampled schools, PIRLS 2001 replace-

Exhibit 9.10: Population Coverage and Exclusions – Trends in IEA's Reading Literacy Study

Country	International Desired Population Coverage	National Desired Population		
		School-Level Exclusions	Within-Sample Exclusions	Overall Exclusions
Greece	100%	2.0%	4.0%	6.0%
Hungary	100%	1.8%	0.0%	1.8%
Iceland	100%	1.8%	2.0%	3.8%
Italy	100%	0.0%	3.4%	3.4%
New Zealand[1]	100%	1.6%	1.3%	2.9%
Singapore	100%	1.3%	0.0%	1.3%
Slovenia	100%	0.0%	0.9%	0.9%
Sweden	100%	2.5%	2.2%	4.7%
United States	100%	0.6%	3.9%	4.5%

[1] The Maori school stratum was not part of the study.

ment schools were used. When available, PIRLS 2001 replacement schools also became Trends in IEA's Reading Literacy Study replacement schools.

This approach was used for all countries, except in Hungary, where all sampled schools did both studies, and in Sweden, where no overlap of school samples was allowed. Summaries of the sample design for each country, including details of population coverage and exclusions, stratification variables, and participation rates, are provided in Appendix B.

9.5.5 Target Population Sizes

Exhibit 9.11 summarizes the number of schools and students in each country's target population, as well as the number of schools and students that participated in the Trends in IEA's Reading Literacy Study. Using the sampling weights computed for each country (see section 9.3), the Trends in IEA's Reading Literacy Study derived an estimate of the student population size, which matched closely the student population size from the sampling frame (see Exhibit 9.11).

Exhibit 9.11: Population and Sample Sizes – Trends in IEA's Reading Literacy Study

Country	Population		Sample			Mean Age
	Schools	Students	Schools	Students	Estimated Student Population	
Greece	4 999	102 927	68	1 109	92 290	9.9
Hungary	2 700	113 594	216	4 707	116 164	9.7
Iceland	140	4 566	65	1 797	4 478	9.8
Italy	7 162	573 571	92	1 590	520 379	9.9
New Zealand[1]	1 925	59 097	73	1 188	58 236	10.0
Singapore	196	50 586	98	3 601	48 566	9.1
Slovenia	443	21 906	75	1 502	22 093	9.8
Sweden	4 040	124 986	148	5 361	114 977	9.8
United States	71 498	3 871 487	85	1 826	3 856 987	10.2

[1] The Maori school stratum was not part of the study.

9.5.6 Sampling Weights and School and Student Participation Rates

Since the sample designs used for PIRLS 2001 and in the Trends in IEA's Reading Literacy studies are similar, the calculation of sampling weights was done in exactly the same way as described in section 9.3.

Participation rates for the Trends in IEA's Reading Literacy Study also were computed in the same way as for PIRLS. Exhibits 9.12 through 9.15 present the school, student, and overall participation rates, and the achieved sample sizes for each participating country. As can be seen from these exhibits, seven of the nine countries met the requirements described in Exhibit 9.4, and belong in Category 1. Because they met the sampling requirements only after including replacement schools, Greece and the United States belong in Category 2. Accordingly, the results for these countries were annotated with an obelisk in the achievement exhibits in the international report. No country was assigned to Category 3.

Exhibit 9.12: School Participation Rates and Sample Sizes – Trends in IEA's Reading Literacy Study

Country	School Participation Before Replacement (Weighted Percentage)	School Participation After Replacement (Weighted Percentage)	Number of Schools in Original Sample	Number of Eligible Schools in Original Sample	Number of Schools in Original Sample That Participated	Total Number of Schools That Participated
Greece	73%	79%	85	85	63	68
Hungary	98%	98%	220	220	216	216
Iceland	93%	93%	70	70	65	65
Italy	89%	100%	92	92	81	92
New Zealand	90%	98%	75	75	67	73
Singapore	100%	100%	98	98	98	98
Slovenia	100%	100%	75	75	75	75
Sweden	96%	100%	150	150	142	148
United States	58%	85%	100	100	54	85

Exhibit 9.13: Student Participation Rates and Sample Sizes – Trends in IEA's Reading Literacy Study

Country	Within School Student Participation (Weighted Percentage)	Number of Sampled Students in Participating Schools	Number of Students Withdrawn from Class/School	Number of Students Excluded	Number of Students Eligible	Number of Students Absent	Number of Students Assessed
Greece	97%	1 195	0	47	1 148	39	1 109
Hungary	97%	4 859	20	0	4 839	132	4 707
Iceland	86%	2 137	14	44	2 079	282	1 797
Italy	97%	1 697	6	56	1 635	45	1 590
New Zealand[1]	95%	1 308	43	19	1 246	58	1 188
Singapore	98%	3 729	46	0	3 683	82	3 601
Slovenia	95%	1 577	0	2	1 575	73	1 502
Sweden	96%	5 706	33	118	5 555	194	5 361
United States	95%	1 980	20	40	1 920	94	1 826

[1] The Maori school stratum was not part of the study.

Exhibit 9.14: School and Student Participation Rates (Weighted) – Trends in IEA's Reading Literacy Study

Country	School Participation Before Replacement	School Participation After Replacement	Student Participation	Overall Participation Before Replacement	Overall Participation After Replacement
Greece	73%	79%	97%	70%	77%
hungary	98%	98%	97%	96%	96%
Iceland	93%	93%	87%	80%	80%
Italy	89%	100%	97%	86%	97%
New Zealand[1]	90%	98%	95%	85%	93%
Singapore	100%	100%	98%	98%	98%
Slovenia	100%	100%	95%	95%	95%
Sweden	96%	100%	97%	93%	97%
United States	58%	85%	95%	55%	81%

[1] The Maori school stratum was not part of the study.

Exhibit 9.15: School and Student Participation Rates (Unweighted) – Trends in IEA's Reading Literacy Study

Country	School Participation Before Replacement	School Participation After Replacement	Student Participation	Overall Participation Before Replacement	Overall Participation After Replacement
Greece	74%	80%	97%	72%	77%
Hungary	98%	98%	97%	96%	96%
Iceland	93%	93%	86%	80%	80%
Italy	88%	100%	97%	86%	97%
New Zealand[1]	89%	97%	95%	85%	93%
Singapore	100%	100%	98%	98%	98%
Slovenia	100%	100%	95%	95%	95%
Sweden	95%	99%	97%	91%	95%
United States	54%	85%	95%	51%	81%

[1] The Maori school stratum was not part of the study.

10

Item Analysis and Review

Ina V.S. Mullis

Michael O. Martin

Ann M. Kennedy

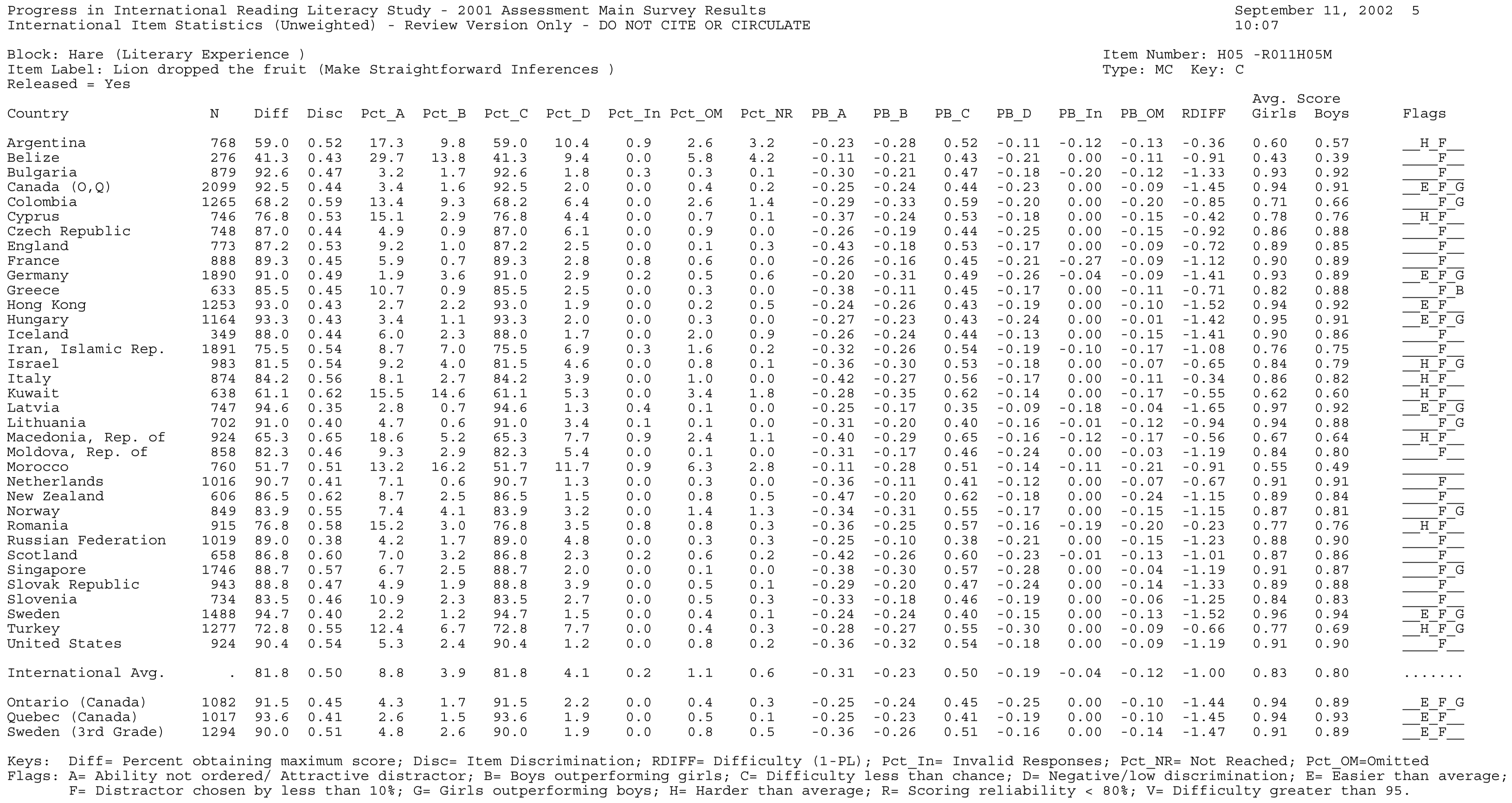
10.1 Overview

Prior to the item response theory (IRT) scaling of the PIRLS 2001 achievement scores, the International Study Center (ISC) reviewed a range of diagnostic statistics to examine and evaluate the psychometric characteristics of each achievement item within and across the 35 countries participating in PIRLS. For constructed-response items, the review included indicators of the reliability of the scoring procedure. The review process was an important step in the quality assurance of the PIRLS 2001 data, screening items for unusual psychometric properties that could signal a problem or error for an item in a particular country. For example, an item uncharacteristically easy or difficult in a country, or with an unusually low discriminating power, could indicate a potential problem with translation or printing. In the rare instances where such items were detected, the country's translation verification documents and printed booklets were examined for flaws or inaccuracies and the items removed from the database for that country. This chapter describes the basic item statistics that were consulted, and provides examples from the assessment to illustrate the review process.

10.2 Statistics for Item Analysis

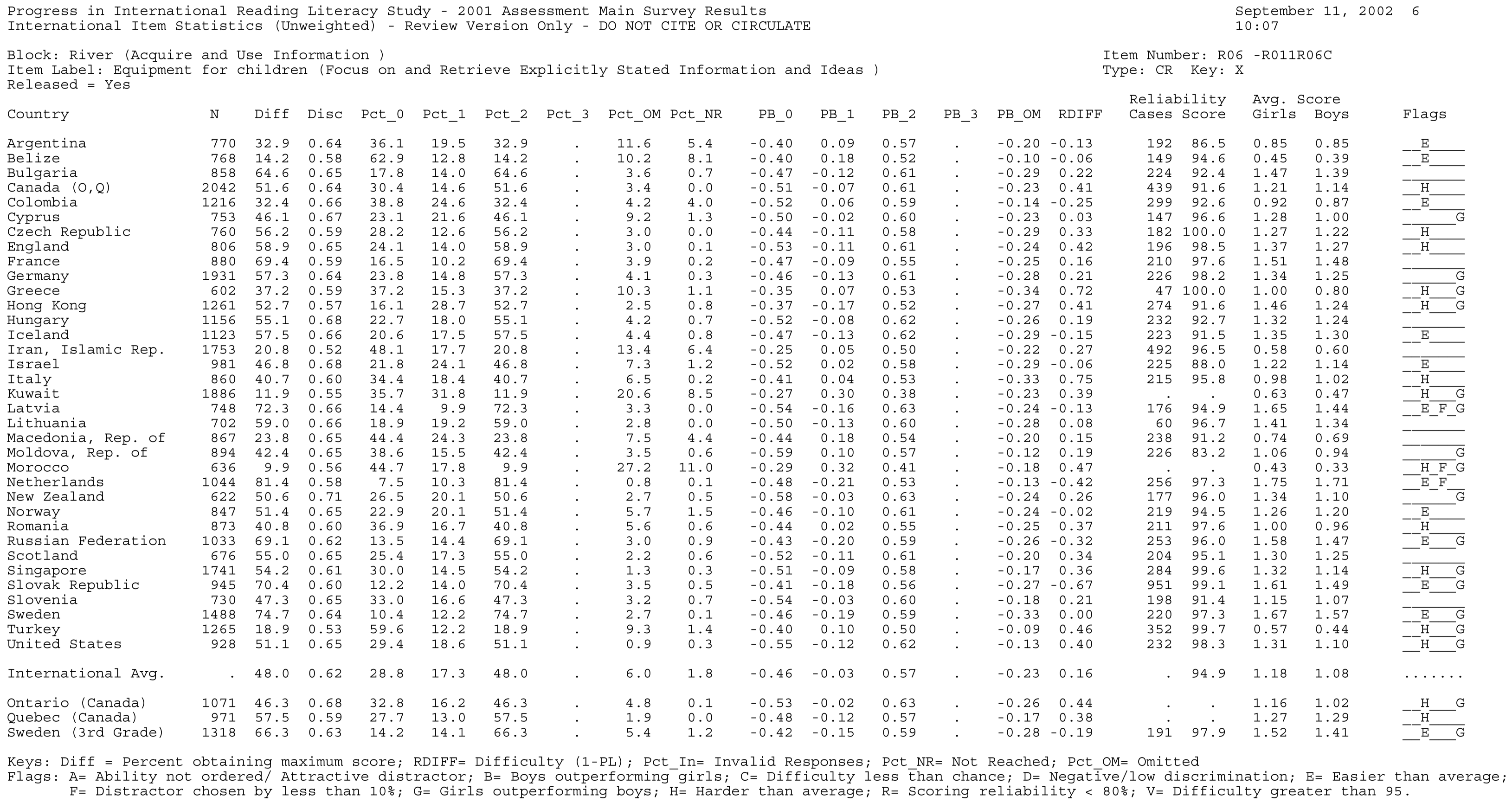
As the first stage in the item review process, the PIRLS ISC computed a set of item statistics for each achievement item, showing the properties of the item in each of the 35 countries participating in PIRLS 2001. Exhibits 10.1 and 10.2 show the statistics calculated for a multiple-choice and a constructed-response item, respectively. Statistics for each item are displayed alphabetically by country, with

Exhibit 10.1: International Item Statistics for Item R011H05M



```
Progress in International Reading Literacy Study - 2001 Assessment Main Survey Results                         September 11, 2002  5
International Item Statistics (Unweighted) - Review Version Only - DO NOT CITE OR CIRCULATE                    10:07

Block: Hare (Literary Experience )                                                                   Item Number: H05 -R011H05M
Item Label: Lion dropped the fruit (Make Straightforward Inferences )                                Type: MC  Key: C
Released = Yes
                                                                                                                 Avg. Score
Country                 N   Diff  Disc  Pct_A  Pct_B  Pct_C  Pct_D  Pct_In Pct_OM  Pct_NR  PB_A   PB_B   PB_C   PB_D   PB_In  PB_OM  RDIFF   Girls  Boys     Flags

Argentina             768   59.0  0.52   17.3    9.8   59.0   10.4    0.9    2.6    3.2   -0.23  -0.28   0.52  -0.11  -0.12  -0.13  -0.36   0.60   0.57    __H_F__
Belize                276   41.3  0.43   29.7   13.8   41.3    9.4    0.0    5.8    4.2   -0.11  -0.21   0.43  -0.21   0.00  -0.11  -0.91   0.43   0.39    ____F__
Bulgaria              879   92.6  0.47    3.2    1.7   92.6    1.8    0.3    0.3    0.1   -0.30  -0.21   0.47  -0.18  -0.20  -0.12  -1.33   0.93   0.92    ____F__
Canada (O,Q)         2099   92.5  0.44    3.4    1.6   92.5    2.0    0.0    0.4    0.2   -0.25  -0.24   0.44  -0.23   0.00  -0.09  -1.45   0.94   0.91    __E_F_G
Colombia             1265   68.2  0.59   13.4    9.3   68.2    6.4    0.0    2.6    1.4   -0.29  -0.33   0.59  -0.20   0.00  -0.20  -0.85   0.71   0.66    ____F_G
Cyprus                746   76.8  0.53   15.1    2.9   76.8    4.4    0.0    0.7    0.1   -0.37  -0.24   0.53  -0.18   0.00  -0.15  -0.42   0.78   0.76    __H_F__
Czech Republic        748   87.0  0.44    4.9    0.9   87.0    6.1    0.0    0.9    0.0   -0.26  -0.19   0.44  -0.25   0.00  -0.15  -0.92   0.86   0.88    ____F__
England               773   87.2  0.53    9.2    1.0   87.2    2.5    0.0    0.1    0.3   -0.43  -0.18   0.53  -0.17   0.00  -0.09  -0.72   0.89   0.85    ____F__
France                888   89.3  0.45    5.9    0.7   89.3    2.8    0.8    0.6    0.0   -0.26  -0.16   0.45  -0.21  -0.27  -0.09  -1.12   0.90   0.89    ____F__
Germany              1890   91.0  0.49    1.9    3.6   91.0    2.9    0.2    0.5    0.6   -0.20  -0.31   0.49  -0.26  -0.04  -0.09  -1.41   0.93   0.89    __E_F_G
Greece                633   85.5  0.45   10.7    0.9   85.5    2.5    0.0    0.3    0.0   -0.38  -0.11   0.45  -0.17   0.00  -0.11  -0.71   0.82   0.88    ____F_B
Hong Kong            1253   93.0  0.43    2.7    2.2   93.0    1.9    0.0    0.2    0.5   -0.24  -0.26   0.43  -0.19   0.00  -0.10  -1.52   0.94   0.92    __E_F__
Hungary              1164   93.3  0.43    3.4    1.1   93.3    2.0    0.0    0.3    0.0   -0.27  -0.23   0.43  -0.24   0.00  -0.01  -1.42   0.95   0.91    __E_F_G
Iceland               349   88.0  0.44    6.0    2.3   88.0    1.7    0.0    2.0    0.9   -0.26  -0.24   0.44  -0.13   0.00  -0.15  -1.41   0.90   0.86    ____F__
Iran, Islamic Rep.   1891   75.5  0.54    8.7    7.0   75.5    6.9    0.3    1.6    0.2   -0.32  -0.26   0.54  -0.19  -0.10  -0.17  -1.08   0.76   0.75    ____F__
Israel                983   81.5  0.54    9.2    4.0   81.5    4.6    0.0    0.8    0.1   -0.36  -0.30   0.53  -0.18   0.00  -0.07  -0.65   0.84   0.79    __H_F_G
Italy                 874   84.2  0.56    8.1    2.7   84.2    3.9    0.0    1.0    0.0   -0.42  -0.27   0.56  -0.17   0.00  -0.11  -0.34   0.86   0.82    __H_F__
Kuwait                638   61.1  0.62   15.5   14.6   61.1    5.3    0.0    3.4    1.8   -0.28  -0.35   0.62  -0.14   0.00  -0.17  -0.55   0.62   0.60    __H_F__
Latvia                747   94.6  0.35    2.8    0.7   94.6    1.3    0.4    0.1    0.0   -0.25  -0.17   0.35  -0.09  -0.18  -0.04  -1.65   0.97   0.92    __E_F_G
Lithuania             702   91.0  0.40    4.7    0.6   91.0    3.4    0.1    0.1    0.0   -0.31  -0.20   0.40  -0.16  -0.01  -0.12  -0.94   0.94   0.88    ____F_G
Macedonia, Rep. of    924   65.3  0.65   18.6    5.2   65.3    7.7    0.9    2.4    1.1   -0.40  -0.29   0.65  -0.16  -0.12  -0.17  -0.56   0.67   0.64    __H_F__
Moldova, Rep. of      858   82.3  0.46    9.3    2.9   82.3    5.4    0.0    0.1    0.0   -0.31  -0.17   0.46  -0.24   0.00  -0.03  -1.19   0.84   0.80    ____F__
Morocco               760   51.7  0.51   13.2   16.2   51.7   11.7    0.9    6.3    2.8   -0.11  -0.28   0.51  -0.14  -0.11  -0.21  -0.91   0.55   0.49    _______
Netherlands          1016   90.7  0.41    7.1    0.6   90.7    1.3    0.0    0.3    0.0   -0.36  -0.11   0.41  -0.12   0.00  -0.07  -0.67   0.91   0.91    ____F__
New Zealand           606   86.5  0.62    8.7    2.5   86.5    1.5    0.0    0.8    0.5   -0.47  -0.20   0.62  -0.18   0.00  -0.24  -1.15   0.89   0.84    ____F__
Norway                849   83.9  0.55    7.4    4.1   83.9    3.2    0.0    1.4    1.3   -0.34  -0.31   0.55  -0.17   0.00  -0.15  -1.15   0.87   0.81    ____F_G
Romania               915   76.8  0.58   15.2    3.0   76.8    3.5    0.8    0.8    0.3   -0.36  -0.25   0.57  -0.16  -0.19  -0.20  -0.23   0.77   0.76    __H_F__
Russian Federation   1019   89.0  0.38    4.2    1.7   89.0    4.8    0.0    0.3    0.3   -0.25  -0.10   0.38  -0.21   0.00  -0.15  -1.23   0.88   0.90    ____F__
Scotland              658   86.8  0.60    7.0    3.2   86.8    2.3    0.2    0.6    0.2   -0.42  -0.26   0.60  -0.23  -0.01  -0.13  -1.01   0.87   0.86    ____F__
Singapore            1746   88.7  0.57    6.7    2.5   88.7    2.0    0.0    0.1    0.0   -0.38  -0.30   0.57  -0.28   0.00  -0.04  -1.19   0.91   0.87    ____F_G
Slovak Republic       943   88.8  0.47    4.9    1.9   88.8    3.9    0.0    0.5    0.1   -0.29  -0.20   0.47  -0.24   0.00  -0.14  -1.33   0.89   0.88    ____F__
Slovenia              734   83.5  0.46   10.9    2.3   83.5    2.7    0.0    0.5    0.3   -0.33  -0.18   0.46  -0.19   0.00  -0.06  -1.25   0.84   0.83    ____F__
Sweden               1488   94.7  0.40    2.2    1.2   94.7    1.5    0.0    0.4    0.1   -0.24  -0.24   0.40  -0.15   0.00  -0.13  -1.52   0.96   0.94    __E_F_G
Turkey               1277   72.8  0.55   12.4    6.7   72.8    7.7    0.0    0.4    0.3   -0.28  -0.27   0.55  -0.30   0.00  -0.09  -0.66   0.77   0.69    __H_F_G
United States         924   90.4  0.54    5.3    2.4   90.4    1.2    0.0    0.8    0.2   -0.36  -0.32   0.54  -0.18   0.00  -0.09  -1.19   0.91   0.90    ____F__

International Avg.      .   81.8  0.50    8.8    3.9   81.8    4.1    0.2    1.1    0.6   -0.31  -0.23   0.50  -0.19  -0.04  -0.12  -1.00   0.83   0.80    .......

Ontario (Canada)     1082   91.5  0.45    4.3    1.7   91.5    2.2    0.0    0.4    0.3   -0.25  -0.24   0.45  -0.25   0.00  -0.10  -1.44   0.94   0.89    __E_F_G
Quebec (Canada)      1017   93.6  0.41    2.6    1.5   93.6    1.9    0.0    0.5    0.1   -0.25  -0.23   0.41  -0.19   0.00  -0.10  -1.45   0.94   0.93    __E_F__
Sweden (3rd Grade)   1294   90.0  0.51    4.8    2.6   90.0    1.9    0.0    0.8    0.5   -0.36  -0.26   0.51  -0.16   0.00  -0.14  -1.47   0.91   0.89    __E_F__

Keys:  Diff= Percent obtaining maximum score; Disc= Item Discrimination; RDIFF= Difficulty (1-PL); Pct_In= Invalid Responses; Pct_NR= Not Reached; Pct_OM=Omitted
Flags: A= Ability not ordered/ Attractive distractor; B= Boys outperforming girls; C= Difficulty less than chance; D= Negative/low discrimination; E= Easier than average;
       F= Distractor chosen by less than 10%; G= Girls outperforming boys; H= Harder than average; R= Scoring reliability < 80%; V= Difficulty greater than 95.
```

