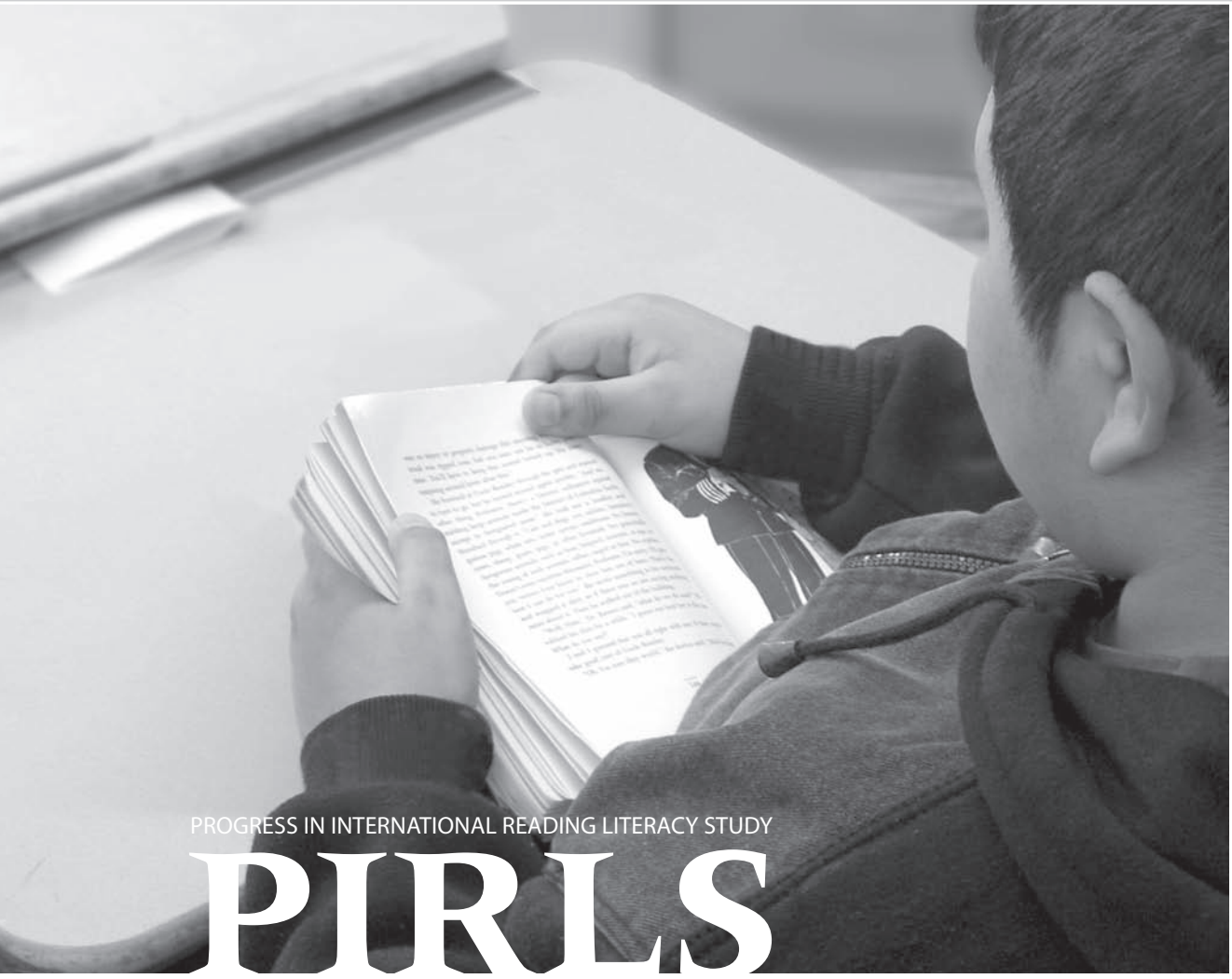
PROGRESS IN INTERNATIONAL READING LITERACY STUDY

PIRLS

Edited by
Michael O. Martin
Ina V.S. Mullis
Ann M. Kennedy

PIRLS 2006 Technical Report

TIMSS & PIRLS
International Study Center
Lynch School of Education, Boston College



PROGRESS IN INTERNATIONAL READING LITERACY STUDY

PIRLS

PIRLS 2006 Technical Report

Edited by

Michael O. Martin

Ina V.S. Mullis

Ann M. Kennedy

TIMSS & PIRLS
International Study Center
Lynch School of Education, Boston College

PIRLS 2006 Technical Report / Edited by Michael O. Martin,
Ina V.S. Mullis, and Ann M. Kennedy

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

Library of Congress
Catalog Card Number: 2007925071

ISBN: 1-889938-46-7

For more information about PIRLS contact:

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

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

email: pirls@bc.edu

http://pirls.bc.edu

Printed and bound in the United States.

Contents

Chapter 1

Overview of PIRLS 2006

Ina V.S. Mullis and Michael O. Martin

1.1 Background

As the recognized pioneer of international assessments, IEA has been conducting comparative studies of students' academic achievement for approximately 50 years. IEA's Progress in International Reading Literacy Study (PIRLS) provides internationally comparative data about students' reading achievement in primary school (the fourth grade in most participating countries). The fourth grade is an important transition point in children's development as readers, because most of them should have learned to read, and are now reading to learn. PIRLS has roots in earlier IEA studies, including the reading component of IEA's six-subject study in 1973 (Thorndike, 1973; Walker, 1976) and IEA's Reading Literacy Study conducted in 1991 (Elley, 1992, 1994) and again in 2001 to provide trends (Martin, Mullis, Gonzalez, & Kennedy, 2003).

PIRLS was inaugurated in 2001 to provide reliable measurement of trends in reading comprehension over time on a 5-year cycle. Thus, PIRLS 2006 is the second in a continuing assessment cycle into the future, whereby PIRLS will be conducted again in 2011 and every 5 years, thereafter. To measure trends, confounding effects due to changes from one assessment to the next must be minimized, implying a certain amount of stability in the measurement process over time. At the same time, the assessment must remain current by allowing the introduction of new curriculum concepts, addressing changes in educational priorities, and incorporating the use of new measurement technology. Thus, while PIRLS 2006 built on PIRLS 2001, it also evolved in

important ways to provide more useful and more comprehensive information to the participating countries.

Because addressing the substantive and policy issues related to better understanding of the achievement results is fundamental to IEA's goal of improving teaching and learning, PIRLS also provides extensive information about the home and school contexts for learning to read. To set the national contexts for reading education, the *PIRLS 2006 Encyclopedia: A Guide to Reading Education in the Forty PIRLS 2006 Countries* (Kennedy, Mullis, Martin, & Trong, 2007) summarizes each country's education system, reading curriculum and instruction in the primary grades, and approaches to teacher education. Also, PIRLS includes an extensive array of questionnaires to collect information from students' parents, teachers, and schools, as well as from the students themselves.

1.2 The Participants in PIRLS 2006

Exhibit 1.1 presents the countries that participated in PIRLS 2006 and in PIRLS 2001. Forty countries and 5 Canadian provinces participated in the 2006 PIRLS assessment. Of these, 26 countries and 2 provinces had trend data from PIRLS 2001.

1.3 The PIRLS 2006 Framework

The underpinnings of the PIRLS 2006 assessment are set forth in the *PIRLS 2006 Assessment Framework and Specifications* (Mullis, Kennedy, Martin, & Sainsbury, 2006). More specifically, the PIRLS 2006 framework describes the two major aspects of reading to be addressed by the PIRLS assessments. PIRLS assesses four processes of reading comprehension: focus on and retrieve explicitly stated information; make straightforward inferences; interpret and integrate ideas and information; and examine and evaluate content, language, and textual elements. The processes are assessed within the two purposes that account for most of the reading done by young students both in and out of school: reading for literary experience and reading to acquire and use information.

To guide questionnaire development, the PIRLS 2006 framework also describes the contexts for learning to read, including national and community contexts, home contexts, school contexts, and classroom contexts. Finally, the framework also presents the basic assessment design and specifications for instrument development.

Exhibit 1.1 Countries Participating in PIRLS 2006 and 2001

Countries	2006	2001
Argentina		●
Austria	●	
Belgium (Flemish)	●	
Belgium (French)	●	
Belize		●
Bulgaria	●	●
Canada, Alberta	●	
Canada, British Columbia	●	
Canada, Nova Scotia	●	
Canada, Ontario	●	●
Canada, Quebec	●	●
Chinese Taipei	●	
Colombia		●
Cyprus		●
Czech Republic		●
Denmark	●	
England	●	●
France	●	●
Georgia	●	
Germany	●	●
Greece		●
Hong Kong SAR	●	●
Hungary	●	●
Iceland	●	●
Indonesia	●	
Iran, Islamic Rep. of	●	●
Israel	●	●
Italy	●	●
[1] Kuwait	●	
Latvia	●	●
Lithuania	●	●
Luxembourg	●	
Macedonia, Rep. of	●	●
Moldova, Rep. of	●	●
Morocco	●	●
Netherlands	●	●
New Zealand	●	●
Norway	●	●
Poland	●	
Qatar	●	
Romania	●	●
Russian Federation	●	●
Scotland	●	●
Singapore	●	●
Slovak Republic	●	●
Slovenia	●	●
South Africa	●	
Spain	●	
Sweden	●	●
Trinidad and Tobago	●	
Turkey		●
United States	●	●

Indicates country participation in that testing cycle ●

1 Although Kuwait participated in PIRLS 2001, the data were not considered comparable for measuring trends.

Chapter 2 of this report describes the updates in the PIRLS framework between 2001 and 2006.

1.4 The PIRLS 2006 Test of Reading Comprehension

The PIRLS 2006 test of reading comprehension was based on 10 passages, 5 literary and 5 informational. Each passage was accompanied by approximately 12 questions, with the assessment comprised of 126 items in total. Two of the literary passages and two of the informational passages had been kept secure from the PIRLS 2001 assessment for the purposes of measuring trend, and these were carried forward for PIRLS 2006. The other three literary and three informational passages were newly developed for the 2006 assessment. The passages were identified and reviewed extensively by representatives of the participating countries. The TIMSS & PIRLS International Study Center conducted an item-writing workshop for country representatives to develop the test questions. So as not to overburden the young students participating in the assessment, PIRLS uses a rotated booklet design and the testing time is limited to 80 minutes (two passages) per student, with an additional 15–30 minutes allotted for a *Student Questionnaire*.

Chapter 2 of this report describes the instrument development process for PIRLS 2006 and provides details about the nature of the passages, items, and scoring guides for the constructed-response questions. The appendix of the *PIRLS 2006 Assessment Framework and Specifications* contains example passages, items, and scoring guides. The *PIRLS 2006 International Report* (Mullis, Martin, Kennedy, & Foy, 2007) includes the outcomes of the scale anchoring analysis conducted to describe students' achievement in terms of the strategies and skills elicited by the assessment, and the appendix contains two of the literary and two of the informational passages from the 2006 assessment.

1.5 The PIRLS 2006 Questionnaires

Building on the foundation provided in PIRLS 2001, the 2006 assessment included five questionnaires to collect data about the educational contexts for learning to read. The students answered questions pertaining to their home and school experiences in learning to read. 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. The

teachers of the sampled students responded to questions 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. The principals of schools responded to questions about enrollment and school characteristics, school organization, staffing, and resources, and the school environment. As an innovation for PIRLS 2006, the National Research Coordinator (NRC) in each country completed an online *Curriculum Questionnaire* providing data on the goals of reading instruction.

Chapter 3 of the *PIRLS 2006 Technical Report* describes the process for developing the background questionnaires and summarizes the topics covered in each of the questionnaires. Chapter 13 describes the analysis of the background questionnaire data. The *PIRLS 2006 International Report* contains the results for the PIRLS background questionnaires including the indices or scales developed for approximately a dozen of the key background factors.

1.6 Sample Design, Implementation, and Participation

As explained in Chapter 4, PIRLS 2006 had as its target population students enrolled in the fourth grade of formal schooling, counting from the first year of primary school defined by UNESCO's International Standard Classification for Education (UNESCO, 1999). Accordingly, the fourth year of formal schooling should be the fourth grade in most countries. To avoid testing very young children, however, PIRLS has a policy that the average age of children in the grade tested should not be below 9.5 years old.

The PIRLS 2006 assessment was administered to carefully drawn probability samples of students from the target population in each country. The basic design of the sample was a two-stage stratified cluster design. The first stage consisted of sampling schools, and the second stage consisted of sampling intact classrooms from the target grade in the sampled schools. Typically, countries sampled 150 schools and one or two intact classrooms. Most countries achieved the minimum acceptable participation rates—85 percent of both the schools and students, or a combined rate (the product of schools' and students' participation) of 75 percent.

Chapter 4 provides details about the PIRLS 2006 sample design, and Chapter 9 describes the procedures used in calculating sampling weights and participation rates.

1.7 Translation Verification

The PIRLS 2006 instruments were prepared in English and translated into 45 languages. Although most countries administered the assessment in just 1 language, 9 countries and the 5 Canadian provinces administered it in 2 languages, Spain administered the assessment in its 5 official languages, and South Africa administered it in 11 languages. To ensure comparability among translated instruments, the IEA Secretariat, with support from the TIMSS & PIRLS International Study Center, orchestrates a rigorous translation, translation verification, and layout verification process.

Chapter 5 contains information about the procedures used in the translation and layout verification process.

1.8 Survey Operations and Quality Assurance

Each participating country and province was responsible for carrying out all aspects of data collection, using standardized procedures developed for the study and explained in specific units of the survey operations manual and in various training manuals. These manuals covered procedures for test security, standardized scripts to regulate the testing sessions, and steps to ensure that the correct students (those sampled) were being assessed. Each country was responsible for conducting quality control procedures and describing this effort in the online Survey Activities Report. In addition, the TIMSS & PIRLS International Study Center, in conjunction with the IEA Secretariat, conducted an independent quality control program. The reports from the Quality Control Monitors indicate that, in general, national centers were able to conduct the data collection efficiently, professionally, and in compliance with international procedures.

Chapter 6 provides an overview of the data collection procedures. A description of the quality assurance program, together with the results of observations of the Quality Control Monitors, is found in Chapter 7.

1.9 The PIRLS 2006 International Database

To ensure the availability of comparable, high-quality data for analysis, PIRLS 2006 took great care in creating the international database. PIRLS 2006 prepared manuals and software for countries to use in creating and checking their data files, and once the data were forwarded to the IEA Data Processing

and Research Center (DPC) in Hamburg, the data underwent an exhaustive cleaning process. Throughout the process, the data were checked and double-checked, and the national centers were contacted regularly and given multiple opportunities to review the data for their countries.

In conjunction with the IEA DPC, the TIMSS & PIRLS International Study Center reviewed item statistics for each achievement item in each country in case there were poorly performing items. Also, the scoring reliability data were checked for the constructed-response items, including the within-country, cross-country, and trend reliability data. In general, the items exhibited very good psychometric properties in all countries, and the scoring reliability was satisfactory (around 90% in most cases).

Chapter 8 of this report describes the procedures used by countries to check their national data, and the series of editing and documentation steps taken by the IEA DPC in creating the international database. Chapter 10 describes the process of reviewing the item statistics, and includes the scoring reliability results. The PIRLS 2006 International Database is publicly available via the TIMSS and PIRLS website, and is accompanied by the *PIRLS 2006 User Guide for the International Database* (Foy & Kennedy, 2008).

1.10 Scaling and Reporting the Student Achievement Data

As described in Chapter 11, the primary approach to reporting the PIRLS 2006 achievement data was based on item response theory (IRT) scaling methods. Student reading achievement was summarized using a family of 2- and 3-parameter IRT models for dichotomously scored items, and generalized partial credit models for constructed-response items with two or three available score points. The PIRLS reading achievement scales were designed to provide reliable measures of student achievement common to both the 2001 and 2006 assessments, based on the metric established originally in 2001. For more accurate estimation of results for subpopulations of students, the PIRLS scaling made use of plausible-value technology. In addition to the scale for reading achievement overall, IRT scales were created to measure changes in achievement in the two purposes of reading and two overarching reading processes.

To provide richly descriptive information about what performance on the PIRLS reading scale means in terms of the reading skills that students have and comprehension processes and strategies they can bring to bear, PIRLS identified four points on the scale for use as international benchmarks of

student achievement. A scale anchoring analysis was conducted to interpret the PIRLS scale scores and analyze achievement at the international benchmarks.

Chapter 11 presents in-depth information about scaling the PIRLS 2006 reading achievement data. Further information about the scale-anchoring analysis is found in Chapter 12. Chapter 12 also covers the procedures for estimating sampling variance and calculating the standard errors provided together with the statistics in the *PIRLS 2006 International Report*. The *PIRLS 2006 International Report* contains the analysis results.

References

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.

Foy, P. & Kennedy, A.M. (Eds.). (2008). *PIRLS 2006 user guide for the international database*. Chestnut Hill, MA: Boston College.

Kennedy, A.M., Mullis, I.V.S., Martin, M.O., & Trong, K.L. (2007). *PIRLS 2006 encyclopedia: A guide to reading education in the forty PIRLS 2006 countries*. Chestnut Hill, MA: Boston College.

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., Kennedy, A.M., Martin, M.O., & Sainsbury, M. (2006). *PIRLS 2006 assessment framework and specifications* (2nd ed.). Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., & Foy, P. (2007). *PIRLS 2006 international report: IEA's progress in international reading literacy study in primary schools in 40 countries*. Chestnut Hill, MA: Boston College.

Thorndike, R. L. (1973). *Reading comprehension education in fifteen countries. International studies in evaluation III*. Stockholm: Almqvist and Wiksell.

UNESCO Institute for Statistics. (1999). *Operational manual for ISCED 1997: International standard classification of education*.

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

Chapter 2

Developing the PIRLS 2006 Reading Assessment and Scoring Guides

Ann M. Kennedy and Marian Sainsbury

2.1 Overview

Development of the PIRLS 2006 reading assessment began early in 2003 and continued until August 2005, when the international version of the assessment was finalized for data collection. The development was a collaborative process involving the PIRLS Reading Coordinator, the Reading Development Group, the PIRLS 2006 National Research Coordinators (NRCs) from the participating countries, the PIRLS Item Development Task Force, and TIMSS & PIRLS International Study Center staff.[1] The *PIRLS 2006 Framework and Specifications* (Mullis, Kennedy, Martin, & Sainsbury, 2006) provided the foundation for the assessment.

PIRLS 2006 was the second cycle of PIRLS, and was structured in a way that included new material that had a recognizable continuity with the previous test. In order to measure trends, the assessment was composed of passages and questions from PIRLS 2001, as well as new passages and items. The main purpose of the development process detailed here was to identify new passages and develop the accompanying items in a way that would continue and expand the range of the assessment model established in 2001. A timeline of the test development process is provided in Exhibit 2.1.

The NRCs were responsible for submitting and approving reading passages for the assessment and were directly involved in developing test items and scoring

1 Marian Sainsbury of the National Foundation for Educational Research in England served as the PIRLS Reading Coordinator. The Item Development Task Force included the PIRLS Reading Coordinator, staff from the TIMSS & PIRLS International Study Center, and Patricia Donahue of Educational Testing Service. Members of the Reading Development Group, National Research Coordinators, and TIMSS & PIRLS International Study Center staff are acknowledged in Appendix A.

guides for constructed-response items. A field test was conducted in March–April 2005 that provided information about the measurement properties of potential passages and items across the countries. Based on the field-test results, the passages and items were selected and finalized for main data collection.

Exhibit 2.1 Overview of the Test Development Process

Date	Group and Activity
March 2003	TIMSS & PIRLS International Study staff begins initial search for PIRLS 2006 passages and sends a call for passages to National Research Coordinators.
September 2003	National Research Coordinators recommend updates to the framework and begin passage review.
January 2004	Reading Development Group reviews draft PIRLS assessment framework, reviews passages and recommends initial passage pool for field-test item development, and reviews draft item and scoring guide development manual.
February 2004	National Research Coordinators give final approval of the PIRLS assessment framework, select final passages for field-test item development, and participate in an item and scoring guide development workshop.
August 2004	Reading Development Group reviews field-test item pool and scoring guides.
November 2004	National Research Coordinators finalize selection of field-test item pool and scoring guides.
March 2005	National Research Coordinators are trained in applying field-test scoring guides for constructed-response items.
March-April 2005	PIRLS 2006 field test is administered.
July 2005	Reading Development Group reviews field-test results and recommends selection for main data collection.
August 2005	National Research Coordinators review field-test results and select operational passages and items.
October-December 2005	PIRLS 2006 data collection is conducted in Southern Hemisphere countries.
November 2005	Southern Hemisphere National Research Coordinators are trained in applying scoring guides for constructed-response items.
March 2006	Northern Hemisphere National Research Coordinators are trained in applying scoring guides for constructed-response items.
March-June 2006	PIRLS 2006 data collection is conducted in Northern Hemisphere countries.

2.2 Updating the PIRLS 2006 Assessment Framework

The PIRLS 2006 assessment framework was based on the *PIRLS 2001 Framework and Specifications* (Campbell, Kelly, Mullis, Martin, & Sainsbury, 2001). The TIMSS & PIRLS International Study Center initiated discussions of any necessary updates for PIRLS 2006 of the PIRLS 2001 framework among the NRCs and PIRLS Reading Development Group. These updates to the framework

were intended to reflect the findings from PIRLS 2001, as well as current reading research since the development of the initial framework. The process began with a review of the existing framework at the first meeting of the NRCs in September 2003, resulting in minor amendments to the definition of reading literacy and expanding the discussion and description of text types used in the assessment. NRCs also suggested an extended discussion of the interaction between purpose and text type.

The PIRLS Reading Development Group met in January 2004 to review the framework in light of the recommendations from the NRCs. In general, there was a consensus among the groups. The definition of reading literacy was reworded to underscore the importance of the variety of contexts in which reading takes place. Further adaptations included elaboration of the terms "read to learn" and "communities of readers" from the definition. The definition follows.

> For PIRLS, reading literacy is defined as the ability to understand and use those written language forms required by society and/or valued by the individual. Young readers can construct meaning from a variety of texts. They read to learn, to participate in communities of readers in school and everyday life, and for enjoyment. (Mullis, Kennedy, Martin, & Sainsbury, 2006, p. 3)

Modifications to the framework's references were implemented in February 2004 following advice from the NRCs and Reading Development Group. The TIMSS & PIRLS International Study Center conducted a literature search of research articles, reports, and papers published since the publication of the 2001 framework that were relevant to the purposes for reading, processes of reading comprehension, and contexts for learning to read. A separate reference section was added to highlight reading research conducted using data from IEA studies.

The PIRLS 2006 framework was initially published in 2004, prior to the administration of the PIRLS 2006 field test. A second edition was published in February 2006 to present example reading test blocks (passages and corresponding questions) that represented the set of test blocks in the 2006 assessment. For this purpose, the appendix containing examples of two reading passages and their corresponding questions was updated with examples from the 2006 field test. The appendix also includes scoring guides for constructed-response questions.

2.3 The PIRLS 2006 Assessment Framework

The *PIRLS 2006 Assessment Framework and Specifications* contains a detailed description of the PIRLS 2006 assessment of reading comprehension. In brief, the PIRLS 2006 framework defines the two major aspects of students' reading literacy—purposes for reading and processes of comprehension. Reading for literary experience and reading to acquire and use information are the two major purposes that account for the majority of reading experiences of young children. Readers make meaning of texts in a variety of ways, depending not only on the purpose for reading, but also on the difficulty of the text and the reader's prior knowledge. PIRLS looks at four processes of comprehension: focus on and retrieve explicitly stated information; make straightforward inferences; interpret and integrate ideas and information; and examine and evaluate content, language, and textual elements. These processes are the basis for developing comprehension questions in the reading assessment. Exhibit 2.2 shows the target percentages of the reading assessment devoted to reading purposes and processes.

Exhibit 2.2 Percentages of Reading Assessment Devoted to Reading Purposes and Processes

Purposes for Reading	
Literary Experience	50%
Acquire and Use Information	50%
Processes of Comprehension	
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%

2.4 PIRLS 2006 Assessment Design

The PIRLS 2006 assessment design, also elaborated in the *PIRLS 2006 Framework and Specifications*, builds on PIRLS 2001, in which there were four literary and four informational test blocks. Several factors influenced the test booklet design used for PIRLS 2006 data collection. However, based on research analyses conducted by Germany and the TIMSS & PIRLS International Study Center using the PIRLS 2001 data,[2] NRCs requested that scaling of the PIRLS assessment be done for processes of comprehension, as well as by purposes

2 Bos, W., Lankes, E.M., Prenzel, M., Schwippert, K., Walther, G., & Valtin, R. (Hrsg.). (2003). *Ergebnisse aus IGLU: Scülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich.* New York: Waxmann.

Mullis, I.V.S., Martin, M.O., & Gonzalez, E.J. (2004). *PIRLS international achievement in the processes of reading comprehension: Results from PIRLS 2001 in 35 countries.* Chestnut Hill, MA: Boston College.

for reading. To support the creation of two reading process scales, the total assessment time required needed to increase, and the booklet design expanded to include additional test booklets.

The decision to report reading achievement scale scores by process as well as by purpose, in combination with the desire to include a range of texts within each reading purpose, made it necessary to increase PIRLS 2006 to include five literary and five informational test blocks. Each of the 10 test blocks included a reading passage and its accompanying questions. As shown in Exhibit 2.3, half of the test blocks were devoted to measuring literary purposes (labeled L1-L5), and the other half were focused on acquiring and using information (labeled I1-I5). Since four of the PIRLS 2001 test blocks were kept secure and carried forward for measuring for trends in 2006, development efforts in 2006 focused on the six remaining blocks.

Exhibit 2.3 PIRLS 2006 Student Booklet Design

Literary Block Number	Literary Title	Informational Block Number	Informational Title
L1	Lump of Clay	I1	Antarctica
L2	Flowers	I2	Leonardo
L3	To be developed	I3	To be developed
L4	To be developed	I4	To be developed
L5	To be developed	I5	To be developed

2.5 Finding and Selecting Passages

Development of the PIRLS 2006 reading assessment involved selecting passages from existing sources representative of the types of materials likely to be read by children in the fourth grade, writing items based on these texts, and devising scoring guides for constructed-response items. These new materials were to reflect the broad approaches established for PIRLS 2001, while refreshing and expanding the range of texts and devising items that brought out the qualities of each passage.

The task of selecting passages for an international assessment is a demanding one. In this case, it was desirable that the new texts have features in line with the framework and maintaining a recognizable continuity with the secure passages from PIRLS 2001. In undertaking all stages of this task, collaboration among the participating countries was a central part of the work.

Based on the need for 6 new text blocks of passages and items, it was decided to develop 12 such blocks for the field test. A call for passages was sent out to all NRCs, with a request for submission of both literary and informational reading passages to be used as the foundation for the development of test items. Research coordinators were asked to submit passages with the following characteristics:

- Suitable for fourth-grade students in content, interest, and reading ability;
- Well written in terms of depth and complexity to allow questioning across the processes and strategies defined in the PIRLS 2006 framework; and
- Sensitive to cultural groups to avoid specific cultural references, wherever possible.

The text of the passages, written in or translated into English, had to be continuous and not exceed 1,200 words. Examples of literary text include short stories, narrative extracts, traditional tales, fables, myths, and play scripts. Informational texts include textbook or expository passages, biographies, and persuasive writing and could include charts, tables, or diagrams.

To begin with, the TIMSS & PIRLS International Study Center received over 50 reading passages from NRCs in the following countries: Canada, the Czech Republic, England, France, Germany, Hong Kong SAR, Iran, Italy, New Zealand, and Singapore. Passages were circulated and reviewed at the first meeting of the coordinators. Of the passages reviewed, they chose four literary and six informational texts to be revised and edited for the field test. After examining themes and content of the selected passages, the coordinators agreed to continue to find and submit additional suitable passages that varied in style and structure.

The PIRLS Reading Development Group convened for the first time in January 2004 to review passages selected by the NRCs, as well as those additional passages submitted afterwards. Reading Development Group members made suggestions for minor text revisions for consistency of language and drafted a preliminary list of possible questions for each passage. In total, the Reading Development Group recommended eight literary and eight informational passages for the next stage of development, an item-writing workshop at the second meeting of NRCs. The passages differed by content, style, and length. The

literary passages ranged in length from 797 words to 1,127 words and included a mixture of traditional and contemporary stories with an array of characters and story plots. Informational passages ranged in length from 693 to 985 words and represented a wide range of topics and informational text structures. For each of the 16 passages presented, the Item Development Task Force constructed a text map highlighting the passage's central ideas and key features.

2.6 Developing Items and Scoring Guides

In February 2004, the TIMSS & PIRLS International Study Center convened a meeting of the NRCs to review the set of 16 passages recommended by the Reading Development Group and to write items and scoring guides for those considered most suitable. From the 16, the coordinators selected a subset of 13 passages—seven literary and six informational—for which they would create items and scoring guides for constructed-response items and recommended that a seventh informational passage be identified before the field test.

The workshop began with basic training in developing reading items. As a basis for the training, the TIMSS & PIRLS International Study Center provided NRCs with *Item-Writing Guidelines for the PIRLS 2006 Field Test* (2004). The guidelines were reviewed and discussed, including general issues for writing items and scoring guides, a system for documenting and classifying items for review, and procedures for reviewing items and scoring guides once they were written. The following is a summary of item-writing guidelines for each passage:

- Write items totaling at least 18-20 score points (approximately 12–13 items) per passage.
- Write questions that match the purpose of the passage as classified for PIRLS 2006, paying close attention to writing questions that cover the range of the four PIRLS comprehension processes.
- Write questions relevant to the central ideas in the passage, making sure that answering a question correctly depends on having read the passage.
- Each item should be independent of the other items (no item should provide "clues" to the correct response for another item).
- For each question, consider the timing, grade appropriateness, difficulty level, potential sources of bias, and ease of translation.

- For multiple-choice questions, ask direct questions, making sure there is one and only one correct answer and provide plausible distracters.
- Develop a unique, tailored scoring guide for each constructed-response item. Write a full-credit answer to each question in terms of language, knowledge, and skills of a typical fourth-grade student.

Guidelines were reviewed for constructing unique scoring guides for 1-, 2-, and 3-point constructed-response items. For each item, scoring criteria were to be as specific as possible in order to standardize scoring decisions across countries, as well as provide for a range of responses within each score level. The guidelines emphasized the features required within each scoring guide for each score level:

- A general statement about the nature of comprehension, which is characteristic of responses at that level,
- Specific content of students' responses that may be considered evidence of an appropriate inference, and
- Examples of the various types of plausible student responses.

The NRCs divided into eight groups with three to six people per group. Each of the groups was assigned at least one literary and one informational passage. At least two groups worked on each passage, in order to maximize the number and variety of items drafted for each passage. Each group reviewed their sets of items and made final revisions before supplying an electronic file of all items to the TIMSS & PIRLS International Study Center. By the end of the meeting, coordinators had drafted a total of 277 items, as well as scoring guides for all constructed-response items.

The TIMSS & PIRLS International Study Center combined and organized the draft items by passage, keeping items with similar topics or themes together and distributed the items in March 2004 to the Item Development Task Force for review. The review was based on a simple rating system of 1 to 3, ranging from exceptional or requiring minimal revision to requiring extensive modifications. Task Force members were invited to comment on individual items, as well as provide a general overview of the set of items for a particular passage. The Task Force met for 3 days in June 2004 to review the draft item pool and make suggestions for revisions and recommendations of items that should be retained for the field test. In addition to the item review, the Task Force drafted items for a new informational passage that had been identified and approved by

NRCs after their second meeting. Throughout the months of June and July, the Task Force continued to refine the field-test items for review by the Reading Development Group.

In August 2004, the Reading Development Group convened to evaluate the 14 passages and accompanying items developed to date, and recommend the 12 most suitable blocks for the field test (see Exhibit 2.4 for a list of titles). Reading Development Group members made minor suggestions for edits to the passages and questions, primarily to refine the intent of questions or how individual constructed-response items should be scored. Another focus of the Reading Development Group meeting was to ensure proper balance in terms of the processes measured, item types, and total number of points across the passages.

Exhibit 2.4 PIRLS 2006 Field Test Passages

Literary Title	Informational Title
Shiny Straw	Spacewalking
The Fox and the Rooster	Sharks
Fly Eagle	Searching for Food
Unbelievable Night	Chocolate Then and Now
Growing Money	Day Hiking
Dolphin Rescue	Ice Age Cave

There was a NRC meeting in November 2004 before conducting the field test to finalize and approve the field-test materials for production and administration. The NRCs generally endorsed the versions of the assessment blocks as revised by the Reading Development Group, with some proposed alterations to two of the informational passages. Other minor changes were made to items in the literary blocks to improve clarity of the questions and options. Following the meeting, the TIMSS & PIRLS International Study Center implemented the suggested changes and provided the final international version of the PIRLS 2006 field-test booklets to the NRCs on December 1, 2004.

2.7 Conducting the PIRLS 2006 Field Test

In preparation for the assessment data collection in 2006, PIRLS conducted a full-scale field test between March and April 2005. The field-test data provided a good basis for evaluating the measurement properties of the new assessment

blocks developed in the 2006 round. The 12 assessment blocks developed for the field test were paired into six booklets, each booklet containing one literary and one informational block. To ensure the target sample size for a minimum of 200 student responses per item, participating countries sampled between 25 and 40 schools, depending upon class sizes. In total, nearly 50,000 students from more than 1,200 schools in 42 countries participated in the field test.

2.8 Piloting Items for Scoring Guides

The PIRLS 2006 constructed-response items permit a wide range of student responses. Young students express their understanding in various ways, and, in some cases, the justification for an answer can take different forms. Because scoring these varying responses must be carried out consistently, extensive training in applying the scoring guides required collecting student responses to use as training materials. These student responses also were helpful for refining the scoring guides.

To refine the scoring guides for constructed-response items and prepare scoring training materials, seven English-speaking countries agreed to administer field-test booklets to a small selection of classrooms. In total, approximately 200 student responses to each of the constructed-response items were collected from the Canadian provinces of Ontario and Quebec, England, New Zealand, Scotland, Singapore, South Africa, and the United States.

The PIRLS Item Development Task Force met at the TIMSS & PIRLS International Study Center in February 2005 to evaluate the student responses and make adjustments to the scoring criteria to accommodate appropriate responses not accounted for in the existing scoring guides. The Task Force scored the student responses according to the guides and grouped the responses to each item by score category. Next, the Task Force reassessed the sets of scored responses and reconciled scores for responses not easily categorized. To create sets of student responses for training scorers, the Task Force selected from among the scored responses from the item pilot.

2.9 Scoring Training for Constructed-response Items from the PIRLS 2006 Field Test

In March 2005, the TIMSS & PIRLS International Study Center held a meeting for the NRCs and their scoring managers who would implement the constructed-response scoring in each participating country. The majority of the meeting

consisted of a 4-day training session in the application of the scoring guides for constructed-response items for the field test. For each item, a set of training materials was provided. The training materials included 8–10 anchor papers and a set of 8–10 practice papers for each of the 76 constructed-response items, arranged by assessment block. Sets of anchor and practice papers contained student responses collected from the item pilot.

Following the review of the text for each scoring guide, the participants were provided with a set of anchor papers comprised of at least three example student responses for each of the score-point categories for an item. Rationales for the score assigned to each anchor paper were included for the set and discussed during the presentation of each example. Upon completion of the review of anchor papers for an item, the NRCs and scorers read unscored student responses in the practice paper sets and participated in an open discussion of rationales for scoring the practice items.

2.10 Selecting Final Reading Passages for the PIRLS 2006 Data Collection

The Reading Development Group met in July 2005 to study the results of the PIRLS 2006 field test and recommend three literary and three informational blocks for inclusion in the main data collection. Criteria for recommendations for the assessment included desirable overall passage statistics and individual item statistics in addition to well-suited blocks representative of the reading experiences of fourth graders internationally. Item statistics were used to evaluate the effectiveness of the items and identify items requiring revisions. The group was generally pleased with the measurement characteristics of the items and proposed 6 of the 10 blocks be presented to the NRCs for inclusion in PIRLS 2006 and two that would be used in the PIRLS framework as examples of the PIRLS assessment.

The NRCs reviewed the item statistics from the field test at their fifth meeting in August 2005. An extensive discussion of the field-test results and the qualities of the secured PIRLS 2001 trend blocks, which also would appear in the 2006 assessment, led NRCs to adopt the recommendations of the Reading Development Group, with the provision that one literary block recommended for the framework replace another block recommended for the assessment. Based on item statistics, a number of individual items were identified by the coordinators and modified to improve clarity and accuracy of student responses.

Following the review, TIMSS & PIRLS International Study Center staff implemented edits to the blocks. The new assessment blocks developed for PIRLS 2006 were combined with the secure blocks from the 2001 assessment, providing an overall assessment that would allow the calculation of trends over 5 years, as well as containing new material. Finalized assessment materials were made available to the NRCs on August 15, 2005, in preparation for the main data collection, which began in Southern Hemisphere countries in October 2005. Exhibit 2.5 lists the PIRLS 2006 passage titles by block.

Exhibit 2.5 PIRLS 2006 Student Booklet Design

Literary Block Number	Literary Title	Informational Block Number	Informational Title
L1	Lump of Clay (2001)	I1	Antarctica (2001)
L2	Flowers (2001)	I2	Leonardo (2001)
L3	Shiny Straw (2006)	I3	Day Hiking (2006)
L4	Fly Eagle (2006)	I4	Sharks (2006)
L5	Unbelievable Night (2006)	I5	Searching for Food (2006)

The PIRLS 2006 assessment included 126 items across the 10 assessment blocks, comprising a total of 167 score points. The numbers of multiple-choice and constructed-response items by reading purpose are presented in Exhibit 2.6. The two question formats—constructed response and multiple choice—were evenly represented in the total number of items, with 64 multiple-choice items and 62 constructed-response items in the assessment. The total number of items and score points were distributed equally between the two purposes for reading.

Exhibit 2.6 PIRLS 2006 Assessment Item Specifications

	Number of Multiple-choice Items	Number of Constructed-response Items			Total Number of Items	Total Number of Score Points
		1 pt.	2 pts.	3 pts.		
Literary	34	13	13	4	64	85
Informational	30	15	14	3	62	82
Total	**64**	**28**	**27**	**7**	126	167

Exhibit 2.7 presents the portion of the assessment devoted to each of the four processes of reading comprehension. The distribution of actual score points across the processes approximates the distribution established in the PIRLS 2006 framework. Equal proportions among the first two and last two processes support the reporting of separate scales for two processes of comprehension: retrieval and straightforward inferencing and interpreting, integrating, and evaluating.[3]

Exhibit 2.7 Distribution of Score Points Across Reading Processes

PIRLS 2006 Processes of Reading Comprehension	Number of Score Points	Percentage of Total Score Points
Focus on and retrieve explicitly stated information	36	22
Make straightforward inferences	47	28
Interpret and integrate ideas and information	61	37
Examine and evaluate content, language, and textual elements	23	14
Total	**167**	**100**

2.11 Finalizing the PIRLS 2006 Scoring Guides for Constructed-response Items

In October 2005, the PIRLS Item Development Task Force met to review and revise the constructed-response scoring guides and sets of training materials in response to changes made to items after the field test. Most constructed-response items required minor changes. Only two constructed-response items and their scoring guides, each from a different assessment block, were modified significantly. After the meeting, a small number of classes from England and Scotland were administered a pilot test consisting of the two assessment blocks, and the student responses were used to rewrite the scoring guides and provide examples for the scoring training materials for the main data collection.

In addition, the Task Force reviewed all the examples and training materials for coherence and consistency, in light of responses from the field test and to ensure that the characteristic patterns of student response were covered by the guides.

These final versions of the scoring guides and training materials from the PIRLS 2006 field test were combined with those from the 2001 passages

3 Retrieval and straightforward inferencing will combine items from the focus on and retrieve explicitly stated information and make straightforward inferences comprehension processes. Similarly, interpreting, integrating, and evaluating will be based on items from the interpret and integrate ideas and information and examine and evaluate content, language, and textual elements processes.

that now were included in the 2006 assessment. As in the field test, training materials, which included 8–10 anchor papers and 8–10 practice papers for each of the 62 constructed-response items, were arranged by assessment block. All scoring guides and training materials then were introduced to the NRCs and their scoring managers in two intensive training sessions in November 2005 for the Southern Hemisphere and March 2006 for the Northern Hemisphere. Discussion of the student responses in the training materials allowed the participants in the training sessions to become confident with the distinctions between the various levels of scoring categories and to pass this knowledge on to their scoring teams.

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.

Mullis, I.V.S., Kennedy, A.M., Martin, M.O., & Sainsbury, M. (2006), *PIRLS 2006 assessment framework and specifications* (2nd ed.), Chestnut Hill, MA: Boston College.

TIMSS & PIRLS International Study Center. (2004). *Item-writing guidelines for the PIRLS 2006 field test*. Chestnut Hill, MA: Boston College.

Chapter 3

Developing the PIRLS 2006 Background Questionnaires

Ann M. Kennedy

3.1 Overview

A major goal of PIRLS is to examine home and school factors associated with students' reading achievement and the PIRLS framework contains a section addressing the contexts for learning and teaching reading. Because measuring trends in students' reading literacy is an important focus of PIRLS, the PIRLS 2006 contextual framework was similar to the framework used in 2001. This chapter describes the updates made to the *Framework and Specifications for PIRLS Assessment 2001* (Campbell, Kelly, Mullis, Martin, & Sainsbury, 2001) with regard to the contexts for reading, in addition to the changes made to the PIRLS 2001 background questionnaires to be aligned with reporting plans for PIRLS 2006. In particular, a curriculum questionnaire was planned for PIRLS 2006 to collect information about the reading curriculum for primary grades in each of the participating countries to be included in the *PIRLS 2006 Encyclopedia* (Kennedy, Mullis, Martin, & Trong, 2007).

3.2 PIRLS 2006 Contextual Framework

The relationships among the home, school, and national and community environments that help to shape the development of reading literacy among young children are discussed in the *PIRLS 2006 Assessment Framework and Specifications* (Mullis, Kennedy, Martin, & Sainsbury, 2006). The process of

reviewing and updating the contextual framework for PIRLS 2006 began at the first meeting of the National Research Coordinators (NRCs) in September 2003. In general, the NRCs were pleased with the existing PIRLS 2001 framework and specifications and recommended making only minor modifications based on the results from PIRLS 2001, as well as recent issues of interest related to research in reading literacy. The TIMSS & PIRLS International Study Center asked NRCs, and Questionnaire Development Group members, in particular, to submit their comments and suggestions within a few months in order to meet a scheduled publication date of August 2004 for the revised framework set.

The TIMSS & PIRLS International Study Center received from the NRCs and Questionnaire Development Group members the following suggestions for revisions that were incorporated into the PIRLS 2006 contextual framework.