Exhibit 10.2: International Item Statistics for Item R011R06C



```
Progress in International Reading Literacy Study - 2001 Assessment Main Survey Results                                                   September 11, 2002  6
International Item Statistics (Unweighted) - Review Version Only - DO NOT CITE OR CIRCULATE                                              10:07

Block: River (Acquire and Use Information )                                                                                      Item Number: R06 -R011R06C
Item Label: Equipment for children (Focus on and Retrieve Explicitly Stated Information and Ideas )                              Type: CR  Key: X
Released = Yes
                                                                                                                                   Reliability   Avg. Score
Country                   N    Diff  Disc  Pct_0  Pct_1  Pct_2  Pct_3  Pct_OM Pct_NR    PB_0    PB_1    PB_2    PB_3  PB_OM  RDIFF   Cases Score   Girls  Boys        Flags

Argentina               770  32.9  0.64   36.1   19.5   32.9    .     11.6    5.4    -0.40   0.09    0.57    .    -0.20 -0.13     192  86.5    0.85   0.85       __E____
Belize                  768  14.2  0.58   62.9   12.8   14.2    .     10.2    8.1    -0.40   0.18    0.52    .    -0.10 -0.06     149  94.6    0.45   0.39       __E____
Bulgaria                858  64.6  0.65   17.8   14.0   64.6    .      3.6    0.7    -0.47  -0.12    0.61    .    -0.29  0.22     224  92.4    1.47   1.39       _______
Canada (O,Q)           2042  51.6  0.64   30.4   14.6   51.6    .      3.4    0.0    -0.51  -0.07    0.61    .    -0.23  0.41     439  91.6    1.21   1.14       __H____
Colombia               1216  32.4  0.66   38.8   24.6   32.4    .      4.2    4.0    -0.52   0.06    0.59    .    -0.14 -0.25     299  92.6    0.92   0.87       __E____
Cyprus                  753  46.1  0.67   23.1   21.6   46.1    .      9.2    1.3    -0.50  -0.02    0.60    .    -0.23  0.03     147  96.6    1.28   1.00       ______G
Czech Republic          760  56.2  0.59   28.2   12.6   56.2    .      3.0    0.0    -0.44  -0.11    0.58    .    -0.29  0.33     182 100.0    1.27   1.22       __H____
England                 806  58.9  0.65   24.1   14.0   58.9    .      3.0    0.1    -0.53  -0.11    0.61    .    -0.24  0.42     196  98.5    1.37   1.27       __H____
France                  880  69.4  0.59   16.5   10.2   69.4    .      3.9    0.2    -0.47  -0.09    0.55    .    -0.25  0.16     210  97.6    1.51   1.48       _______
Germany                1931  57.3  0.64   23.8   14.8   57.3    .      4.1    0.3    -0.46  -0.13    0.61    .    -0.28  0.21     226  98.2    1.34   1.25       ______G
Greece                  602  37.2  0.59   37.2   15.3   37.2    .     10.3    1.1    -0.35   0.07    0.53    .    -0.34  0.72      47 100.0    1.00   0.80       __H___G
Hong Kong              1261  52.7  0.57   16.1   28.7   52.7    .      2.5    0.8    -0.37  -0.17    0.52    .    -0.27  0.41     274  91.6    1.46   1.24       __H___G
Hungary                1156  55.1  0.68   22.7   18.0   55.1    .      4.2    0.7    -0.52  -0.08    0.62    .    -0.26  0.19     232  92.7    1.32   1.24       _______
Iceland                1123  57.5  0.66   20.6   17.5   57.5    .      4.4    0.8    -0.47  -0.13    0.62    .    -0.29 -0.15     223  91.5    1.35   1.30       __E____
Iran, Islamic Rep.     1753  20.8  0.52   48.1   17.7   20.8    .     13.4    6.4    -0.25   0.05    0.50    .    -0.22  0.27     492  96.5    0.58   0.60       _______
Israel                  981  46.8  0.68   21.8   24.1   46.8    .      7.3    1.2    -0.52   0.02    0.58    .    -0.29 -0.06     225  88.0    1.22   1.14       __E____
Italy                   860  40.7  0.60   34.4   18.4   40.7    .      6.5    0.2    -0.41   0.04    0.53    .    -0.33  0.75     215  95.8    0.98   1.02       __H____
Kuwait                 1886  11.9  0.55   35.7   31.8   11.9    .     20.6    8.5    -0.27   0.30    0.38    .    -0.23  0.39       .     .    0.63   0.47       __H___G
Latvia                  748  72.3  0.66   14.4    9.9   72.3    .      3.3    0.0    -0.54  -0.16    0.63    .    -0.24 -0.13     176  94.9    1.65   1.44       __E_F_G
Lithuania               702  59.0  0.66   18.9   19.2   59.0    .      2.8    0.0    -0.50  -0.13    0.60    .    -0.28  0.08      60  96.7    1.41   1.34       _______
Macedonia, Rep. of      867  23.8  0.65   44.4   24.3   23.8    .      7.5    4.4    -0.44   0.18    0.54    .    -0.20  0.15     238  91.2    0.74   0.69       _______
Moldova, Rep. of        894  42.4  0.65   38.6   15.5   42.4    .      3.5    0.6    -0.59   0.10    0.57    .    -0.12  0.19     226  83.2    1.06   0.94       ______G
Morocco                 636   9.9  0.56   44.7   17.8    9.9    .     27.2   11.0    -0.29   0.32    0.41    .    -0.18  0.47       .     .    0.43   0.33       __H_F_G
Netherlands            1044  81.4  0.58    7.5   10.3   81.4    .      0.8    0.1    -0.48  -0.21    0.53    .    -0.13 -0.42     256  97.3    1.75   1.71       __E_F__
New Zealand             622  50.6  0.71   26.5   20.1   50.6    .      2.7    0.5    -0.58  -0.03    0.63    .    -0.24  0.26     177  96.0    1.34   1.10       ______G
Norway                  847  51.4  0.65   22.9   20.1   51.4    .      5.7    1.5    -0.46  -0.10    0.61    .    -0.24 -0.02     219  94.5    1.26   1.20       __E____
Romania                 873  40.8  0.60   36.9   16.7   40.8    .      5.6    0.6    -0.44   0.02    0.55    .    -0.25  0.37     211  97.6    1.00   0.96       __H____
Russian Federation     1033  69.1  0.62   13.5   14.4   69.1    .      3.0    0.9    -0.43  -0.20    0.59    .    -0.26 -0.32     253  96.0    1.58   1.47       __E___G
Scotland                676  55.0  0.65   25.4   17.3   55.0    .      2.2    0.6    -0.52  -0.11    0.61    .    -0.20  0.34     204  95.1    1.30   1.25       _______
Singapore              1741  54.2  0.61   30.0   14.5   54.2    .      1.3    0.3    -0.51  -0.09    0.58    .    -0.17  0.36     284  99.6    1.32   1.14       __H___G
Slovak Republic         945  70.4  0.60   12.2   14.0   70.4    .      3.5    0.5    -0.41  -0.18    0.56    .    -0.27 -0.67     951  99.1    1.61   1.49       __E___G
Slovenia                730  47.3  0.65   33.0   16.6   47.3    .      3.2    0.7    -0.54  -0.03    0.60    .    -0.18  0.21     198  91.4    1.15   1.07       _______
Sweden                 1488  74.7  0.64   10.4   12.2   74.7    .      2.7    0.1    -0.46  -0.19    0.59    .    -0.33  0.00     220  97.3    1.67   1.57       __E___G
Turkey                 1265  18.9  0.53   59.6   12.2   18.9    .      9.3    1.4    -0.40   0.10    0.50    .    -0.09  0.46     352  99.7    0.57   0.44       __H___G
United States           928  51.1  0.65   29.4   18.6   51.1    .      0.9    0.3    -0.55  -0.12    0.62    .    -0.13  0.40     232  98.3    1.31   1.10       __H___G

International Avg.        .  48.0  0.62   28.8   17.3   48.0    .      6.0    1.8    -0.46  -0.03    0.57    .    -0.23  0.16       .  94.9    1.18   1.08       .......

Ontario (Canada)       1071  46.3  0.68   32.8   16.2   46.3    .      4.8    0.1    -0.53  -0.02    0.63    .    -0.26  0.44       .     .    1.16   1.02       __H___G
Quebec (Canada)         971  57.5  0.59   27.7   13.0   57.5    .      1.9    0.0    -0.48  -0.12    0.57    .    -0.17  0.38       .     .    1.27   1.29       __H____
Sweden (3rd Grade)     1318  66.3  0.63   14.2   14.1   66.3    .      5.4    1.2    -0.42  -0.15    0.59    .    -0.28 -0.19     191  97.9    1.52   1.41       __E___G

Keys: Diff = Percent obtaining maximum score; RDIFF= Difficulty (1-PL); Pct_In= Invalid Responses; Pct_NR= Not Reached; Pct_OM= Omitted
Flags: A= Ability not ordered/ Attractive distractor; B= Boys outperforming girls; C= Difficulty less than chance; D= Negative/low discrimination; E= Easier than average;
       F= Distractor chosen by less than 10%; G= Girls outperforming boys; H= Harder than average; R= Scoring reliability < 80%; V= Difficulty greater than 95.
```

the international average for each statistic at the bottom. For countries testing in more than one language, statistics are presented separately by language group. For all items, regardless of item format, statistics included the number of students in each country that responded, the difficulty level (the percentage of students that answered the item correctly), and the discrimination index (the point-biserial correlation between success on the item and a total score).[1] Also provided is an estimate of the item's difficulty using a Rasch one-parameter IRT model. The international means of the item difficulties and item discriminations serve as guides to the overall statistical properties of the items.

For multiple-choice items, statistics included the percentage of students that chose each option, as well as the percentage of students that omitted or did not reach the item, and the point-biserial correlation between the response to each option and the total score. For constructed-response items (which could have one, two, or three score levels) statistics included the difficulty and discrimination of each score level. Constructed-response item displays also provide information about the reliability with which the item was scored in each country, with the total number of double-scored cases and the percent exact agreement between the scorers.

For all items, the item-analysis includes the average score for male and female students. This is the average score received by girls and boys on a scale ranging from zero to the maximum possible score point for the item. For multiple-choice items or 1-point constructed-response items, this statistic also represents the average difficulty of the item for girls and boys.

Detailed descriptions of the statistics provided in Exhibits 10.1 and 10.2 are listed below in order of appearance in the displays:

N: This is the number of students to whom the item was administered. If a student did not reach an item in the achievement booklet, the item was considered "Not Administered" for the purpose of the item analysis.[2]

Diff: Item difficulty is the percentage of students providing a fully correct response to the item. In the case of constructed-response items worth more than one point, this is the percentage of students receiving the maximum score. For the computation of this statistic, "Not Reached" items were treated as "Not Administered".

[1] For the purpose of computing the discrimination index, the total score was the percentage of the items presented that a student answered correctly.

[2] In calculating item statistics and in item parameter estimation for scaling, items not reached by a student were treated as if they had not been administered. In estimating student proficiency, however, not reached items were treated as answered incorrectly.

Disc: Item discrimination is the correlation between a correct response to the item and the total score on all of the items in the test booklet.[3] Items exhibiting good measurement properties should have a moderately positive correlation.

Pct_A, Pct_B, Pct_C, and Pct_D: Used for multiple-choice items only (see Exhibit 10.1), each column indicates the percentage of students choosing the particular response option for the item (A, B, C, or D). Not-reached items were excluded from the denominator for these calculations.

Pct_0, Pct_1, Pct_2, and Pct_3: Used for constructed-response items only (see Exhibit 10.2), each column indicates the percentage of students scoring at the particular score level, up to and including the maximum score level for the item. Not-reached items were excluded from the denominator for these calculations.

Pct_In: Used for multiple-choice items only, this is the percentage of students that provided an invalid response to a multiple-choice item. Typically, invalid responses were the result of students selecting more than one response option for the same item.

Pct_OM: This is the percentage of students who, having reached the item, did not provide a response. Not reached items were excluded from the denominator when calculating this statistic.

Pct_NR: This is the percentage of students that did not reach the item in their booklets. An item was coded as not reached when there was no evidence of a response to any subsequent items in the booklet and the response to the item preceding it was omitted.

PB_A, PB_B, PB_C, and PB_D: Used for multiple-choice items only, these are the correlation between choosing each of the response options A, B, C, or D and the total score. Items with good psychometric properties have near-zero or negative correlations for the distracter options (the incorrect options) and moderately positive correlations for the correct option.

PB_0, PB_1, PB_2, and PB_3: Used for constructed-response items only, these present the correlation between the score levels on the item (0, 1, 2, or 3) and the score on the test booklet. For items with good measurement properties, the correlation coefficients should change from negative to positive as the score level increases.

[3] For constructed-response items, the discrimination is the correlation between the number of score points and total score.

PB_OM: This is the correlation between a binary variable – indicating an omitted response to the item – and the total score. This correlation should be negative or near zero.

PB_In: Used for multiple-choice items only, this presents the correlation between an invalid response to the item (usually caused by selecting more than one response option) and the total score. This correlation also should be negative or near zero.

RDIFF: This is an estimate of the item's difficulty based on a Rasch one-parameter IRT model. The difficulty estimate is expressed in the logit metric (with a positive logit indicating a difficult item) and was scaled so that the average Rasch item difficulty was zero within each country.

Reliability – Cases:
To provide a measure of the reliability of the scoring of the constructed-response items, those items in approximately one-quarter of the test booklets in each country were scored by two independent scorers. This column indicates the number of times the item was double-scored in each country.

Reliability – Score:
This column contains the percentage of exact agreement between the two independent scorers.

As an aid to reviewers, the item-analysis display includes a series of "flags" signaling the presence of one or more conditions that might indicate a problem with an item. The following conditions are flagged for each country:

- Item difficulty exceeds 95 percent
- Item difficulty is less than 25 percent for four-option multiple-choice items
- One or more of the distracter percentages is less than 10 percent
- One or more of the distracter percentages is greater than the percentage for the correct answer, or the point-biserial correlation for one or more of the distracters exceeds zero
- Item discrimination (i.e., the point-biserial for the correct answer) is less than 0.2
- Item discrimination does not increase with each score level (for constructed-response items with more than one score level)
- The Rasch difficulty estimate is above the average across all items

- The Rasch difficulty estimate is below the average across all items

- Difficulty levels on the item differ significantly for males and females

- Scoring reliability is less than 80 percent (for constructed-response items only).

Although not all of these conditions necessarily indicate a problem, the flags are a useful way to draw attention to potential sources of concern.

10.2.1 Item-by-Country Interaction

Although countries are expected to exhibit some variation in performance across items, in general, countries with high average performance on the achievement test as a whole should perform relatively well on each of the items, and low-scoring countries should do less well on each of items. When this does not occur (i.e., when a high-scoring country has low performance on an item on which other countries are doing well), there is said to be an item-by-country interaction. When large, such item-by-country interactions may be a sign of an item that is flawed in some way, and measures should be taken to address the problem.

To assist in detecting sizeable item-by-country interactions, the International Study Center produced a graphical display for each item showing the average probability across all countries of a correct response for a student of average proficiency internationally, compared with the probability of a correct response by a student of average proficiency in each country. Exhibit 10.3 provides an example of a PIRLS item-by-country interaction display.

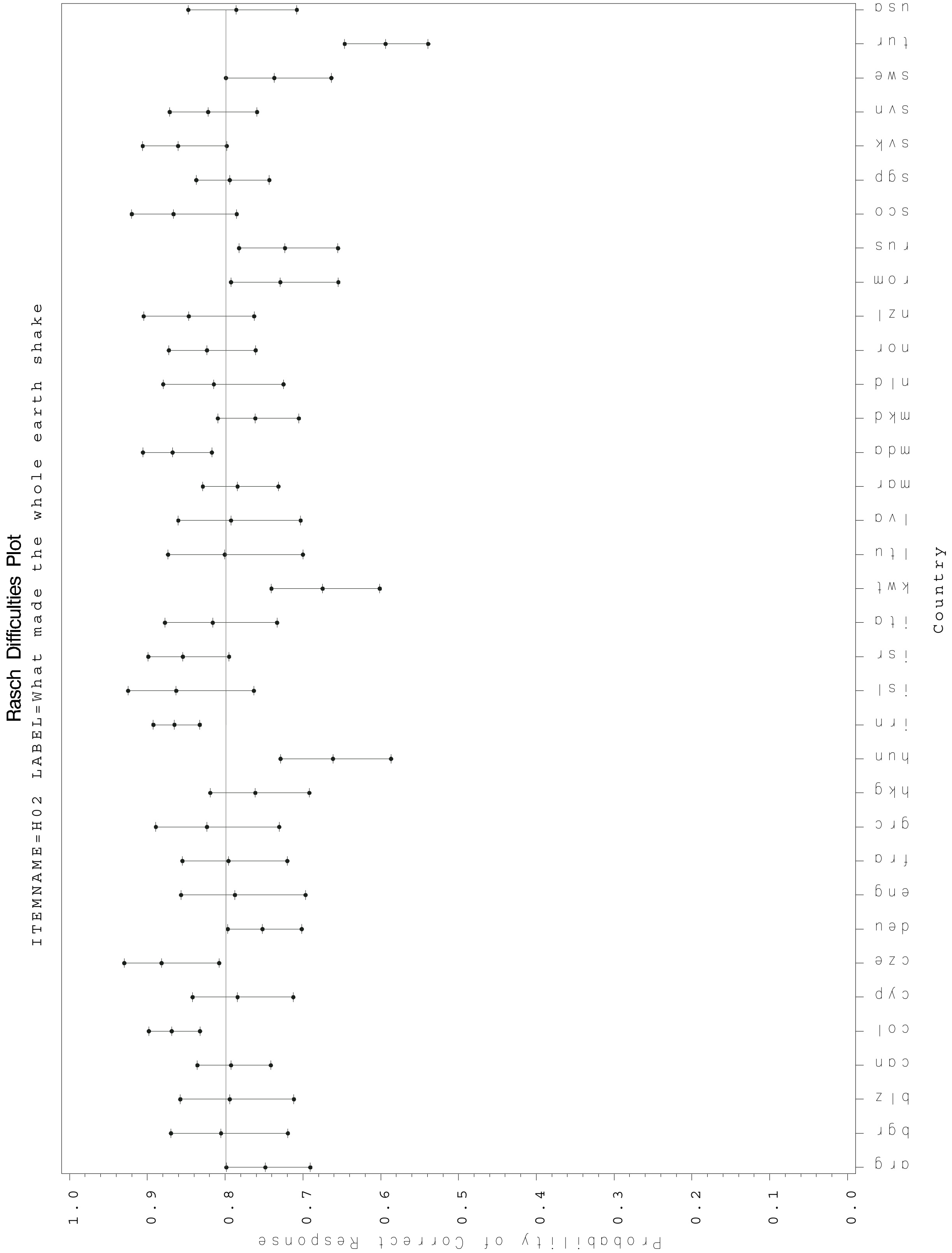
The probability for each country is presented as a 95 percent confidence interval, which includes a built-in Bonferroni correction for multiple comparisons. The limits for the confidence interval are computed as follows:

$$\text{UpperLimit} = \bar{1} - \frac{e^{RDIFF_{ik} - SE_{RDIFF_{ik}} * Z_b}}{1 + e^{RDIFF_{ik} - SE_{RDIFF_{ik}} * Z_b}}$$

$$\text{LowerLimit} = \bar{1} - \frac{\mathrm{e}^{\mathrm{RDIFF}_{\mathrm{ik}} + SE_{RDIFF_{ik}} * Z_b}}{1 + e^{RDIFF_{ik} + SE_{RDIFF_{ik}} * Z_b}}$$

where $RDIFF_{ik}$ is the Rasch difficulty of item *k* within country i; $SE_{RDIFF_{ik}}$ is the standard error of the difficulty of item *k* in country *i*; and Z_b is the critical value from the *Z* distribution, corrected for multiple comparisons using the Bonferroni procedure.

Exhibit 10.3: Example Item-by-Country Interaction Display for Item R011H02M



10.3 Scoring Reliability for Constructed-Response Items

Half of the items in the PIRLS assessment were constructed-response items, comprising nearly two-thirds of the score points for the assessment.[4] An essential requirement for use of such items is that they be reliably scored by all participants. That is, a particular student response should receive the same score, regardless of the scorer. In conducting PIRLS, measures taken to ensure that the constructed-response items were scored reliably in all countries included developing scoring guides for each constructed-response question (which provided descriptions of acceptable responses for each score point value),[5] and providing extensive training in the application of the scoring guides. Scoring procedures for organizing and monitoring the scoring sessions were outlined in the PIRLS *Survey Operations Manual*.

10.3.1 Within-Country Scoring Reliability

To gather and document information about the agreement among scorers, a random sample of at least 200 students' responses to each item (approximately 25% of the total responses) was selected by the National Research Coordinators to be scored independently by two scorers. A measure of agreement between scorers (the percentage of times the scores of the two scorers agreed exactly) was calculated for each item in each country, and was examined as part of the item review process. Items with percentage agreement less than 70 percent were flagged for further examination. The average and range of the exact percent of agreement across all items is presented (Exhibit 10.4) for each country. The average of exact percent agreement across items was high – on average, across countries, exact percent agreement was 93 percent. All countries had an average exact percent agreement above 83 percent.

[4] For details on the development of the PIRLS assessment items, see Chapter 2.

[5] Discussion of the development of the scoring guides for constructed-response items is provided in Chapter 2.

Exhibit 10.4: Within-Country Constructed-Response Scoring Reliability

Countries	Correctness Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent of Agreement	
		Minimum	Maximum
Argentina	86	71	95
Belize	92	86	97
Bulgaria	83	60	99
Canada (O,Q)	87	66	99
Colombia	83	65	100
Cyprus	96	86	100
Czech Republic	97	82	100
England	96	81	100
France	96	87	100
Germany	89	71	100
Greece	98	92	100
Hong Kong, SAR	88	61	97
Hungary	94	80	100
Iceland	86	70	99
Iran, Islamic Rep. of	95	90	99
Israel	91	83	97
Italy	94	68	100
Kuwait	–	–	–
Latvia	92	64	99
Lithuania	88	68	100
Macedonia, Rep. of	94	85	98
Moldova, Rep. of	94	83	99
Morocco	–	–	–
Netherlands	90	67	100
New Zealand	97	89	100
Norway	92	81	99
Romania	94	76	100
Russian Federation	98	91	100
Scotland	93	76	100
Singapore	99	98	100
Slovak Republic	99	99	100
Slovenia	92	67	100
Sweden	94	86	100
Turkey	99	98	100
United States	97	89	100
International Avg.	93	79	99

* A dash (–) indicates data not available

10.3.2 Cross-Country Scoring Reliability Study

To gather information about how consistently the scoring guides were applied across countries, the International Study Center conducted a cross-country reliability study in which a sample of student responses was scored independently by two English-proficient scorers from each participating country. Taking into consideration available resources and other feasibility issues, the cross-country scoring reliability study was conducted in English, using a core set of 200 student responses to each of 25 constructed-response questions from half of the assessment blocks – two literary and two informational.

The core set of 5,000 responses comprised student responses from Canada, England, Scotland, and the United States. A total of 55 scorers from 28 PIRLS countries participated in the study.[6] Scoring for this study took place shortly after the within-country scoring reliability activities were completed. Using the same scoring guides from the national within-country scoring activities, each scorer was asked to assign a score to each student response in the set. Each student response to an individual question resulted in 1,485 possible comparisons among scorers. When aggregated across all 200 student responses to the item, there were a total of 297,000 comparisons, provided a score was assigned by all 55 scorers.

Exhibit 10.5 shows the percentage of paired scorers that were in exact agreement across all responses to each of the items used in the reliability study. The extent of agreement varied across items. On average, across all items, 85 percent of all possible paired-scorer combinations were in exact agreement on the assigned score.

[6] Only one scorer proficient in English was available in Macedonia. In the Russian Federation, resources permitted only a portion of the English-language responses to be scored.