- Expand the section on national and community contexts to include the emphasis on literacy in a country.
- Include more references to the home context, such as home resources and students' literacy activities outside of school.
- Separate school and classroom contexts to differentiate between the influencing factors of these environments.
- Add a section to address homework and both formal and informal assessment of performance in reading within classroom contexts.
- Update references to include current research since PIRLS 2001.

3.3 The PIRLS 2006 Background Questionnaires

In order to measure trends and collect baseline information about key factors related to students' home and school environments, PIRLS 2006 administered questionnaires to students, parents, teachers, and school principals. Additionally, PIRLS 2006 included a newly constructed curriculum questionnaire that provided information about the national context. Based on the contexts for learning to read, as defined in the PIRLS framework, the information from the five questionnaires complements the fourth-grade students' reading achievement results.

- The *Student Questionnaire* collected information about literacy-related activities and resources both at home and in school.

- The *Learning to Read Survey* (home) asked parents or primary caregivers to reflect on literacy-related activities and resources at home and their perceptions of support provided by the school environment.
- The *Teacher Questionnaire* asked about the structure and content of reading instruction in the classroom, as well as within the school as a whole. It also obtained information about the teacher's preparations for teaching reading at the fourth grade.
- The *School Questionnaire* gathered information from the school principal about the school's reading curriculum and instructional policies in addition to the school's demographics and resources.
- The *Curriculum Questionnaire*, newly created in 2006, focused on the nature of the development and implementation of a nationally (or regionally) defined reading curriculum in primary schools within each participating country.

3.3.1 Updating the PIRLS 2006 Background Questionnaires

Updating the PIRLS 2001 background questionnaires for PIRLS 2006 was a collaborative effort among the TIMSS & PIRLS International Study Center, PIRLS 2006 NRCs, the Questionnaire Development Group, and the IEA Data Processing and Research Center (DPC). The process of review and revision began in February 2004, in preparation for the 2005 PIRLS field test. Results from the field-test administration prompted further discussions and refinement of the questionnaires for the PIRLS 2006 data collection.

At the second meeting of the PIRLS 2006 NRCs in February 2004, NRCs thoroughly reviewed the contents of each questionnaire and shared comments about the usefulness of items and response categories, in light of the reporting of trend results for the 2006 survey. In order to minimize the burden on the respondents, NRCs were asked to recommend removing items before adding new ones.

The Questionnaire Development Group met in August 2004 to review drafts of the 2006 questionnaires that incorporated the changes from the previous National Research Coordinator meeting. These draft questionnaires emphasized coverage of questions across the contexts, described in the recently published PIRLS 2006 framework. Another primary objective of the Questionnaire Development Group meeting was to initiate the construction of the *Curriculum*

Questionnaire. The expert panel worked to define and outline topics that were comparable across the education systems of the participating countries and information which could be readily provided at a national level. This outline was developed with the intention of providing readily comparable facts about each country's curriculum that could be displayed in tables throughout the *PIRLS 2006 International Report*, as well as more detailed contextual information to be used in the *PIRLS 2006 Encyclopedia*.

NRCs met in November 2004 for a final review of the PIRLS 2006 student, parent, teacher, and school questionnaires before the administration of the field test during March and April 2005. In this review, the NRCs recommended minor wording or formatting changes. Since the *Curriculum Questionnaire* would not undergo a field test, NRCs continued its development, based on the Questionnaire Development Group's outline and recommendations.

Following the field-test administration, countries sent data files to the IEA DPC for data cleaning, verification, and formatting before sending the data to the TIMSS & PIRLS International Study Center. The TIMSS & PIRLS International Study Center staff then prepared data almanacs to present the results for the student, parent, teacher, and school questionnaires. For each item in the questionnaire, unweighted statistics were displayed for every country, as well as for the international average. Displays for categorical variables included columns with the percentages of respondents in each category and the corresponding average student reading achievement scores. Displays for numeric variables included the mean, mode, minimum, and maximum values, and selected percentiles. The data almanacs were used by the TIMSS & PIRLS International Study Center, the Questionnaire Development Group, and NRCs to evaluate the performance and quality of the field-test questionnaire items and make suggestions for revisions for the main PIRLS 2006 data collection.

The review of field-test questionnaire data almanacs began with a meeting of the Questionnaire Development Group in July 2005. The Questionnaire Development Group examined item statistics to determine whether the questions seemed to be functioning well across the countries and whether response options were the most advantageous. As a result, the Questionnaire Development Group proposed a few changes to each of the four questionnaires. Typical changes included removing items, rewording or replacing items, and collapsing or expanding response categories. Additionally, there were suggestions for restructuring item placement and layout for better organization and clarity within the questionnaires.

In August 2005, the NRCs convened to review the field-test data almanacs in light of the recommendations by the Questionnaire Development Group. In general, the NRCs agreed to adopt the Questionnaire Development Group's suggestions with modest modifications and rewordings of items and response options. Immediately following the meeting, the TIMSS & PIRLS International Study Center finalized the questionnaires and provided them to the NRCs so that they could begin translation and verification for the PIRLS 2006 data collection.

3.3.2 Content of the PIRLS 2006 Background Questionnaires

The content of each PIRLS 2006 background questionnaire is summarized below. Exhibits 3.2 through 3.6, which follow the summaries, provide descriptions of the variables within the questionnaires. The variables are grouped and arranged according to their related contextual factors.

Student Questionnaire

Each student in the selected class completed a *Student Questionnaire*. The questionnaire included questions about home resources, languages spoken in the home, students' reading habits both inside and outside of school, students' reading self-concept and their attitudes towards reading, classroom instructional practices related to teaching reading, and school safety.

Learning to Read Survey (Home)

The parents or guardians of each student completed a *Learning to Read Survey*. The questionnaire asked about preparations for primary schooling, including attendance in preschool and literacy-centered activities in the home before the child began school, such as reading books, singing songs, or writing letters or words. Parents answered questions about home resources in addition to information about their highest level of education and employment situations.

Teacher Questionnaire

Teachers of the assessed classes responded to the *Teacher Questionnaire*. The questionnaire focused on reading activities and materials used for reading instruction and the assessment of students' performance in reading. Teachers were asked to refer specifically to the class of students selected for the PIRLS assessment. Teachers also answered questions about their professional preparation and experience in teaching reading.

School Questionnaire

The principal of each school sampled for PIRLS completed a *School Questionnaire*. Principals answered questions about the emphasis on the reading curriculum in the school, the availability and use of materials to teach reading, and whether the school provided programs and services that involve the students and their families. Additionally, the questionnaire asked school principals general questions about their school's demographic characteristics, resources, and environment.

Curriculum Questionnaire

The National Research Coordinator within each country was responsible for completing the *Curriculum Questionnaire*. Questions primarily centered on the defined national or regional curriculum in fourth grade, including what it prescribed and how it is disseminated. NRCs also answered questions about requirements for teachers and how teachers are informed about the reading curriculum. An addendum to the questionnaire asked about country-level policies regarding entry to primary school as they related to the students tested in PIRLS 2006.

Exhibit 3.2 Content of the PIRLS Learning to Read Survey (Home Questionnaire)

Context	Variable(s)
Student Characteristics	Whether, and for how long, child attended kindergarten (or equivalent)
	Age when child began formal schooling
	Child's literacy skills when he/she began formal schooling
Activities Fostering Literacy	Frequency parents engaged in home literacy activities with child during early childhood
	Frequency parents engaged more recently in home literacy activities with child
Language(s) in the Home	Language(s) spoken by child during early childhood
	Language of early childhood home literacy activities
	Language of present day home literacy activities
	Language of children's books in home
	Language used most often when parents speak with their child
Home–School Connection	Time student spent on homework each day
	Parents' opinion of child's school
Social and Cultural Resources	Time spent by parent reading for him/herself at home each week
	Frequency that parent read for his/her own enjoyment
	Parents' attitudes toward reading
Economic Resources	Perception of wealth relative to others
	Number of books in the home
	Number of children's books in the home

Exhibit 3.3 Content of PIRLS School Questionnaire

Context	Variable(s)
School Characteristics	Number of students in school and in grade tested
	Size and type of community in which the school is located
	Percentage of students from economically affluent and disadvantaged homes
	Percentage of students whose first language is not the language of the test and percentage who receive some instruction in this language
	Proportion of students who received free or reduced-price lunch
	Literacy skills of students when they began formal schooling
School Policy and Curriculum	Days per week and year that school was open for instruction
	Total instruction time in a typical day
	Emphasis on language and literacy skills in comparison to other areas of the curriculum
	Whether school had a written statement of the school reading curriculum
	Whether school had a policy to coordinate reading instruction across teachers
	Emphasis on different literacy skills and activities at different grades in primary school
	Whether school had a policy promoting collaboration among teachers
School Environment and Resources	Whether extended instructional time was offered, and if so, how many students participate
	Whether before- or after-school child care was offered, and if so, how many students participate
	Whether provisions were made for students whose mother tongue is not the language of the test
	Number of computers available for instructional purposes
	Material factors affecting school's capacity to provide instruction
	Frequency of scheduled times for teachers to meet and share instructional ideas
	Workspace facilities provided for teachers
	Time principal spent on different tasks and functions
Literacy Resources	Whether school had informal initiatives to encourage students to read
	Whether school had programs for teachers to improve reading instruction
	Emphasis on different types of materials for reading instruction
	Whether school had a library, and the number of books and magazines within it
Community Relations	Availability of literacy and educational programs for students' families
	Frequency of communications with students' families
	Percentage of students' parents who participated in school events
School Climate	Principal's perception of different aspects of school climate
	Principal's perception of the severity of different problems within the school

Exhibit 3.4 Content of the PIRLS Teacher Questionnaire

Context	Variable(s)
Teacher Demographics	Age and gender
	Total number of years teaching and number of years teaching fourth grade
	Whether teacher worked full time or part time
	Teacher's satisfaction with his/her role as a teacher
Class Characteristics	Number of years the teacher had taught this class
	Number of students in class, and how many of those were in the grade tested
	Teacher's perception of class reading level
	Number of students with difficulty understanding spoken language of the test
	Number of students who needed remedial instruction in reading, and how many of those received it
	Whether enrichment reading instruction was available, and how many students received it
Teacher Training and Preparation	Teacher's highest level of formal education
	Type of teacher certification
	Areas of study during training and formal education
	In-service time spent on reading or teaching reading
	Time spent reading various materials for professional development
	Time spent reading for enjoyment at home
Classroom Environment and Structure	Whether other teachers taught the class for a significant portion of time
	Organization of students for reading instruction
Instructional Materials and Technology	Frequency teacher used different resources for reading instruction
	Frequency teacher used different types of text for reading instruction
	Use of reading instructional materials for students at different reading levels
	Availability of computers and the Internet, and student activities on the computer
	Availability, size, and use of classroom library
	Frequency of use of school library
	Availability of specialists for students who had difficulty with reading
	Where teacher prepares materials for instruction
Instructional Strategies and Activities	Percentage of time spent on different instructional and administrative activities
	Time spent on language instruction in a week
	Time spent on reading instruction in a week, formally and informally
	Frequency of reading instruction and activities
	Frequency of different reading activities with students
	Frequency of different activities after students have read something
	Tasks teacher asks students to complete to develop reading comprehension skills or strategies
	Strategies used if a student begins to fall behind in reading

Exhibit 3.4 Content of the PIRLS Teacher Questionnaire *(continued)*

Context	Variable(s)
Homework and Assessment	Frequency teacher assigned reading for homework and how much time was expected to be spent on it
	Emphasis placed on assessment sources to monitor students' progress in reading
	Use of different tools to assess students' progress in reading
	Use of portfolios as part of reading assessment
Home–School Connection	Frequency of communication with parents about students' reading progress

Exhibit 3.5 Content of the PIRLS Student Questionnaire

Context	Variable
Student Characteristics	Age and gender
	Whether student and parents were born in country
Literacy Activities Outside of School	Frequency student engaged in different reading activities
	Types of texts that students read outside of school
	Frequency that student borrowed books from a library and the language of these books
Other Activities Outside of School	Frequency of other activities outside of school (e.g., watching television, playing video games)
	Frequency of computer use in various places
	Frequency of Internet use for various purposes
Literacy Activities in School	Frequency of various reading activities in school
	Frequency of various activities after student has read something in class
Languages in the Home	Language(s) that student spoke before starting school
	Frequency student spoke language of the test at home
Home–School Connection	Frequency of reading assigned for homework and time spent on it each day
	Person who helps student most with reading homework
Student Attitudes	Student's attitudes toward reading
	Student's self-concept regarding his/her reading ability
	Student's attitudes toward school
	Student's reports of problematic behavior by other students at school
Economic Resources	Number of books in the home
	The presence of various socio-economic indicators in the home

Exhibit 3.6 Content of the PIRLS Curriculum Questionnaire

Context	Variable(s)
Demographics and Resources	Age students began primary school
	Number of school days per year
Emphasis on Literacy	Emphasis placed on various reading processes in reading curriculum
	Emphasis placed on various reading purposes in reading curriculum
Governance and Organization of Education System	Highest level of decision-making authority that provides a curriculum covering reading instruction
	Grade-to-grade structure of primary school curriculum
	Whether local authorities had a significant role in reading curriculum development
Curriculum Characteristics and Policies	Year reading curriculum was introduced
	Whether the reading curriculum was being revised
	Whether reading was presented as a part of language instruction or as a separate curriculum area
	Whether the reading curriculum prescribed goals, methods, and materials
	How reading curriculum addressed the issue of students with different levels of ability
	Form(s) in which the reading curriculum was made available to the public
	Total instructional time per week prescribed by curriculum, and percentage devoted to language and reading instruction
	Methods used to evaluate the implementation of the reading curriculum
	Whether there was a policy regarding promotion and retention of students in primary school grades
Teacher Training and Preparation	Requirements and certification process for becoming a primary school teacher
	Whether teachers received preparation on how to teach the reading curriculum in pre-service education
	Help provided to teachers to implement the reading curriculum
	Methods used to communicate changes in reading curriculum to teachers
Home–School Connection	Methods used to communicate changes in reading curriculum to parents

3.4 PIRLS 2006 Encyclopedia

The *PIRLS 2006 Encyclopedia* is a companion publication to the *PIRLS 2006 International Report* (Mullis, Martin, Kennedy, & Foy, 2007), providing a qualitative perspective on the national contexts for reading education. It provides an overview of the characteristics of each participating country, including information collected from the *PIRLS Curriculum Questionnaire*, as well as a detailed chapter for each participant describing reading education. Each NRC was responsible for writing a chapter for the encyclopedia, using an outline provided by the TIMSS & PIRLS International Study Center. The individual

chapters describe organization of the education system (national or regional), provide detail about the reading curriculum for the primary grades, and discuss resources for reading education.

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.

Kennedy, A.M., Mullis, I.V.S., Martin, M.O., & Trong, K.L. (Eds.). (2007). *PIRLS 2006 encyclopedia: A guide to reading education in the forty PIRLS 2006 countries.* Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Kennedy, A.M., Martin, M.O., & Sainsbury, M. (2006). *PIRLS 2006 assessment framework and specifications* (2nd ed.). 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 35 countries.* Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., & Flaherty, C.L. (Eds.). (2002). *PIRLS 2001 encyclopedia: A reference guide to reading education in the countries participating in IEA's Progress in International Reading Literacy Study (PIRLS).* Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., & Foy, P. (2007). *PIRLS 2006 international report: IEA's progress in international reading literacy study in primary schools in 40 countries.* Chestnut Hill, MA: Boston College.

Chapter 4

PIRLS 2006 Sample Design

Marc Joncas

4.1 Overview

This chapter describes the PIRLS 2006 sample design, which consists of a set of specifications for the target and survey populations, sampling frames, survey units, sample selection methods, sampling precision, and sample sizes. The sample design is intended to ensure that the PIRLS 2006 survey data provide accurate and economical estimates of national student populations. Since measuring trends is a central goal of PIRLS, the sample design also aims to provide accurate measures of changes in student achievement from 2001 to 2006. In addition to the sample design, the PIRLS 2006 sampling activities also include estimation procedures for sample statistics and procedures for measuring sampling error. These other components are described in Chapters 9 and 12, respectively. The basic PIRLS sample design has two stages: schools are sampled with probability proportional to size at the first stage, and one or two intact classes of students from the target grade are sampled at the second stage.

All participants followed the uniform sampling approach specified by the PIRLS 2006 sample design, with minimum deviations. This ensured that high quality standards were maintained for all participants, avoiding the possibility that differences between countries in survey results could be attributable to the use of different sampling methodologies. This uniform approach also facilitated an efficient approval process of the national designs by the international project team.

The PIRLS National Research Coordinator (NRC) of each participating country was responsible for implementing the sample design, including documenting every step of the sampling procedure for approval by the TIMSS & PIRLS International Study Center and Statistics Canada prior to implementation. To support NRCs in their sampling activities, a series of manuals (the *School Sampling Manual (*PIRLS, 2004*),* the *Survey Operations Procedures* (PIRLS*,* 2005b*)*, and the *School Coordinator Manual* (PIRLS, 2005a) and sampling software (IEA, 2005)) were provided. In addition to these materials, Statistics Canada consulted with each country throughout the process.

4.2 PIRLS 2006 Target Population

PIRLS is a study of student achievement in reading comprehension in primary school, and is targeted at the grade level in which students are at the transition from learning to read to reading to learn, which is the fourth grade in most countries. The formal definition of the PIRLS target population makes use of UNESCO's International Standard Classification of Education (ISCED) in identifying the appropriate target grade:

> …all students enrolled in the grade that represents four years of schooling, counting from the first year of ISCED Level 1, providing the mean age at the time of testing is at least 9.5 years. For most countries, the target grade should be the fourth grade, or its national equivalent.

ISCED Level 1 corresponds to primary education or the first stage of basic education, and should mark the beginning of "systematic apprenticeship of reading, writing, and mathematics" (UNESCO, 1999). By the fourth year of Level 1, students have had 4 years of formal instruction in reading, and are in the process of becoming independent readers.

In IEA studies, the above definition corresponds to what is known as the *international desired target population*. Each participating country was expected to define its *national desired population* to correspond as closely as possible to this definition (i.e., its fourth grade of primary school). In order to measure trends, it was critical that countries that participated in PIRLS 2001, the previous cycle of PIRLS, choose the same target grade for PIRLS 2006 that was used in

PIRLS 2001. Information about the target grade in each country is provided in Chapter 9.

Although countries were expected to include all students in the target grade in their definition of the population, sometimes it was not possible to include all students who fell under the definition of the international desired target population. Consequently, occasionally a country's *national desired target population* excluded some section of the population, based on geographic or linguistic constraints. For example, Lithuania's national desired target population included only students in Lithuanian-speaking schools, representing approximately 93 percent of the international desired population of students in the country.

Working from the national desired population, each country had to operationalize the definition of its population for sampling purposes and define their *national defined population.* While this national defined target population should ideally coincide with the national desired target population, in reality, there may be some regions or school types that cannot be included. All students in the desired population who were 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 of students. Exclusions (which had to be kept to a minimum) could occur at the school level, within the sampled schools, or both. Although countries were expected to do everything possible to maximize coverage of the national desired population, *school-level exclusions* sometimes were necessary. Keeping within the 95 percent limit, school-level exclusions could include schools that:

- were geographically remote,
- had very few students,
- had a curriculum or structure different from the mainstream education system, or
- were specifically for students with special needs.

The difference between these school-level exclusions and those at the previous level is that these schools were included as part of the sampling frame (i.e., the list of schools to be sampled). They then were eliminated on an individual basis if it was not feasible to include them in the testing.

In many education systems, students with special educational needs are included in ordinary classes. Due to this fact, another level of exclusions is necessary to reach an effective target population—the population of students who ultimately will be tested. These are called *within-school exclusions* and pertain to students who are unable to be tested for a particular reason but are part of a regular classroom. There are three types of within-school exclusions, which are explained below.

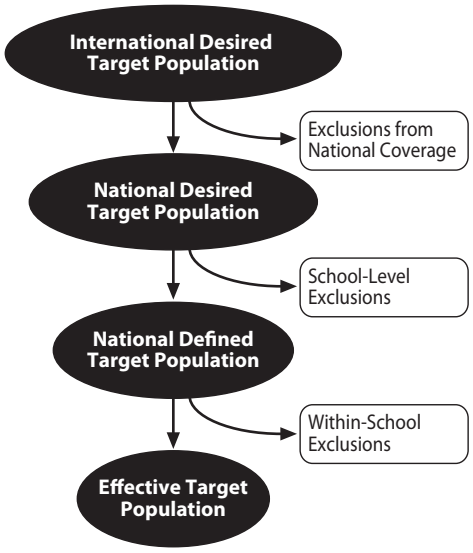
- **Intellectually disabled students:** These are students who are considered in the professional opinion of the school principal, or by other qualified staff members, to be intellectually disabled or who have been tested psychologically as such. This includes students who are emotionally or mentally unable to follow even general test instructions. Students should not be excluded solely because of poor academic performance or normal disciplinary problems.
- **Functionally disabled students:** These are students who are permanently, physically disabled in such a way that they cannot perform in the PIRLS testing situation. Functionally disabled students who are able to respond should be included in the testing.
- **Non-native language speakers:** These are students who are unable to read or speak the language(s) of the test and would be unable to overcome the language barrier of the test. Typically, a student who has received less than 1 year of instruction in the language(s) of the test should be excluded, but this definition may need to be adapted in different countries.

Students eligible for within-school exclusion were identified by staff at the schools and could still be administered the test if the school did not want the student to feel out of place during the assessment (though the data from these students were not included in any analyses). Again, it was important to ensure that this population was as close to the national desired target population as possible.

If combined, school-level and within-school exclusions exceeded 5 percent of the national desired target population, results were annotated in the *PIRLS 2006 International Report* (Mullis, Martin, Kennedy, & Foy, 2007). Target population coverage and exclusion rates are displayed for each country in Chapter 9. Descriptions of the countries' school-level and within-school exclusions can be found in Appendix B.

In any study that utilizes sampling, the population that ultimately participates usually differs slightly from the target population, with some portion of the target population being excluded from the study. A major objective of the PIRLS sampling strategy was to ensure that the effective target population, the population actually sampled by PIRLS, was as close as possible to the international desired population, and to document clearly all excluded populations. Exhibit 4.1 illustrates the relationship between successively more refined definitions of the target population and the excluded populations at each stage.

Exhibit 4.1 Relationship Between the Desired Populations and Exclusions



4.3 Sample Design

Once the survey population was defined, the next step involved building a sampling frame in which all sampling units within the national defined target population have a known probability of being sampled. In PIRLS 2006, however, it is important to note that in addition to gathering data on sampled students, a large amount of information also was gathered about their classes and schools, which required other types of sampling units. The intrinsic, hierarchical nature of these nested units necessitated the creation of a sampling frame by stages.

Therefore, a two-stage stratified cluster sample design was used, with schools as the first stage and intact classes as the second.[1]

4.3.1 Sampling Precision and Sample Size

Because PIRLS is fundamentally a study of reading comprehension among fourth-grade students, the precision of survey estimates of student characteristics was of primary importance. However, PIRLS reports extensively on school, teacher, and classroom characteristics also, so it is necessary to have sufficiently large samples of schools and classes. The PIRLS standard for sampling precision requires that all student samples should have an effective sample size of at least 400 students for the main criterion variable, which is reading achievement. 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.

Given that sampling error, when using simple random sampling, can be expressed as $SE_{SRS} = S/\sqrt{n}$ where S gives the population standard deviation and n the sample size, a simple random sample of 400 students would yield a 95 percent confidence interval for an estimate of a student-level mean of plus and minus 10 percent of its standard deviation (1.96 times $1/\sqrt{400}$ times S). Because the PIRLS achievement scale has a standard deviation of 100 points, this translates into a ±10 point confidence interval (or a standard error estimate of approximately 5 points). Similarly, sample estimates of student-level percentages would have confidence intervals of approximately ±5 percentage points.

Notwithstanding these precision requirements, PIRLS required that all student sample sizes should not be less than 4,000 students. This was necessary to ensure adequate sample sizes for analyses where the student population was broken down into many subgroups. Furthermore, since PIRLS planned to conduct analyses at the school and classroom level in addition to the student level, all school sample sizes were required to be not less than 150 schools, unless a complete census fails to reach this minimum. Under simple random sampling assumptions, a sample of 150 schools yields a 95 percent confidence interval for an estimate of a school-level mean that is plus and minus 16 percent of a standard deviation.

Although the PIRLS sampling precision requirements are such that they would be satisfied by a simple random sample of 400 students, student samples chosen using multi-stage cluster designs, such as the PIRLS 2006 school-and-class design, typically require much larger student samples to achieve the same

1 Because their large population size, it was necessary to include a preliminary sampling stage in the United States and the Russian Federation, where regions were sampled first, and then schools.

level of precision. Because students in the same school, and even more so in the same class, tend to be more like each other than like other students in the population, sampling a single class of 30 students will yield less information per student than a random sample of students drawn from across all students in the population. PIRLS uses the intra-class correlation, a statistic indicating how much students in a group are similar on an outcome measure, and a related measure known as the design effect, to adjust for this "clustering" effect in planning sample sizes.

For countries taking part in PIRLS for the first time in 2006, we used the following mathematical formulas to estimate how many schools should be sampled to achieve an acceptable level of sampling precision.

$$Var_{PPS} = Deff * Var_{SRS} = \frac{Deff * S^2}{n} = \frac{\left[1+\rho(mcs-1)\right] * S^2}{n} = \frac{\left[1+\rho(mcs-1)\right] * S^2}{a * mcs}$$

Where *Deff* is a compensation factor for using a sample selection method that differs from a simple random sample (also called design effect). S^2 gives the variance of the population, ρ measures the intra-class correlation between clusters, *mcs* corresponds to the average number of sampled students per class (assuming one class per school), and *a* gives the number of schools to sample. Incorporating the precision requirements into this equation gives the number of schools required as:

(1)
$$a = 400 * \frac{\left[1+\rho(mcs-1)\right]}{mcs}$$

For planning purposes, the intra-class correlation coefficient was usually set to 0.3 if no other information was available. For example, with a MCS of 20 students and a ρ of 0.3, equation (1) gives 134 schools.

Equation (1) is a model for determining how many schools would be required for the PIRLS 2006 sample under the assumption that the standard error of the criterion variable (student reading achievement) reflects only sampling variance—the usual situation in sample surveys. However, because of its complex matrix-sampling assessment design, standard errors in PIRLS include an imputation error component in addition to the usual sampling error component (see Chapter 11). To keep the standard error within the prescribed

precision limits, the number of schools determined by equation (1) have to be increased, as shown in equation (2):

(2)

$$a_{imp} = (400 * 0.5) / mcs$$

Continuing the example for a country with a MCS of 20 students, according to equation (2), 10 schools would have to be added to the 134 schools from equation (1), for a total of 144 schools.

For PIRLS 2006 countries that also had participated in PIRLS 2001, the standard error estimates computed from the 2001 data were reviewed to ensure that the student samples had been large enough to meet the precision requirements in 2001 and would be sufficiently precise to measure trends to 2006. For the several countries falling somewhat short of the sampling requirements not met in 2001, the school sample size for 2006 was increased, using as a rule of thumb that sampling error is inversely proportional to the square root of the sample size. For example, if the sample size in 2001 yielded a standard error of 7 points for an estimate of a mean, the sample size in 2006 was increased by a factor of 2 to provide a standard error of 5 points $((7/5)^2=2)$. Intra-class correlation coefficients also were calculated for countries that participated in PIRLS 2001. These coefficients were presented in the *PIRLS 2006 School Sampling Manual* (PIRLS, 2004).

4.3.2 Stratification

Stratification is the grouping of sampling units into smaller sampling frames according to information found on the initial sampling frame prior to sampling, and may be employed to improve the efficiency of the sample design, to sample sections of the population at different rates, or to ensure adequate representation of specific groups in the sample. The stratification by itself can take two forms: explicit or implicit.

Explicit stratification physically creates smaller sampling frames from which samples of schools and classes will ultimately be drawn. In PIRLS, this type of stratification is used when the usual proportional allocation (i.e., students in certain regions or types of school are represented in the sample in proportion to their distribution in the population) may not result in adequate representation of some groups in the sample. For example, if a country wanted to make generalizations regarding the reading achievement of private sector

students, the sampling frame could be split into two strata—public and private sector schools. The sample could then be allocated between the two strata to achieve the desired level of precision in each. In most countries in PIRLS 2006, the sample allocation among strata was proportional to the number of students found in each stratum. However, it could be noted in passing that, even without any stratification, the PIRLS samples represented the different groups found in the population, on average.

Implicit stratification only requires that the sampling frame is sorted according to some variable(s) prior to sampling and can be nested within explicit stratification. By combining the sorting of the frame with a systematic sampling of the units, we get a sample where units are in the same proportions as those found at the population level. When schools from the same implicit stratum tend to have similar behavior, in terms of reading achievement, implicit stratification will produce more reliable estimates.

In the basic PIRLS 2006 sample design, all schools in the sampling frame for a country were sorted according to some measure of their size (MOS—see next section). If implicit stratification was used, then the sorting by MOS was done within each stratum using a serpentine approach—high to low for the first stratum, followed by low to high for the next, and so on (see example in Exhibit 4.2).

Exhibit 4.2 MOS Sort Order Across Implicit Strata

Implicit Stratum	Sort Order of MOS
1. Rural – Public	High to Low
2. Rural – Private	Low to High
3. Urban – Public	High to Low
4. Urban – Private	Low to High

This way of sorting sampling units optimizes the chances of choosing replacement schools with a MOS close to the original sampled schools they are meant to replace.

4.3.3 Replacement Schools

Ideally, response rates to study samples should always be 100 percent, and although the PIRLS 2006 participants worked hard to achieve this goal, it was anticipated that a 100 percent participation rate would not be possible in all

countries. To avoid sample size losses, the PIRLS sampling plan identified, *a priori*, replacement schools for each sampled school. Therefore, if an originally selected school refused to participate in the study, it was possible to replace it with a school that already was identified prior to school sampling. Each originally selected school had up to two pre-assigned replacement schools. In general, the school immediately following the originally selected school on the ordered sampling frame and the one immediately preceding it were designated as replacement schools. Replacement schools always belonged to the same explicit stratum, although they could come from different implicit strata if the originally selected school was either the first or last school of an implicit stratum.

The main objective for having replacement schools in PIRLS 2006 was to ensure adequate sample sizes for analysis of sub-population differences. Although the use of replacement schools did not eliminate the risk of bias due to nonresponse, employing implicit stratification and ordering the school sampling frame by size increased the chances that any sampled school's replacements would have similar characteristics. This approach maintains the desired sample size while restricting replacement schools to strata where nonresponse occurred. Since the school frame is ordered by school size, replacement schools also tended to be of the same size as the school they were meant to replace. For the field test, replacement schools were used to make sure sample sizes were large enough to validate new items, and no more than one replacement school was assigned per originally selected school.

4.4 Sample Selection

The school sampling selection method used in PIRLS 2006 is a classic approach that can be found in most sampling textbooks (e.g., Cochran, 1997). The method is usually referred to as a systematic *probability proportional-to-size* (PPS) technique. This sampling method is a natural match with the hierarchical nature of the sampling units, with classes of students nested within schools. Even if a country had a list from which students could be selected directly, the sampling frame for most of the countries participating in PIRLS was first made of schools. From these sampled schools, lists of classes were created and sampled. For each sampled class, a list of students was created.

4.4.1 Sampling Schools

In order to draw school samples representative of the student population, NRCs were asked to provide vital information about the schools within the sampling frame. The following data were required for each school:

- A *measure of size* (MOS) (e.g., the average student enrollment in the fourth grade, the number of classrooms in the fourth grade, or the total student enrollment in the school);
- The expected number of sampled students per class, also called *minimum cluster size* (MCS). This was required if the number of classrooms in the fourth grade couldn't be provided and was calculated as the ratio of the total number of students to the total number of classes for schools having more than one class in the fourth grade; and
- Any variables describing school characteristics to be used for stratification purposes, such as type of school, degree of urbanization, or sex of students served by the school.

Schools were sampled using systematic random sampling with probability proportional to their MOS. For example, if school A had a MOS value twice as large as school B, then School A had twice the chance of being in the sample compared to school B. Similarly, if region A had a MOS value twice as large as region B, then region A had twice the chance of being in the sample.

To implement the school sampling, schools in each explicit stratum were sorted in order by the implicit stratification variables and within these by the MOS. The measures of size are accumulated from school to school, and a running total, the cumulative measure of size, is recorded next to each school. The cumulative MOS is an indicator of the size of the population of sampling elements (students). Dividing the cumulative MOS by the number of schools to sample gives the sampling interval.

With systematic PPS sampling, it is possible for a large sampling unit to be selected more than once if its size is greater than the sampling interval. To avoid this situation, all such units were automatically selected by changing their MOS to the sampling interval of the associated explicit stratum.

Some schools have so few students that their selection using probability proportional to their size (MOS) becomes problematic. Since the selection of these schools depends on their size, a difference between the number of expected students when drawing the sample and the number of students actually

found in the field can substantially contribute to the sampling error. To lessen the impact of this eventuality, any schools with fewer expected students than the average minimum cluster size (MCS) for the explicit stratum were sampled with equal probabilities. For example, if the MCS was 30 students and there were 28 schools with less than 30 students for a total of 476 students, the MOS of these small schools was changed to 476/28 = 17. By doing this, the overall size of the explicit stratum stayed the same but all small schools had an equal chance of being selected.

The MCS also was used to define very small schools. Whenever a school had an expected number of students less than one quarter of the average MCS, the school was labeled as a very small school. These schools could be excluded, as long as they did not exceed 2 percent of the national desired target population and the overall exclusion rate did not exceed 5 percent.

4.4.1 Sampling Classes

For all participants to PIRLS 2006 but two (Morocco and Singapore),[2] intact student classes were the second and final sampling stage, with no student subsampling. This means that all students within sampled classes participated in PIRLS 2006, with the exception of excluded students and students absent the day of the assessment. Classes were selected with equal probability of selection using systematic random sampling. Within each sampled school, all fourth-grade classes were listed, and one or two classes were sampled, using a random start (different in each sampled school). This method, combined with the PPS sampling method for schools, results in a self-weighted student sample under the following conditions: a) there is a perfect correlation between the school MOS reported in the sampling frame and the actual school size; b) the same number of classes is selected in each school and c) the MCS is the same for all schools. Given that these conditions were never totally met, student sampling weights varied somewhat from school to school (see Chapter 9 for details about sampling weights).

Within sampled schools, some classes have so few students that it was unreasonable to go through the sampling process and end up with these small classes. Furthermore, small classes tend to increase the risk of unreliable survey estimates. To avoid these problems, a class smaller than half the specified MCS was combined with another class from the same school prior to class sampling.

2 Two classes per school were selected using systematic PPS sampling in Singapore, and then 19 students were sampled within each class. One class per school was selected using PPS sampling in Morocco, with 25 students (all student if less than 25 students in the class) were sampled within each class.

4.5 Selecting Field-test Samples

Prior to the main data collection, which was conducted October–November 2005 in Southern Hemisphere countries and April–May 2006 in Northern Hemisphere countries, PIRLS 2006 conducted a full-scale field test in April 2005 in all participating countries. The field-test sample size was approximately 30 schools in each country. Countries were required to draw their field-test samples using the same random sampling procedures that they employed for the main sample. This ensured that field-test samples approximated closely the main samples, while reducing the burden on schools, the field-test and main data collection samples were drawn simultaneously, so that a school could be selected for either the field test or the main data collection, but not both. For example, if 150 schools were needed for the main data collection and another 30 schools needed for the field test, a larger sample of 180 schools was selected using the sampling method described earlier. A systematic subsample of 30 schools then was selected from the 180 schools and assigned to the field test, leaving 150 schools for data collection.[3]

References

Cochran, W. G. (1997). *Sampling techniques*. New York: John Wiley.

IEA. (2005). *WinW3S: Within-school sampling software manual*. Hamburg: IEA Data Processing and Research Center.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., & Foy, P. (2007). *PIRLS 2006 international report: IEA's progress in international reading literacy study in primary schools in 40 countries.* Chestnut Hill, MA: Boston College, PIRLS.

TIMSS & PIRLS International Study Center. (2004). *PIRLS 2006 school sampling manual*. Chestnut Hill, MA: Boston College.

TIMSS & PIRLS International Study Center. (2005a). *PIRLS 2006 school coordinator manual*. Chestnut Hill, MA: Boston College.

TIMSS & PIRLS International Study Center. (2005b). *PIRLS 2006 survey operations procedures unit 1: Contacting schools and sampling classes*. Chestnut Hill, MA: Boston College.

UNESCO Institute for Statistics. (1999). *Operational manual for ISCED-1997 : International standard classification of education*.

3 In countries where it was necessary to conduct a complete census of all schools, or where the NRC believed that the sampling frame used to draw the combined sample was not appropriate for the data collection, separate sampling frames were provided for the field test and main data collection. In such situations, no attempt was made to minimize the overlap. This issue is discussed in more detail in Appendix B.

Chapter 5

Translation and Translation Verification of the PIRLS Reading Assessment and Questionnaires

Barbara Malak and Kathleen L. Trong

5.1 Overview

The PIRLS 2006 reading assessment, background questionnaires, and procedural manuals were developed in English, the working language of the International Association for the Evaluation of Education Achievement (IEA). Using this English international version, participants translated the materials into their target language(s) and adapted them to be appropriate for their cultural context. Throughout this translation and adaptation process, the overarching purpose was to create a set of instruments that was comparable to the originals in terms of reading difficulty and accessibility, while still allowing each country to adapt the materials to their national needs. Guidelines for translating the materials were described in the *Survey Operations Procedures Unit 2: Preparing Materials for the PIRLS Assessment* (TIMSS & PIRLS International Study Center, 2005), developed by the TIMSS & PIRLS International Study Center.

Since high-quality translations were essential to the success of PIRLS 2006, these translated texts were subjected to a stringent international translation verification process. This process was intended to make certain that the translated materials were equivalent to the international version through direct comparisons of the two. Each participating country was asked to submit materials for verification prior to both the field test and main data collection.

5.2 Identifying the Target Language

For most participating countries, identifying the language that would be used for testing, or the target language, was simple since they have one dominant language that is used in public and private arenas. However, schools in some countries provided instruction in more than one language.[1] Thus, such a country would have prepared test instruments in more than one language. In other cases, while there may be one language of instruction, there are other languages that are prominent in other parts of society. For example, most students in Singapore are taught in English, but the language used at home is primarily Chinese and also may be Tamil or Malay. Therefore, while Singapore prepared the reading assessment in English, the *Learning to Read Survey* was provided in all four languages so that parents could respond in whichever language they were most comfortable.

In total, the PIRLS 2006 data collection materials were translated into 44 languages, with English used the most often (8 participants), followed by French and Arabic (4 participants in each). Of the 45 participants, 15 administered the reading assessment in more than one language (ranging from 2 up to 11). Exhibit 4.1 shows the languages used by each participant for the various instruments.

5.3 PIRLS Instruments to Be Translated

For PIRLS 2006, the following materials required translation:

- Reading assessment passages, items, and directions;
- Questionnaires for students, teachers, school, and home;
- Manuals for preparing for the assessment within schools, and administering the assessment; and
- Scoring guides for constructed-response items, where necessary.

Of these, the components of the reading assessment and the questionnaires were verified. Participants who tested in English also were required to go through the verification steps. Although they had not translated the instruments, the materials were reviewed for national adaptations and comparable layout.

5.3.1 Reading Assessment

The PIRLS 2006 reading assessment is comprised of 13 booklets, one of which is distributed to each student. A booklet contains two "blocks," each of which

1 Further discussion of languages of instruction in the PIRLS 2006 countries is available in the *PIRLS 2006 Encyclopedia: A Guide to Reading Education in the Forty PIRLS 2006 Countries.* (Kennedy, Mullis, Martin, & Trong, 2007).

contains a story or article followed by a series of questions pertaining to the text passage. There are 10 blocks in total (5 for each reading purpose),[2] which are systematically rotated throughout the booklets. Most blocks appear in three different booklets, with the exception of the PIRLS 2006 Reader. This is a full-color, magazine-style booklet that contains two passages that only appear in the Reader. The questions that are associated with these passages are located in an accompanying booklet, called Booklet R.

Each test block was translated once and then used to create the various booklets. The same was true for the directions that were included in the beginning of each booklet. These also were translated a single time and then distributed throughout the booklets. The National Research Coordinator (NRC) for each country was provided with the electronic files necessary to facilitate the translation of the blocks and the subsequent creation of the booklets.

5.3.2 Questionnaires

In addition to the PIRLS reading assessment, four questionnaires were translated and administered to gather information about the home and school contexts for learning to read.[3] Separate questionnaires were developed for the participating students, their parents/caregivers, their teachers, and principals of their schools. As with the reading assessment, NRCs were provided with the electronic files necessary to create a translated version of each of these questionnaires.

5.4 Translators and Reviewers

All study participants were strongly encouraged to hire an experienced translator who would be well suited to the task of working with the PIRLS materials.

Qualifications for translators included:

- An excellent knowledge of English;
- The target language as a native language;
- Some experience translating literary texts;
- Experience in the country cultural context; and if possible,
- Experience with students in the target population, and
- Familiarity with test development.

2 PIRLS assesses students' reading literacy for two purposes—reading for literary experience and reading to acquire and use information. For more information about the PIRLS test, please refer to Chapter 2.