Exhibit 10.5: Cross-Country Constructed-Response Scoring Reliability

Purpose	Item Label	Total Valid Comparisons*	Exact Percent Agreement
Literary Experience	Unreleased C01	275496	99%
	Unreleased C02	275444	89%
	Unreleased C03	275548	93%
	Unreleased C06	275341	98%
	Unreleased C08	275496	92%
	Unreleased C10	275548	66%
	Unreleased C11	275444	72%
	Hare H03	275600	90%
	Hare H04	275393	93%
	Hare H07	275444	79%
	Hare H08	275086	84%
	Hare H09	275236	84%
	Hare H10	273661	73%
Acquire and Use Information	Unreleased A01	296892	96%
	Unreleased A03	296676	98%
	Unreleased A04	296676	90%
	Unreleased A07	296892	87%
	Unreleased A08	296623	80%
	Unreleased A09	296784	81%
	Unreleased A11	296191	80%
	Pufflings N07	274724	78%
	Pufflings N08	274724	83%
	Pufflings N10	273947	84%
	Pufflings N12	274673	76%
	Pufflings N13	274621	73%
		Average Percent Agreement	85%

* Values for items differ slightly due to a small number of missing responses.

10.4 Item Analysis for the Trends in IEA's Reading Literacy Study

The review of the item statistics for each of the nine countries participating in the Trends in IEA's Reading Literacy Study followed the PIRLS approach. Statistics calculated for the trend study items were the same as those used in PIRLS (as described in Section 10.2). An example item statistics display for a trend study item is presented in Exhibit 10.6. Different from the PIRLS item statistics, the trend item statistics include countries' statistics for both 1991 and 2001. In reviewing the item statistics, comparisons in performance were made across countries within a year, as well as within countries across years.

Exhibit 10.6: International Item Statistics for Trend Item E25

```
 Progress in International Reading Literacy Study - 2001 (10YTS) Results                                                                   September 24, 2002  7
International Item Statistics (Unweighted) - Review Version                                                                                17:49
For Internal Review Only: DO NOT CITE OR CIRCULATE

Reporting Category: Expository Prose (WALRUS )                                                     Item Number: E25 -AEWALRU5
Item Label: HOW WALRUS GETS UP ON ICE                                                              Type: MC  Key: C

* * * * * 1991 * * * * *                                                                                                                     Avg. Score
Country                N    Diff  Disc  Pct_A  Pct_B  Pct_C  Pct_D  Pct_In Pct_OM  Pct_NR  PB_A   PB_B   PB_C   PB_D   PB_In  PB_OM  RDIFF   Girls  Boys      Flags
Greece                3373  75.0  0.34    5.3    2.8   75.0   15.6    0.0    1.2    3.8    -0.25  -0.16   0.34  -0.12   0.00  -0.08  -0.39   0.74   0.76      __E_F._
Hungary               2307  75.6  0.35    6.4    1.4   75.6   15.9    0.0    0.7   22.4    -0.25  -0.09   0.35  -0.13   0.00  -0.04  -0.19   0.75   0.76      ____F._
Iceland               3471  82.7  0.35    4.2    0.7   82.7   11.7    0.0    0.6   12.1    -0.24  -0.13   0.35  -0.11   0.00  -0.02  -0.58   0.86   0.79      __E_F.G
Italy                 2097  73.3  0.23    1.9    1.1   73.3   23.1    0.0    0.5    5.6    -0.19  -0.17   0.23  -0.02   0.00  -0.11   0.12   0.77   0.70      __H_F.G
New Zealand           2859  74.3  0.44    6.5    1.9   74.3   16.9    0.0    0.4    4.3    -0.29  -0.22   0.44  -0.21   0.00   0.00  -0.05   0.77   0.72      __H_F.G
Singapore             7277  69.2  0.38    7.7    1.2   69.2   21.8    0.0    0.1    0.7    -0.27  -0.18   0.38  -0.20   0.00   0.00  -0.06   0.71   0.67      __H_F.G
Slovenia              2932  68.1  0.42    6.3    1.9   68.1   23.3    0.0    0.3   10.6    -0.24  -0.15   0.42  -0.21   0.00  -0.05  -0.05   0.71   0.65      __H_F.G
Sweden                3688  88.8  0.34    3.1    0.4   88.8    7.3    0.0    0.4   13.8    -0.21  -0.14   0.34  -0.15   0.00  -0.04  -0.64   0.91   0.87      __E_F.G
United States         6278  80.6  0.41    5.9    1.2   80.6   11.8    0.0    0.5    1.9    -0.29  -0.17   0.41  -0.22   0.00  -0.07  -0.33   0.81   0.80      ____F._

International Avg.       .  76.4  0.36    5.3    1.4   76.4   16.4    0.0    0.5    8.4    -0.25  -0.16   0.36  -0.15   0.00  -0.05  -0.24   0.78   0.75      .......

* * * * * 2001 * * * * *                                                                                                                     Avg. Score
Country                N    Diff  Disc  Pct_A  Pct_B  Pct_C  Pct_D  Pct_In Pct_OM  Pct_NR  PB_A   PB_B   PB_C   PB_D   PB_In  PB_OM  RDIFF   Girls  Boys      Flags
Greece                1061  77.7  0.30    9.0    1.0   77.7   11.5    0.0    0.8    4.3    -0.30  -0.08   0.30  -0.07   0.00  -0.04  -0.21   0.78   0.77      ____F._
Hungary               3750  82.0  0.29    4.8    1.7   82.0   10.7    0.0    0.7   20.3    -0.23  -0.19   0.29  -0.11   0.00  -0.02  -0.59   0.81   0.83      __E_F.B
Iceland               1663  84.9  0.29    3.1    0.9   84.9   10.6    0.0    0.4    7.4    -0.24  -0.16   0.29  -0.11   0.00  -0.11  -0.62   0.85   0.85      __E_F._
Italy                 1505  67.8  0.21    2.1    0.5   67.8   28.8    0.0    0.7    5.3    -0.19  -0.11   0.21  -0.07   0.00  -0.06   0.46   0.71   0.65      __H_F.G
New Zealand           1157  73.9  0.46    7.1    1.7   73.9   16.4    0.0    0.9    2.5    -0.33  -0.27   0.46  -0.20   0.00  -0.07   0.04   0.76   0.71      __H_F.G
Singapore             3564  68.4  0.42    9.3    1.7   68.4   20.5    0.0    0.1    0.9    -0.30  -0.18   0.42  -0.21   0.00  -0.02  -0.05   0.72   0.65      __H_F.G
Slovenia              1415  74.0  0.36    5.0    1.3   74.0   19.4    0.0    0.3    5.8    -0.26  -0.11   0.36  -0.22   0.00  -0.06  -0.12   0.76   0.72      ____F._
Sweden                4488  86.1  0.31    4.5    0.9   86.1    7.8    0.0    0.8   14.5    -0.29  -0.13   0.31  -0.15   0.00  -0.05  -0.52   0.86   0.86      __E_F._
United States         1817  80.5  0.40    5.2    1.0   80.5   12.9    0.0    0.4    0.4    -0.27  -0.18   0.40  -0.23   0.00  -0.03  -0.28   0.81   0.80      ____F._

International Avg.       .  77.3  0.34    5.6    1.2   77.3   15.4    0.0    0.6    6.8    -0.27  -0.16   0.34  -0.15   0.00  -0.05  -0.21   0.79   0.76      .......


Keys:  Diff= Percent obtaining maximum score; Disc= Item Discrimination; RDIFF= Difficulty (1-PL); Pct_In= Invalid Responses; Pct_NR= Not Reached; Pct_OM=Omitted
Flags: A= Ability not ordered/ Attractive distractor; B= Boys outperforming girls; C= Difficulty less than chance; D= Negative/low discrimination; E= Easier than average;
       F= Distractor chosen by less than 10%; G= Girls outperforming boys; H= Harder than average; V= Difficulty greater than 95.
```

10.4.1 Item-by-Country Interactions for the Trends in IEA's Reading Literacy Study

The international Study Center also produced item-by-country interaction displays for each item in the trend study, showing the results from 1991 and 2001 separately in each display. An example of an item-by-country interaction display for a trend item is presented in Exhibit 10.7. Confidence intervals for 1991 and 2001 within a country appear side-by-side in the display to compare performance from one administration to the next. At the same time, the display can be used to detect item-by-country interactions across all countries. The procedure for computing the 95 percent confidence interval limits for the probability for each country is presented in Section 10.2.1.

10.5 Item Review Procedures

The International Study Center thoroughly reviewed the item statistics for all participating countries to ensure that items were performing comparably across countries. In particular, items with the following problems were considered for possible deletion from the international database:

- An error was detected during PIRLS 2001 translation verification but was not corrected before test administration.

- Data checking revealed a multiple-choice item with more or fewer options than in the international version.

- The item analysis showed the item to have a negative biserial, or, for an item with more than one score point, a non-monotonic relationship between score level and total score.

- The item-by-country interaction results showed a very large negative interaction for a particular country.

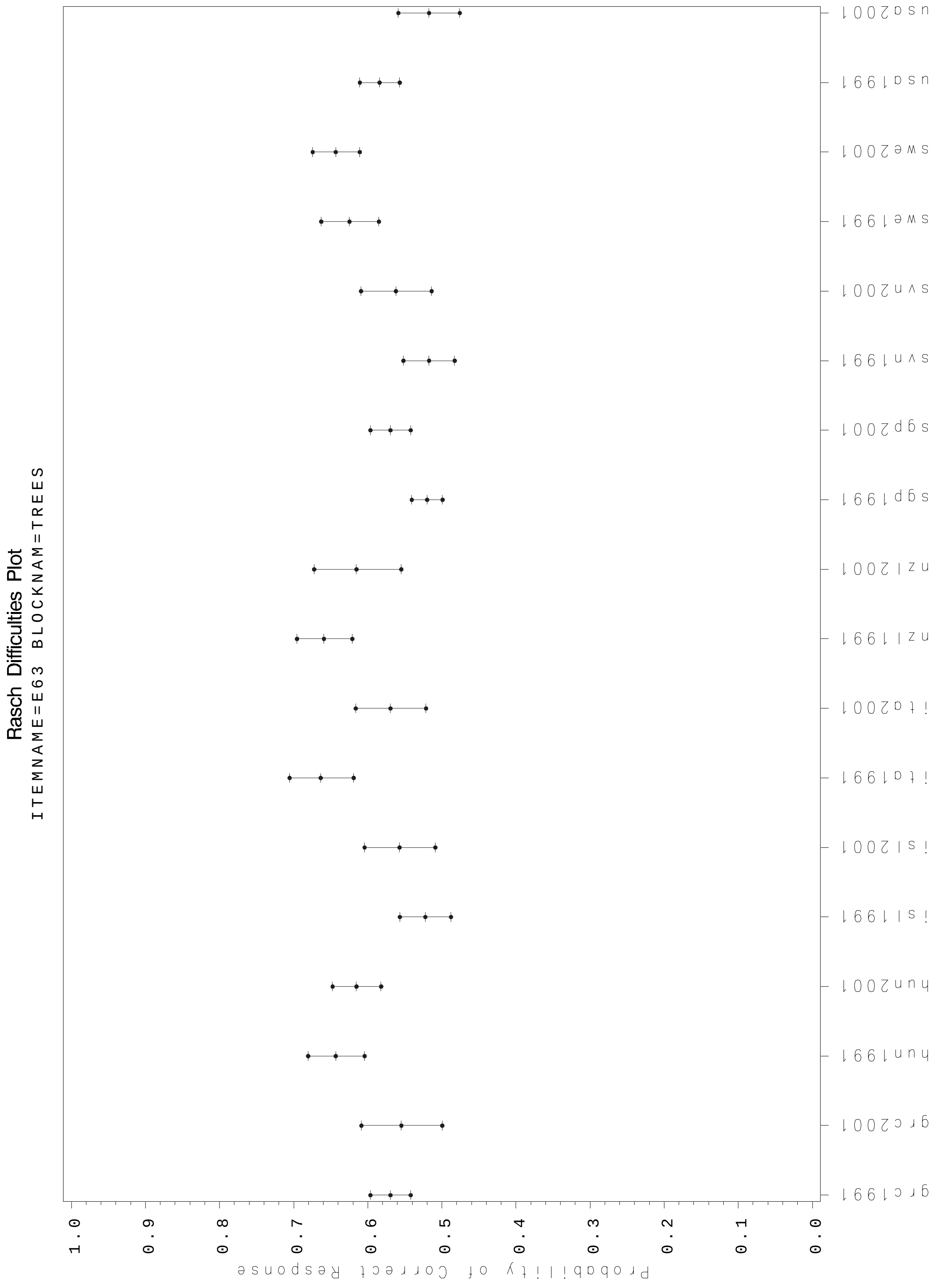
- For constructed-response items, the within-country scoring reliability data showed an agreement of less than 70 percent.

- For Trends in IEA's Reading Literacy Study items, an item performed substantially differently in 1991 compared to 2001, or an item was not included in the 1991 assessment for a particular country.

When the item statistics indicated a problem with an item, the documentation from the translation verification[7] was used as an aid in checking the test booklets. If a question remained about potential translation or cultural issues, however, then the National Research Coordinator (NRC) was consulted before deciding how the item should be treated. If a problem could be detected by the International Study Center (such as a negative point-biserial for a correct answer or too few options for a multiple-choice item), the item was deleted from the international scaling.

[7] See chapter 5 for a description of the process for translating and verifying the PIRLS data-collection instruments.

Exhibit 10.7: Example Item-by-Country Interaction Display for Trend Item E63



The checking of the PIRLS 2001 achievement data involved 98 items for 35 countries (approximately 3,500 item-country combinations), and resulted in the detection of very few items that were inappropriate for international comparisons. Just two items had to be deleted from the international database, one for Cyprus and one for the Russian-speaking part of Moldova (see Appendix C). The checking of the Trends in IEA's Reading Literacy Study data involved 66 items for 9 countries. The items were deleted for all countries, and several items were identified in individual countries as inappropriate for international comparisons. Appendix C provides a list of deleted items as well as a list of recodes made to constructed-response item codes.

11

Scaling the PIRLS Reading Assessment Data

Eugenio J. Gonzalez

11.1 Overview

To achieve its goal of broad coverage of the reading purposes and processes specified in the assessment framework,[1] the PIRLS 2001 assessment included a range of reading passages and items arranged into eight 40-minute assessment blocks. Each student participating in the assessment completed one student booklet made up of just two of these blocks, keeping individual student response burden to a minimum. PIRLS used a matrix-sampling design[2] to assign assessment blocks to student booklets so that a comprehensive picture of the reading achievement of fourth-grade students in each country could be assembled from the components completed by individual students. PIRLS relied on Item Response Theory (IRT) scaling to combine the student responses to provide accurate estimates of reading achievement in the student population in each country. The PIRLS IRT scaling also uses multiple imputation or "plausible values" methodology to obtain proficiency scores in reading for all students, even though each student responded to only a part of the assessment item pool.

This chapter first reviews the psychometric models and the multiple imputation or "plausible values" methodology used in scaling the PIRLS 2001 data, and then describes how this approach was applied to the PIRLS 2001 data and to the data from IEA's Trends in Reading

[1] The PIRLS 2001 assessment framework is described in Campbell, Kelly, Mullis, Martin, & Sainsbury (2001).

[2] The PIRLS 2001 achievement test design is described in Chapter 2.

Literacy Study. The PIRLS scaling was conducted at the PIRLS International Study Center (ISC) at Boston College, with software and psychometric support from Educational Testing Service.[3]

11.2 PIRLS 2001 Scaling Methodology[4]

The scaling approach used by PIRLS was developed originally by Educational Testing Service for use in the U.S. National Assessment of Educational Progress. It is based on psychometric models that were first used in the field of educational measurement in the 1950s, and have become popular since the 1970s for use in large-scale surveys, test construction, and computer adaptive testing.[5] This approach also has been used to scale IEA's TIMSS data.

Three distinct scaling models, depending on item type and scoring procedure, were used in the analysis of the PIRLS 2001 assessment data. Each is a "latent variable" model that describes the probability that a student will respond in a specific way to an item in terms of the respondent's proficiency, which is an unobserved or "latent" trait, and various characteristics (or "parameters") of the item. A three-parameter model was used with multiple-choice items, which were scored as correct or incorrect, and a two-parameter model for those constructed-response items with just two response options – which also were scored as correct or incorrect. Since each of these item types has just two response categories, they are known as dichotomous items. A partial credit model was used with polytomous constructed-response items (i.e., those with more than two score points).

11.2.1 Two- and Three-Parameter IRT Models for Dichotomous Items

The fundamental equation of the three-parameter (3PL) model gives the probability that a person whose proficiency on a scale k is characterized by the unobservable variable θ will respond correctly to item i:

Equation 1

$$P\left(x_i = 1 \middle| \theta_k, a_i, b_i, c_i\right) = c_i + \frac{\left(1 - c_i\right)}{1.0 + \exp\left(-1.7 a_i \left(\theta_k - b_i\right)\right)}$$

where

x_i is the response to item i, 1 if correct and 0 if incorrect;

[3] PIRLS is indebted to Matthias von Davier, Ed Kulick, and John Barone of Educational Testing Service for their advice and support.

[4] This section describing the PIRLS scaling methodology has been adapted with permission from the *TIMSS 1999 Technical Report* (Yamamoto & Kulick, 2000).

[5] For a description of IRT scaling see Birnbaum (1968); Lord and Novick (1968); Lord (1980); Van Der Linden and Hambleton (1996). The theoretical underpinning of the imputed value methodology was developed by Rubin (1987), applied to large-scale assessment by Mislevy (1991), and studied further by Mislevy, Johnson and Muraki (1992), and Beaton and Johnson (1992). The procedures used in PIRLS have been used in several other large-scale surveys, including Trends in International Mathematics and Science Study (TIMSS), the U.S. National Assessment of Educational Progress (NAEP), the U.S. National Adult Literacy Survey (NALS), the International Adult Literacy Survey (IALS), and the International Adult Literacy and Life Skills Survey (IALLS).

θ_k is the proficiency of a person on a scale k (note that a person with higher proficiency has a greater probability of responding correctly);

a_i is the slope parameter of item i, characterizing its discriminating power;

b_i is the location parameter for the item, characterizing its difficulty;

c_i is the lower asymptote parameter for the item, reflecting the chances of respondents of very low proficiency selecting the correct answer.

The probability of an incorrect response to the item is defined as:

Equation 2

$$P_{i0} \equiv P\left(x_i = 0 \middle| \theta_k, a_i, b_i, c_i\right) = 1 - P_{i1}\left(\theta_k\right)$$

The two-parameter (2PL) model was used for the short constructed-response items that were scored as correct or incorrect. The form of the 2PL model is the same as Equation 1, with the c_i parameter fixed at zero.

11.2.2 The IRT Model for Polytomous Items

In PIRLS 2001, constructed-response items requiring an extended response were scored for partial credit (with 0, 1, 2, and 3 as the possible score levels). These polytomous items were scaled using a generalized partial credit model (Muraki, 1992). The fundamental equation of this model gives the probability that a person with proficiency θ_k on scale k will have, for the i-th item, a response x_i that is scored in the l-th of m_i ordered score categories (see Equation 3), where:

m_i is the number of response categories for item i;

x_i is the response to item i, possibilities ranging between 0 and m_i-1;

θ_k is the proficiency of person on a scale k;

a_i is the slope parameter of item i, characterizing its discrimination power;

b_i is the location parameter of item i, characterizing its difficulty;

$d_{i,l}$ is category l threshold parameter.

Equation 3

$$P\left(x_i = l \middle| \theta_k, a_i, b_i, d_{i,l}, \ldots d_{i,m_i-l}\right) = \frac{\exp\left[\mathop{\vphantom{\sum}}\limits_{v=0}^{l} 1.7a_i\left(\theta_k - b_i + d_{i,v}\right)\right]}{\mathop{\vphantom{\sum}}\limits_{g=0}^{m_i-l} \exp\left[\mathop{\vphantom{\sum}}\limits_{v=0}^{g} 1.7a_i\left(\theta_k - b_i + d_{i,v}\right)\right]} = P_{il}\left(\theta_k\right)$$

Indeterminacy of model parameters of the polytomous model are resolved by setting $d_{i,0}$ =0, and setting the sum of the threshold parameters equal to 0.

For all of the IRT models there is a linear indeterminacy of the values of item parameters and proficiency parameters (i.e., mathematically equivalent but different values of item parameters can be estimated on an arbitrarily linearly transformed proficiency scale). This linear indeterminacy can be resolved by setting the origin and unit size of the proficiency scale to arbitrary constants, (such as a mean of 500 and a standard deviation of 100). The indeterminacy is most apparent when the scale is set for the first time.

IRT modeling relies on a number of assumptions, the most important being conditional independence. Under this assumption, item response probabilities depend only on θ_k (a measure of person proficiency) and the specified parameters of the item, and are assumed unaffected by the demographic characteristics or unique experiences of the respondents, the data collection conditions, or the other items presented in the test. Under this assumption, the joint probability of a particular response pattern x across a set of n items is given by:

$$P\left(x|\theta_k, itemparameters\right) = \prod_{i=1}^{n}\prod_{l=0}^{m_i-1} P_{il}\left(\theta_k\right)^{u_{il}}$$

where $P_{il}(\theta_k)$ is of the form appropriate to the type of item (dichotomous or polytomous), m_i is equal to 2 for the dichotomously scored items, and u_{il} is an indicator variable defined by:

$$U_{il} = \begin{matrix} 1 \text{ if response } x_i \textit{ is in category}_l \\ 0 \text{ otherwise} \end{matrix}$$

Replacing the hypothetical response pattern with the real scored data, the above function can be viewed as a likelihood function to be maximized by a given set of item parameters. Once items were calibrated in this manner, a likelihood function for the proficiency θ_k was induced from student responses to the calibrated items. This likelihood function for the proficiency θ_k is called the posterior distribution of the θ*s* for each respondent.

11.2.3 Proficiency Estimation Using Plausible Values

Most cognitive skills testing is concerned with accurately assessing the performance of individual respondents for the purposes of diagnosis, selection, or placement. Regardless of the measurement model used, whether classical test theory or item response theory, the accuracy of these measurements can be improved – that is, the amount of measurement error can be reduced – by increasing the number of items given to the individual. Thus, it is common to see achievement tests designed to provide information on individual students that contain more than 70 items. Since the uncertainty associated with each θ in such tests is negligible, the distribution of θ or the joint distribution of θ with other variables can be approximated using individual θ's.

For the distribution of proficiencies in large populations, however, more efficient estimates can be obtained from a matrix-sampling design like that used in PIRLS 2001. This design solicits relatively few responses from each sampled respondent while maintaining a wide range of content representation when responses are aggregated across all respondents. With this approach, however, the advantage of estimating population characteristics is more efficiently offset by the inability to make precise statements about individuals. The uncertainty associated with individual θ estimates becomes too large to be ignored. In this situation, aggregations of individual student scores can lead to seriously biased estimates of population characteristics (Wingersky, Kaplan, & Beaton, 1987).

Plausible values methodology was developed as a way to address this issue by using all available data to estimate directly the characteristics of student populations and subpopulations, and then generating multiple imputed scores (called plausible values) from these distributions, which can be used in analyses with standard statistical software. A detailed review of plausible values methodology is given by Mislevy (1991).[6]

What follows is a brief overview of the plausible values approach. Let y represent the responses of all sampled students to background questions or background data of sampled students collected from other sources, and let θ represent the proficiency of interest. If θ were known for all sampled students, it would be possible to compute a statistic t(θ,y) – such as a sample mean or sample percentile point – to estimate a corresponding population quantity T.

Because of the latent nature of the proficiency, however, θ values are not known even for sampled respondents. One solution to this problem is to follow Rubin (1987) by considering θ as "missing data" and approximate $t(\theta,y)$ by its expectation given (x,y), the data that actually were observed, as follows:

Equation 4

$$t^*(x,y) = E\left[t\left(\underline{\theta},\underline{y}\right)\middle|\underline{x},\underline{y}\right]$$

$$= \int t\left(\underline{\theta},\underline{y}\right)p\left(\underline{\theta}\middle|\underline{x},\underline{y}\right)d\theta$$

It is possible to approximate t^* using random draws from the conditional distribution of the scale proficiencies given the student's item responses x_j, the student's background variables y_j, and model parameters for the student. These values are referred to as "imputations" in the sampling literature, and as "plausible values" in large-scale surveys such as TIMSS, NAEP, NALS, and IALLS. The value of θ for any respondent that would enter into the computation of t is thus replaced by a randomly selected value from his or her conditional distribution. Rubin (1987) proposed repeating this process several times so that the uncertainty associated with imputation can be quantified. For example, the average of multiple estimates of t, each computed from a different set of plausible values, is a numerical approximation of t^* of Equation 4; the

[6] Along with theoretical justifications, Mislevy presents comparisons with standard procedures; discusses biases that arise in some secondary analyses; and offers numerical examples.