3 For more information on the PIRLS questionnaires, please refer to Chapter 3.

Exhibit 5.1 PIRLS 2006 Main Survey Languages

Country	Language	Instruments				
		Student Test	Student Questionnaire	Parent Questionnaire	Teacher Questionnaire	School Questionnaire
Austria	German	•	•	•	•	•
Belgium (Flemish)	Flemish	•	•	•	•	•
Belgium (French)	French	•	•	•	•	•
Bulgaria	Bulgarian	•	•	•	•	•
Canada, Alberta[1]	English	•	•	•	•	•
	French	•	•	•	•	•
Canada, British Columbia	English	•	•	•	•	•
	French	•	•	•	•	•
Canada, Nova Scotia	English	•	•	•	•	•
	French	•	•	•	•	•
Canada, Ontario	English	•	•	•	•	•
	French	•	•	•	•	•
Canada, Québec	English	•	•	•	•	•
	French	•	•	•	•	•
Chinese Taipei	Chinese Mandarin	•	•	•	•	•
Denmark	Danish	•	•	•	•	•
England	English	•	•	•	•	•
France	French	•	•	•	•	•
Georgia	Georgian	•	•	•	•	•
Germany	German	•	•	•	•	•
Hong Kong, SAR	Modern Standard Chinese	•	•	•	•	•
Hungary	Hungarian	•	•	•	•	•
Iceland	Icelandic	•	•	•	•	•
Indonesia	Indonesian	•	•	•	•	•
Iran	Farsi	•	•	•	•	•
Israel	Hebrew	•	•	•	•	•
	Arabic	•	•	•		
Italy	Italian	•	•	•	•	•
Kuwait	Arabic	•	•	•	•	•
Latvia	Latvian	•	•	•	•	•
	Russian	•	•	•		
Lithuania	Lithuanian	•	•	•	•	•
Luxembourg	German	•	•	•	•	•
	French			•		
	Portuguese			•		
Macedonia	Macedonian	•	•	•	•	•
	Albanian	•	•	•	•	
Moldova	Romanian	•	•	•	•	•
	Russian	•	•	•	•	•
Morocco	Arabic	•	•	•	•	•
Netherlands	Dutch	•	•	•	•	•
New Zealand	English	•	•	•	•	•
	Maori	•	•	•	•	•
Norway	Bokmål	•	•	•	•	•
	Nynorsk	•	•	•	•	•
Qatar	Arabic	•	•	•	•	•
Poland	Polish	•	•	•	•	•

Exhibit 5.1 PIRLS 2006 Main Survey Languages *(continued)*

Country	Language	Instruments				
		Student Test	Student Questionnaire	Parent Questionnaire	Teacher Questionnaire	School Questionnaire
Romania	Romanian	•	•	•	•	•
	Hungarian	•				
Russian Federation	Russian	•	•	•	•	•
Scotland	English	•	•	•	•	•
Singapore	English	•	•	•	•	•
	Chinese			•		
	Malay (Bahasa Melayu)			•		
	Tamil			•		
Slovak Republic	Slovak	•	•	•	•	•
	Hungarian	•	•	•		
Slovenia	Slovenian	•	•	•	•	•
South Africa[2]	Afrikaans	•	•	•	•	•
	English	•	•	•	•	•
	isiZulu	•	•	•		
	isiXhosa	•	•	•		
	Sepedi	•	•	•		
	Sesotho	•	•	•		
	Setswana	•	•	•		
	isiNdebele*	•	•	•		
	Siswati*	•	•	•		
	Tshivenda*	•	•	•		
	Xitsonga*	•	•	•		
Spain	Spanish (Castilian)	•	•	•	•	•
	Catalonian	•	•	•	•	•
	Galician	•	•	•	•	•
	Basque	•	•	•	•	•
	Valencian	•	•	•	•	•
Sweden	Swedish	•	•	•	•	•
Trinidad & Tobago	English	•	•	•	•	•
United States	English	•	•		•	•

1 Please note that the participating Canadian provinces administered the assessment separately, but all used the same set of translated materials.

2 Please note that the South African languages with asterisks (*) were not internationally verified.

A reviewer was responsible for reviewing the translation, paying particular attention to the readability of the texts for the target population. Participants were asked to hire a reviewer with the following qualifications:

- An excellent knowledge of English;
- The target language as a native language;
- Experience in the country and cultural context; and
- Experience with students (in target grade if possible).

5.5 Translation and Adaptation Guidelines

To ensure that appropriate translations and adaptations were made when the PIRLS instruments were produced, the TIMSS & PIRLS International Study Center provided basic guidelines for these processes in the *Survey Operations Procedures Unit 2*, which was distributed to all NRCs. These guidelines are summarized in the list below.

- Translated passages should have the same register (language level and degree of formality) as the source text.
- Translated passages should have correct grammar and usage: subject/verb agreement, prepositions, verb tenses, etc.
- Translated passages should neither clarify nor omit text from the source text, nor should information be added that is not given in the source text.
- Translated passages 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 the country/cultural context.

For countries administering the PIRLS instruments in English, these guidelines are applicable to any changes made to the text to adapt the American English of the international version to the variant of English that is appropriate for their context.

5.5.1 Adaptations in Passages and Items

The equivalence of materials across countries is a key aspect of the PIRLS assessment. However, it also is important to consider the cultural spectrum of the participating systems, and allow for adaptations that are appropriate for their situations. NRCs were encouraged to keep these to a minimum and to only make changes that were vital to students' understanding of the text. These alterations included changes in vocabulary, expressions, and names of people and places.

Words or phrases within a text could be altered if a participant believed that the term's unfamiliarity would inhibit students' abilities to read the passage. When making these changes, it was important to make sure that the meaning

and difficulty level of the text remained unchanged. For instance, an "apartment" in American English would be changed to a "flat" in British English. Such changes also were necessary in order to follow national conventions, such as measurement units or date formats in the various countries. For example, "feet" could be changed to "meters" or quotation marks replaced with dashes. For the unit conversions, the TIMSS & PIRLS International Study Center provided participants with a list of all instances of measurement units used in the reading assessment and their appropriate conversions (in most cases, to the whole numbers). This was done to standardize those adaptations made across participating countries.

The PIRLS reading assessment is comprised of a collection of authentic passages that have been contributed by participating systems. If the passage contained a nonfictional name or place that was central to the meaning of the passage, this could not be changed. Otherwise, participants were permitted to adapt the names to those that would be more familiar to their students.

5.5.2 Adaptations in Questionnaires

The questionnaires involved a required set of adaptations for each participant. Some of the items in the questionnaires contained words or phrases that needed to be translated according to country-specific contexts and usage. Thus, the international version of the questionnaires contained words and phrases placed in carets (< >), indicating that the text within the carets should be adapted. For example, <tutor> in the international version of the questionnaires was replaced by <support teacher> in the Norwegian version. Items that asked parents and teachers about levels of education completed utilized the ISCED-1997[4] system. The international versions of the questionnaires provided the generic ISCED levels in carets, to be replaced with the educational terms appropriate for each country. For example, <ISCED 3> was replaced with the term "high school" in the United States version of the questionnaires. NRCs were provided with the *Operational Manual for ISCED-1997* (UNESCO, 1999) to assist them in determining the equivalent educational level for each item.

The TIMSS & PIRLS International Study Center provided participants with a detailed description of the intention of each required adaptation in order to clarify the terms used and help translators choose the appropriate corresponding term. In regards to the ISCED levels, the TIMSS & PIRLS International Study Center also provided participants with a cross-referencing list to ensure that the same educational level adaptations were made across different items and questionnaires.

4 ISCED (International Standard Classification of Education) was developed by UNESCO for cross-national comparisons. *The Operational Manual for ISCED-1997* describes the levels of education in that system.

In addition to these required adaptations, participants were allowed to add items to the questionnaires if there were pertinent issues related to reading in their country not addressed by the international items. Participants were encouraged to add items only to the end of the questionnaire to avoid influencing the responses to the international items in any way. The country-specific items were required to appear in the same form as the rest of the questionnaire and required approval from the TIMSS & PIRLS International Study Center.

5.6 Documenting National Adaptations

All deviations from the international version of the reading assessment or questionnaires were documented on the National Adaptation Forms. For each instrument, a form was completed that listed any changes made and, in the rare cases of not administered questions, the rationale behind these decisions. These forms were updated after each stage of the verification process.

5.7 Translating the PIRLS Materials

Each translator and reviewer was given the international version of the set of PIRLS 2006 materials that were being translated. Each also was given information to familiarize them with PIRLS and the translation procedures and the National Adaptation Forms that would be used to document all adaptations.

The translator used these materials to translate each of the instruments, following the adaptation guidelines that were described earlier in this chapter. If more than one translator was employed for a target language, then whichever translator worked with a passage also translated the corresponding questions. During translation, translators were instructed to document any changes made to the original text in an electronic version of the National Adaptation Forms.

This translated set of materials then was given to the reviewer, whose purpose was to make sure that the translations were at an appropriate level for the target population. The reviewer's suggestions were then incorporated into the materials by the translator, and the forms were updated accordingly.

Countries that also participated in PIRLS 2001 used an unchanged version of the blocks that carried over from the previous cycle in order to accurately measure trend. In some cases, improvements were made to the translations from 2001. In these cases, changes were carefully documented and were referenced during data analysis. If the text changes seemed to have dramatically altered

the performance of any item, then this item was not included in trend analyses for this participant.

5.8 Verification of Translation and Layout of PIRLS Instruments

Once the instruments had been translated and reviewed, the text of the directions, assessment blocks, and questionnaires were submitted for international translation and layout verification. This process was managed by the IEA Secretariat in Amsterdam, who enlisted the assistance of two independent translation companies to verify translations for each of the countries: Bowne Global Solutions (Luton, England) and Capstan Linguistic Quality Control (Brussels, Belgium). Of the 45 participants in PIRLS 2006, all except 2 submitted materials for verification for the field test. All participants submitted instruments for verification before the main data collection.

5.8.1 International Translation Verifiers

The international translation verifiers for PIRLS 2006 were required to have the target language as their first language, have formal credentials as translators working in English, be educated at the university level, and, if possible, live and work in the country where the verification was being carried out (or in close contact with this country). Verifiers received general information about the study and the design of the instruments 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 international version.

5.8.2 International Translation Verification

The main task of the translation verifiers was to evaluate the accuracy of the translation and adequacy of the national adaptations (reported in the National Adaptation Forms). Their instructions emphasized the importance of maintaining the meaning and difficulty level in test passages and related questions, as well as questions included in each of the four questionnaires. Verifiers also were asked to pay attention to correspondence between the reading passages and test questions. Specifically, verifiers had to ensure the following:

- The translation has not affected the meaning or difficulty of the text.
- The test questions have not been made easier or more difficult when translated/adapted.

- No information has been omitted or added in the translated text.
- No question related to the passage was omitted.
- The questionnaires contain all and correct questions.
- The order of questions and response options to questions are the same as in the international version.

The verifiers documented any errors or suggested changes using the "Track Changes" function in Microsoft® Word. Additionally, for the 28 participating countries who also were a part of PIRLS 2001, a comparison was made between those blocks that were being used to measure trends in 2001 to the blocks used in PIRLS 2006. To help NRCs understand the comparability of the translated text with the international version, verifiers were asked to assign a "severity code" to any deviations. The severity codes ranged from 1 (major change or error) to 4 (acceptable change) as follows:

- **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 indicated by the question, an incorrect translation which changes the meaning or difficulty of the passage or question, and the questions being in the incorrect order.
- **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, and discrepancies in the headers and footers of the document.
- **Suggestion for Alternative**: The translation may be adequate, but the verifier suggests a different wording.
- **Acceptable Change:** The change was acceptable and appropriate but was not documented. An example would be the Southern Hemisphere changing a reference to winter from January to July.

The instruments were returned to the NRC of each participating system with the verifier's suggestions. The NRC was responsible for reviewing translation verifier's suggestions and revising the instruments.

5.8.3 International Layout Verification

Verified texts were then used to generate the booklets and questionnaires in their final form, utilizing the appropriate layout and graphics. Completed instruments were then submitted, along with updated National Adaptation Forms, to the TIMSS & PIRLS International Study Center for international layout verification.

During international layout verification, each booklet was reviewed in its print-ready form. The TIMSS & PIRLS International Study Center compared each of the translated booklets to the international version, documenting any discrepancies between the two. During this verification, it was recognized that the materials may not be exactly identical, due to the changes in text length that often occurred during translation. However, the international versions were created with this in mind, and extra space was provided in the margins of the pages to facilitate the use of a longer text without extensive changes to the layout of the instrument. For countries that also participated in PIRLS 2001, the booklets from the previous cycle were compared to the newly submitted instruments to make sure that they were identical. The verifier's comments and suggested changes were returned to the NRC, along with permission to print and administer the materials once they had been revised.

5.8.4 Quality Control Monitor Review

Quality Control Monitors (QCMs) from each participating country were hired by the IEA to document the quality of the PIRLS 2006 assessment administration, including that of the assessment materials.[5] An important part of the QCMs' responsibilities included review of the booklets and questionnaires used during test administration. The QCMs compared the final version of the booklets with the international translation verifier's comments to ensure that their suggestions had been incorporated appropriately into the materials. The QCMs' report with this information was then delivered to the TIMSS & PIRLS International Study Center.

5 For more information on the PIRLS Quality Control program, please see Chapter 7.

References

Hastedt, D., Gonzalez, E.J., & Kennedy, A.M. (2003). PIRLS survey operations procedures. In M.O. Martin, I.V.S. Mullis, and A.M Kennedy (Eds.), *PIRLS 2001 technical report*. Chestnut Hill, MA: Boston College.

Kennedy, A.M., Mullis, I.V.S., Martin, M.O., & Trong, K.L. (Eds.). (2007). *PIRLS 2006 encyclopedia: A guide to reading education in the forty PIRLS 2006 countries*. Chestnut Hill, MA: Boston College.

TIMSS & PIRLS International Study Center. (2005). *PIRLS 2006 Survey operations procedures unit 2: Preparing materials for the PIRLS assessment*. Chestnut Hill, MA: Boston College.

UNESCO Institute for Statistics. (1999). *Operational manual for ISCED-1997: International standard classification of education.*

Chapter 6

PIRLS Survey Operations Procedures

Juliane Barth, Ann M. Kennedy, and Kathleen L. Trong

6.1 Overview

Conducting PIRLS 2006 was an ambitious enterprise in each country that required the careful coordination of schools, staff, and materials by the National Research Coordinator (NRC). In order to assist the NRCs and synchronize activities internationally, a standardized set of survey operations procedures was developed for each country to follow.

The design of the survey operations procedures was a collaborative effort between the TIMSS & PIRLS International Study Center, the IEA Secretariat, the IEA Data Processing and Research Center (DPC), and Statistics Canada. Procedures used successfully in PIRLS 2001, previous TIMSS studies, and other IEA studies, as well as feedback received from the countries that participated in the PIRLS 2006 field test, were used as a basis for developing these procedures.

Survey operations procedures included contacting schools and sampling classes, preparing materials for data collection, administering the assessment, scoring the assessment, and creating the data files. Procedures for quality control and attaining feedback on survey activities also were provided. Guidelines for each of these activities, outlined in subsequent sections of this chapter, were described in an international set of materials that was provided to each NRC.

6.2 Responsibilities of the National Research Coordinator

The NRC for each country was responsible for coordinating PIRLS survey activities at the national level. This included acting as the contact person for all those involved in PIRLS within the country, as well as being the representative of the country at the international level. With guidance from organizations that directed PIRLS and experts from within the country, the NRC ultimately made all of the national decisions regarding PIRLS, adapting procedures as necessary to make them appropriate for their national context.

6.3 Documentation and Software

Each NRC was provided with a comprehensive set of manuals and software to guide them through the survey operations procedures. Each of these is described below.

- The *School Sampling Manual* (TIMSS & PIRLS International Study Center, 2004) defines the PIRLS 2006 target population and sampling goals and describes the procedures for the sampling of schools.
- The *Survey Operations Procedures Units* are a series of documents that provided a framework for the survey operations. These were organized and distributed chronologically according to the activity and were meant to be used in conjunction with other more specialized manuals.
    - *Unit 1—Contacting Schools and Sampling Classes* (TIMSS & PIRLS International Study Center, 2005e)
    - *Unit 2—Preparing Materials for Data Collection* (TIMSS & PIRLS International Study Center, 2005f)
    - *Unit 3—Administering the PIRLS 2006 Assessment* (TIMSS & PIRLS International Study Center, 2005g)
    - *Unit 4—Scoring the PIRLS 2006 Assessment* (TIMSS & PIRLS International Study Center, 2005h)
    - *Unit 5—Creating the PIRLS 2006 Data Files* (TIMSS & PIRLS International Study Center, 2005i)
- The *School Coordinator Manual* (TIMSS & PIRLS International Study Center, 2005d) describes the steps to be taken by the School Coordinator, which included being responsible for all testing materials

and survey tracking forms, organizing the test administration, and returning the completed testing materials to the NRC.

- The *Windows Within-school Sampling Software and Manual* (IEA, 2005d) helps the NRC to randomly select the PIRLS classes in each sampled school, prepares the survey tracking forms, assigns test booklets to students, and prints labels for the test booklets and questionnaires.
- The *Test Administrator Manual* (TIMSS & PIRLS International Study Center, 2005j) describes the procedures for the Test Administrator to follow during testing, including the timing of and the script used to administer the test, as well as how materials should be returned to the School Coordinator.
- The *International and National Quality Control Manuals* (TIMSS & PIRLS International Study Center, 2005a, 2005b) describe the procedures that quality control monitors should follow when observing testing sessions, as well as the materials they should collect as part of quality control.
- The *Scoring Guides for Constructed-response Items* (TIMSS & PIRLS International Study Center, 2005c) provide detailed and explicit guides used to score each constructed-response item.
- The *Trend Scoring and Reliability Scoring Software and Manual* (TSRS) (IEA, 2005b) is used to ensure consistent scoring over time. This program incorporates a database for countries that participated in PIRLS 2001 and contains a sample of student responses from the PIRLS 2001 data collection. The software allows PIRLS 2006 scorers to rescore the 2001 student response sample, train scorers, and document the reliability of the scoring process over time.
- The *Cross-country Scoring and Reliability Software and Manual* (CCSRS) (IEA, 2005a) is used to document the reliability of scoring across countries. The program incorporates a database containing a sample of student responses to constructed-response questions, collected from English-speaking countries and enables every country to score a common set of student responses.
- The *Windows Data Entry Manager Software and Manual* (IEA, 2005c) captures all PIRLS 2006 responses using keyboard data entry and performs a number of validation checks on the data entered.

- The *Data Correction Software* (DCS) enables national center staff to detect and correct logical inconsistencies in the PIRLS background data.

6.4 Contacting Schools and Sampling Classes

One of the essential first steps in PIRLS 2006 was to establish good working relationships with the schools that have been sampled to participate in the study (for more information on sampling procedures, please refer to Chapter 4). NRCs were responsible for contacting these schools and encouraging participation in the assessment, which often involved obtaining support from national or regional education authorities, depending on the national context.

6.4.1 School Coordinators

Once a school agreed to participate, a School Coordinator was identified and trained by staff at the national center. This person was responsible for all PIRLS activities within that particular school and often was a teacher or staff member. In some cases, a School Coordinator was a member of the national center staff and was responsible for several schools in an area. School Coordinators were provided with the *School Coordinator Manual*, describing their responsibilities in detail and encouraging them to contact the NRC if they had any questions.

The responsibilities of the School Coordinator included providing information about their school; coordinating the date, time, and place for testing; distributing teacher and school questionnaires; obtaining parental permission (if necessary); and identifying and training a Test Administrator. They also ensured that all testing materials were received and kept secure until administration and returned the completed materials to the national center.

6.4.2 Survey Tracking Forms

A large part of the School Coordinator's activities involved providing information about the classes and students in their school. To do this in an organized manner, survey tracking forms were used. Most of these forms were generated by the *Windows Within-school Sampling Software* (WinW3S), completed by schools, and returned to the national centers. The forms were extremely useful in the facilitation of sampling and data collection and were retained for the purpose of data entry verification.

A Class Listing Form was provided to each School Coordinator who listed all of the eligible classes in the target population at that school and provided

details about these classes. From this information, a Class Sampling Form was produced by WinW3S for each school, indicating which classes in the school were selected as part of the sample. A Student Listing Form was created for these sampled classes so that the School Coordinator could list all of the students' names and their information (including exclusion codes, which are discussed in Chapter 9) and return this form to the national center. In addition, a Student Tracking Form was used to document the participation status of each student in the tested classes, and a Teacher Tracking Form was used to document the completion of the *Teacher Questionnaire*.

6.5 Administering the PIRLS 2006 Assessment

Distributing materials to the schools required careful organization and planning on the part of the NRC. Each sampled student was assigned 1 of 13 achievement booklets in a systematic rotation so that each achievement block within the booklets was given to an equal number of students in each country. Each student also was assigned a *Student Questionnaire* and a *Learning to Read Survey* for his or her parent to complete. These materials were packaged for each sampled class. In addition, a *Teacher Questionnaire* was sent for each teacher listed on the Teacher Tracking Form and a *School Questionnaire* for the principal. The packaged materials were sent to the School Coordinator, who confirmed receipt of all instruments, prior to the testing date. The *School Questionnaire* and *Teacher Questionnaire* then were distributed, while the other instruments were kept in a secure room until the testing date.

6.5.1 Test Administrators

The PIRLS 2006 assessment was conducted by the Test Administrator for each class. This person was chosen and trained by the School Coordinator, although in many cases, the School Coordinator also acted as the Test Administrator. Each Test Administrator was provided with the *Test Administrator Manual*, which outlined his or her responsibilities. The Test Administrator was responsible for distributing materials to the appropriate students, leading students through the assessment, and timing the sessions accurately. Following the assessment, they administered the *Student Questionnaire* and distributed the *Learning to Read Survey* for the students' parents.

6.5.2 Timing of the Testing Sessions

The administration of PIRLS 2006 consisted of two sessions, a test administration session and a student questionnaire session. The test administration session concerned the achievement booklets, which contained two parts. This was followed by the completion of the *Student Questionnaire*. The time allotted for each of these sections was standardized across countries, with 40 minutes allowed for each part of the achievement booklet. However, if all of the students finished after 30 minutes, the section could be ended sooner. Test Administrators were required to document the starting and ending time of each section on the Test Administration Form. The timing of the sessions was as follows:

- Preparation for Part 1: approximately 10 minutes, including instructions and booklet distribution
- Achievement booklet, Part 1: 40 minutes
- Break: approximately 15 minutes
- Preparation for Part 2: approximately 5 minutes
- Achievement booklet, Part 2: 40 minutes
- *Student Questionnaire*: at least 20 minutes
- Distribution of *Learning to Read Survey*: approximately 5 minutes

6.5.3 Documenting Participation

In addition to the information about the school and its students collected by the School Coordinator, the Test Administrator also used the *Student Tracking Form* during testing. This form was used to distribute the booklets to the correct students and to document student participation.

The School Coordinator used this information to calculate the participation rate. If this was below 90 percent in any class, it was the coordinator's responsibility to hold a makeup session for the absent students before returning all testing materials and survey tracking forms to the national center.