Equation 5

$$P\left(\theta_j \middle| x_j, y_j, \Gamma, \Sigma\right) \propto P\left(x_j \middle| \theta_j, y_j, \Gamma, \Sigma\right) P\left(\theta_j \middle| y_j, \Gamma, \Sigma\right) = P\left(x_j \middle| \theta_j\right) p\left(\theta_j \middle| y_j, \Gamma, \Sigma\right)$$

variance among them reflects uncertainty due to not observing $\underline{\theta}$. It should be noted that this variance does not include the variability of sampling from the population.

Plausible values are not test scores for individuals in the usual sense, but rather are imputed values that may be used to estimate population characteristics correctly. When the underlying model is correctly specified, plausible values will provide consistent estimates of population characteristics – even though they are not generally unbiased estimates of the proficiencies of the individuals with whom they are associated.[7]

Plausible values for each respondent j are drawn from the conditional distribution $P(\theta_j | x_j, y_j, \Gamma, \Sigma)$,where Γ is a matrix of regression coefficients for the background variables, and Σ is a common variance matrix for residuals. Using standard rules of probability, the conditional probability of proficiency can be represented as Equation 5, where θ_j is a vector of scale values, $P(x_j | \theta_j)$ is the product over the scales of the independent likelihoods induced by responses to items within each scale, and $P(\theta_j | y_j, \Gamma, \Sigma)$ is the multivariate joint density of proficiencies of the scales, conditional on the observed value y_j of background responses and parameters Γ and Σ. Item parameter estimates are fixed, and regarded as population values in the computations described in this equation.

11.2.4 Conditioning

A multivariate normal distribution was assumed for $P(\theta_j | x_j, y_j, \Gamma, \Sigma)$, with a common variance, Σ, and with a mean given by a linear model with regression parameters, Γ. Since, in large-scale studies like PIRLS, there are many hundreds of background variables, it is customary to conduct a principal components analysis to reduce the number to be used in Γ. Typically, components representing 90 percent of the variance in the data are selected. These principal components are referred to as the conditioning variables, and denoted as y^c. The following model is then fit to the data:

Equation 6

$$\theta = \Gamma' y^c + \varepsilon$$

In Equation 6, ε is normally distributed with mean zero and variance Σ. As in a regression analysis, Γ is a matrix each of whose columns is the effects for each scale, and Σ is the matrix of residual variance between scales.

[7] For further discussion, see (Mislevy, Beaton, Kaplan, & Sheehan, 1992).

Note that, in order to be strictly correct for all functions Γ of θ, it is necessary that $P(\theta|y)$ be correctly specified for all background variables in the survey. Estimates of functions Γ involving background variables not conditioned on in this manner are subject to estimation error due to misspecification. The nature of these errors was discussed in detail in Mislevy (1991). In PIRLS 2001, however, principal component scores based on nearly all background variables were used. Those selected variables were chosen to reflect high relevance to policy, and to education practices. The computation of marginal means and percentile points of θ for these variables is nearly optimal.

The basic method for estimating Γ and Σ with the expectation and maximization (EM) procedure is described in Mislevy (1985) for a single scale case. The EM algorithm requires the computation of the mean θ, and variance Σ, of the posterior distribution in Equation 6.

11.2.5 Generating Proficiency Scores

After completing the EM algorithm, plausible values for all sampled students are drawn from the joint distribution of the values of Γ in a three-step process. First, a value of Γ is drawn from a normal approximation to $P(\Gamma,\Sigma|x_j,y_j)$ that fixes Σ at the value $\hat{\Sigma}$ (Thomas, 1993). Second, conditional on the generated value of Γ (and the fixed value of $\Sigma=\hat{\Sigma}$), the mean θ, and variance Σ_j^p of the posterior distribution in Equation 6 are computed using the methods applied in the EM algorithm. In the third step, the proficiency values are drawn independently from a multivariate normal distribution with mean θ and variance Σ_j^p. These three steps are repeated five times, producing five imputations of θ for each sampled respondent.

For respondents with an insufficient number of responses, the Γ and Σs described in the previous paragraph are fixed. Hence, all respondents – regardless of the number of items attempted – are assigned a set of plausible values.

The plausible values can then be employed to evaluate an arbitrary statistic T as follows:

1. Using the first vector of plausible values for each respondent, evaluate T as if the plausible values were the true values of θ. Denote the result T_1.

2. As in step 1 above, evaluate the sampling variance of T, or $Var(T_1,)$, with respect to respondents' first vectors of plausible values.

3. Carry out steps 1 and 2 for the second through fifth vectors of plausible values, thus obtaining T_u and Var_u for u=2, . . ., M, where M is the number of imputed values.

4. The best estimate of T obtainable from the plausible values is the average of the five values obtained from the different sets of plausible values:

$$T. = \frac{\sum_{\cup} T_u}{5}$$

5. An estimate of the variance of $T.$ is the sum of two components: an estimate of $Var(T_u)$ obtained as in step 4 and the variance among the T_u:

$$Var(T.) = \sum_{u} \frac{VAR_{ut}}{M} + \left(1 + M^{-1}\right) \frac{\Sigma\left(T_u - T.\right)^2}{M-1}$$

The first component in V_M reflects uncertainty due to sampling respondents from the population; the second reflects uncertainty due to the fact that sampled respondents' θs are not known precisely, but only indirectly through x and y.

11.2.6 Working with Plausible Values

Plausible values methodology was used in PIRLS 2001 to ensure the accuracy of estimates of the proficiency distributions for the PIRLS population as a whole, and particularly for comparisons between subpopulations. A further advantage of this method is that the variation between the five plausible values generated for each respondent reflects the uncertainty associated with proficiency estimates for individual respondents. However, retaining this component of uncertainty requires that additional analytical procedures be used to estimate respondents' proficiencies, as follows:

If θ values were observed for all sampled respondents, the statistic $(t\text{-}T)/\ U^{1/2}$ would follow a t-distribution with d degrees of freedom. Then the incomplete-data statistic $(t^*\text{-}T)/(Var(t^*))^{1/2}$ is approximately t-distributed, with degrees of freedom (Johnson & Rust, 1993) given by:

$$v = \frac{1}{\dfrac{f_M^{\ 2}}{M-1} + \dfrac{\left(1 - f_M\right)^2}{d}}$$

where d is the degrees of freedom for the complete-data statistic, and f is the proportion of total variance due to not observing θ values:

$$f_M = \frac{\left(1 + M^{-1}\right) B_M}{V_M}$$

where B_M is the variance among M imputed values and V_M is the final estimate of the variance of T. When B is small relative to U^*, the reference distribution for the incomplete-data statistic differs little from the reference distribution for the corresponding complete-data statistics. If, in addition, d is large, the normal approximation can be used instead of the t-distribution.

For k-dimensional t, such as the k coefficients in a multiple regression analysis, each U and U^* is a covariance matrix, and B is an average of squares and cross-products rather than simply an average of squares. In this case, the quantity $(T\text{-}t^*)V^{-1}\ (T\text{-}t^*)'$ is approximately F distributed with degrees of freedom equal to k and ν, with ν defined as above but with a matrix generalization of f_M :

$$f = \frac{\left(1 - M^{-1}\right) Trace\left(BV^{-1}\right)}{k}$$

A chi-square distribution with k degrees of freedom can be used in place of the above quantity $(T\text{-}t^*)V^{-1}\ (T\text{-}t^*)'$ for the same reason that the normal distribution can approximate the t distribution.

Statistics t^*, the estimates of ability conditional on responses to cognitive items and background variables, are consistent estimates of the corresponding population values T, as long as background variables are included in the conditioning variables. The consequences of violating this restriction are described by Beaton & Johnson (1990), Mislevy (1991), and Mislevy & Sheehan (1987). To avoid such biases, the PIRLS 2001 analyses included effectively all background variables in the conditioning.

11.3 Implementing the Scaling Procedures for the PIRLS 2001 Assessment Data

The application of IRT scaling and plausible value methodology to the PIRLS 2001 assessment data involved three major tasks: calibrating the achievement test items (estimating model parameters for each item), creating principal components from the questionnaire data for use in conditioning, and generating IRT scale scores (proficiency scores) for reading overall, and for each of two reading purposes (reading for literary experience and reading to acquire and use information).

11.3.1 Calibrating the PIRLS 2001 Test Items

As shown in Exhibit 11.1, the PIRLS achievement test design consisted of a total of eight reading blocks (a block consisting of a text passage to be read followed by a set of questions about the passage) distributed across nine student booklets and a PIRLS Reader. Each block was developed to assess one of the two reading purposes specified in the framework: reading for literary experience, or reading to acquire and use information. The literary blocks are designated L1, L2, L3, and L4 – and the information blocks I1, I2, I3, and I4. Each student booklet, as well as the Reader, contained two blocks, which were chosen according to a matrix-sampling scheme that kept the number of booklets as low as possible while maximizing the number of times blocks were paired together in a booklet. Each sampled student completed one of the nine student booklets or the Reader.

Exhibit 11.1: Distribution of Literary and Information Blocks Across Booklets*

	Reading Achievement Overall							
	Reading for Literary Experience				Reading to Acquire and Use Information			
Booklet	L1	L2	L3	L4	I1	I2	I3	I4
Booklet 1	X	X						
Booklet 2		X	X					
Booklet 3			X		X			
Booklet 4					X	X		
Booklet 5						X	X	
Booklet 6	X						X	
Booklet 7	X				X			
Booklet 8		X				X		
Booklet 9			X				X	
Reader				X				X

* An 'X' in a cell indicates that the block in that column was assigned to the booklet in that row.

The booklets and Reader were distributed among the students in each sampled class according to a scheme that ensured comparable random samples of students responded to each block. Because blocks L1 through L3 and I1 through I3 each appear in three booklets, but blocks L4 and I4 appear only in the Reader, the assignment plan ensured that the Reader was assigned after every third booklet. Effectively, this meant that each block was administered to approximately 1/4 of the student sample.

Following the PIRLS framework, IRT scales for reporting student reading achievement were constructed for reading overall (both reading purposes combined) as well as separately for reading for literary experience, and for reading to acquire information.

The first step in constructing these scales was to estimate the IRT model item parameters for each item on each of the scales. This item calibration was conducted using the commercially available Parscale software (Muraki & Bock, 1991; version 3.5). The entire PIRLS student sample (146,490 students from 35 countries) was used in the calibration runs. However, to ensure that the data from each country contributed equally to the item calibration, the student sampling weights within each country were scaled to add to 500, so that – for the purposes of item parameter estimation only – each country had a weighted sample size of 500 students.

Three separate item calibrations were run: one for the overall reading scale, and one for each of the literary and information scales. All items were included in the calibration of the overall reading scale. Interim reading scores[8] for use in generating conditioning variables were produced as a by-product of this calibration. For the calibration run for the reading for literary experience scale, only those items from the literary blocks and only those students completing a booklet containing a literary block (121,228 students) were included. Similarly, only the items from the information blocks and only the students responding to information items (121,065 students) were included in the calibration for the information scale. Exhibits D.1 through D.3 in Appendix D present the item parameters for the three calibration runs.

11.3.2 Omitted and Not-Reached Responses

Apart from missing data on items that by design were not administered to a student, missing data could also occur because a student did not answer an item – whether because the student did not know the answer, omitted it by mistake, or did not have time to attempt the item. An item was considered not reached when (within part 1 or part 2 of the booklet) the item itself and the item immediately preceding were not answered, and there were no other items completed in the remainder of the booklet.

[8] Because each student responded to only a subset of the assessment item pool, these interim scores, known as expected a posterior or EAP scores, were not sufficiently reliable for reporting PIRLS results. The plausible value proficiency estimates were used for this purpose.

In PIRLS 2001, not-reached items were treated differently in estimating item parameters and in generating student proficiency scores. In estimating the values of the item parameters, items that were considered not to have been reached by students were treated as if they had not been administered. This approach was optimal for parameter estimation. However, because the time allotment for the PIRLS tests was generous – enough for even marginally able respondents to attempt all items – not-reached items were considered as incorrect responses when student proficiency scores were generated.

11.3.3 Evaluating Fit of IRT Models to the PIRLS 2001 Data

After the calibration runs were completed, checks were performed to verify that the item parameters obtained from Parscale adequately reproduced the observed distribution of responses across the proficiency continuum. The fit of the IRT models to the PIRLS 2001 data was examined by comparing the theoretical item response function curves generated using the item parameters estimated from the data with the empirical item response functions calculated from the posterior distributions of the θs for each respondent who received the item.

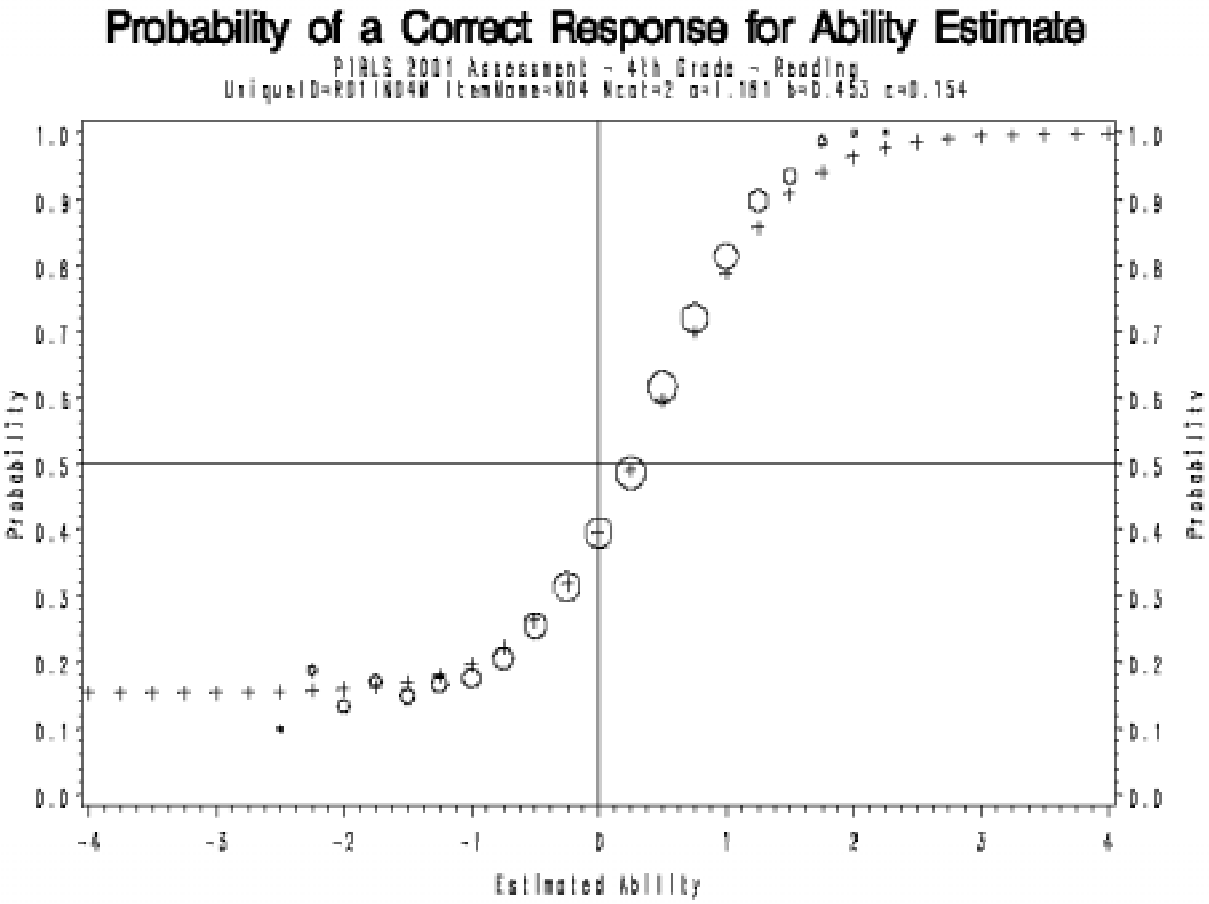
Exhibit 11.2 shows a plot of the empirical and theoretical item response functions for a dichotomous item. In the plot, the horizontal axis represents the proficiency scale, and the vertical axis represents the probability of a correct response. Values from the theoretical curve based on the estimated item parameters are shown as crosses. Empirical results are represented by circles. The centers of the circles represent the empirical proportions correct. The size of the circles is proportional to the sum of the posteriors at each point on the proficiency scale for all of those who received the item; this is related to the number of respondents contributing to the estimation of that empirical proportion correct. Exhibit 11.3 contains a plot of the empirical and theoretical item response functions for a polytomous item. As for the dichotomous item plot above, the horizontal axis represents the proficiency scale, but the vertical axis represents the probability of having a response fall in a given score category. The interpretation of the small circles is the same as in Exhibit 11.2. For items where the model fits the data well, the empirical and theoretical curves are close together.

11.3.4 Variables for Conditioning the PIRLS 2001 Data

Because there were so many background variables that could be used in conditioning, PIRLS followed the practice established in other large-scale studies of using principal components analysis to replace the original variables with a smaller number of principal components that explain most of their common variance. Principal components for the PIRLS 2001 student background data were constructed as follows:

- For categorical variables (questions with a small number of fixed response options), a “dummy coded” variable was created for each response option, with a value of one if the option was chosen and zero otherwise. If a student omitted or was

Exhibit 11.2: PIRLS 2001 Reading Assessment Example Item Response Function Dichotomous Item



not administered a particular question, all dummy coded variables associated with that question were assigned the value zero.

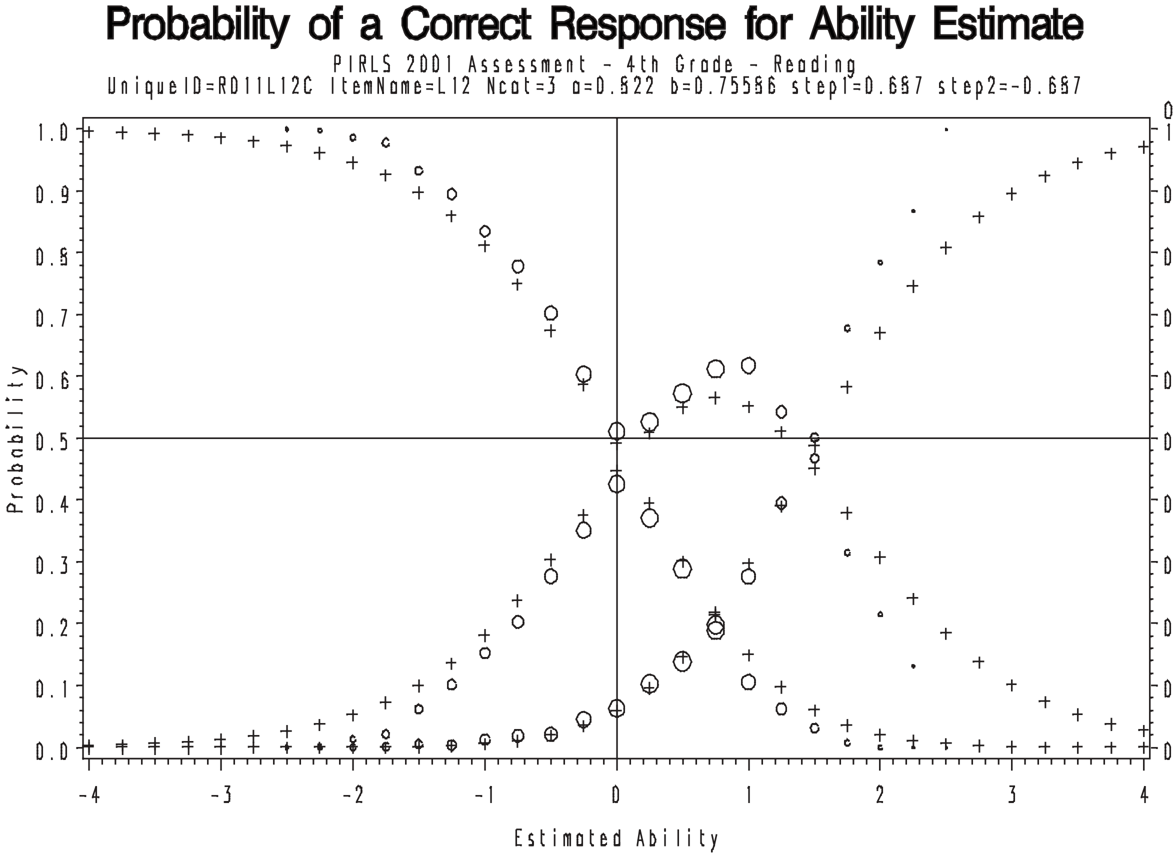
- Background variables with numerous response options (such as year of birth, or number of people who live in the home) were recoded using criterion scaling.[9] This was done by replacing each response option with the mean interim (EAP) score of the students choosing that option.

- Separately for each PIRLS country, all the dummy-coded and criterion-scaled variables were included in a principal components analysis. Those principal components accounting for 90 percent of the variance of the background variables were retained for use as conditioning variables.[10] Because the principal components analysis was performed separately for each country, the number of principal components required to account for 90 percent of the variance in the background variables varied from country to country. Exhibit 11.4 shows

[9] The process of generating criterion scaled variables is described in Beaton(1969).

[10] Exceptions were Belize, Latvia, and Lithuania – where component accounting for only 80% of the variance were selected.

Exhibit 11.3: PIRLS 2001 Reading Assessment Example Item Response Function Polytomous Item



the total number of variables that were used in the principal component analysis and the number of principal components needed to account for 90 percent of the variance in the background variables within each country.

In addition to the principal components, student gender (dummy coded), the language of the test (dummy coded), an indicator of the classroom in the school to which the student belonged (criterion scaled), and an optional, country-specific variable (dummy coded) were included as conditioning variables.

11.3.5 Generating IRT Proficiency Scores for the PIRLS 2001 Data

Educational Testing Service's MGROUP program (ETS, 1998; version 3.1)[11] was used to generate the IRT proficiency scores. This program takes as input the students' responses to the items they were given, the item parameters estimated at the calibration stage, and the conditioning variables, then generates the plausible values that represent student proficiency in reading as output. Three MGROUP runs were conducted,

[11] The MGROUP program was provided by ETS under contract to the PIRLS International Study Center at Boston College.

one for reading overall, and one each for reading for literary experience and reading to acquire and use information.

Plausible values generated by the MGROUP program are initially on the same scale as the item parameters used to estimate them. This scale metric is generally not useful for reporting purposes, since it is somewhat arbitrary, ranges between approximately -3 and +3, and has a mean of zero across all countries. For reporting PIRLS results, a scale metric was selected such that the combined proficiency distribution for fourth grade students across all PIRLS countries had a mean of 500 and a standard deviation of 100. The same metric (mean of 500 and standard deviation of 100) was also used for the literary and information scales.

Although practically all of the plausible values on the new metric were between 0 and 1000, there were a few outliers with values outside this range. These were recoded so that plausible values below 5 were set to 5, and plausible values above 995 were set to 995. This truncation did not have a noticeable effect on the distribution of the plausible values.

11.3.6 Implementing the Scaling Procedures for the Trends in IEA's Reading Literacy Study Data

In conjunction with the PIRLS 2001 assessment, IEA offered countries that had participated in the 1991 IEA Reading Literacy Study at the fourth grade the opportunity to measure trends over a ten-year period by re-administering the 1991 test at the same time as the PIRLS data collection was taking place. Nine of the 35 PIRLS countries took part in this Trends in IEA's Reading Literacy Study. The IRT scaling methodology used with the PIRLS 2001 data was applied also in scaling the trend study data. From a scaling perspective, the challenge was to place the 1991 data and the 2001

Exhibit 11.4: Number of Principal Components Selected to Account for the Variance in PIRLS 2001 Background Variables

Country	Number of Principal Components
Argentina	291
Belize	221
Bulgaria	287
Canada (O,Q)	291
Colombia	305
Cyprus	290
Czech Republic	288
England	265
France	282
Germany	305
Greece	284
Hong Kong, SAR	294
Hungary	282
Iceland	291
Iran, Islamic Rep. of	304
Israel	295
Italy	298
Kuwait	265
Latvia	223
Lithuania	272
Macedonia, Rep. of	296
Moldova	293
Morocco	160
Netherlands	280
New Zealand	280
Norway	293
Romania	287
Russia	288
Scotland	279
Singapore	303
Slovak Republic	297
Slovenia	226
Sweden	297
Turkey	287
United States	163

data on the same scale so that changes in average student reading literacy in the participating countries over the ten-year period could be accurately described.