Once the materials had been returned to the national center, the NRC verified the materials, checking that all survey tracking forms had been completed. The national center verified that testing materials were returned for each student listed on the Student Tracking Form, and that the recorded participation status matched the information in the test instruments. Information recorded on the survey tracking forms was then recorded in

Windows Within-school Sampling Software (WinW3S). The software was used to check the data for missing and/or inconsistent information and for verification of the data entry process at a later stage, in conjunction with the data entry software, *Windows Data Entry Manager* (WinDEM).

6.5.4 Quality Control

During the test administrations, 10 percent of schools were visited by an International Quality Control Monitor. These monitors were hired by the IEA to verify the quality of the materials and adherence to the test administration procedures in each country. During their school visits, they noted any changes made to the standardized administration script, timing, or procedures and interviewed the School Coordinator about his or her experiences with the PIRLS 2006 assessment. They also were responsible for a final verification of the translated achievement booklets. These were examined while reviewing the comments made by the international translation verifier, and the extent to which the verifier's suggested changes had been integrated was documented. These responsibilities were described in the *International Quality Control Monitor Manual*, and training was provided by staff from the IEA Secretariat and the TIMSS & PIRLS International Study Center.

Additionally, countries were asked to conduct their own quality control procedures in another 10 percent of sampled schools, based on the international program. To assist them, countries were provided with the *National Quality Control Observer Manual*, which was used to train their observers and modified to suit their national system.

6.6 Scoring the PIRLS 2006 Assessment

Scoring the PIRLS 2006 instruments in a reliable manner was critical to the quality of the results. To prepare for this task, NRCs were provided with suggestions on how to organize staff and materials. They also were given guidelines on how to select and train scorers to accurately and reliably score the constructed-response achievement items. NRCs were encouraged to employ scorers who were attentive to detail and familiar with education, particularly those with a background in reading instruction.

At international meetings, NRCs were trained to score each of the constructed-response items in the PIRLS 2006 assessment. At these training sessions (which were discussed in Chapter 2), each scoring guide was reviewed

together with examples of student responses that had already been scored according to the guide. The examples were chosen to represent a range of response types, intended to demonstrate the guides as clearly as possible. Following this, NRCs practiced applying the scoring guide to a different set of student responses that had not yet been scored. The scores NRCs gave to these practice papers were shared with the group and any discrepancies discussed. Following the training, NRCs were given a set of the correct scores for the practice papers along with rationales.

NRCs used this information to train their scoring staff on how to apply the *PIRLS 2006 Scoring Guides*. In some cases, NRCs created their own anchor and practice papers from student responses collected from the field test in their country.

In order to demonstrate the quality of the PIRLS 2006 data, it was important to document the reliability of the scoring process within countries, over time, and across countries.

To establish the reliability of the scoring *within each country*, NRCs were required to have a random sample of at least 200 student responses to each item scored independently by two different scorers. The double-scored booklets were selected randomly by *Windows Within-school Sampling Software*, indicated on the cover page of the test booklet. The degree of agreement between the scores assigned by the two independent scorers is a measure of the reliability of the scoring process. The scoring procedure, recommended by the TIMSS & PIRLS International Study Center, interspersed the scoring of the reliability sample with the normal scoring activity, with both taking place simultaneously in a systematic manner.

To measure the reliability of the scoring process *over time* (trend scoring), PIRLS 2006 took steps to document that the constructed-response questions that were carried over from PIRLS 2001 have been scored in the same way in both assessments. For this purpose, following the PIRLS 2001 data collection, countries that participated in this assessment sent samples of their administered and scored test booklets to the IEA DPC. These were digitally scanned and stored for later use in PIRLS 2006. Using this approach, the student responses from 2001 could be rescored by the 2006 scoring staff as a reliability check. The responses were made available to the scorers by the *Trend Scoring Reliability Software* (TSRS). This software allowed student responses to each of the items to be scored electronically. NRCs were asked to have at least two independent

scorers rescore all student responses presented by the software, totaling approximately 200 responses per item. Half of the items had to be scored before the normal scoring activity for PIRLS 2006 began. If the agreement of the scorers fell below 85 percent, retraining of the scorers was required and previously entered scores were disregarded and were scored again, as long as none of the items scored violated the agreement criteria. As soon as the 85 percent agreement criteria agreement was reached on the scored items, the second half of the TSRS student responses could be completed. Whereas the first half of the items were scored before the normal scoring activity for PIRLS 2006 took place, the second half of the items were scored at the same time as the PIRLS 2006 scoring.

In order to measure the reliability of the scoring process *across countries*, NRCs had to have at least two members of their PIRLS 2006 scoring staff score approximately 200 student responses to constructed-response items in English. Student responses to one half of the items were collected throughout the field test from English-speaking countries. The student responses to the second half of the items were taken from English-speaking countries' booklets from PIRLS 2001. Again, the *Cross-country Scoring Reliability Software* scanned the student responses and made them available to the scorers. The program allowed scorers to score the responses electronically by item. The cross-country scoring took place after the normal PIRLS 2006 scoring activity. The degree of agreement between scorers from the various countries may be taken as a measure of cross-country scoring reliability.

6.7 Creating the PIRLS 2006 Data Files

As described earlier in this chapter, the IEA DPC provided an integrated computer program for keyboard data entry and data verification known as WinDEM. The program worked in conjunction with WinW3S, so that it was not necessary to reenter tracking information that had been recorded in WinW3S. WinDEM was primarily used for the entry of data from test booklets and questionnaires. The software also offered data and file management capabilities, a convenient checking and editing mechanism, interactive error detection, and reporting and quality-control procedures. Detailed information and operational instructions were provided in the manual that accompanied the software.

One of the benefits of using WinDEM was that it incorporated the international codebooks describing all variables and their characteristics, thus ensuring that the data files that were produced fulfilled the PIRLS 2006 rules and standards for data entry. Data entry training was provided to NRCs and their national center staff at various stages of the project, including an extensive 4-day training seminar before the field test and before the main data collection.

During the PIRLS 2006 assessment, data were gathered from students, parents, teachers, and school principals. These data were recorded into WinDEM data files as follows:

- *School background data file* contained principals' responses recorded from the *School Questionnaire*.
- *Teacher background data file* contained responses recorded from the *Teacher Questionnaire*.
- *Student background data file* contained responses recorded from the *Student Questionnaire*.
- *Student achievement data file* contained responses recorded from the test booklets.
- *Constructed-response scoring reliability data file* contained the within-country scoring reliability data for the constructed-response questions.

Quality control throughout the data entry process is essential in maintaining accurate data. Therefore, NRCs were responsible for performing periodic reliability checks on the data entry and for applying a series of data verification checks provided in WinDEM. NRCs had to ensure that all data files submitted to the IEA DPC followed the international format and had passed all verification checks. As part of this process, NRCs required their data entry staff to double enter at least 5 percent of each instrument type to ensure puncher reliability and retrain staff if agreement fell below 1 percent. Additionally, the data verification module of WinDEM identified any problems with identification codes and out-of-range and otherwise invalid codes. NRCs also were asked to verify the integrity of the linkage between the students, parents, teachers, and schools entered into the WinDEM data files and the tracking information for those specified in WinW3S. When all data files had passed the WinDEM quality control checks, they were submitted to the IEA DPC along with data documentation for further checking and processing.

6.8 Survey Activities Questionnaire

As a structured way to obtain feedback about the survey operations procedures from NRCs, the *Survey Activities Questionnaire* was administered. This consisted of a series of questions concerning each of the various survey activities, how the NRCs conducted them, and space for any comments or suggestions they had. This questionnaire was available online for the NRCs to complete as each of the survey activities was concluding. This format enabled the respondents to reflect on their experiences immediately and to more accurately provide information that can be used to improve survey operations in the future.

6.9 PIRLS 2006 Field Test

The PIRLS 2006 field test was a smaller administration of the PIRLS 2006 assessment, involving approximately 1,200 students from each country. It was conducted from March to April 2005 in each of the 40 participating countries and involved 12 newly developed blocks (6 for each reading purpose). One primary goal of the field test was to gather data on the newly developed items in order to analyze their statistical properties. These analyses were used to select six blocks to include (along with secure blocks from PIRLS 2001) in the PIRLS 2006 assessment. Another goal of the field test was to practice conducting the survey operations procedures. This allowed the NRCs and their staff members to become acquainted with the data collection activities and refine their national operations. The field test gave NRCs a basis from which to improve the procedures for the PIRLS 2006 data collection. The field test resulted in some modifications to survey operations procedures and contributed significantly to the successful administration of PIRLS 2006.

References

IEA. (2005a). *Cross-country scoring and reliability manual*. Hamburg: IEA Data Processing and Research Center.

IEA. (2005b). *Trend scoring and reliability scoring manual*. Hamburg: IEA Data Processing and Research Center.

IEA. (2005c). *WinDEM: Windows data entry manager manual*. Hamburg: IEA Data Processing and Research Center.

IEA. (2005d). *WinW3S: Windows within-school sampling software manual*. Hamburg: IEA Data and Research Processing Center.

References *(continued)*

TIMSS & PIRLS International Study Center. (2004). *PIRLS 2006 school sampling manual*. Chestnut Hill, MA: Boston College.

TIMSS & PIRLS International Study Center. (2005a). *PIRLS 2006 international quality control monitor manual*. Chestnut Hill, MA: Boston College.

TIMSS & PIRLS International Study Center. (2005b). *PIRLS 2006 national quality control observer manual*. Chestnut Hill, MA: Boston College.

TIMSS & PIRLS International Study Center. (2005c). *PIRLS 2006 scoring guides for constructed-response items*. Chestnut Hill, MA: Boston College.

TIMSS & PIRLS International Study Center. (2005d). *PIRLS 2006 school coordinator manual*. Chestnut Hill, MA: Boston College.

TIMSS & PIRLS International Study Center. (2005e). *PIRLS 2006 survey operations procedures unit 1: Contacting schools and sampling classes*. Chestnut Hill, MA: Boston College.

TIMSS & PIRLS International Study Center. (2005f). *PIRLS 2006 survey operations procedures unit 2: Preparing materials for data collection*. Chestnut Hill, MA: Boston College.

TIMSS & PIRLS International Study Center. (2005g). *PIRLS 2006 survey operations procedures unit 3: Administering the PIRLS assessment*. Chestnut Hill, MA: Boston College.

TIMSS & PIRLS International Study Center. (2005h). *PIRLS 2006 survey operations procedures unit 4: Scoring the PIRLS assessment*. Chestnut Hill, MA: Boston College.

TIMSS & PIRLS International Study Center. (2005i). *PIRLS 2006 survey operations procedures unit 5: Creating the PIRLS data files*. Chestnut Hill, MA: Boston College.

TIMSS & PIRLS International Study Center. (2005j). *PIRLS 2006 test administrator manual*. Chestnut Hill, MA: Boston College.

Chapter 7

Quality Assurance in the PIRLS 2006 Data Collection

Ieva Johansone and Ann Kennedy

7.1 Overview

Quality assurance in large-scale international surveys such as PIRLS is extremely important for making valid comparisons of student achievement across many countries. In order to ensure the quality of the PIRLS data, considerable effort was made in developing standardized materials and survey operations procedures (for more information on survey operations procedures, please see Chapter 6.) In its commitment to high quality standards, the TIMSS & PIRLS International Study Center developed an ambitious program to monitor and document data collection activities in participating countries. To implement this program, an international Quality Control Monitor (QCM) in each of the participating countries was selected by the IEA Secretariat in cooperation with the national center.

The TIMSS & PIRLS International Study Center conducted an extensive, 2-day QCM training on observing PIRLS 2006 testing sessions and documenting test administration procedures in 15 classrooms. The QCMs were introduced to the PIRLS 2006 survey operations procedures, including data collection in the schools. Each QCM received the necessary materials for completing their tasks, including a copy of the *PIRLS 2006 International Quality Control Monitor Manual*, Classroom Observation Record*, PIRLS 2006 Survey Operations Procedures Units 1–3*, *School Coordinator Manual*, and *Test Administrator Manual*.

The major task of the international QCMs was to conduct site visits to a random sample of 15 schools during the test administration in their countries. Where necessary, the QCMs were permitted to recruit one or more assistants in order to effectively cover the territory and testing timetable. A total of 103 international QCMs and their assistants were trained across the 45 participants in PIRLS 2006.[1] Altogether, these monitors observed 669 testing sessions. The results of the QCM observations are reported in Section 7.2.

In addition to the international and national quality control programs, the National Research Coordinators (NRCs) were asked to complete the *Survey Activities Questionnaire* about their experiences with the PIRLS 2006 survey operations procedures and the quality of the assessment materials. The main purpose of the questionnaire was to gather opinions and information to be used to further improve the quality of the survey activities and materials for future PIRLS cycles. Section 7.3 summarizes information that reflects the quality of the PIRLS 2006 survey materials and procedures within the participating countries.

7.2 Quality Control Observations of the PIRLS 2006 Test Administration

For each testing session observed, QCMs completed the PIRLS 2006 Classroom Observation Record. The observation record was organized into the four sections, listed below, in order to facilitate accurate recording of the test administration's major activities.

Section A: Preliminary Activities of the Test Administrator

Section B: Assessment Session Activities

Section C: Summary Observations

Section D: Interview with the School Coordinator

7.2.1 Preliminary Activities of the Test Administrator

Section A of the Classroom Observation Record addresses preparation for the testing session. QCMs 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.

Exhibit 7.1 summarizes the results for Section A. In nearly all testing sessions, Test Administrators observed the proper preparatory procedures. For those few deviations that occurred, the QCMs provided reasonable explanations

1 PIRLS 2006 was conducted in 40 countries, including Belgium with 2 education systems (Flemish and French) and Canada with 5 provinces (Alberta, British Colombia, Nova Scotia, Ontario, and Quebec)—45 participants in total.

for almost all the discrepancies. For example, QCMs sometimes noted that a student had left school and/or a new student had joined the class, which was not documented on the list. In fact, this was the main reason for information about student test instruments not corresponding exactly to the Student Tracking Forms.

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

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?	98%	2%	0%
Had the Test Administrator familiarized himself or herself with the test administration script prior to the testing?	97%	2%	1%
Did the student identification information on the test booklets and student questionnaires correspond with the Student Tracking Form?	90%	3%	7%
Was there adequate seating space for the students to work without distractions?	97%	1%	2%
Was there adequate room for the Test Administrator to move around during the testing to ensure that student were following directions correctly?	98%	1%	1%
Did the Test Administrator have a stop watch or timer for accurately timing the testing session?	92%	6%	2%

7.2.2 Assessment Session Activities

Exhibits 7.2 through 7.4 present the QCM reports about the activities conducted during the assessment sessions. During each session, the achievement test was administered in two parts with a short break in between followed by another short break and the administration of the *Student Questionnaire*. Section B of the Classroom Observation Record addressed the activities that took place during the actual assessment session, including following the Test Administrator script, distributing and collecting test booklets, and making announcements during the testing sessions.

Activities during the first part of the testing session are presented in Exhibit 7.2. One of the most important standardizations for the assessment administration was the fact that the test administrator's script was followed in all participating countries. The QCMs reported that in almost all of their observations, 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 additions for clarification of procedures, rather than revisions or deletions.

Primarily because students had completed Part 1 before the allotted time had elapsed, the total testing time for the first part was not equal to the time allowed in 9 percent of the sessions. After 40 minutes, the Test Administrator instructed students to close their test booklets and announced the break to be followed by the second part of the test. In 97 percent of the cases, the Test Administrator made sure that students stopped working immediately. In most sessions, the room then was either secured or supervised during the break. When asked whether the break between parts was equal or less than 15 minutes, QCMs interpreted the question literally. As a result, QCMs gave a negative answer to this question, unless the break was "exactly" 15 minutes. The QCMs reported that the break between parts ranged from no break at all (in one case) to about half an hour.

Exhibit 7.3 summarizes the QCMs' observations during the second part of the testing session. In 92 percent of the sessions, the time spent to restart the testing session was 5 minutes or less. Similar to the timing of Part 1, in 14 percent of the classrooms, the testing session in Part 2 was shorter than the allotted 40 minutes because students had finished the achievement test early.

Exhibit 7.2 Assessment 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	91%	8% (Minor changes) 0% (Major)	1%
Distributing the materials	93%	4% (Minor) 1% (Major)	1%
Reading the directions	82%	16% (Minor) 1% (Major)	1%
Reading the examples	88%	10% (Minor) 1% (Major)	1%
If the Test Administrator made changes to the script, how would you describe them?			
Additions	21%	26%	53%
Revisions	10%	31%	59%
Deletions	4%	35%	61%
Did the Test Administrator distribute the test booklets according to the booklet assignment on the Student Tracking Form?	98%	1%	1%
Did the Test Administrator record attendance correctly on the Student Tracking Form?	98%	1%	1%
Did the total testing time for Part 1 equal the time allowed?	90%	9%	1%
Did the Test Administrator announce "you have 5 minutes left" prior to the end of Part 1?	94%	6%	0%
Were there any other time remaining announcements made during Part 1?	14%	85%	1%
At the end of Part 1, did the Test Administrator make sure all students had closed their booklets?	97%	2%	1%
Was the total time for the break equal to or less than 15 minutes?	73%	23%	4%
Were the booklets left unattended or unsecured during the break?	13%	85%	2%

Exhibit 7.3 Assessment 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?	92%	4%	4%
Was the total time for testing in Part 2 correct as indicated in the script?	85%	14%	1%
Did the Test Administrator announce "you have 5 minutes left" prior to the end of Part 2?	87%	12%	1%
Were there any other time remaining announcements made during Part 2?	10%	86%	4%
At the end of Part 2, did the Test Administrator collect the test booklets one at a time from each student?	94%	6%	0%
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*?	83%	14%	3%
Did the Test Administrator accurately read the script to end the testing and signal a break?	68%	22% (Minor changes) 3% (Major changes)	7%
If there were changes, how would you describe them?			
Additions	12%	24%	64%
Some minor changes	15%	20%	65%
Omissions	10%	25%	65%
Did the Test Administrator distribute the *Student Questionnaires* and give directions as specified in the script?	84%	6%	10%
Did the students ask for additional time to complete the questionnaire?	37%	51%	12%
Did the Test Administrator distribute a *Learning to Read Survey* to each student who participated in the testing?	60%	32%	8%
At the end of the session, prior to dismissing the students, did the Test Administrator thank the students for participating in the study?	88%	4%	8%

About 68 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, only 3 percent reported major changes. Most had made additions or other minor changes, such as paraphrasing the directions. In 37 percent of the QCM observations, the students requested additional time to complete the questionnaire, which in all cases was granted.

Exhibit 7.4 provides observations on student compliance with instructions and the alignment of the scripted instructions with their implementation.

The results show that in almost all of the sessions, the students complied well or very well with the instructions to stop working between parts of the test, and, in most cases, the dismissal of the students was orderly or very orderly.

Exhibit 7.4 Student Cooperation at the End of the Assessment Sessions

Question	Very Well	Well	Fairly Well	Not Well at All	Not Answered
When the Test Administrator ended Part 1, how well did the student comply with the instruction to stop work?	87%	11%	1%	0%	1%
When the Test Administrator ended Part 2, how well did the student comply with the instruction to stop work?	89%	9%	1%	0%	1%

Question	Very Orderly	Somewhat Orderly	Not Orderly at All	Not Answered
How orderly was the dismissal of the students?	77%	13%	3%	7%

7.2.3 General Observations

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

The results presented in Exhibits 7.5 and 7.6 show that, for most testing sessions, no problems were observed. In 99 percent of the cases, Test Administrators addressed students' questions, as instructed in the *Test Administrator Manual*.

QCMs reported evidence of students attempting to cheat on the test in only 2 percent of testing sessions. However, when asked to explain the situation, QCMs generally indicated that students were merely looking around at their neighbors to see whether or not their test booklets were different. Because the PIRLS 2006 test design involves 13 different booklets, students were unlikely to have the same booklet as their neighbors.

In the few sessions where a defective test instrument was detected, the Test Administrator nearly always replaced the instrument appropriately. All cases of a student refusing to take the test happened prior to the testing and were due mostly to the fact that parental permission for participation was denied.

In 13 percent of the observed testing sessions, a student left the room for an "emergency" during the testing session. In such cases, Test Administrators were instructed to collect the student's test booklet, and give it back after he or she returned. However, in many cases, the student had already completed the test and, thus, did not want to receive his or her test booklet back after returning to the classroom. In three cases, a student got sick and did not return to the testing at all, and, in all the remaining cases, students were instructed to close their booklets and leave them on their tables while being out of the classroom.

The QCMs reported that there were no cases where students were not orderly and cooperative at all during the testing sessions. In the very few cases where students' order or cooperation was less than perfect or very good, problems mostly appeared during the *Student Questionnaire* administration because students were obviously tired. In such cases, the Test Administrators managed to control the situation. The QCMs reported that the overall quality of all the testing sessions was good, very good, or, in 54 percent of the cases, excellent.

Exhibit 7.5 General Observations

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?	97%	2%	1%
Did the Test Administrator address students' questions appropriately?	99%	1%	0%
Did you see any evidence of students attempting to cheat on the tests (e.g., by copying from a neighbor)?	2%	97%	1%
Were any defective test booklets detected and replaced before the testing began?	2%	97%	1%
Were any defective test booklets detected and replaced after the testing began?	2%	96%	2%
If any defective test booklets were replaced, did the Test Administrator replace them appropriately?	46%	12%	42%
Did any students refuse to take the test either prior to the testing or during the testing?	5%	93%	2%
If a student refused, did the Test Administrator accurately follow the instructions for excusing the student (collect the test booklet and record the incident on the Student Tracking Form)?	32%	16%	53%
Did any students leave the room for an "emergency" during the testing?	13%	85%	2%
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)?	60%	31%	9%

Exhibit 7.6 Observations of Student Behavior

Question	Extremely	Moderately	Somewhat	Hardly	Not Answered
To what extent would you describe the students as orderly and cooperative?	76%	20%	2%	0%	2%

	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?	93%	2%	2%	2%	1%

	Excellent	Very good	Good	Fair	Poor	Not Answered
In general, how would you describe the overall quality of the testing session?	54%	33%	8%	3%	0%	2%

7.2.4 Interview with the School Coordinator

The QCMs 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 test administration, the responsiveness of the NRC to queries, the necessity for makeup sessions, and, as a validation of within-school sampling procedures, the organization of classes in the school.

The results presented in Exhibit 7.7 show that PIRLS 2006 was an administrative success in the eyes of School Coordinators. Mistakes that did occur tended to be minor and could be remedied prior to testing. There were only a few cases where there were items missing in the shipment of the test materials, and, in all such cases, they were resolved before the testing date. By and large, the of School Coordinators (82 percent) reported that the NRCs were responsive to their questions or concerns.

Sixty-three percent of the School Coordinators reported that they were able to collect the completed *Teacher Questionnaires* prior to the student testing. It was estimated that the *Teacher Questionnaire* would take about 30 minutes to complete. About half of the School Coordinators indicated that the estimate of 30 minutes was approximately correct, while 29 percent reported that the questionnaire took longer, and 7 percent said that it took less time to complete.

In 51 percent of the cases, School Coordinators indicated that students were given special instructions, motivational talks, or incentives by a school official or the classroom teacher prior to testing.

In 21 percent of the observed schools, the School Coordinator anticipated that a makeup session would be needed, and almost all of them were sure that a makeup session would be conducted.

Because the sampling of classes requires a complete list of all classes in the school at the target grade, QCMs were asked to verify that the class list did indeed include all classes. Although a significant number of School Coordinators reported that this was not so, the additional comments show that they were very confused by the question itself. Almost all of them commented that they sent a list of all classes to the national center, but only one or two classes were selected to participate. Therefore, there are students at the grade level who did not have a chance to participate. In one case, the School Coordinator reported that there was a class with students who had special needs at this grade level, which indicates that this class had been excluded from the testing at the very beginning of the sampling process. The confusion about this question will require some extra explanation and revision for future cycles of PIRLS studies.

The results in Exhibit 7.8 suggest that the majority of School Coordinators believed that the testing session went very well and that school staff had positive attitudes towards the PIRLS testing. The fact that 89 percent of respondents said they would be willing to serve as a School Coordinator in future international assessments may also be attibuted to these positive attitudes.

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?	84%	9%	7%
Did you receive the correct shipment of the following items?			
School Coordinator Manual	82%	11%	7%
Test Administrator Manual	81%	12%	7%
Student Tracking Forms	88%	5%	7%
Test booklets	83%	10%	7%
Student Questionnaires	83%	10%	7%
Learning to Read Surveys	85%	8%	7%
Teacher Questionnaires	91%	4%	5%
School Questionnaire	90%	4%	6%
Test Administration Form	83%	10%	7%
Teacher Tracking Form	75%	18%	7%
Envelopes or boxes addressed to the national center for the purpose of returning the materials after the assessment	76%	17%	7%
Was the National Coordinator responsive to your questions or concerns?	82%	3%	15%
Were you able to collect completed *Teacher Questionnaire(s)* prior to the test administration?	63%	34%	3%
Was the estimated time of 30 minutes to complete the *Teacher Questionnaires* a correct estimate?	47%	29% (Took longer) 7% (Took less time)	17%
Were you able to collect the completed *School Questionnaire* prior to the test administration?	61%	35%	4%
Were you satisfied with the accommodations (testing room) you were able to arrange for the testing?	97%	1%	2%
Do you anticipate that a makeup session will be required at your school?	21%	75%	4%
If you anticipate a makeup session, do you intend to conduct one?	91%	3%	6%
Did the students receive any special instructions, a motivational talk, or incentives to prepare them for the assessment?	51%	45%	4%
Is this a complete list of the classes in this grade in this school?	81%	6%	13%
To the best of your knowledge, are there any students in this grade level who are not in any of these classes?	17%	80%	3%
To the best of your knowledge, are there any students in this grade level in more than one of these classes?	1%	96%	3%
If there was another international assessment, would you be willing to serve as a School Coordinator?	89%	7%	4%

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?	84%	13%	0%	3%

	Positive	Neutral	Negative	Not Answered
Overall, how would you rate the attitude of the other school staff members towards the PIRLS testing?	74%	21%	2%	3%

	Worked Well	Needs Improvement	Not Applicable
Overall, do you feel the *PIRLS 2006 School Coordinator Manual* worked well or does it need improvement?	74%	15%	11%

7.3 Survey Activities Questionnaire

The *Survey Activities Questionnaire* was designed to elicit information about NRCs' experiences in preparing for and conducting the PIRLS 2006 data collection, with a focus on identifying and selecting samples, translating the test instruments, assembling and printing the test materials, packing and shipping the test materials, scoring constructed-response items, entering and verifying data, implementing the national quality assurance program, and suggesting improvements in the process. This section reports information gathered from the *Survey Activities Questionnaire*, reflecting the quality of the PIRLS 2006 survey materials and procedures in the participating countries.

To make this data collection more efficient, the questionnaire was administered to coordinators online. Out of 45 PIRLS 2006 participants, only the coordinator for Moldova did not complete the questionnaire.

7.3.1 Sampling

The *Survey Activities Questionnaire* involved some questions about sampling schools and classes.

Exhibit 7.9 shows that 40 countries were able to select their samples using the manuals provided by the TIMSS & PIRLS International Study Center. Three countries answered that their sample was selected by Statistics Canada, even if

they actually sampled classes themselves. In one case (Qatar), no school or class sampling was necessary because the PIRLS' sample included the entire target population. Almost all the countries used the Within-school Sampling Software provided by the IEA Data Processing and Research Center (DPC) to select classes. In the two cases where the sampling software was not used, countries chose to use their own software because they felt their experience using this software would make the process more efficient.

Eight NRCs encountered organizational constraints in their systems that necessitated deviations from the sample design. In each case, the Statistics Canada sampling expert was consulted to ensure that the altered design remained compatible with the PIRLS standards.

Exhibit 7.9 Results of the Survey Activities Questionnaire — 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 TIMSS & PIRLS International Study Center?	40	4	1
Did you use the Within-School Sampling Software provided by the IEA Data Processing and Research Center to select classes or students?	42	2	1
Were there any conditions or organizational constraints that necessitated deviations from the basic PIRLS sampling design?	8	36	1

7.3.2 Translating the Test Instruments

Exhibit 7.10 reports NRCs' answers to some of the questions about translating the test instruments. In translating the test passages and items, NRCs generally reported using their own staff or a combination of their staff and outside experts. The majority used their own staff for translating the background questionnaires. Almost all NRCs reported that they had gone through the process of external translation verification of passages, items, and background questionnaires organized by the IEA Secretariat. Luxembourg reported that to improve response rates they also administered the *Learning to Read Survey* for parents in French and Portuguese, even though only the German version was submitted for verification.

Exhibit 7.10 Results of the Survey Activities Questionnaire — Translating the Test Instruments

Question	Own Staff	Outside Translator(s)	Outside Reviewer(s)	Combination	Not Answered
Did you use your own staff or outside experts to translate the passages and items?	12	8	1	23	1
Did you use your own staff or outside experts to translate the background questionnaires?	23	2	0	19	1

	Yes	No	Not Answered
Did you go through the process of external translation verification of the passages and items by the IEA?	43	0	2
Did you go through the process of external translation verification of the background questionnaires by the IEA?	43	1	1

7.3.3 Assembling and Printing the PIRLS 2006 Instruments

The NRCs were asked to answer some questions about assembling and printing the test materials, as well as issues related to checking the materials and securely storing them.

The results in Exhibit 7.11 show that almost all NRCs were able to assemble the test booklets according to the instructions provided and that all countries went through the process of external layout verification of the test booklets by the TIMSS & PIRLS International Study Center. All countries except one conducted the recommended quality control checks during the printing process. In the one case, the NRCs did not conduct quality assurance procedures during the printing process due to a shortage of time. Eleven countries detected errors during the printing process that were fixed before sending the tests for administration.

All countries but one reported having followed procedures to protect the security of the tests during assembly and printing. One country was concerned that the potential exists for a breach of security because information was exchanged via email. However, steps are taken by using password protected secure sites developed by the IEA DPC for sharing data files between the NRCs and the IEA DPC, IEA Secretariat, and the TIMSS & PIRLS International Study Center.

Exhibit 7.11 Results of the Survey Activities Questionnaire — Assembling and Printing the PIRLS 2006 Instruments

Question	Yes	No	Not Answered
Were you able to assemble the test booklets according to the instructions provided by the TIMSS & PIRLS International Study Center?	43	1	1
Did you go through the process of external layout verification of the test booklets by the TIMSS & PIRLS International Study Center?	44	0	1
Did you conduct the quality assurance procedures for checking the test booklets during the printing process?	43	1	1
Were any errors detected during the printing process?	11	33	1
If errors were detected, what was the nature of the errors?			
Poor print quality	5	4	36
Pages missing	6	4	35
Page order	4	6	35
Upside down pages	7	0	38
Did you follow procedures to protect the security of the tests during the assembly and printing process?	43	1	1
Did you discover any potential breaches of security?	1	43	1

7.3.4 Packing and Shipping the Testing Materials

Some questions in the questionnaire addressed the extent to which NRCs detected errors in the testing materials as they were being packed for shipping to School Coordinators. However, as shown in Exhibit 7.12, very few errors were found in any of the materials, and NRC reported that these were remedied.

Exhibit 7.12 Results of the Survey Activities Questionnaire — Packing and Shipping the Testing 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	34	3	5	3
Supply of *Student Questionnaires*	37	2	3	3
Supply of *Learning to Read Surveys*	40	2	0	3
Student Tracking Forms	37	2	3	3
Teacher Tracking Forms	41	1	0	3
Test Administrator Manual	38	1	3	3
School Coordinator Manual	41	1	0	3
Supply of *Teacher Questionnaires*	41	1	0	3
School Questionnaire	40	2	0	3
Test booklet ID labels	35	5	2	3
Sequencing of booklets or questionnaires	40	2	0	3
Return labels	42	0	0	3
Self-addressed postcards for test dates	42	0	0	3

7.3.5 Scoring Constructed-response Items

The *Survey Activities Questionnaire* collected information from the NRCs about preparation for scoring the constructed-response items as well as the actual implementation of this complex task. The scoring process was an ambitious effort, requiring recruiting and training scoring staff to score student responses, including independent double scoring of a representative sample of responses to verify scoring reliability.

Exhibit 7.13 indicates that almost all NRCs understood the procedures for scoring the reliability sample, as explained in the *Survey Operations Manual*. In one case, it turned out that the scoring and questions related to data entry were answered by the data manager instead of the NRC by mistake. Thus, this person was not informed about scoring procedures.

Three NRCs reported that their own staff scored the constructed-response items, 17 reported that teachers did the scoring, 6 reported that university students were employed, and 16 reported that a combination of various professionals scored the constructed-response items.

Thirty-eight countries reported that they completed the cross-country reliability scoring, as instructed by the TIMSS & PIRLS International Study Center. Three countries had some time- and money-related problems in completing the task. Two countries could not find two English-speaking scorers, and, thus, only one person did the cross-country reliability scoring.

Only the trend countries that participated in PIRLS 2001 were asked to perform the trend reliability scoring, and almost all of them completed this task, as instructed by the TIMSS & PIRLS International Study Center. One country used a different software than the one provided by the IEA DPC. Three countries had failed to scan their PIRLS 2001 test booklets and, thus, did not have the student answers to use for scoring purposes. One country could not overcome some technical problems, and two countries did not complete the trend reliability scoring due to financial problems.

Exhibit 7.13 Results of the Survey Activities Questionnaire — Scoring Constructed-response Items

Question	Own Staff	Teachers	University Students	Combination of Scorers	Other	Not Answered
Who primarily scored your constructed-response items?	3	17	6	16	2	1

Question	Yes	No	Not Answered
Do you understand the procedure for scoring the within-country reliability sample, as explained by the TIMSS & PIRLS International Study Center?	43	1	1
Did you perform the Cross-country Reliability Scoring, as described by the TIMSS & PIRLS International Study Center?	38	5	2
Did you perform the Trend Reliability Scoring, as described by the TIMSS & PIRLS International Study Center?	21	7	17

7.3.6 Data Entry and Verification

Exhibit 7.14 shows that two thirds of the NRCs reported that they entered the data from a percentage of test booklets twice as a verification procedure. The estimated proportion of booklets to be entered twice ranged from 5 to 30 percent, with one country reporting that they re-entered 100 percent of the data. All NRCs established a secure storage area for the returned tests after data entry.

Exhibit 7.14 Results of the Survey Activities Questionnaire — Data Entry and Verification

Question	Yes	No	Not Answered
Did you enter a percentage of test booklets twice as a verification procedure?	30	13	2
Did you use the Windows Data Entry Manager software provided by the IEA Data Processing and Research Center to enter your test instrument data?	38	5	2
Where the returned tests stored in a secure area after scoring and data entry until the original documents could be discarded?	44	0	1

7.3.7 National Quality Assurance Program

As part of the national quality assurance activities, NRCs were required to send National Quality Control Observers to 10 percent of the participating schools to observe the test administration and document compliance with prescribed procedures. The last section of the *Survey Activities Questionnaire* addressed preparation for and implementation of the national quality assurance program.

As shown in Exhibit 7.15, all the national centers used the *National Quality Control Monitor Manual* provided by the TIMSS & PIRLS International Study Center in order to conduct their quality assurance program. Six NRCs reported that an external agency would conduct the classroom observations, 19 reported that a member of their staff would do so, and 7 reported that a combination of staff and external agency people would conduct the observations. Eleven NRCs reported that other professionals, such as inspectors, retired teachers, mathematics and science supervisors, or ministry representatives were recruited to conduct the quality assurance observations.

Exhibit 7.15 Results of the Survey Activities Questionnaire — National Quality Assurance Program

Question	An External Agency	Members of the National Center	A Combination of Observers	Other	Not Answered
Who conducted the classroom observations?	6	19	7	11	2

Question	Yes	No	Not Answered
When conducting your own quality assurance program, did you use the *National Quality Control Monitor Manual* provided by the TIMSS & PIRLS International Study Center?	44	0	1

References

TIMSS & PIRLS International Study Center. (2005a). *PIRLS 2006 international quality control monitor manual*. Chestnut Hill, MA: Boston College.

TIMSS & PIRLS International Study Center. (2005b). *PIRLS 2006 national quality control monitor manual*. Chestnut Hill, MA: Boston College.

TIMSS & PIRLS International Study Center. (2005d). *PIRLS 2006 school coordinator manual*. Chestnut Hill, MA: Boston College.

TIMSS & PIRLS International Study Center. (2005e). *PIRLS 2006 survey operations procedures unit 1: Contacting schools and sampling classes*. Chestnut Hill, MA: Boston College.

TIMSS & PIRLS International Study Center. (2005f). *PIRLS 2006 survey operations procedures unit 2: Preparing materials for data collection*. Chestnut Hill, MA: Boston College.

TIMSS & PIRLS International Study Center. (2005g). *PIRLS 2006 survey operations procedures unit 3: Administering the PIRLS assessment*. PIRLS, Chestnut Hill, MA: Boston College.

TIMSS & PIRLS International Study Center. (2005h). *PIRLS 2006 survey operations procedures unit 4: Scoring the PIRLS assessment*. Chestnut Hill, MA: Boston College.

TIMSS & PIRLS International Study Center. (2005i). *PIRLS 2006 survey operations procedures unit 5: Creating the PIRLS data files*. Chestnut Hill, MA: Boston College.

TIMSS & PIRLS International Study Center. (2005j). *PIRLS 2006 test administrator manual*. Chestnut Hill, MA: Boston College.

Chapter 8

Creating and Checking the PIRLS International Database

Juliane Barth and Oliver Neuschmidt

8.1 Overview

The PIRLS 2006 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 in 40 countries. Creating the PIRLS 2006 database and ensuring its integrity was a complex endeavor requiring close coordination and cooperation among the staff at the Data Processing and Research Center (DPC), the TIMSS & PIRLS International Study Center at Boston College, Statistics Canada, and the national centers of 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 (NRCs) and data managers of PIRLS participants, and the data checking and database creation procedures implemented by the IEA DPC in collaboration with the TIMSS & PIRLS International Study Center and Statistics Canada.

8.2 Software for Data File Creation

The IEA DPC went to great lengths to ensure that the data received from the PIRLS 2006 participants were of high quality and were internationally comparable. The foundation for quality assurance was laid before the first data arrived at the DPC by providing the PIRLS countries with software designed to standardize a range of operational and data related tasks.

- The WinW3S: Within-school Sampling Software for Windows (WinW3S) (IEA, 2005a) performed the within-school sampling operations adhering strictly to the sampling rules defined by the Statistics Canada and TIMSS & PIRLS International Study Center. The software also created all necessary tracking forms and stored student- and teacher-specific tracking form information (such as student's age, gender, and participation status).
- The WinDEM: Windows Data Entry Manager program (IEA, 2005b) enabled key entry of all PIRLS test and questionnaire data in a standard, internationally defined format. The software also includes a range of checks for data verification.

8.3 Data Entry at the National Centers

Each PIRLS 2006 national center was responsible for transcribing the information from the achievement booklets and questionnaires into computer data files. As described in Chapter 6, the IEA DPC supplied national research centers with the WinDEM software and manual (IEA, 2005b) to assist with data entry. The IEA DPC also provided countries with codebooks describing the structure of the data. The codebooks contained information about the variable names used for each variable in the survey instruments, and about field lengths, field locations, labels, valid ranges, default values, and missing codes. In order to facilitate data entry, the codebooks and data files were structured to match the test instruments and international version of the questionnaires. This meant that for each survey instrument there was a corresponding codebook, which served as a template for creating the corresponding survey instrument data file. The IEA DPC conducted a 3-day training seminar for the data managers from participating countries on the use of the WinW3S, WinDEM, and the codebooks.

The TIMSS & PIRLS International Study Center provided each NRC with the survey operations procedures, including general instructions about the

within-school sampling, translation and verification of test instruments, test administration, scoring procedures, and data entry and verification procedures (PIRLS 2005).

The national center in each country gathered data from tracking forms that were used to record information on students selected to participate in the study, as well as their schools, and teachers. Information from tracking forms was entered with help of WinW3S. The responses from the student achievement booklets as well as student, parents, teacher, and school questionnaires were entered into computer data files created from the codebook templates.

8.4 Data Checking and Editing at the National Centers

Before sending the data to the IEA DPC for further data processing, countries were responsible for checking the data files with the checks incorporated in WinDEM and specifically prepared for PIRLS 2006 and for undertaking corrections as necessary. The checks were mandatory for all countries:

- The structure of the data files conforms to the specifications in the international codebooks;
- The data values of categorical variables conform to the range validation criteria specified in the international codebooks;
- There are no duplicate records in the data file;
- There are no column shifts in the data file;
- The availability of the data is consistent with the corresponding indicator variables; and
- All participating schools, teachers, and students that have been selected are represented in the data files in accordance with the information in the survey tracking forms.

8.5 Submitting Data Files and Data Documentation to the IEA DPC

The following data files were used during data entry and submitted to the IEA DPC:

- The WinW3S database contained sampling information, as well as tracking form information (such as student's age, gender, and participation status), from all sampled students, teachers, and schools.

- The student background data file contained data from the *Student Questionnaire.*
- The parent (home) background data files contained data from the *Learning to Read Survey.*
- The student achievement data file contained student responses to the assigned test booklets.
- In order to check the reliability of the constructed-response item scoring, the constructed-response items were scored independently by a second scorer in a random sample of 100 booklets per type.[1] WinW3S defined the random sample. The responses from these booklets were stored in a reliability scoring file.
- The teacher background data files contained data from the *Teacher Questionnaire.*
- The school data file contained data from the *School Questionnaire*.

In addition to the submission of their survey data files to the IEA DPC, countries were requested to provide detailed data documentation. This included copies of all original survey tracking forms, copies of the national versions of translated test booklets and questionnaires, and National Adaptation Forms documenting all country-specific adaptations to the test instruments (for a description of authorized adaptations, see Chapter 5).

Countries also were asked to submit 100 test booklets of each type, which had been selected for the double scoring of constructed-response items. These booklets will be used to document the trend reliability of the scoring process between PIRLS 2006 and future cycles of the study.