The Reading Literacy data collected in 1991 were scaled, at that time, using a one-parameter IRT model known as the Rasch model.[12] However, the two- and three-parameter models with conditioning and plausible values used in scaling the PIRLS data were preferred also for scaling the trend data – partly for consistency with the PIRLS approach, but mainly because they were likely to be a better fit to the data (important when trying to detect possibly small changes in achievement between 1991 and 2001). Under the Rasch model, items may vary in difficulty, but are assumed to have the same discriminating power, and to not be answerable by guessing. The two- and three-parameter models feature an extra item parameter that accounts for differences among items in discriminating power, and the three-parameter model introduces a third parameter that models guessing behavior. The extra parameters mean that these models can more accurately account for the differences among items in their ability to discriminate between students of high and low ability, and are more effective than the simpler Rasch models in reducing errors due to model misspecification. Specification errors are apparent when data predicted on the basis of the model do not match the observed data. The difference between the observed data and those generated by the model is directly proportional to the degree of model misspecification.

One disadvantage of the one- and two-parameter models, compared with the one-parameter Rasch model, is that because more item parameters must be estimated, larger amounts of data – and consequently larger sample sizes – are required to obtain the same degree of confidence in the estimated item parameters. However, the trend database is more than large enough to provide the required level of confidence.

As with the PIRLS 2001 data, the application of IRT scaling and plausible value methodology to the trend study data involved three major tasks: calibrating the items of the Reading Literacy test using the combined data from 1991 and 2001, creating principal components from the questionnaire data for use in conditioning, and creating IRT scale scores (proficiency scores) for the required reading scales.

11.3.7 Calibrating the 1991 Reading Literacy Test Items

By comparison with the PIRLS assessment, the design of the 1991 Reading Literacy test was relatively simple, consisting of a single test booklet administered to all sampled students. This test booklet contained a total of 65 test items addressing three different text types: narrative texts (22 items), expository texts (21 items), and documents (22 items).

[12] The analysis of the 1991 data is described in Elley (1994).

Scales for reporting student achievement in reading literacy were constructed for reading overall (using all 65 items), and of the three text types – narrative texts, expository texts, and documents. The data from each of the nine countries consisted of student responses to the test items collected from nationally-representative samples of students at two points in time: 1991 and 2001.

The first step in constructing the trend study reading scales was to estimate the IRT model item parameters for each item on each of the scales. As with the PIRLS data, the item calibration was conducted using the commercially available Parscale software (Muraki & Bock, 1991; version 3.5). The data from 1991 and 2001 were combined for the calibration runs. A total of 59,761 student records were used in the calibration of the test items. To ensure that the data from each country contributed equally to the item calibration, and that data from 1991 and from 2001 contributed equally, the student sampling weights within each country for each data collection were scaled to add to 500 – so that, for the purposes of item parameter estimation, each country had a weighted sample size of 1000 students, 500 from 1991 and 500 from 2001.

Four separate item calibrations were run: one for the overall reading scale, and one for each of the text types – narrative texts, expository texts, and documents. All items and all students were included in the calibration of the overall reading scale. As in the PIRLS 2001 scaling, interim reading scores for use in generating conditioning variables were produced as a by-product of this calibration. Only the narrative items were included in the calibration run for the narrative scale, only the expository items for the expository scale, and only the documents items for the documents scale. Since all students responded to all items, all students were included in the calibration of each of the three scales. Exhibits D.4 through D.7 in Appendix D present the item parameters for the four calibration runs.

After the calibration was completed, checks were performed to verify that the item parameters obtained from the Parscale runs adequately reproduced the observed distribution of responses across the proficiency continuum.

11.3.8 Variables for Conditioning the Reading Literacy Trend Data

Similar to the procedure followed in conditioning the PIRLS 2001 data, principal components analysis was used to summarize the background questionnaire data collected during the 1991 and 2001 administrations of the 1991 Reading Literacy test. Identical procedures for coding the questionnaire variables prior to extracting principal components were followed. As before, those components accounting for 90 percent of the variance in the background variables were retained for conditioning.

Because the principal component analysis was performed separately for each country and for each data-collection year, the number of principal components necessary to account for 90 percent of the variance varied from country to country. Exhibit 11.5 shows the total number of variables that were used in the principal component analysis as well as the number of principal

Exhibit 11.5: Number of Principal Components Selected to Account for the Variance in Trends in IEA's Reading Literacy Study Background Variables

Country	Number of Principal Components	
	1991	2001
Greece	124	117
Hungary	122	129
Iceland	128	122
Italy	121	117
New Zealand	121	116
Singapore	123	124
Slovenia	125	120
Sweden	121	113
United States	119	119

components selected to account for 90 percent of the background variance within each country.

As with the PIRLS 2001 data, student gender (dummy coded), the language of the test (dummy coded), an indicator of the classroom in the school to which the student belonged (criterion scaled), and an optional, country-specific variable (dummy coded) were included as conditioning variables in addition to the principal components.

11.3.9 Generating IRT Proficiency Scores for the Trends in IEA's Reading Literacy Study

As with the PIRLS 2001 data, the MGROUP program (ETS, 1998; version 3.1) was used to generate the IRT proficiency scores for the trend study data. Four MGROUP runs were conducted: one for reading overall, and one each for the narrative, expository, and documents reading scales.

Because the data from 1991 and 2001 were combined during item calibration, the plausible values generated by the MGROUP program for each of the two data collections were on the same scale, and could be compared directly for purposes of analysis and reporting. To facilitate reporting, the original metric of the plausible values, which had a range of approximately from –3 to +3 with a mean of zero over all countries and across both data collections, was rescaled so that the mean of the 2001 data across all countries was 500 and the standard deviation was 100. This transformation was then applied to the 1991 data also, so that the 1991 and 2001 data had the same metric. This metric (mean of 500 and standard deviation of 100) also was used for the narrative, expository, and documents scales.

As with PIRLS 2001, practically all of the plausible values on the new metric were between 0 and 1000, with few outliers with values outside this range. Outliers were recoded so that plausible values below 5 were set to 5, and plausible values above 995 were set to 995. This truncation did not have a noticeable effect on the distribution of the plausible values.

References

Beaton, A.E. (1969). Scaling criterion of questionnaire items. *Socio-Economic Planning Sciences*, 2, 355-362.

Beaton, A.E., & Johnson, E.G. (1990). The average response method of scaling. *Journal of Educational Statistics*, 15, 9-38.

Beaton, A.E., & Johnson, E.G. (1992). Overview of the scaling methodology used in the National Assessment. *Journal of Educational Measurement*, 26(2), 163-175.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley Publishing.

Elley, W.B. (Ed.). (1994). *The IEA study of reading literacy: Achievement and instruction in 32 school systems*. Oxford, England: Elsevier Science Ltd.

Johnson, E.G., & Rust, K.F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics*.

Lord, F.M.,& Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mislevy, R.J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, 80, 993-97.

Mislevy, R.J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56, 177- 196.

Mislevy, R.J., Beaton, A., Kaplan, B.A., & Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29(2), 133-161.

Mislevy, R.J., Johnson, E.G. & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, 17(2), 131- 154.

Mislevy, R.J., & Sheehan, K. (1987). Marginal estimation procedures. In A. E. Beaton (Ed.), *Implementing the new design: The NAEP 1983-84 technical report* (pp. 293-360). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159-176.

Muraki, E., & Bock, R.D. (1991). *PARSCALE*: *Parameter scaling of rating data*. Chicago, IL: Scientific Software, Inc.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Rubin, D.B. (1991). EM and beyond. *Psychometrika*, 56, 241-254.

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics*, 2, 309-22.

Van Der Linden, W.J. & Hambleton, R. (1996). *Handbook of modern item response theory*. New York. Springer-Verlag.

Wingersky, M., Kaplan, B.A., & Beaton, A.E. (1987). Joint estimation procedures. In A. E. Beaton (Ed.), *Implementing the new design: The NAEP 1983-84 technical report* (pp.285-92). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Yamamoto, K., & Kulick, E. (2000). Scaling Methodology and Procedures for the TIMSS Mathematics and Science Scales. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS 1999 technical report*. Chestnut Hill, MA: Boston College.

12

Statistical Analysis and Reporting of the PIRLS Data

Eugenio J. Gonzalez

Ann M. Kennedy

12.1 Overview

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

12.2 Estimation of Sampling and Imputation Variance

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

12.2.1 Estimating Sampling Variance

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

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

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

Exhibit 12.1: Number of Sampling Zones Used in Each Country

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

Construction of Sampling Zones

To apply the JRR technique used in PIRLS 2001, the sampled schools are paired and assigned to a series of groups known as sampling zones. This was done at Statistics Canada by working through the list of sampled schools in the order in which they were selected and assigning the first and second schools to the first sampling zone, the third and fourth schools to the second zone, and so on. In total, 75 zones were used, allowing for 150 schools per country. When more than 75 zones were constructed, they were collapsed to keep the total number to 75.

Sampling zones were constructed within design domains, or explicit strata. Where there was an odd number of schools in an explicit stratum, either by design or because of school nonresponse, the students in the remaining school were randomly divided to make up two “quasi” schools for the purposes of calculating the jackknife standard error. Each zone then consisted of a pair of schools or “quasi” schools. Exhibit 12.1 shows the number of sampling zones used in each country.

Computing Sampling Variance Using the JRR Method

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

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

where H is the number of pairs in the sample for the country. The term $t(S)$ corresponds to the statistic for the whole sample (computed with any specific weights that may have been used to compensate for the unequal probability of selection of the different elements in the sample or any other post-stratification weight). The element $t(J_h)$ denotes the same statistic using the hth jackknife replicate. This is computed using all cases except those in the hth zone of the sample; for those in the hth zone, all cases associated with one of the randomly selected units of the pair are removed, and the elements associated with the other unit in the zone are included twice. In practice, this is effectively accomplished by recoding to zero the weights for the cases of the element of the pair to be excluded from the replication, and multiplying by two the weights of the remaining element within the hth pair.

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

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

Within each zone, the members of the pair of schools are assigned an indicator (u_i), coded randomly to 1 or 0 so that one of them has a value of 1 on the variable u_i, and the other a value of 0. This indicator determines whether the weights for the elements in the school in this zone are to be doubled or zeroed. The replicate weights $\left(W_h^{g,i,j}\right)$ for the elements in a school assigned to zone h is computed as the product of k_h times their overall sampling weight, where k_h can take values of 0, 1, or 2 depending on whether the school is to be omitted, be included with its usual weight, or have its weight doubled for the computation of the statistic of interest. In PIRLS 2001, the replicate weights were not permanent variables, but were created temporarily by the sampling variance estimation program as a useful computing device.

To create replicate weights, each sampled student was first assigned a vector of 75 weights $W_h^{g,i,j}$ where h takes values from 1 to 75. The value of $W_O^{g,i,j}$ is the overall sampling weight, which is simply the product of the final school weight, the appropriate final classroom weight, and the appropriate final student weight, as described in Chapter 9.

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

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

where the variable k_h for an individual i takes the value $k_{hi} = 2^* u_i$ if the record belongs to zone h, and $k_{hi} = 1$ otherwise.

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

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

12.2.2 Estimating Imputation Variance

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

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

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

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

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

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

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

12.2.3 Combining Sampling and Imputation Variance

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

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

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

Exhibits 12.2 through 12.4 show basic summary statistics for reading achievement in the PIRLS 2001 assessment, for reading overall, as well as for reading for literary and informational purposes. Each exhibit presents the student sample size, the mean and standard deviation, averaged across the five plausible values, the jackknife standard error for the mean, and the overall standard errors for the mean including imputation error. Exhibits 12.5 through 12.8 provide comparable statistics for the 1991 and 2001 data from IEA's trends in reading literacy study.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

12.3 Reporting Student Achievement in Reading

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

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

National averages were computed as the average of the weighted means for each of the five plausible values. The weighted mean for each plausible value was computed as follows:

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

where:

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

$pv_{l,j}$ is the l-th plausible value for the j-th student

$W^{i,j}$ is the weight associated with the j-th student in class i,

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

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

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

where

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

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

K is the number of countries.

12.3.1 Achievement Differences Across Countries

A basic aim of the PIRLS 2001 international report is to provide fair and accurate comparisons of student achievement across the participating countries. Most of the exhibits in the PIRLS reports summarize student achievement by means of a statistic such as a mean or percentage, and each statistic is accompanied by its standard error, which is a measure of the uncertainty due to student sampling and the imputation process. In comparisons of performance across countries, standard errors can be used to assess the statistical significance of the difference between the summary statistics.

The multiple comparison charts presented in the PIRLS 2001 international report facilitate the comparison of the average performance of a country with that of other participating countries. Reading achievement means were considered significantly different if the absolute difference between them, divided by the standard error of the difference, was greater than the critical value of 1.96, corresponding to a test of significance with 95% confidence. For differences between countries, which can be considered as independent samples, the standard error of the difference in means was computed as the square root of the sum of the squared standard errors of each mean:

$$se_{diff} = \sqrt{se_1^2 + se_2^2}$$

where se_1 and se_2 are the standard errors of the means. Exhibit 12.9 shows the PIRLS 2001 means and standard errors used in the calculation of statistical significance for the PIRLS international report.

The significance tests reported in these charts have NOT been adjusted for multiple comparisons. Although adjustments such as the Bonferroni procedure guard against misinterpreting the outcome of multiple simultaneous significance tests, and have been used in previous IEA studies,[4] the results vary depending on the number of countries

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

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

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

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

12.3.2 Comparing Achievement with the International Mean

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

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

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

where

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

K is the number of countries

se_{j}^{2} is the sampling standard error for country j

se_{k}^{2} is the sampling standard error for country k

The imputation component of the standard error was computed by taking the square root of the imputation variance calculated as follows

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

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

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

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

12.3.3 International Benchmarks of Reading Achievement

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

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

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

* Fourth-grade in most countries.

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

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

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

Exhibit 12.11: International Benchmarks of Fourth-grade Reading Achievement

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

* Fourth-grade in most countries.

12.3.4 Gender Differences

PIRLS reported gender differences in student achievement in reading overall, as well as in the two reading purposes. Gender differences were presented in an exhibit showing the percentages of males and females and their mean reading achievement in each country, together with an indication of whether the male-female difference in reading achievement was statistically significant. Because in most countries males and females attend the same schools, the samples of males and females cannot be treated as independent for the purpose of statistical significance testing. Accordingly, PIRLS used a jackknife procedure applicable to correlated samples for estimating the standard error of the male-female difference. This involved computing the average difference between boys and girls once for each of the 75 replicate samples, and five more times, once for each plausible value, as described earlier in this chapter.

12.3.5 Reporting Student Performance on Individual Items

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

When computing the percent correct for individual example items, student responses were classified in the following way: for multiple-choice items, a response to item j was classified as correct (C_j) when the correct option was selected; incorrect (W_j) when the incorrect option or no option was selected; invalid (I_j) when two or more options were selected; not reached (R_j) when it was assumed that the student stopped working on the test before reaching the question; and not administered (A_j) when the question was not included in the student's booklet or had been mistranslated or misprinted. For a particular score level of a constructed-response item, student responses to item j were classified as correct (C_j) when the corresponding number of points was obtained; incorrect (W_j) when the wrong answer or an answer worth less than the maximum points was given; invalid (N_j) when the response was not legible or interpretable or was simply left blank; not reached (R_j) when it was determined that the student stopped working on the test before reaching the question; and not administered (A_j) when the question

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

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

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

was not included in the student's booklet or had been mistranslated or misprinted. The percent correct for an item (P_j) was computed as:

$$P_j = \frac{c_j}{c_j + w_j + i_j + r_j + n_j}$$

where c_j, w_j, i_j, r_j and n_j are the weighted counts of the correct, wrong, invalid, not reached, and not interpretable responses to item *j*, respectively.

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

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

12.4 Describing International Benchmarks of Student Achievement[5]

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

12.4.1 Identifying the Anchor Items

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

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

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

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

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

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

Anchoring Criteria

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Computing the Item Percent Correct at Each Level

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

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

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

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

Identifying Anchor Items

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

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

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

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

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

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

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

12.4.2 Review of Anchor Items Development of Anchor Level Descriptions

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

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

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

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

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

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

12.5 Reporting Questionnaire Data

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

12.5.1 Reporting Student Questionnaire Data

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

12.5.2 Reporting Teacher Questionnaire Data

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

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

12.5.3 Reporting Parents' Questionnaire Data

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

12.5.4 Reporting School Questionnaire Data

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

12.5.5 Reporting Response Rates for Background Questionnaire Data

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

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

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

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

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

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

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

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

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

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

12.5.6 Development of the PIRLS International Report

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

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

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

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

References

Gonzalez, E.G. & Gregory, K.D. (2000). Reporting student achievement in mathematics and science. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS 1999 technical report*. Chestnut Hill, MA: Boston College.

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

Johnson, E.G., & Rust, K.F. (1992). Population references and variance estimation for NAEP data. *Journal of Educational Statistics*, 17, 175–190.

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

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

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

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

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

A

Acknowledgements

The design and development of PIRLS 2001 was achieved through the collaborative efforts of individuals around the world. Staff from the national research centers in each participating country, members of advisory committees, the International Association for the Evaluation for Educational Achievement (IEA), funding agencies, and the International Study Center (ISC) at Boston College worked together to develop and implement the PIRLS 2001 assessment. This appendix acknowledges the individuals and organizations for their contributions. Given that the development and implementation of PIRLS 2001 has spanned approximately four years and has involved so many people and organizations, this list may not include all who contributed. Any omission is inadvertent. PIRLS 2001 also acknowledges the students, teachers, and school principals who contributed their time and effort to the study. This report would not be possible without them.

Funding Agencies

Funding for the development of PIRLS 2001 was provided by the National Center for Education Statistics of the U.S. Department of Education (NCES), The World Bank, and the participating countries. Valena Plisko, Eugene Owen, Dawn Nelson and Lawrence Ogle of NCES and Vincent Greaney of the World Bank were instrumental in making PIRLS 2001 possible and for ensuring the quality of the study. Each participating country was responsible for funding national project costs and implementing PIRLS in accordance with the international procedures.

Management and Operations

PIRLS 2001 was conducted under the auspices of the IEA. The study is directed by Ina V.S. Mullis and Michael O. Martin, and managed centrally by the staff of the International Study Center at Boston College, Lynch School of Education. The PIRLS International Study Center worked closely with organizations that were responsible for particular aspects of the study, the PIRLS advisory committees, and representatives of the participating countries.

In the IEA Secretariat, Hans Wagemaker was responsible for overseeing fundraising and country participation. Barbara Malak, also of the IEA Secretariat, was responsible for managing the ambitious translation verification effort and for recruiting international quality control monitors. Statistics Canada worked with countries to ensure that the international sampling procedures were followed, adapted the international design to national conditions, documented the national samples, and computed sampling weights. The National Foundation for Educational Research in England and Wales had major responsibility for developing the reading test, including collecting reading passages, developing items and scoring guides, and conducting scoring training. The IEA Data Processing Center was responsible for processing and verifying the data from the 35 countries, and for constructing the international database

IEA Secretariat

Alejandro Tiana, IEA Chair
Hans Wagemaker, Executive Director
Barbara Malak, Manager Membership Relations
Juriaan Hartenberg, Financial Manager

PIRLS and TIMSS International Study Center at Boston College

Ina V.S. Mullis, Co-Director
Michael O. Martin, Co-Director
Eugenio J. Gonzalez, Director of Operations and Data Analysis
Ann Kennedy, PIRLS Project Coordinator
Cheryl Flaherty, PIRLS Research Associate
Teresa A. Smith, TIMSS Science Coordinator
Robert A. Garden, TIMSS Mathematics Coordinator
Steven J. Chrostowski, TIMSS Project Coordinator
Ebru Erberber, TIMSS Research Associate
Monica Guidi, TIMSS Research Assistant
Maria José Ramirez, Graduate Assistant
Alka Arora, Graduate Assistant
Joseph Galia, Statistician Programmer
Isaac Li, Statistician Programmer
José R. Nieto, Publications Manager
Mario Pita, Data Graphics Specialist
Betty Hugh, Data Graphics Specialist
Susan Messner, Data Graphics Specialist
Marcie Petras, Manager of Office Administration
Christine Hoage, Manager, Finance
Rita Holmes, Administrative Coordinator
Laura Misas, Administrative Coordinator

IEA Data Processing Center

Pierre Foy, Senior Researcher
Ursula Itzlinger, Senior Researcher
Juliane Barth, Research Assistant

Statistics Canada

Marc Joncas, Senior Methodologist

National Foundation for Educational Research in England and Wales

Chris Whetton, Assistant Director
Marian Sainsbury, Principal Research Officer
Jenny Bradshaw, Senior Research Officer
Anne Kispal, Senior Research Officer
Jane Sowerby, Senior Research Officer
Jenny Phillips, Research Officer

Educational Testing Services

Jay Campbell, PIRLS Reading Coordinator
Ed Kulick, Psychometric Scaling Consultant
Matthias von Davier, Psychometric Scaling Consultant

Westat

Keith Rust, Sampling Referee

American Institutes for Research

Eugene Johnson, Psychometric Design Consultant

PIRLS Advisory Committees

The PIRLS Reading Development Group (RDG) contributed their invaluable expertise to developing and reviewing the framework and reading test. The RDG also worked with the PIRLS Reading Coordinator, NFER, and ISC staff members to develop the descriptions of achievement at international benchmarks. The PIRLS Questionnaire Development Group (QDG) helped develop the PIRLS questionnaires, including writing items and reviewing drafts of all questionnaires.

PIRLS Reading Development Group

Marilyn Binkley
National Center for Education Statistics
United States

Karl Blueml
Vienna School Board
Austria

Sue Horner
Qualifications and Curriculum Authority
England

Pirjo Linnakylä
University of Jyväskylä
Finland

Martine Rémond
Institut National de la Recherche Pédagogique
France

William Tunmer
Massey University
New Zealand

Tan See Keen
Ministry of Education
Singapore

PIRLS Questionnaire Development Group

Annette Lafontaine
Université de Liège
Belgium

Michael Marshall
University of British Columbia
Canada

Ivana Prochazkova
Institute for Information on Education
Czech Republic

Monica Rosén
Göteborg University
Sweden

Graham Ruddock
National Foundation for Educational Research
in England and Wales
England

Maurice Walker
Ministry of Education
New Zealand

National Research Coordinators

The PIRLS 2001 National Research Coordinators (NRCs) were responsible for the crucial task of implementing the study in their countries. They participated in every aspect of the work to ensure that the study was of high quality. All the PIRLS 2001 NRCs and their staff members are to be commended for their professionalism and their dedication in conducting all aspects of the project.

Argentina

Lilia Toranzos
Ministerio de Educación

Belize

Rosalind Bradley
Denise Robateau
Belize Teachers' Training College

Bulgaria

Georgi Bishkov
Felyanka Kaftandjieva
University of Sofia

Canada

Francine Jaques
Robert Deschênes
Education Quality & Accountability Office (EQAO)

Louis-Philippe Gaudreault
Jean-Louis Lebel
Ministère de l'Éducation

Michael Marshall
University of British Columbia

Colombia

Martha Rocha
Claudia Saenz
Servicio Nacional de Pruebas

Cyprus

Mary Koutselini
Constantinos Papanastasiou
University of Cyprus

Czech Republic

Iveta Kramplova
Ivana Prochazkova
Institute for Information on Education

England

Liz Twist
National Foundation for Educational Research (NFER)

France

Marc Colmant
Ministère de l'Éducation Nationale

Germany

Wilfried Bos
Knut Schwippert
Eva-Maria Lankes
University of Hamburg

Greece

Georgia Kontogiannopoulou-Polydorides
Costas Basbas
University of Athens

Hong Kong, SAR

Tse Shek-Kam
The University of Hong Kong

Hungary

Péter Vari
Emese Felvégi
National Institute of Public Education
Centre for Evaluation Studies

Iceland

Einar Gudmundsson
University of Iceland

Islamic Republic of Iran

Abdol'azim Karimi
Ministry of Education

Israel

Elite Olshtain
Hebrew University

Ruth Zuzovsky
Tel Aviv University

Italy

Gabriella Pavan de Gregorio
National Institute for the Evaluation
of the Educational System (CEDE)

Kuwait

Mansour Hussein
Ministry of Education

Latvia

Ieva Johansone
University of Latvia

Lithuania

Aiste Mackeviciute
Ministry of Education and Science

Republic of Macedonia

Bojana Naceva
Pedagogical Institute of Macedonia

Moldova

Ilie Nasu
Ministry of Education & Science

Morocco

Abdellah Belachkar
Ministère de L'Education Nationale

The Netherlands

Mieke Van Diepen
University of Nijmegen

New Zealand

Megan Chamberlain
Maurice Walker
Ministry of Education

Norway

Finn Egil Tønnessen
Ragnar Gees Solheim
Stavanger College

Romania

Gabriela Noveanu
Institute for Educational Sciences

Russian Federation

Galina Kovalyova
Institute of General Secondary Education

Scotland

Liz Levy
Brian Semple
Scottish Office Education and Industry Department

Singapore

Siow-Chin Ng
Ministry of Education

Slovak Republic

Zuzana Lukačková
SPU–National Institute for Education

Slovenia

Marjeta Doupona-Horvat
Center for IEA Studies
Educational Research Institute

Sweden

Monica Rosén
Göteborg University

Turkey

Yurdanur Atlıoğlu
Educational Research and Development
Directorate (ERDD)

United States of America

Laurence Ogle
Dawn Nelson
National Center for Education Statistics (NCES)

B

Sample Implementation

B.1 Introduction

For each country participating in PIRLS 2001, this appendix describes the target population definition (where necessary), the extent of coverage and exclusions, the use of stratification variables, and any deviations from the general PIRLS sample design.

B.2 ARGENTINA

B.2.1 Coverage and Exclusions

School-level exclusions consisted of very small schools (MOS less than 8), and special schools (schools for disabled children and remedial classrooms).

B.2.2 Sample Design

- Explicit stratification by province (province 02 versus all other provinces), for a total of two strata
- Implicit stratification by province (25 provinces), urbanization (rural/urban), and school type (public/private), for a total of 72 strata
- Small schools sampled with equal probabilities (small school definitions differ by province)

Exhibit B.1: Allocation of School Sample in Argentina

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Province 02	30	0	28	2	0	0
All Other Provinces	120	0	105	3	0	12
Total	150	0	133	5	0	12

B.3 BELIZE

B.3.1 Coverage and Exclusions

School-level exclusions consisted of very small schools (MOS less than 10).

B.3.2 Sample Design

- No explicit stratification
- Implicit stratification by school type (public/private), and region (six regions) among public schools, for a total of seven strata
- Schools sampled with equal probabilities

Exhibit B.2: Allocation of School Sample in Belize

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Belize	150	0	119	1	0	30
Total	150	0	119	1	0	30

B.4 BULGARIA

B.4.1 Coverage and Exclusions

School-level exclusions consisted of special schools (educable mentally disabled students, permanent physically or functionally disabled students, students with criminal behavior) and very small schools (MOS less than 8).

B.4.2 Sample Design

- Explicit stratification by school size (large schools, small schools), for a total of two strata

- No implicit stratification

- Schools in the ‘‘Small Schools’’ stratum sampled with equal probabilities

Exhibit B.3: Allocation of School Sample in Bulgaria

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Large Schools	154	0	148	0	0	6
Small Schools	23	1	22	0	0	0
Total	177	1	170	0	0	6

B.5 CANADA

B.5.1 Coverage and Exclusions

Only Ontario and Quebec participated in the study. All other provinces and Territories are excluded from national coverage.

School-level exclusions consisted of private schools, native schools, special schools and very small schools (MOS less than 10) for Ontario; and special schools, Northern schools, non-ministry schools, and very small schools (MOS less than 10) for the province of Quebec.

Within-school exclusions consisted of disabled students and non-native speakers in both provinces.

B.5.2 Sample Design

- Explicit stratification by province (two provinces), language (French/English), school size (very large schools, large schools), for a total of seven strata
- Explicit stratum for one specific school district in Ontario (Rainbow District)
- Implicit stratification by school type in Quebec (public/private), for a total of eight strata
- Very large schools sampled with equal probabilities in both provinces
- Extra sample of schools in order to meet national objectives

Exhibit B.4: Allocation of School Sample in Canada

Explicit Stratum		Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
				Sampled	1st Replacement	2nd Replacement	
Ontario	Rainbow District	2	0	2	0	0	0
	English schools	120	0	102	8	4	6
	Very Large French schools	4	0	3	0	0	1
	Large French schools	76	0	71	0	0	5
Quebec	French schools	100	0	100	0	0	0
	Very Large English schools	4	0	4	0	0	0
	Large English schools	81	0	77	1	0	3
Total		387	0	359	9	4	15

B.6 COLOMBIA

B.6.1 Coverage and Exclusions

School level exclusions consisted of Amazonian and Orinoquian regions (isolated regions), and evening schools (older student population).

Within-school exclusions consisted of disabled students.

B.6.2 Sample Design

- Explicit stratification by urbanization (rural/urban), for a total of two strata
- Implicit stratification by school type (public/non-public), for a total of four strata
- Two classrooms sampled per selected school
- Small schools (MOS less than 20) sampled with equal probabilities

Exhibit B.5: Allocation of School Sample in Colombia

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Rural	59	0	43	12	3	1
Urban	91	0	76	12	1	2
Total	150	0	119	24	4	3

B.7 CYPRUS

B.7.1 Coverage and Exclusions

There were no reported school-level exclusions.

B.7.2 Sample Design

- Explicit stratification by district, for a total of four strata
- Implicit stratification by urbanization (rural/urban), for a total of eight strata
- School sampled with equal probabilities

Exhibit B.6: Allocation of School Sample in Cyprus

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Nicosia	55	0	54	1	0	0
Lanarka	43	0	42	1	0	0
Limassol	36	0	36	0	0	0
Pafos	16	0	16	0	0	0
Total	150	0	148	2	0	0

B.8 CZECH REPUBLIC

B.8.1 Coverage and Exclusions

School-level exclusions consisted of schools for functionally and mentally disabled students, and Polish language schools.

B.8.2 Sample Design

- No explicit stratification

- Implicit stratification by school type (complete basic school/only primary level), for a total of two strata

Exhibit B.7: Allocation of School Sample in Czech Republic

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Czech Republic	150	2	135	6	0	7
Total	150	2	135	6	0	7

B.9 ENGLAND

B.9.1 Coverage and Exclusions

School-level exclusions consisted of special schools and very small schools (MOS less than 8).

Within-school exclusions consisted of special needs pupils within schools.

B.9.2 Sample Design

- Explicit stratification by school size (large/small), for a total of two strata

- Implicit stratification by school type (primary, junior/middle, independent) and school performance (six levels), for a total of 25 strata

- Schools in the "Small Schools" stratum sampled with equal probabilities

Exhibit B.8: Allocation of School Sample in England

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Small Schools	25	0	14	9	0	2
Large Schools	125	0	74	29	5	17
Total	150	0	88	38	5	19

B.10 FRANCE

B.10.1 Coverage and Exclusions

School-level exclusions consisted of overseas territories (TOM), private schools "without contract," French schools in foreign countries (Guyanne and La Reunion), specialized schools, and very small schools (MOS less than 4).

B.10.2 Sample Design

- Explicit stratification by school size (large/small), for a total of two strata

- Implicit stratification by school type (public, public ZEP, private), for a total of six strata

- Schools in the "Small Schools" stratum sampled with equal probabilities

- Two classrooms sampled per selected school

Exhibit B.9: Allocation of School Sample in France

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Large Schools	100	0	92	5	0	3
Small Schools	50	0	48	0	0	2
Total	150	0	140	5	0	5

B.11 GERMANY

B.11.1 Coverage and Exclusions

School-level exclusions consisted of schools for disabled students and very small schools (definition varies by state).

Within-school exclusions consisted of disabled students within schools and non-native speakers.

B.11.2 Sample Design

- Explicit stratification by state (16 states), for a total of 16 strata
- Implicit stratification by school type (primary, special education), for a total of 32 strata
- Small schools sampled with equal probabilities (small schools defined by numbers shown in parentheses in table below)
- Two classrooms sampled per selected school
- Extra sample of schools in order to meet national objectives

Exhibit B.10: Allocation of School Sample in Germany

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Baden-Württemberg (less than 21)	25	0	24	1	0	0
Bayern (less than 24)	5	0	5	0	0	0
Berlin (less than 46)	25	0	24	0	0	1
Branderburg (less than 25)	25	0	25	0	0	0
Bremen (less than 25)	25	0	23	1	0	1
Hamburg (less than 23)	25	0	25	0	0	0
Hessen (less than 23)	2	0	2	0	0	0
Mecklenburg-Vorpommern (less than 19)	3	0	2	0	0	1
Niedersachsen (less than 22)	15	1	14	0	0	0
Nordrhein-Westfalen (less than 24)	35	0	34	0	0	1
Rheinland-Pfalz	8	0	8	0	0	0
Saarland (less than 46)	6	0	6	0	0	0
Sachsen (less than 22)	2	0	2	0	0	0
Sachsen-Anhalt (less than 20)	7	0	7	0	0	0
Schleswig-Holstein (less than 19)	4	0	4	0	0	0
Thüringen (less than 46)	4	0	4	0	0	0
Total	216	1	209	2	0	4

B.12 GREECE

B.12.1 Coverage and Exclusions

School-level exclusions consisted of students taught in foreign languages only, schools for students with special needs, and very small schools (MOS less than 3).

Within-school exclusions consisted of non-native language speakers.

B.12.2 Sample Design

- Explicit stratification by school type (public, private) and school size within public schools (small, large), for a total of three strata
- Implicit stratification by school type (public/private), urbanization (rural/urban) within public schools and region (7 regions) within public urban schools, for a total of 17 strata
- Schools in the "Small Public Schools" stratum sampled with equal probabilities

Exhibit B.11A: Allocation of School Sample in Greece

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Large Public Schools	132	0	110	6	4	12
Small Public Schools	29	0	17	0	2	10
Private Schools	9	0	6	0	0	3
Total	170	0	133	6	6	25

Trends in IEA's Reading Literacy Study

B.12.3 Sample Design

Sampled every second PIRLS school, same target grade

Exhibit B.11B: Allocation of School Sample in Greece (Trend)

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Large Public Schools	66	0	50	5	0	11
Small Public Schools	15	0	11	0	0	4
Private Schools	4	0	2	0	0	2
Total	85	0	63	5	0	17

B.13 HONG KONG, SAR

B.13.1 Coverage and Exclusions

School-level exclusions consisted of international schools and very small schools (MOS less than 9).

B.13.2 Sample Design

- No explicit stratification

- Implicit stratification by gender (boys, girls, mixed), school type (whole day, non-whole day) within mixed schools and district (18 districts) for mixed schools, for a total of 38 strata

Exhibit B.12: Allocation of School Sample in Hong Kong, SAR

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Hong Kong, SAR	150	0	115	29	3	3
Total	150	0	115	29	3	3

B.14 HUNGARY

B.14.1 Coverage and Exclusions

School-level exclusions consisted of very small schools (MOS less than 12).

B.14.2 Sample Design

- Explicit stratification by urbanization (cities and towns, villages) and village size (four levels) within villages, for a total of four strata

- Implicit stratification by urbanization (Budapest, county seats, towns, villages) within cities and towns, counties (19 counties) within cities and towns and regions (seven regions) within villages, for a total of 67 strata

- Extra sample of schools in order to meet national objectives

Exhibit B.13: Allocation of School Sample in Hungary

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Cities and Towns	100	0	98	0	0	2
Villages: 0-999	30	0	29	0	0	1
Villages: 1000-2999	30	0	30	0	0	0
Villages: 3000-4999	30	0	30	0	0	0
Villages: 5000-19999	30	0	29	0	0	1
Total	220	0	216	0	0	4

Trends in IEA's Reading Literacy Study

B.14.3 Target Population

The target population consisted of students in grade 3.

B.14.4 Sample design

- Sampled a 3rd grade class in each participating PIRLS school
- Allocation of school sample unchanged (see table C13 above)

B.15 ICELAND

B.15.1 Coverage and Exclusions

School-level exclusions consisted of very small schools (MOS less than 5).

Within-school exclusions consisted of disabled students.

B.15.2 Sample Design

- Implicit stratification by region (nine regions), for a total of nine strata
- All schools and all classrooms in the sample

Exhibit B.14A: Allocation of School Sample in Iceland

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Iceland	140	0	133	0	0	7
Total	140	0	133	1	0	7

Trends in IEA's Reading Literacy Study

B.15.3 Sample design

- Sampled every second PIRLS school, same target grade
- All classrooms in the sample

Exhibit B.14B: Allocation of School Sample in Iceland (Trend)

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Iceland	70	0	65	0	0	5
Total	70	0	65	0	0	5

B.16 ISLAMIC REPUBLIC OF IRAN

B.16.1 Coverage and Exclusions

School-level exclusions consisted of mentally and physically disabled students

Within-school exclusions consisted of disabled students.

B.16.2 Sample Design

- Explicit stratification by school size (large/small) and school type (public/private), for a total of four strata

- No implicit stratification

- Two classrooms sampled per selected school in the ''Large schools'' strata

- Schools in the ''Small schools'' strata sampled with equal probabilities

Exhibit B.15: Allocation of School Sample in Islamic Republic of Iran

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Small Schools, Public	39	0	38	1	0	0
Small Schools, Private	16	0	15	1	0	0
Large Schools, Public	105	0	103	2	0	0
Large Schools, Private	24	0	24	0	0	0
Total	184	0	180	4	0	0

B.17 ISRAEL

B.17.1 Coverage and Exclusions

School-level exclusions consisted of special education schools, extreme Orthodox Jewish schools, East Jerusalem Arab schools teaching the Jordanian curriculum, and very small schools (MOS less than 13).

Within-school exclusions consisted of disabled students.

B.17.2 Sample Design

- Explicit stratification by school type (Hebrew religious, Hebrew secular, Arab), for a total of three strata

- Implicit stratification by socioeconomic status (three levels), for a total of nine strata

- Five sampled Jordanian schools were excluded from data collection. As a result, all Jordanian schools (21 with 2,114 students) were identified on the school sampling frame and added to the excluded population

Exhibit B.16: Allocation of School Sample in Israel

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Hebrew, Religious	40	0	38	0	1	1
Hebrew, Secular	70	0	68	0	0	2
Arab	40	0	38	1	1	0
Total	150	0	144	1	2	3

B.18 ITALY

B.18.1 Coverage and Exclusions

There were no reported school-level exclusions.

Within-school exclusions consisted of disabled students and non-native language speakers.

B.18.2 Sample Design

- No explicit stratification
- Implicit stratification by regions (20 regions) and urbanization (capital city, other towns), for a total of 40 strata

Exhibit B.17A: Allocation of School Sample in Italy

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Italy	184	0	164	15	5	0
Total	184	0	164	15	5	0

Trends in IEA's Reading Literacy Study

B.18.3 Sample Design

- Sampled every second PIRLS school, same target grade

Exhibit B.17B: Allocation of School Sample in Italy (Trend)

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Italy	92	0	81	9	2	0
Total	92	0	81	9	2	0

B.19 KUWAIT

B.19.1 Coverage and Exclusions

There were no reported school-level exclusions.

There were no reported within-school exclusions.

B.19.2 Sample Design

- Explicit stratification by region (five regions) and gender (girls/boys), for a total of ten strata
- No implicit stratification
- Sampled all schools in strata 5, 6, 9 and 10

- Schools sampled with equal probabilities

- Two classrooms sampled per selected school

Exhibit B.18: Allocation of School Sample in Kuwait

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Al-Asima, Boys Schools	15	0	10	0	0	5
Al-Asima, Girls Schools	15	0	15	0	0	0
Hawali, Boys Schools	15	0	10	1	0	4
Hawali, Girls Schools	15	0	15	0	0	0
Al-Farwaniya, Boys Schools	15	0	15	0	0	0
Al-Farwaniya, Girls Schools	15	0	15	0	0	0
Al-Ahmadi, Boys Schools	15	0	13	1	0	1
Al-Ahmadi, Girls Schools	15	0	15	0	0	0
Al-Jahra, Boys Schools	15	0	10	0	0	5
Al-Jahra, Girls Schools	15	0	15	0	0	0
Total	150	0	133	2	0	15

B.20 LATVIA

B.20.1 Coverage and Exclusions

School-level exclusions consisted of special schools, Lithuanian, Polish, Ukrainian and Byelorussian schools, and very small schools (MOS less than 6).

B.20.2 Sample Design

- Explicit stratification by school size (small, large, very large) and language (Latvian, Russian), for a total of five strata

- Implicit stratification by regions (five regions), for a total of 23 strata

- Schools in "Very large schools" and "Small schools" strata sampled with equal probabilities

- Because some schools had the possibility of being sampled twice for each language group, the school weights on the school-level file were re-calibrated to compensate for this effect

- One school sampled twice, once for each language group

Exhibit B.19: Allocation of School Sample in Latvia

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Small Schools, Latvian	25	1	21	2	0	0
Large Schools, Latvian	73	0	68	4	1	1
Very Large Schools, Latvian	4	0	4	0	0	0
Small Schools, Russian	4	0	3	0	0	1
Large Schools, Russian	42	0	37	1	0	4
Total	148	1	133	7	1	6

B.21 LITHUANIA

B.21.1 Coverage and Exclusions

Coverage in Lithuania was restricted to students whose language of instruction is Lithuanian. School-level exclusions consisted of very small schools (MOS less than 4).

B.21.2 Sample Design

- No explicit stratification
- No implicit stratification
- 49 schools were treated as replacement schools because they had at least one classroom with no chance of being sampled, due to an inaccurate count of classrooms in the school

Exhibit B.20: Allocation of School Sample in Lithuania

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Lithuania	150	0	84	58	4	4
Total	150	0	84	58	4	4

B.22 REPUBLIC OF MACEDONIA

B.22.1 Coverage and Exclusions

School-level exclusions consisted of special schools and Turkish and Serbian schools.

B.22.2 Sample Design

- Explicit stratification by school size (large, very large), for a total of two strata
- Implicit stratification by language (Albanian/Macedonian) and urbanization (rural/urban), for a total of eight strata
- Schools in "Very large schools" stratum sampled with equal probabilities
- Because some schools had the possibility of being sampled twice for each language group, the school weights on the school-level file were re-calibrated to compensate for this effect
- Eight schools sampled twice, once for each language group

Exhibit B.21: Allocation of School Sample in Republic of Macedonia

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Large Schools	120	0	115	1	0	4
Very Large Schools	30	0	30	0	0	0
Total	150	0	145	1	0	4

B.23 MOLDOVA

B.23.1 Coverage and Exclusions

School-level exclusions consisted of foreign language schools and very small schools (MOS less than 6).

B.23.2 Sample Design

- No explicit stratification
- Implicit stratification by language (Romanian, Russian, mixed) and by region (12 regions), for a total of 14 strata

- Small schools (MOS less than 26) sampled with equal probabilities

- Nine schools were treated as replacement schools because they had at least one classroom with no chance of being sampled, due to an inaccurate count of classrooms in the school

Exhibit B.22: Allocation of School Sample in Moldova

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Moldova	150	0	133	16	1	0
Total	150	0	133	16	1	0

B.24 MOROCCO

B.24.1 Coverage and Exclusions

School-level exclusions consisted of very small schools (MOS less than 5).

B.24.2 Sample Design

- Explicit stratification by school type (public/ private), for a total of two strata

- Implicit stratification by regions (16 regions) and urbanization (rural/urban), for a total of 33 strata

- Schools in the "Private schools" stratum sampled with equal probabilities

- Small schools (MOS less than 30) sampled with equal probabilities

Exhibit B.23: Allocation of School Sample in Morocco

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Private Schools	8	0	6	0	0	2
Public Schools	150	0	111	0	0	39
Total	158	0	117	0	0	41

B.25 THE NETHERLANDS

B.25.1 Coverage and Exclusions

School-level exclusions consisted of special schools.

Within-school exclusions consisted of non-native language speakers.

B.25.2 Sample Design

- No explicit stratification
- Implicit stratification by mean student weight (three levels) and by urbanization (five levels), for a total of 15 strata
- Small schools (MOS less than 23) sampled with equal probabilities

Exhibit B.24: Allocation of School Sample in The Netherlands

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
The Netherlands	150	0	80	32	22	16
Total	150	0	80	32	22	16

B.26 NEW ZEALAND

B.26.1 Target Population

Children scheduled to begin secondary school in 2005 (four years of formal schooling)

B.26.2 Coverage and Exclusions

School-level exclusions consisted of correspondence schools, special schools, Rudolph Steiner schools, and very small schools (MOS less than 4).

Within-school exclusions consisted of special needs students.

B.26.3 Sample Design

- Explicit stratification for Maori immersion schools and urbanization (rural/urban), for a total of three strata
- Implicit stratification by socioeconomic status indicator (low, middle, high, NA), for a total of nine strata

Exhibit B.25A: Allocation of School Sample in New Zealand

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Maori Schools	6	0	5	0	1	0
Major Urban Locations	103	0	94	8	1	0
Other Locations	47	0	45	1	1	0
Total	156	0	144	9	3	0

Trends in IEA's Reading Literacy Study

B.26.4 Coverage and Exclusions

Schools in the "Maori schools" stratum were excluded from the Trends in IEA's Reading Literacy study because they were not part of the 1991 Reading Literacy study.

B.26.5 Sample Design

Sampled every second PIRLS school, same target grade

Exhibit B.25B: Allocation of School Sample in New Zealand (Trend)

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Major Urban Locations	51	0	46	4	1	0
Other Locations	24	0	21	0	1	2
Total	75	0	67	4	2	2

B.27 NORWAY

B.27.1 Coverage and Exclusions

School-level exclusions consisted of Sami language schools.

Within-school exclusions consisted of non-native language speakers.

B.27.2 Sample Design

- Explicit stratification by language (Bokmal/Nynorsk), by count of classrooms (three levels), by economic status in municipalities (four levels), and by immigration status (two levels), for a total of 44 strata

- Implicit stratification by counties (19 counties), for a total of 1 115 strata

- Two classrooms sampled per selected school

- One explicit stratum had no participating schools, it was added to the exclusion population

- Alternate method for identifying replacement schools

- The jackknife zones ignore the last two levels of explicit stratification to reduce the number of single-school zones

Exhibit B.26: Allocation of School Sample in Norway

Explicit Stratum				Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
						Sampled	1st Replacement	2nd Replacement	
Bokmal	No Class	Low Expenditures	No Immigrants	2	0	2	0	0	0
			Immigrants	2	0	2	0	0	0
		Medium Expenditures	No Immigrants	3	0	2	0	0	1
			Immigrants	3	0	3	0	0	0
		High Expenditures	No Immigrants	2	0	2	0	0	0
			Immigrants	2	0	2	0	0	0
		Four Largest Cities	No Immigrants	2	0	1	0	0	1
			Immigrants	2	0	0	1	0	1
	One Class	Low Expenditures	No Immigrants	3	0	2	1	0	0
			Immigrants	7	0	7	0	0	0
		Medium Expenditures	No Immigrants	3	0	3	0	0	0
			Immigrants	6	0	5	1	0	0
		High Expenditures	No Immigrants	2	0	0	0	1	1
			Immigrants	2	0	2	0	0	0
		Four Largest Cities	No Immigrants	2	0	2	0	0	0
			Immigrants	2	2	0	0	0	0
	Two+ Class	Low Expenditures	No Immigrants	5	0	4	1	0	0
			Immigrants	31	0	25	2	0	4
		Medium Expenditures	No Immigrants	2	0	1	0	0	1
			Immigrants	10	0	7	1	0	2
		High Expenditures	Immigrants	2	0	1	0	0	1
		Four Largest Cities	No Immigrants	2	0	1	1	0	0
			Immigrants	20	0	17	1	0	2
Nynorsk	No Class	Low Expenditures	No Immigrants	2	0	1	0	0	1
			Immigrants	2	0	2	0	0	0
		Medium Expenditures	No Immigrants	2	0	2	0	0	0
			Immigrants	2	0	2	0	0	0
		High Expenditures	No Immigrants	2	0	0	1	0	1
			Immigrants	2	0	1	0	0	1
		Four Largest Cities	No Immigrants	2	0	1	1	0	0
			Immigrants	1	0	1	0	0	0
	One Class	Low Expenditures	No Immigrants	2	0	2	0	0	0
			Immigrants	3	0	2	0	0	1
		Medium Expenditures	No Immigrants	2	0	0	1	0	1
			Immigrants	3	0	3	0	0	0
		High Expenditures	No Immigrants	2	0	1	0	0	1
			Immigrants	2	0	0	2	0	0
		Four Largest Cities	Immigrants	1	0	1	0	0	0
	Two+ Class	Low Expenditures	No Immigrants	2	0	1	0	0	1
			Immigrants	5	0	5	0	0	0
		Medium Expenditures	Immigrants	2	0	1	1	0	0
		High Expenditures	No Immigrants	2	0	1	0	0	1
			Immigrants	2	0	1	0	0	1
		Four Largest Cities	Immigrants	2	0	0	0	1	1
Total				162	2	119	15	2	24

B.28 ROMANIA

B.28.1 Coverage and Exclusions

School-level exclusions consisted of special schools and very small schools (MOS less than 8).

B.28.2 Sample Design

- Explicit stratification by school size (small rural schools, large schools), for a total of two strata

- Implicit stratification by regions (seven regions) and by urbanization (Rural/Urban) within "Large schools" stratum, for a total of 21 strata

- All sampled pseudo-classrooms were ignored, they are added to the excluded population

- Schools in "Small rural schools" stratum sampled with equal probabilities

Exhibit B.27: Allocation of School Sample in Romania

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Large Schools	120	0	117	0	0	3
Small Rural Schools	30	0	27	0	0	3
Total	150	0	144	0	0	6

B.29 RUSSIAN FEDERATION

B.29.1 Target Population

The target population consisted of students in grade 3 in stream I and of students in grade 4 in stream II.

B.29.2 Coverage and Exclusions

School-level exclusions consisted of schools for students with special needs, schools where the language of instruction is not Russian, and very small schools (MOS less than 6).

Within-school exclusions consisted of disabled students and non-native language speakers.

B.29.3 Sample Design

- Preliminary sampling of 45 regions from a frame of 89 regions, 17 regions large enough to be sampled with certainty

- No explicit stratification (the explicit strata in table C28 correspond to the primary sampling units)

- Implicit stratification by school size (small, large), by urbanization (six levels), and by school type (Primary, Basic, Secondary), for a total of 1,094 strata

- Generally, four schools sampled per region, more schools sampled in some certainty regions

- Schools in “Small Schools” strata sampled with equal probabilities

Exhibit B.28: Allocation of School Sample in Russian Federation

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Sankt-Petersburg*	6	0	6	0	0	0
Archangelsk_obl	4	0	4	0	0	0
Komi	4	0	4	0	0	0
Karelia	4	0	4	0	0	0
Moscow*	10	0	10	0	0	0
Moscow_obl*	8	0	8	0	0	0
Voronezh_obl	4	0	4	0	0	0
Tula_obl	4	0	4	0	0	0
Bransk_obl	4	0	4	0	0	0
Yaroslavl_obl	4	0	4	0	0	0
Tambov_obl	4	0	4	0	0	0
Rasan_obl	4	0	4	0	0	0
Kaluga_obl	4	0	4	0	0	0
Bashkortostan*	8	0	8	0	0	0
Tatarstan*	6	0	6	0	0	0
N_Novgorod_obl*	4	0	4	0	0	0
Samara_obl*	4	0	4	0	0	0
Perm_obl*	4	0	4	0	0	0
Saratov_obl	4	0	4	0	0	0
Orenburg_obl	4	0	4	0	0	0
Udmurtia	4	0	4	0	0	0
Kirov_obl	4	0	4	0	0	0
Pensa_obl	4	0	4	0	0	0
Marii_Al	4	0	4	0	0	0
Krasnodar_kr*	6	0	6	0	0	0
Rostov_obl*	6	0	6	0	0	0
Dagestan*	6	0	6	0	0	0
Stavropol_kr*	4	0	4	0	0	0
Volvograd_obl	4	0	4	0	0	0
Alania	4	0	4	0	0	0
Sverdlovsk_obl*	6	0	6	0	0	0
Chelyabinsk_obl*	4	0	4	0	0	0
Hanty_Mansii_ok	4	0	4	0	0	0
Tumen_obl	4	0	4	0	0	0
Krasnoyarsk_obl*	4	0	4	0	0	0
Kemerovo_obl*	4	0	4	0	0	0
Irkutsk_obl*	4	0	4	0	0	0
Altay_kr	4	0	4	0	0	0
Novosibirsk_obl	4	0	4	0	0	0
Omsk_obl	4	0	3	1	0	0
Chita_obl	4	0	4	0	0	0
Tyva	4	0	4	0	0	0
Primorsk_kr	4	0	4	0	0	0
Saha	4	0	4	0	0	0
Magadan_obl	4	0	4	0	0	0
Total	206	0	205	1	0	0

1 Strata marked with (*) were large enough to be selected with certainty

B.30 SCOTLAND

B.30.1 Coverage and Exclusions

School-level exclusions consisted of special schools, Gaelic schools, and very small schools (MOS less than 7).

Within-school exclusions consisted of special needs students.

B.30.2 Sample Design

- No explicit stratification

- Implicit stratification by Education Authority, for a total of 29 strata

Exhibit B.29: Allocation of School Sample in Scotland

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Scotland	150	0	113	5	0	32
Total	150	0	113	5	0	32

B.31 SINGAPORE

B.31.1 Coverage and Exclusions

School-level exclusions for both PIRLS and the 10-year Trend Study consisted of religious, private, and special (handicapped) schools.

B.31.2 Sample Design

All schools in the sample

Exhibit B.30A: Allocation of School Sample in Singapore

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Singapore	196	0	196	0	0	0
Total	196	0	196	0	0	0

Trends in IEA's Reading Literacy Study

B.31.3 Target Population

The target population consisted of students in grade 3.

B.31.4 Sample Design

Sampled every second PIRLS school

Exhibit B.30B: Allocation of School Sample in Singapore (Trend)

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Singapore	98	0	98	0	0	0
Total	98	0	98	0	0	0

B.32 SLOVAK REPUBLIC

B.32.1 Coverage and Exclusions

School-level exclusions consisted of foreign language schools and very small schools (MOS less than 6).

B.32.2 Sample Design

- No explicit stratification
- Implicit stratification by region (eight regions), by school type (comprehensive, primary), and by language (Slovak, Hungarian), for a total of 26 implicit strata
- Small schools (MOS less than 24) sampled with equal probabilities

Exhibit B.31: Allocation of School Sample in Slovak Republic

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Slovak Republic	150	0	130	19	1	0
Total	150	0	130	19	1	0

B.33 SLOVENIA

B.33.1 Target Population

The target population consisted of students in grade 3.

B.33.2 Coverage and Exclusions

School-level exclusions consisted of schools where the language of instruction is Italian, and very small schools (MOS less than 5).

Within-school exclusions consisted of children taught in English (temporary residents).

B.33.3 Sample Design

- Explicit stratification by school size (very large schools, large schools), for a total of two strata

- Implicit stratification by urbanization (five levels), for a total of ten strata

- Schools in "Very large schools" sampled selected with equal probabilities

Exhibit B.32: Allocation of School Sample in Slovenia

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Large Schools	138	0	136	1	0	1
Very Large Schools	12	0	11	0	0	1
Total	150	0	147	1	0	2

Trends in IEA's Reading Literacy Study

B.33.4 Sample Design

Sampled every second PIRLS school, same target grade

Exhibit B.32B: Allocation of School Sample in Slovenia (Trend)

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Large Schools	69	0	69	0	0	0
Very Large Schools	6	0	6	0	0	0
Total	75	0	75	0	0	0

B.34 SWEDEN

B.34.1 Coverage and Exclusions

School-level exclusions consisted of special schools for disabled students, Non-Swedish speaking schools, hospital and refugee schools, and very small schools (MOS less than 9 in public schools and MOS less than 5 in independent schools).

Within-school exclusions consisted of disabled students and non-native language speakers.

B.34.2 Sample Design

- Explicit stratification by school composition (grade 4 only, grades 3 and 4), school type (public/independent), and school size (large, very large) within independent schools, for a total of six strata
- No implicit stratification
- Schools in "Very Large Schools" stratum sampled with equal probabilities
- Small schools sampled with equal probabilities
- All classrooms sampled in selected schools

Exhibit B.33A: Allocation of School Sample in Sweden

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Independent, Grade 4 Only, Very Large Schools	1	0	1	0	0	0
Independent, Both Grades, Very Large Schools	2	0	2	0	0	0
Independent, Grade 4 Only	2	0	2	0	0	0
Independent, Both Grades	25	1	20	2	0	2
Public, Grade 4 Only	12	0	12	0	0	0
Public, Both Grades	108	0	105	2	0	1
Total	150	1	142	4	0	3

Trends in IEA's Reading Literacy Study

B.34.3 Target Population

The target population consisted of students in grade 3.

B.34.4 Sample Design

- Independent sample of 150 schools, but same sample design as in PIRLS (there is no overlap between PIRLS and Trends in IEA's Reading Literacy Study school samples)

Exhibit B.33B: Allocation of School Sample in Sweden (Trend)

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Independent, Both grades, Very Large Schools	2	0	2	0	0	0
Independent, Grade 3 Only	3	0	3	0	0	0
Independent, Both Grades	25	0	20	3	0	2
Public, Grade 3 Only	12	0	10	2	0	0
Public, Both Grades	108	0	107	1	0	0
Total	150	0	142	6	0	2

B.35 TURKEY

B.35.1 Coverage and Exclusions

School-level exclusions consisted of schools for handicapped, schools with combined classes, schools with a bussing system (remote), and very small schools (MOS less than 16).

B.35.2 Sample Design

- Explicit stratification by school type (private, public), for a total of two strata

- Implicit stratification by region (81 regions) within public schools, for a total of 82 strata

- Schools in the "Private schools" stratum sampled with equal probabilities

- Small schools (MOS less than 40) in the "Public Schools" stratum sampled with equal probabilities

Exhibit B.34: Allocation of School Sample in Turkey

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Private Schools	4	0	4	0	0	0
Public Schools	150	0	150	0	0	0
Total	154	0	154	0	0	0

B.36 UNITED STATES

B.36.1 Coverage and Exclusions

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

Within-school exclusions consisted of disabled students unable to take the assessment and English language learners.

B.36.2 Sample Design

- An additional sampling stage was added prior to sampling schools. Fifty-two PSUs were drawn at this stage following systematic probability proportional to size sampling procedures. Extremely large PSUs were selected with certainty. Sorting of schools within PSUs was done prior to sample the schools.

- Explicit stratification of PSUs by area status (metropolitan/non-metropolitan) within the non-certainty PSUs

- Implicit stratification of PSUs by 1990-1997 change in population, percent minorities, percent unemployed, and per capita income within the non-certainty PSUs

- Further explicit stratification of schools within sampled PSUs by school type (public/private)

- Further implicit stratification of schools within sampled PSUs by PSU and minority status (high, low) for public schools, and by religious denomination (Catholic, other religions, non-sectarian), and PSU for private schools

- The stratification shown in the table below was used for the computation of school participation adjustments (the last two levels of stratification were combined in order to derive the jackknife zones).

Exhibit B.35A: Allocation of School Sample in United States

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Public, Certainty PSUs	46	0	28	11	2	5
Private, Certainty PSUs	20	0	15	4	0	1
Public, Non-Certainty PSUs	104	0	63	16	9	16
Private, Non-Certainty PSUs	30	0	19	4	3	4
Total	200	0	125	35	14	26

Trends in IEA's Reading Literacy Study

B.36.3 Sample Design

Sampled every second PIRLS school, same target grade

Exhibit B.35B: Allocation of School Sample in United States (Trend)

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Public, Certainty PSUs	23	0	16	4	0	3
Private, Certainty PSUs	10	0	3	5	1	1
Public, Non-Certainty PSUs	52	0	26	12	6	8
Private, Non-Certainty PSUs	15	0	9	3	0	3
Total	100	0	54	24	7	15

C

Country Adaptations to Items and Item Scoring

C.1 PIRLS 2001 Items To Be Deleted in Countries

Cyprus

N13 (incorrect scoring scheme used)

Moldova (Russian Only)

H06 (mistranslated item)

C.2 PIRLS 2001 Items Needing Constructed-Response Category Recoding

All Countries

F12 (Recode 3 into 2)
N13 (Recode 2 into 1)

C.3 Trends in IEA's Reading Literacy Study Items To Be Deleted in Countries (1991 & 2001)

All Countries

D48 ADTEMPR02 (two apparently correct answers)

Greece

D42 ADBUSES03 (was not administered in 1991)

Hungary

E52 AEMARM01 (performed substantially different from 1991 to 2001)

Iceland

N60 ANGRAPA05 (unexpectedly difficult given rest of items on test and performance of other countries)

Italy

N20 ANNODOG06 (unexpectedly difficult given rest of items on test and performance of other countries)

Singapore

N17 ANNODOG03 (performed substantially different from 1991 to 2001)

N06 ANBIRD04 (was deleted for Singapore in 1991)

Slovenia

D13 ADMARIA02 (unexpectedly difficult given rest of items on test and performance of other countries)

Sweden

N33 ADSHARK03 (unexpectedly difficult given rest of items on test and performance of other countries)

D13 ADMARIA02 (performed substantially different from 1991 to 2001)

D

Parameters for IRT Analyses of PIRLS Achievement Data

Exhibit D.1: IRT Parameters for Analyses of PIRLS Overall Reading Achievement

Scale Name	Item	Slope (aj)	S.E. (aj)	Location (bj)	S.E. (bj)	Guessing (cj)	S.E. (cj)	Step 1 (dj1)	S.E. (dj1)	Step 2 (dj2)	S.E. (dj2)	Step 3 (dj3)	S.E. (dj3)
Reading	R011A01C	0.827	0.031	-1.226	0.044	0.000	0.000						
Reading	R011A02M	1.108	0.073	0.264	0.047	0.246	0.021						
Reading	R011A03C	0.773	0.029	-0.858	0.038	0.000	0.000						
Reading	R011A04C	0.775	0.021	0.000	0.020	0.000	0.000	0.987	0.033	-0.987	0.030		
Reading	R011A05M	1.054	0.062	-0.956	0.083	0.263	0.038						
Reading	R011A06M	1.014	0.060	-1.112	0.092	0.255	0.041						
Reading	R011A07C	0.696	0.018	-0.374	0.017	0.000	0.000	0.106	0.049	0.050	0.045	-0.157	0.035
Reading	R011A08C	0.626	0.020	-0.683	0.029	0.000	0.000	0.255	0.048	-0.255	0.035		
Reading	R011A09C	0.724	0.023	0.085	0.020	0.000	0.000	0.518	0.034	-0.518	0.031		
Reading	R011A10M	1.384	0.068	0.140	0.030	0.116	0.016						
Reading	R011A11C	0.889	0.033	0.153	0.024	0.000	0.000						
Reading	R011L01M	0.526	0.046	-2.497	0.419	0.000	0.179						
Reading	R011L02M	0.779	0.078	0.805	0.064	0.236	0.024						
Reading	R011L03C	0.655	0.026	-0.288	0.033	0.000	0.000						
Reading	R011L04C	0.548	0.014	0.515	0.022	0.000	0.000	1.522	0.044	-0.901	0.047	-0.621	0.064
Reading	R011L05M	1.269	0.092	0.745	0.033	0.207	0.015						
Reading	R011L06C	0.743	0.029	0.247	0.027	0.000	0.000						
Reading	R011L07M	0.787	0.066	0.655	0.057	0.162	0.023						
Reading	R011L08C	0.816	0.025	0.723	0.020	0.000	0.000	0.659	0.027	-0.659	0.035		
Reading	R011L09M	0.951	0.055	-0.740	0.082	0.214	0.036						
Reading	R011L10C	0.769	0.025	0.767	0.021	0.000	0.000	0.144	0.030	-0.144	0.037		
Reading	R011L11M	0.932	0.064	-0.011	0.069	0.231	0.029						
Reading	R011L12C	0.822	0.026	0.756	0.020	0.000	0.000	0.687	0.027	-0.687	0.036		
Reading	R011N01M	0.835	0.062	-0.281	0.096	0.285	0.035						
Reading	R011N02M	0.804	0.073	0.306	0.083	0.290	0.030						
Reading	R011N03M	1.034	0.066	-0.643	0.082	0.294	0.035						
Reading	R011N04M	1.181	0.073	0.453	0.036	0.154	0.017						
Reading	R011N05M	1.344	0.081	0.240	0.038	0.223	0.019						
Reading	R011N06M	1.768	0.126	0.875	0.025	0.184	0.011						
Reading	R011N07C	0.607	0.022	0.642	0.025	0.000	0.000	0.254	0.038	-0.254	0.044		

Exhibit D.1: IRT Parameters for Analyses of PIRLS Overall Reading Achievement (continued)

Scale Name	Item	Slope (aj)	S.E. (aj)	Location (bj)	S.E. (bj)	Guessing (cj)	S.E. (cj)	Step 1 (dj1)	S.E. (dj1)	Step 2 (dj2)	S.E. (dj2)	Step 3 (dj3)	S.E. (dj3)
Reading	R011N08C	0.628	0.021	0.235	0.023	0.000	0.000	0.712	0.038	-0.712	0.039		
Reading	R011N09M	1.171	0.064	-0.422	0.055	0.192	0.027						
Reading	R011N10C	0.866	0.039	1.052	0.037	0.000	0.000						
Reading	R011N11M	1.058	0.076	0.446	0.047	0.205	0.021						
Reading	R011N12C	0.801	0.024	0.494	0.018	0.000	0.000	-0.032	0.032	0.032	0.034		
Reading	R011N13C	0.625	0.030	0.095	0.034	0.000	0.000						
Reading	R011R01M	0.812	0.055	0.100	0.067	0.165	0.027						
Reading	R011R02M	1.268	0.087	0.618	0.034	0.217	0.016						
Reading	R011R03M	0.800	0.052	-1.104	0.127	0.240	0.051						
Reading	R011R04C	0.876	0.031	-1.000	0.037	0.000	0.000						
Reading	R011R05C	1.129	0.037	-0.590	0.025	0.000	0.000						
Reading	R011R06C	0.604	0.015	-0.143	0.021	0.000	0.000	-0.545	0.046	0.545	0.043		
Reading	R011R07C	0.967	0.034	0.196	0.022	0.000	0.000						
Reading	R011R08C	0.813	0.025	0.348	0.017	0.000	0.000	0.304	0.029	-0.304	0.030		
Reading	R011R09C	0.643	0.020	-0.029	0.021	0.000	0.000	0.134	0.040	-0.134	0.036		
Reading	R011R10C	0.364	0.012	0.372	0.028	0.000	0.000	0.936	0.074	0.593	0.066	-1.529	0.079
Reading	R011R11C	0.640	0.020	0.007	0.018	0.000	0.000	0.404	0.050	0.168	0.044	-0.571	0.040
Reading	R011C01C	1.411	0.044	-0.075	0.018	0.000	0.000						
Reading	R011C02C	0.845	0.032	0.397	0.025	0.000	0.000						
Reading	R011C03C	1.322	0.042	-0.385	0.021	0.000	0.000						
Reading	R011C04M	1.346	0.080	0.458	0.031	0.174	0.015						
Reading	R011C05M	0.948	0.075	0.014	0.079	0.358	0.029						
Reading	R011C06C	1.163	0.038	-0.030	0.020	0.000	0.000						
Reading	R011C07M	1.175	0.068	-0.281	0.057	0.259	0.026						
Reading	R011C08C	0.651	0.018	0.353	0.019	0.000	0.000	-0.259	0.038	0.259	0.040		
Reading	R011C09M	1.297	0.086	0.692	0.030	0.158	0.014						
Reading	R011C10C	0.671	0.018	0.390	0.015	0.000	0.000	0.153	0.039	-0.172	0.045	0.020	0.044
Reading	R011C11C	0.812	0.025	0.269	0.019	0.000	0.000	0.698	0.030	-0.698	0.030		
Reading	R011C12M	1.000	0.081	0.431	0.057	0.269	0.024						
Reading	R011C13M	0.942	0.075	0.484	0.055	0.212	0.024						
Reading	R011F01M	1.496	0.073	-0.342	0.038	0.194	0.021						
Reading	R011F02M	0.667	0.049	-0.565	0.127	0.185	0.046						
Reading	R011F03M	0.956	0.054	-0.456	0.069	0.188	0.031						
Reading	R011F04M	1.276	0.069	-0.705	0.058	0.247	0.030						
Reading	R011F05M	1.009	0.062	-0.180	0.063	0.233	0.027						
Reading	R011F06C	0.831	0.030	-0.240	0.027	0.000	0.000						
Reading	R011F07C	0.495	0.014	0.532	0.024	0.000	0.000	-0.769	0.052	0.769	0.055		
Reading	R011F08C	1.092	0.036	-0.095	0.021	0.000	0.000						
Reading	R011F09C	1.061	0.028	-0.458	0.017	0.000	0.000	0.068	0.031	-0.068	0.024		
Reading	R011F10C	0.912	0.034	-1.077	0.040	0.000	0.000						
Reading	R011F11M	0.680	0.054	0.252	0.083	0.126	0.032						
Reading	R011F12C	0.658	0.020	0.755	0.022	0.000	0.000	-0.273	0.037	0.273	0.043		

Exhibit D.1: IRT Parameters for Analyses of PIRLS Overall Reading Achievement (continued)

Scale Name	Item	Slope (aj)	S.E. (aj)	Location (bj)	S.E. (bj)	Guessing (cj)	S.E. (cj)	Step 1 (dj1)	S.E. (dj1)	Step 2 (dj2)	S.E. (dj2)	Step 3 (dj3)	S.E. (dj3)
Reading	R011F13M	1.037	0.071	0.207	0.055	0.222	0.025						
Reading	R011H01M	0.690	0.059	-1.454	0.226	0.356	0.072						
Reading	R011H02M	0.919	0.054	-1.430	0.107	0.166	0.048						
Reading	R011H03C	0.320	0.015	0.873	0.049	0.000	0.000	0.741	0.066	-0.741	0.081		
Reading	R011H04C	0.868	0.032	-1.105	0.041	0.000	0.000						
Reading	R011H05M	1.109	0.058	-1.014	0.069	0.161	0.034						
Reading	R011H06M	0.724	0.045	-0.600	0.095	0.101	0.037						
Reading	R011H07C	0.608	0.020	-0.550	0.027	0.000	0.000	0.292	0.048	-0.292	0.037		
Reading	R011H08C	0.788	0.031	0.006	0.027	0.000	0.000						
Reading	R011H09C	0.726	0.022	-0.681	0.026	0.000	0.000	0.080	0.045	-0.080	0.034		
Reading	R011H10C	0.644	0.016	0.475	0.017	0.000	0.000	-0.134	0.049	1.126	0.048	-0.992	0.046
Reading	R011H11M	1.289	0.069	-0.438	0.051	0.182	0.027						
Reading	R011M01M	1.320	0.076	-0.513	0.057	0.313	0.027						
Reading	R011M02M	1.188	0.069	-1.096	0.078	0.294	0.037						
Reading	R011M03M	1.288	0.074	0.332	0.034	0.183	0.016						
Reading	R011M04C	0.838	0.034	0.791	0.030	0.000	0.000						
Reading	R011M05M	1.187	0.066	-0.413	0.057	0.254	0.027						
Reading	R011M06C	1.071	0.029	-0.343	0.016	0.000	0.000	0.274	0.028	-0.274	0.022		
Reading	R011M07C	1.095	0.036	-0.550	0.025	0.000	0.000						
Reading	R011M08M	1.144	0.101	0.852	0.042	0.267	0.017						
Reading	R011M09M	1.128	0.058	-0.530	0.056	0.174	0.027						
Reading	R011M10C	1.184	0.044	-1.353	0.037	0.000	0.000						
Reading	R011M11C	0.839	0.033	0.575	0.027	0.000	0.000						
Reading	R011M12C	0.629	0.020	0.728	0.019	0.000	0.000	0.727	0.037	-0.074	0.042	-0.654	0.056
Reading	R011M13M	0.970	0.075	-0.045	0.080	0.346	0.030						
Reading	R011M14C	0.961	0.034	-0.141	0.024	0.000	0.000						
Reading	R011M01M	1.320	0.076	-0.513	0.057	0.313	0.027						
Reading	R011M02M	1.188	0.069	-1.096	0.078	0.294	0.037						
Reading	R011M03M	1.288	0.074	0.332	0.034	0.183	0.016						
Reading	R011M04C	0.838	0.034	0.791	0.030	0.000	0.000						
Reading	R011M05M	1.187	0.066	-0.413	0.057	0.254	0.027						
Reading	R011M06C	1.071	0.029	-0.343	0.016	0.000	0.000	0.274	0.028	-0.274	0.022		
Reading	R011M07C	1.095	0.036	-0.550	0.025	0.000	0.000						
Reading	R011M08M	1.144	0.101	0.852	0.042	0.267	0.017						
Reading	R011M09M	1.128	0.058	-0.530	0.056	0.174	0.027						
Reading	R011M10C	1.184	0.044	-1.353	0.037	0.000	0.000						
Reading	R011M11C	0.839	0.033	0.575	0.027	0.000	0.000						
Reading	R011M12C	0.629	0.020	0.728	0.019	0.000	0.000	0.727	0.037	-0.074	0.042	-0.654	0.056
Reading	R011M13M	0.970	0.075	-0.045	0.080	0.346	0.030						
Reading	R011M14C	0.961	0.034	-0.141	0.024	0.000	0.000						

Exhibit D.2: IRT Parameters for Analyses of PIRLS Reading Achievement for Literary Purposes

Scale Name	Item	Slope (aj)	S.E. (aj)	Location (bj)	S.E. (bj)	Guessing (cj)	S.E. (cj)	Step 1 (dj1)	S.E. (dj1)	Step 2 (dj2)	S.E. (dj2)	Step 3 (dj3)	S.E. (dj3)
Literary	R011C01C	1.393	0.043	-0.095	0.018	0.000	0.000						
Literary	R011C02C	0.823	0.031	0.390	0.025	0.000	0.000						
Literary	R011C03C	1.286	0.040	-0.419	0.022	0.000	0.000						
Literary	R011C04M	1.311	0.077	0.455	0.031	0.175	0.015						
Literary	R011C05M	0.918	0.070	-0.019	0.079	0.352	0.028						
Literary	R011C06C	1.142	0.037	-0.049	0.020	0.000	0.000						
Literary	R011C07M	1.174	0.067	-0.281	0.055	0.272	0.025						
Literary	R011C08C	0.637	0.018	0.345	0.020	0.000	0.000	-0.259	0.039	0.259	0.040		
Literary	R011C09M	1.225	0.081	0.711	0.032	0.162	0.014						
Literary	R011C10C	0.647	0.017	0.384	0.016	0.000	0.000	0.161	0.040	-0.181	0.046	0.020	0.046
Literary	R011C11C	0.783	0.024	0.256	0.020	0.000	0.000	0.726	0.031	-0.726	0.031		
Literary	R011C12M	0.971	0.077	0.432	0.057	0.271	0.023						
Literary	R011C13M	0.897	0.070	0.494	0.057	0.217	0.023						
Literary	R011F01M	1.517	0.072	-0.346	0.036	0.195	0.020						
Literary	R011F02M	0.697	0.049	-0.474	0.113	0.220	0.039						
Literary	R011F03M	0.971	0.053	-0.448	0.064	0.195	0.028						
Literary	R011F04M	1.282	0.067	-0.710	0.055	0.253	0.027						
Literary	R011F05M	1.002	0.060	-0.171	0.061	0.240	0.026						
Literary	R011F06C	0.846	0.029	-0.239	0.027	0.000	0.000						
Literary	R011F07C	0.506	0.014	0.532	0.024	0.000	0.000	-0.736	0.051	0.736	0.054		
Literary	R011F08C	1.109	0.036	-0.096	0.021	0.000	0.000						
Literary	R011F09C	1.117	0.030	-0.465	0.017	0.000	0.000	0.107	0.030	-0.107	0.023		
Literary	R011F10C	0.925	0.033	-1.091	0.039	0.000	0.000						
Literary	R011F11M	0.694	0.054	0.338	0.076	0.160	0.028						
Literary	R011F12C	0.634	0.019	0.773	0.023	0.000	0.000	-0.287	0.039	0.287	0.045		
Literary	R011F13M	1.000	0.067	0.208	0.055	0.223	0.024						
Literary	R011H01M	0.646	0.049	-1.691	0.221	0.280	0.075						
Literary	R011H02M	0.944	0.055	-1.397	0.103	0.191	0.045						
Literary	R011H03C	0.320	0.015	0.874	0.049	0.000	0.000	0.746	0.066	-0.746	0.081		
Literary	R011H04C	0.882	0.032	-1.113	0.040	0.000	0.000						
Literary	R011H05M	1.176	0.062	-0.971	0.063	0.189	0.031						
Literary	R011H06M	0.808	0.048	-0.463	0.079	0.156	0.031						
Literary	R011H07C	0.631	0.020	-0.549	0.026	0.000	0.000	0.312	0.046	-0.312	0.036		
Literary	R011H08C	0.837	0.031	0.008	0.026	0.000	0.000						
Literary	R011H09C	0.779	0.023	-0.676	0.024	0.000	0.000	0.125	0.043	-0.125	0.032		
Literary	R011H10C	0.639	0.016	0.474	0.017	0.000	0.000	-0.118	0.050	1.132	0.048	-1.014	0.046
Literary	R011H11M	1.216	0.064	-0.467	0.052	0.188	0.026						
Literary	R011M01M	1.231	0.068	-0.591	0.058	0.293	0.026						
Literary	R011M02M	1.127	0.065	-1.200	0.082	0.271	0.037						
Literary	R011M03M	1.268	0.071	0.326	0.034	0.181	0.016						
Literary	R011M04C	0.845	0.033	0.789	0.030	0.000	0.000						
Literary	R011M05M	1.172	0.063	-0.440	0.055	0.251	0.025						

Exhibit D.2: IRT Parameters for Analyses of PIRLS Reading Achievement for Literary Purposes (continued)

Scale Name	Item	Slope (aj)	S.E. (aj)	Location (bj)	S.E. (bj)	Guessing (cj)	S.E. (cj)	Step 1 (dj1)	S.E. (dj1)	Step 2 (dj2)	S.E. (dj2)	Step 3 (dj3)	S.E. (dj3)
Literary	R011M06C	1.103	0.030	-0.360	0.016	0.000	0.000	0.303	0.028	-0.303	0.022		
Literary	R011M07C	1.047	0.034	-0.591	0.027	0.000	0.000						
Literary	R011M08M	1.132	0.098	0.876	0.042	0.270	0.016						
Literary	R011M09M	1.142	0.057	-0.540	0.052	0.181	0.024						
Literary	R011M10C	1.149	0.042	-1.428	0.038	0.000	0.000						
Literary	R011M11C	0.841	0.033	0.573	0.026	0.000	0.000						
Literary	R011M12C	0.624	0.020	0.732	0.020	0.000	0.000	0.753	0.038	-0.081	0.042	-0.672	0.056
Literary	R011M13M	0.952	0.070	-0.062	0.078	0.344	0.029						
Literary	R011M14C	0.931	0.032	-0.162	0.025	0.000	0.000						

Exhibit D.3: IRT Parameters for Analyses of PIRLS Reading Achievement for Informational Purposes

Scale Name	Item	Slope (aj)	S.E. (aj)	Location (bj)	S.E. (bj)	Guessing (cj)	S.E. (cj)	Step 1 (dj1)	S.E. (dj1)	Step 2 (dj2)	S.E. (dj2)	Step 3 (dj3)	S.E. (dj3)
Information	R011A01C	0.788	0.028	-1.337	0.046	0.000	0.000						
Information	R011A02M	1.189	0.075	0.253	0.042	0.254	0.019						
Information	R011A03C	0.748	0.027	-0.935	0.040	0.000	0.000						
Information	R011A04C	0.778	0.021	-0.044	0.020	0.000	0.000	1.007	0.034	-1.007	0.030		
Information	R011A05M	0.975	0.059	-1.043	0.090	0.277	0.035						
Information	R011A06M	0.998	0.060	-1.156	0.092	0.277	0.037						
Information	R011A07C	0.739	0.019	-0.410	0.017	0.000	0.000	0.183	0.048	0.026	0.043	-0.210	0.033
Information	R011A08C	0.640	0.020	-0.728	0.028	0.000	0.000	0.289	0.048	-0.289	0.035		
Information	R011A09C	0.729	0.023	0.048	0.020	0.000	0.000	0.533	0.034	-0.533	0.031		
Information	R011A10M	1.419	0.069	0.129	0.029	0.128	0.015						
Information	R011A11C	0.869	0.033	0.117	0.024	0.000	0.000						
Information	R011L01M	0.587	0.040	-1.963	0.213	0.208	0.071						
Information	R011L02M	0.810	0.076	0.785	0.059	0.235	0.022						
Information	R011L03C	0.672	0.026	-0.273	0.032	0.000	0.000						
Information	R011L04C	0.583	0.015	0.504	0.021	0.000	0.000	1.463	0.041	-0.854	0.044	-0.609	0.060
Information	R011L05M	1.254	0.090	0.736	0.033	0.204	0.015						
Information	R011L06C	0.777	0.030	0.249	0.026	0.000	0.000						
Information	R011L07M	0.822	0.066	0.664	0.052	0.170	0.021						
Information	R011L08C	0.858	0.026	0.707	0.019	0.000	0.000	0.638	0.026	-0.638	0.034		
Information	R011L09M	1.004	0.058	-0.661	0.073	0.240	0.031						
Information	R011L10C	0.825	0.027	0.744	0.019	0.000	0.000	0.150	0.028	-0.150	0.035		
Information	R011L11M	0.967	0.065	0.022	0.062	0.243	0.026						
Information	R011L12C	0.882	0.027	0.732	0.019	0.000	0.000	0.659	0.025	-0.659	0.034		
Information	R011N01M	0.869	0.060	-0.288	0.082	0.281	0.030						
Information	R011N02M	0.842	0.066	0.220	0.070	0.262	0.026						
Information	R011N03M	1.131	0.070	-0.548	0.069	0.331	0.029						
Information	R011N04M	1.304	0.078	0.456	0.031	0.161	0.015						
Information	R011N05M	1.542	0.089	0.244	0.032	0.224	0.016						
Information	R011N06M	1.898	0.129	0.835	0.023	0.178	0.011						
Information	R011N07C	0.659	0.023	0.612	0.023	0.000	0.000	0.266	0.035	-0.266	0.041		
Information	R011N08C	0.657	0.021	0.218	0.022	0.000	0.000	0.711	0.037	-0.711	0.037		
Information	R011N09M	1.232	0.065	-0.407	0.048	0.200	0.023						
Information	R011N10C	0.903	0.041	1.020	0.035	0.000	0.000						
Information	R011N11M	1.083	0.071	0.378	0.041	0.181	0.019						
Information	R011N12C	0.706	0.022	0.499	0.020	0.000	0.000	-0.082	0.036	0.082	0.039		
Information	R011N13C	0.556	0.028	0.050	0.038	0.000	0.000						
Information	R011R01M	0.868	0.056	0.160	0.058	0.175	0.024						
Information	R011R02M	1.423	0.091	0.603	0.030	0.210	0.015						
Information	R011R03M	0.895	0.056	-0.901	0.100	0.283	0.040						
Information	R011R04C	0.938	0.033	-0.908	0.035	0.000	0.000						
Information	R011R05C	1.281	0.042	-0.497	0.023	0.000	0.000						
Information	R011R06C	0.694	0.017	-0.083	0.019	0.000	0.000	-0.435	0.041	0.435	0.037		

Exhibit D.3: IRT Parameters for Analyses of PIRLS Reading Achievement for Informational Purposes (continued)

Scale Name	Item	Slope (aj)	S.E. (aj)	Location (bj)	S.E. (bj)	Guessing (cj)	S.E. (cj)	Step 1 (dj1)	S.E. (dj1)	Step 2 (dj2)	S.E. (dj2)	Step 3 (dj3)	S.E. (dj3)
Information	R011R07C	1.079	0.037	0.226	0.020	0.000	0.000						
Information	R011R08C	0.884	0.027	0.367	0.016	0.000	0.000	0.298	0.027	-0.298	0.028		
Information	R011R09C	0.641	0.021	-0.004	0.021	0.000	0.000	0.123	0.040	-0.123	0.036		
Information	R011R10C	0.372	0.013	0.388	0.028	0.000	0.000	0.919	0.073	0.577	0.064	-1.497	0.077
Information	R011R11C	0.688	0.021	0.035	0.017	0.000	0.000	0.410	0.047	0.148	0.041	-0.558	0.037

Exhibit D.4: IRT Parameters for Analyses of Overall Reading Literacy Achievement

Scale Name	Item	Slope (aj)	S.E. (aj)	Location (bj)	S.E. (bj)	Guessing (cj)	S.E. (cj)
Reading	ADISLAN1	0.623	0.030	-1.422	0.130	0.167	0.052
Reading	ADISLAN2	0.549	0.027	-2.867	0.177	0.163	0.067
Reading	ADISLAN3	0.625	0.038	0.241	0.077	0.187	0.026
Reading	ADISLAN4	0.491	0.027	-1.129	0.159	0.152	0.051
Reading	ADMARIA1	0.491	0.024	-2.430	0.177	0.150	0.061
Reading	ADMARIA2	0.580	0.030	-0.934	0.114	0.125	0.041
Reading	ADMARIA3	0.653	0.028	-1.844	0.123	0.147	0.053
Reading	ADBOTTL1	0.833	0.036	-1.767	0.105	0.184	0.054
Reading	ADBOTTL2	0.921	0.044	-0.030	0.050	0.263	0.020
Reading	ADBOTTL3	0.890	0.037	-2.050	0.094	0.146	0.052
Reading	ADBOTTL4	0.811	0.034	-1.278	0.085	0.161	0.041
Reading	ADBUSES1	0.828	0.024	-1.726	0.038	0.000	0.000
Reading	ADBUSES2	0.778	0.019	-0.598	0.021	0.000	0.000
Reading	ADBUSES3	0.695	0.022	1.302	0.037	0.000	0.000
Reading	ADBUSES4	0.542	0.016	-0.123	0.025	0.000	0.000
Reading	ADCONTE1	0.706	0.025	-2.302	0.061	0.000	0.000
Reading	ADCONTE2	1.271	0.050	-1.907	0.056	0.111	0.037
Reading	ADCONTE3	0.919	0.036	-1.491	0.073	0.138	0.039
Reading	ADTEMPR1	0.542	0.033	-1.114	0.179	0.234	0.058
Reading	ADTEMPR3	0.776	0.055	1.127	0.047	0.233	0.014
Reading	ADTEMPR4	0.711	0.043	0.604	0.054	0.194	0.019
Reading	ADTEMPR5	0.817	0.035	-0.702	0.067	0.156	0.030
Reading	AEPCARD1	0.660	0.031	-2.243	0.150	0.184	0.067
Reading	AEPCARD2	0.792	0.040	-2.815	0.146	0.194	0.076
Reading	AEWALRU1	0.603	0.032	-2.670	0.197	0.216	0.082
Reading	AEWALRU2	0.831	0.039	-2.317	0.125	0.183	0.067
Reading	AEWALRU3	0.793	0.035	-1.150	0.088	0.168	0.041
Reading	AEWALRU4	0.783	0.033	-1.336	0.090	0.145	0.042
Reading	AEWALRU5	0.469	0.025	-1.386	0.169	0.143	0.054
Reading	AEWALRU6	0.806	0.037	-0.228	0.059	0.165	0.025
Reading	AEQSAND1	0.622	0.039	-0.400	0.119	0.277	0.038
Reading	AEQSAND2	0.990	0.038	-0.871	0.052	0.153	0.027
Reading	AEQSAND3	0.929	0.038	-1.183	0.071	0.182	0.036
Reading	AEMARMO1	0.775	0.044	0.274	0.059	0.218	0.022
Reading	AEMARMO2	0.808	0.039	0.292	0.045	0.152	0.018
Reading	AEMARMO3	0.672	0.038	0.539	0.055	0.140	0.020
Reading	AEMARMO4	0.733	0.048	0.743	0.054	0.234	0.018
Reading	AETREES1	0.930	0.042	-0.855	0.070	0.217	0.033
Reading	AETREES2	0.467	0.031	-0.044	0.126	0.123	0.037
Reading	AETREES3	0.626	0.039	0.088	0.088	0.187	0.030
Reading	AETREES4	0.684	0.047	0.764	0.060	0.199	0.020
Reading	AETREES5	0.768	0.047	-0.282	0.093	0.328	0.032

Exhibit D.4: IRT Parameters for Analyses of Overall Reading Literacy Achievement (continued)

Scale Name	Item	Slope (aj)	S.E. (aj)	Location (bj)	S.E. (bj)	Guessing (cj)	S.E. (cj)
Reading	AETREES6	0.138	0.020	0.000	0.844	0.224	0.075
Reading	ANBIRD01	0.951	0.038	-0.183	0.041	0.153	0.018
Reading	ANBIRD02	0.547	0.026	-0.913	0.114	0.122	0.040
Reading	ANBIRD03	0.782	0.033	0.092	0.042	0.092	0.017
Reading	ANBIRD04	1.111	0.045	-1.020	0.054	0.173	0.030
Reading	ANBIRD05	1.165	0.051	-1.873	0.076	0.189	0.049
Reading	ANNODOG1	0.823	0.041	-0.906	0.092	0.280	0.038
Reading	ANNODOG2	0.812	0.039	-0.017	0.055	0.196	0.022
Reading	ANNODOG3	0.909	0.043	-0.819	0.076	0.237	0.035
Reading	ANNODOG4	1.054	0.045	-0.137	0.041	0.221	0.018
Reading	ANNODOG5	0.725	0.036	0.231	0.053	0.136	0.020
Reading	ANNODOG6	1.252	0.055	-0.759	0.049	0.273	0.026
Reading	ANSHARK1	1.114	0.044	-1.103	0.055	0.200	0.030
Reading	ANSHARK2	0.806	0.032	-1.194	0.075	0.130	0.036
Reading	ANSHARK3	0.950	0.043	-0.939	0.070	0.229	0.034
Reading	ANSHARK4	0.793	0.037	-0.568	0.074	0.217	0.030
Reading	ANSHARK5	0.865	0.038	-0.603	0.064	0.203	0.028
Reading	ANGRAPA1	0.938	0.038	-0.790	0.059	0.165	0.028
Reading	ANGRAPA2	1.510	0.053	-0.536	0.028	0.165	0.016
Reading	ANGRAPA3	0.914	0.042	-0.433	0.059	0.225	0.026
Reading	ANGRAPA4	0.774	0.034	-0.645	0.072	0.150	0.031
Reading	ANGRAPA5	1.331	0.058	-0.687	0.044	0.263	0.024
Reading	ANGRAPA6	0.972	0.041	-0.315	0.047	0.176	0.021

Exhibit D.5: IRT Parameters for Analyses of Reading Literacy for Narrative Purposes

Scale Name	Item	Slope (aj)	S.E. (aj)	Location (bj)	S.E. (bj)	Guessing (cj)	S.E. (cj)
Narrative	ANBIRD01	0.908	0.036	-0.199	0.044	0.132	0.020
Narrative	ANBIRD02	0.564	0.025	-0.902	0.100	0.101	0.037
Narrative	ANBIRD03	0.728	0.030	0.084	0.045	0.072	0.018
Narrative	ANBIRD04	1.174	0.043	-1.043	0.045	0.110	0.027
Narrative	ANBIRD05	1.228	0.051	-1.814	0.067	0.156	0.048
Narrative	ANNODOG1	0.747	0.035	-1.091	0.105	0.193	0.046
Narrative	ANNODOG2	0.759	0.037	-0.052	0.063	0.170	0.025
Narrative	ANNODOG3	0.886	0.040	-0.897	0.078	0.182	0.038
Narrative	ANNODOG4	1.034	0.044	-0.131	0.043	0.213	0.019
Narrative	ANNODOG5	0.680	0.035	0.238	0.060	0.125	0.023
Narrative	ANNODOG6	1.273	0.056	-0.769	0.049	0.257	0.027
Narrative	ANSHARK1	1.117	0.040	-1.162	0.050	0.126	0.030
Narrative	ANSHARK2	0.851	0.030	-1.176	0.061	0.092	0.031
Narrative	ANSHARK3	0.974	0.042	-0.965	0.067	0.182	0.035
Narrative	ANSHARK4	0.740	0.032	-0.713	0.078	0.139	0.034
Narrative	ANSHARK5	0.866	0.034	-0.693	0.060	0.133	0.029
Narrative	ANGRAPA1	1.032	0.040	-0.752	0.051	0.151	0.027
Narrative	ANGRAPA2	1.696	0.058	-0.541	0.024	0.135	0.015
Narrative	ANGRAPA3	0.942	0.041	-0.471	0.057	0.190	0.026
Narrative	ANGRAPA4	0.818	0.035	-0.617	0.067	0.141	0.031
Narrative	ANGRAPA5	1.594	0.068	-0.643	0.035	0.259	0.021
Narrative	ANGRAPA6	1.011	0.039	-0.363	0.043	0.134	0.021

Exhibit D.6: IRT Parameters for Analyses of Reading Literacy for Expository Purposes

Scale Name	Item	Slope (aj)	S.E. (aj)	Location (bj)	S.E. (bj)	Guessing (cj)	S.E. (cj)
Expository	AEPCARD1	0.675	0.031	-2.175	0.140	0.172	0.066
Expository	AEPCARD2	0.834	0.042	-2.661	0.139	0.195	0.076
Expository	AEWALRU1	0.703	0.035	-2.384	0.155	0.198	0.075
Expository	AEWALRU2	0.951	0.044	-2.116	0.103	0.173	0.063
Expository	AEWALRU3	0.853	0.033	-1.156	0.069	0.115	0.036
Expository	AEWALRU4	0.853	0.034	-1.271	0.077	0.129	0.040
Expository	AEWALRU5	0.503	0.025	-1.315	0.146	0.130	0.050
Expository	AEWALRU6	0.796	0.036	-0.264	0.061	0.141	0.026
Expository	AEQSAND1	0.574	0.032	-0.643	0.129	0.179	0.045
Expository	AEQSAND2	1.116	0.041	-0.797	0.045	0.152	0.025
Expository	AEQSAND3	1.003	0.040	-1.116	0.065	0.174	0.036
Expository	AEMARMO1	0.778	0.045	0.326	0.061	0.228	0.022
Expository	AEMARMO2	0.872	0.042	0.363	0.042	0.173	0.017
Expository	AEMARMO3	0.712	0.040	0.594	0.053	0.154	0.019
Expository	AEMARMO4	0.745	0.049	0.820	0.055	0.247	0.018
Expository	AETREES1	0.949	0.041	-0.897	0.068	0.175	0.035
Expository	AETREES2	0.502	0.032	-0.008	0.118	0.128	0.037
Expository	AETREES3	0.610	0.038	0.050	0.095	0.169	0.032
Expository	AETREES4	0.681	0.045	0.736	0.062	0.187	0.021
Expository	AETREES5	0.720	0.045	-0.378	0.108	0.293	0.038
Expository	AETREES6	0.580	0.125	3.205	0.308	0.300	0.014

Exhibit D.7: IRT Parameters for Analyses of Reading Literacy for Document Purposes

Scale Name	Item	Slope (aj)	S.E. (aj)	Location (bj)	S.E. (bj)	Guessing (cj)	S.E. (cj)
Document	ADISLAN1	0.766	0.037	-1.100	0.107	0.230	0.046
Document	ADISLAN2	0.607	0.030	-2.566	0.175	0.191	0.075
Document	ADISLAN3	0.673	0.037	0.238	0.068	0.176	0.024
Document	ADISLAN4	0.564	0.030	-0.946	0.137	0.166	0.049
Document	ADMARIA1	0.492	0.025	-2.348	0.188	0.165	0.066
Document	ADMARIA2	0.623	0.033	-0.805	0.113	0.144	0.043
Document	ADMARIA3	0.677	0.033	-1.669	0.143	0.202	0.063
Document	ADBOTTL1	1.111	0.047	-1.468	0.070	0.194	0.043
Document	ADBOTTL2	1.031	0.043	-0.099	0.041	0.214	0.019
Document	ADBOTTL3	1.127	0.047	-1.751	0.074	0.166	0.048
Document	ADBOTTL4	0.954	0.035	-1.183	0.060	0.119	0.033
Document	ADBUSES1	0.990	0.028	-1.522	0.030	0.000	0.000
Document	ADBUSES2	0.935	0.022	-0.504	0.018	0.000	0.000
Document	ADBUSES3	0.765	0.023	1.283	0.033	0.000	0.000
Document	ADBUSES4	0.638	0.017	-0.070	0.022	0.000	0.000
Document	ADCONTE1	0.830	0.027	-2.033	0.048	0.000	0.000
Document	ADCONTE2	1.616	0.065	-1.703	0.042	0.108	0.032
Document	ADCONTE3	1.039	0.036	-1.416	0.051	0.085	0.029
Document	ADTEMPR1	0.620	0.036	-0.932	0.148	0.253	0.052
Document	ADTEMPR3	0.927	0.057	1.095	0.039	0.239	0.012
Document	ADTEMPR4	0.777	0.043	0.623	0.048	0.197	0.018

TYPOGRAPHY
Set in Apollo, Frutiger, and Meridien,
designed by Adrian Frutiger.

PHOTOGRAPHS
Getty Images

COVER & BOOK DESIGN
José R. Nieto

PRODUCTION COORDINATOR
Betty Hugh

LAYOUT
Betty Hugh and Mario A. Pita