8.6 Creating National Data Files for Within-country Analysis

Once the data were entered into data files at the national center, the data files were submitted to the IEA DPC for checking and input into the international database. This process is generally referred to as data cleaning. 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 by the IEA DPC:

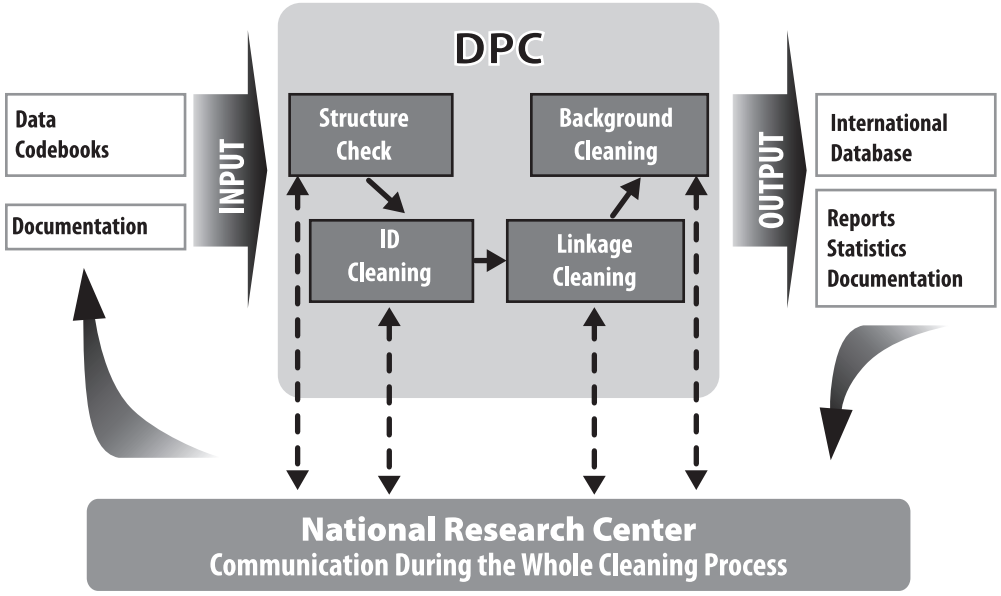
1 Booklet 9 and the Reader were exceptions, as they included only released texts from PIRLS 2006.

- Before use of real data, all data-cleaning programs were thoroughly tested using simulated data sets containing all possible problems and inconsistencies.
- All incoming data and documents were documented into a specific database. The date of arrival was recorded, along with any specific issues meriting attention.
- 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.
- All corrections to a country's data files were listed in a country-specific cleaning report.
- Occasionally, it was necessary to make changes to a country's data files. Every "manual" correction was logged using a specially developed editing program, which recorded all changes and allowed IEA DPC staff to undo changes, or to redo the whole manual cleaning process automatically at a later stage of the cleaning.
- Data Correction Software was developed at the IEA DPC and distributed among the participating countries to assist them in identifying and correcting inconsistencies between variables in the background questionnaire files.
- Once data cleaning was completed for a country, all cleaning steps were repeated from the beginning to detect any problems that might have been inadvertently introduced during the cleaning process.
- All national adaptations that countries recorded in their documentation were verified against the structure of the national data files. All deviations from the international data structure that were detected were recorded in a National Adaptation Database. The content of this database is available for data analysts as a supplement to the *PIRLS 2006 User Guide for the International Database* (Foy & Kennedy, 2008).

The main objective of the process of data checking and editing at the IEA DPC was to ensure that the data adheres to international formats, that school, teacher, parents and student information could be linked between different survey files, and that the data accurately and consistently reflected

the information collected within each country. Exhibit 8.1 presents a graphical representation of PIRLS data processing.

Exhibit 8.1 Overview of Data Processing at the IEA DPC



The program-based data cleaning consisted of the following steps:

- Documentation and structure check,
- Valid range check,
- Identification variable (ID) cleaning,
- Linkage check, and
- Resolving inconsistencies in background questionnaire 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: National Adaptations Forms, Student Tracking 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 IEA DPC looked for differences between the international file structure and the national file structure. 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 the nature of such changes differed across the countries. Some countries administered the questionnaires without any changes apart from translation, whereas other countries inserted items or options within existing international variables or added national variables. To keep track of any adaptations, National Adaptation Forms were used to adapt the codebooks, and, where necessary, the IEA DPC modified the structure of the country's data to ensure comparability with the structure of the international codebooks.

As part of this standardization process, since direct correspondence between the data entry instruments and the data files was no longer necessary, the file structure was rearranged from a booklet-oriented model designed to facilitate data entry to an item-oriented layout more suited to data analysis. Variables created purely for verification purposes during data entry were dropped at this time, and a provision was added for new variables necessary for analysis and reporting (i.e., reporting variables, derived variables, sampling weights, and achievement scores).

After each data file matched the international standard, as specified in the international codebooks, a series of standard cleaning rules were applied to the files. This was conducted using the set of programs developed at the IEA DPC that could identify and, in many cases, correct inconsistencies in the data. Each problem was recorded in a database, identified by a unique problem number, together with a description of the problem and the action taken.

Problems that could not be addressed were reported to the responsible NRC so that original data collection instruments and tracking forms could be checked to trace the source of the discrepancies. Wherever possible, staff at the IEA DPC suggested a remedy, and data files then were updated to reflect the solutions. After all automatic updates had been applied, remaining corrections to the data files were modified directly by keyboard, using a specially developed editing program.

8.6.2 Valid Range Check

"Valid range" indicates the range of the values considered to be correct and meaningful for a specific variable. For example, students' gender had two valid values: "1" for girls and "2" for boys. All other values were considered invalid. There also were questions in the school and teacher background questionnaires where the respondent wrote in a number—the principal was asked to supply the school enrollment, for example. For such variables, valid ranges may vary from country to country, and the broad ranges were set as acceptable 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. Data cleaning at the IEA 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 the countries.

8.6.3 Identification Variable (ID) Cleaning

Each record in a data file should have a unique identification number (ID). 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 data file. In the rare case that records contained different data apart from the ID numbers, and it was not possible to detect which records contained the "true data", both records were removed from the data files. However, the IEA DPC made every effort to keep such losses to a minimum.

The ID cleaning focused on the student background questionnaire file, because most of the critical variables were present in this file type. Apart from the unique ID, there were variables pertaining to 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 were essential in resolving any anomalies, as was close cooperation with National Research Coordinators. The information about participation and exclusion was sent to Statistics Canada, where it was used to calculate participation rates, exclusion rates, and sampling weights.

8.6.4 Linkage Check

In PIRLS, data about students and their homes, schools, and teachers appear in several data files. It is crucial that the records from these files are linked to each other correctly to obtain meaningful results. Therefore, another important check

run at the IEA DPC is 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 student 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.

8.6.5 Resolving Inconsistencies in Background Questionnaires

All background questionnaire data were checked for consistency among the responses given. The number of inconsistent and implausible responses in background files varied from country to country, but considering the complexities involved, no country submitted data completely free of inconsistent responses. Inconsistencies were addressed on a question-by-question basis, using available documentation to make an informed decision. For example, question number 1 in the *School Questionnaire* asked for the total school enrollment (number of students) in all grades, while question 2 asked for the enrollment in the fourth grade only. Clearly, the number given should not exceed the number given for 1. All such inconsistencies that were detected were flagged, and the NRCs were asked to investigate. Those cases that could not be corrected and where the data made no sense were recoded to "Omitted".

Occasionally, filter questions with "Yes" or "No" answers were used to direct respondents to a particular section of the questionnaire. These filter questions and the following dependent questions were subjected to the following cleaning rule: If the answer to the filter question was "No" and yet the dependent questions were answered, then the filter question was recoded to "Yes". During data entry, dependent variables were not treated differently from others. However, a special missing code was applied ("Not applicable") to dependent variables during data processing.

Split variable checks were applied to questions where the answer was coded into several variables. For example, question 21 in the *Student Questionnaire* asked students to respond "Yes" or "No" to each item in a list of home possessions. Occasionally, students responded to the "Yes" boxes, but left the "No" boxes blank. Since in these cases it was clear that no response meant "No", these were recoded accordingly.

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 described as student 23 in class 5 in school 122).

For further details about the standard cleaning procedures, please refer to the *General Cleaning Documentation PIRLS 2006* (IEA, 2007).

8.6.6 National Cleaning Documentation

National Research Coordinators received a detailed report of all problems identified in their data. This included documentation of any data problems detected by the cleaning programs and the steps applied to resolve them. NRCs also received a record of all deviations from the international data collection instruments and the international file structure

Additionally, the IEA DPC provided each NRC with revised data files incorporating all agreed upon edits, updates, and structural modifications. The revised files included a range of new variables that could be used for analytic purposes. For example, the student files included nationally standardized reading scores that could be used in preliminary national analyses to be conducted before the PIRLS 2006 International Database became available.

8.7 Handling of Missing Data

When the PIRLS data were entered using WinDEM, two types of entries were possible: valid data values and missing data values. Missing data can be assigned a value of omitted or not administered 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, four missing codes are used:

- Not administered: the respondent was not administered the actual item, and thus 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).

8.8 Data Products

8.8.1 Data Almanacs and Item Statistics

Each country received a set of data almanacs, or summary statistics, produced by the TIMSS & PIRLS International Study Center. These contained weighted summary statistics for each participating country on each variable included in the survey instruments. These data almanacs were sent to the participating countries for review. When necessary, they were accompanied by specific questions about the data presented in them. They were also used by the TIMSS & PIRLS International Study Center during the data review and in the production of the reporting exhibits.

Each country also received a set of preliminary national item and reliability statistics for review purposes. The item statistics contained summary information about items characteristics, such as the classical item difficulty index, the classical item discrimination index, the Rasch item difficulty, and the Rasch mean square fit index. The reliability statistics contained summary statistics about the percent of agreement between scorers on the scores assigned to the item.

8.8.2 Versions of the National Data Files

Building the international database was an iterative process. The IEA DPC provided NRCs with revised versions of their country's data files whenever a major step in data processing was completed. This also guaranteed that the NRCs had a chance to review their data and run their own checks to validate the data files. Several versions of the data files were sent to each country before the PIRLS 2006 International Database was made available. Each country received its own data only. The first version was sent as soon as the data could be regarded as 'clean' concerning identification codes and linkage issues. These first files contained nationally standardized achievement scores calculated by the IEA DPC using a Rasch-based scaling method. Documentation, with a list of the cleaning checks and corrections made in the data, was included to enable the National Research Coordinator to review the cleaning process.

Updated versions of data almanacs were posted at regular intervals on the Internet by the TIMSS & PIRLS International Study Center for statistical review. A second version of the data files was sent to the NRCs when the weights and the international achievement scores were available and had been merged to the files. A third version was sent after all exhibits of the international report

had been verified and final updates to the data files had been implemented, to enable the NRCs to validate the results presented in the report.

8.9 The PIRLS 2006 International Database

The international database incorporates all national data files. Data processing at the IEA DPC 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, parents, teacher, or principal; and
- Sampling weights and student achievement scores are available for international comparisons.

In a joint effort of the IEA DPC and the TIMSS & PIRLS International Study Center at Boston College, a national adaptations database was constructed to document all adaptations to background questionnaires, including a description of how the adaptations were addressed in the international database, such as recoding requirements. The information contained in this database is provided in Supplement 2 of the *PIRLS 2006 User Guide for the International Database* (Foy & Kennedy, 2008). This accompanying documentation listing all national deviations from the international version of the background instruments will help analysts interpret the results correctly.

References

Foy, P., & Kennedy, A.M. (Eds.). (2008). *PIRLS 2006 user guide for the international database.* Chestnut Hill, MA: Boston College.

IEA. (2005a). WinW3S: Within-school sampling software for Windows. [Computer software and manual.] Hamburg: IEA Data Processing and Research Center.

IEA. (2005b). WinDEM: Data entry manager for Windows. [Computer software and manual.] Hamburg: IEA Data Processing and Research Center.

IEA. (2007). *General cleaning documentation, PIRLS 2006.* Hamburg: IEA Data Processing and Research Center.

TIMSS & PIRLS International Study Center. (2005). *PIRLS 2006 survey operations procedures.* Chestnut Hill, MA: Boston College.

Chapter 9

PIRLS 2006 Sampling Weights and Participation Rates

Marc Joncas

9.1 Overview

Rigorous sampling of schools and students was a key component of the PIRLS 2006 project. Implementing the sampling plan was the responsibility of the National Research Coordinator (NRC) in each participating country. NRCs were supported in this endeavor by the PIRLS 2006 sampling consultants—staff from Statistics Canada and the Sampling Unit of the IEA Data Processing and Research Center (DPC)—who conducted the school sampling for most countries and trained the NRCs in selecting probability samples of students and using the *WinW3S: Within-school Sampling Software for Windows (WinW3S)* software provided by the IEA DPC (2005). As an essential part of their sampling activities, NRCs were responsible for providing detailed documentation describing their national sampling plans (sampling data, school sampling frames and school sample selections). The documentation for each PIRLS participant was reviewed and completed by the sampling consultants, including details on coverage and exclusion levels, stratification variables, sampling, participation rates, and variance estimates. The TIMSS & PIRLS International Study Center at Boston College, jointly with the PIRLS 2006 sampling consultants at Statistics Canada and the PIRLS 2006 Sampling Referee, Dr. Keith Rust of Westat, Inc., used this information to evaluate the quality of the samples.

This chapter gives a summary of the major characteristics of the national samples, along with a description of how sampling weights and participation rates are calculated. School and student participation rates for each country also are presented. More detailed 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.

9.2 Sampling Implementation

9.2.1 Target Population

As described in Chapter 4, the international desired target population for PIRLS 2006 was the grade that represented 4 years of schooling, counting from the first year of primary or elementary schooling, unless this would result in an average student age of less than 9.5 years. Exhibit 9.1 presents the grade identified as the target grade for sampling by each country, together with the number of years of formal schooling the grade represents and the average age of the students in that grade that were sampled for PIRLS. With few exceptions, the PIRLS 2006 target population in each country did indeed represent the fourth year of formal schooling. However, in England, New Zealand, Scotland, and Trinidad and Tobago children begin primary school at age 5, and therefore these countries assessed students in the fifth year of schooling. Their students were still among the youngest in PIRLS 2006 (9.9 to 10.3 years old). Because of issues related to the language of instruction, Luxembourg and South Africa also tested the fifth grade, even though it meant that their students were older. In Luxembourg, the assessment was conducted in German, which is the language of reading instruction but usually is either the student's second language or a foreign language. In an attempt to conduct the assessment in each student's language of instruction, South Africa tested in 11 different languages.

9.2.2 Population Coverage and Exclusions

Exhibit 9.2 summarizes the population coverage and exclusions for PIRLS 2006. National coverage of the international desired target population was generally comprehensive. All but Georgia, Lithuania, and Moldova sampled from 100 percent of their international desired population. Since coverage was below 100 percent, the results for these countries were footnoted in the PIRLS 2006 international report.

Exhibit 9.1 PIRLS 2006 National Grade Definitions

Country	Country's Name for Grade Tested	Years of Formal Schooling	Mean Age of Students Tested
Austria	Grade 4	4	10.3
Belgium Flemish	Grade 4 primary education	4	10.0
Belgium French	Grade 4	4	9.9
Bulgaria	Grade 4	4	10.9
Canada (Alberta)	Grade 4	4	9.9
Canada (British Columbia)	Grade 4	4	9.8
Canada (Nova Scotia)	Grade 4	4	10.0
Canada (Ontario)	Grade 4	4	9.8
Canada (Quebec)	2nd Year of 2nd Cycle	4	10.1
Chinese Taipei	Elementary school, Grade 4	4	10.1
Denmark	4th Form	4	10.9
England	Year 5	5	10.3
France	Cours Moyen 1	4	10.0
Georgia	Grade 4	4	10.1
Germany	Grade 4	4	10.5
Hong Kong SAR	Primary 4	4	10.0
Hungary	Grade 4	4	10.7
Iceland	Grade 4	4	9.8
Indonesia	Grade 4	4	10.4
Iran, Islamic Rep. Of	4th of Primary School	4	10.2
Israel	Grade 4	4	10.1
Italy	Grade 4 (IV Elementare)	4	9.7
Kuwait	Grade 4	4	9.8
Latvia	Grade 4	4	11.0
Lithuania	Grade 4	4	10.7
Luxembourg	Upper Primary Year 5	5	11.4
Macedonia, Rep of	Grade 4	4	10.6
Moldova, Rep. Of	Grade IV	4	10.9
Morocco	Grade 4 primary	4	10.8
Netherlands	Grade 4	4	10.3
New Zealand	Year 5	5	10.0
Norway	Grade 4	4	9.8
Poland	Grade 4	4	9.9
Qatar	Grade 4	4	9.8
Romania	Grade 4	4	10.9
Russian Federation	4th grade fro 4-year primary school; 3rd grade for 3-year primary school	3 or 4	10.8
Scotland	Primary 5 (P5)	5	9.9
Singapore	Primary 4	4	10.4
Slovak Republic	Grade 4	4	10.4
Slovenia	Grade 3 of 8-year elementary school; Grade 4 of 9-year elementary school	3 or 4	9.9
South Africa	Grade 4	4	10.9
Spain	Grade 4	4	9.9
Sweden	Grade 4	4	10.9
Trinidad and Tobago	Standard 3	5	10.1
United States	Grade 4	4	10.1
Iceland (5)	Grade 5	5	10.8
Norway (5)	Grade 5	5	10.8

Exhibit 9.2 Coverage of PIRLS 2006 Target Population

Countries	International Desired Population		National Desired Population		
	Country Coverage	Notes on Coverage	School-level Exclusions	Within-sample Exclusions	Overall Exclusions
Austria	100%		1.4%	3.8%	5.1%
Belgium (Flemish)	100%		6.1%	1.1%	7.1%
Belgium (French)	100%		3.7%	0.3%	3.9%
Bulgaria	100%		2.2%	4.3%	6.4%
Canada, Alberta	100%		2.0%	5.2%	7.1%
Canada, British Columbia	100%		2.2%	5.5%	7.6%
Canada, Nova Scotia	100%		0.2%	3.8%	4.0%
Canada, Ontario	100%		1.6%	6.8%	8.3%
Canada, Quebec	100%		2.4%	1.2%	3.6%
Chinese Taipei	100%		1.8%	1.1%	2.9%
Denmark	100%		0.5%	5.7%	6.2%
England	100%		1.6%	0.9%	2.4%
France	100%		3.4%	0.4%	3.8%
Georgia	80%	Students taught in Georgian	2.4%	5.0%	7.3%
Germany	100%		0.4%	0.3%	0.7%
Hong Kong SAR	100%		3.0%	0.9%	3.9%
Hungary	100%		2.3%	1.4%	3.7%
Iceland	100%		1.3%	2.5%	3.8%
Indonesia	100%		3.2%	0.0%	3.2%
Iran, Islamic Rep. of	100%		2.9%	0.9%	3.8%
Israel	100%		17.5%	6.1%	22.5%
Italy	100%		0.1%	5.2%	5.3%
Kuwait	100%		0.3%	0.0%	0.3%
Latvia	100%		4.3%	0.5%	4.7%
Lithuania	93%	Students taught in Lithuanian	0.9%	4.2%	5.1%
Luxembourg	100%		0.9%	3.0%	3.9%
Macedonia, Rep. of	100%		4.6%	0.3%	4.9%
Moldova, Rep. of	91%	Moldova less Predniestrian – Moldovan Republic	0.6%	0.0%	0.6%
Morocco	100%		1.1%	0.0%	1.1%
Netherlands	100%		3.5%	0.1%	3.6%
New Zealand	100%		1.4%	3.9%	5.3%
Norway	100%		1.0%	2.8%	3.8%
Poland	100%		0.9%	4.2%	5.1%
Qatar	100%		0.7%	0.7%	1.4%
Romania	100%		2.4%	0.0%	2.4%
Russian Federation	100%		6.8%	1.0%	7.7%
Scotland	100%		1.4%	0.9%	2.3%
Singapore	100%		0.9%	0.0%	0.9%
Slovak Republic	100%		1.8%	1.9%	3.6%
Slovenia	100%		0.2%	0.5%	0.8%
South Africa	100%		4.2%	0.1%	4.3%
Spain	100%		1.3%	4.0%	5.3%
Sweden	100%		2.4%	1.5%	3.9%
Trinidad and Tobago	100%		0.7%	0.0%	0.7%
United States	100%		3.2%	2.8%	5.9%

Within the national desired population, it was possible to exclude certain types of schools, such as very small or very remote schools, and certain types of students, such as those with a disability that prevented them from participating in the assessment. For the most part, school-level exclusions consisted of schools for students with disabilities and very small or remote schools. However, occasionally schools were excluded for other reasons, as documented in Appendix B. Within-school exclusions generally consisted of disabled students, or students who could not be assessed in the language of the test (Appendix B gives more details about the exclusions for each participant to PIRLS 2006). For most participants, the overall percentage of excluded students (combining school and within-school levels) was less than 5 percent. However, for Belgium (Flemish), Bulgaria, Denmark, Georgia, the Russian Federation, the United States, and the Canadian provinces of Alberta, British Columbia, and Ontario, exclusions accounted for between 5 and 10 percent of the desired population, and only for Israel did exclusions exceed 10 percent. Results for participants with more than 5 percent exclusions were annotated in the international report. Note that some PIRLS participants had no within-school exclusions.

9.2.3 General Sampling Approach

The basic sample design used in PIRLS 2006 is known as a two-stage stratified cluster design,[1] with the first stage consisting of a sample of schools, and the second stage consisting of a sample of intact classrooms from the target grade in the sampled schools. While all participants adopted this basic two-stage design, four countries, with approval from the PIRLS sampling consultants, added an extra sampling stage. The Russian Federation and the United States introduced a preliminary sampling stage, (first sampling regions in the case of the Russian Federation and primary sampling units consisting of metropolitan areas and counties in the case of the United States). Morocco and Singapore also added a third sampling stage; in these cases sub-sampling students within classrooms rather than selecting intact classes.

For countries participating in PIRLS 2006, school stratification was used to enhance the precision of the survey results. Many participants employed explicit stratification, where the complete school sampling frame was divided into smaller sampling frames according to some criterion, such as region, to ensure a predetermined number of schools sampled for each stratum. For example, Austria divided its sampling frame into nine regions to ensure proportional representation by region (see Appendix B for stratification information for each country). Stratification also could be done implicitly, a procedure by which

1 See Chapter 4 for a description of the sample design.

schools in a sampling frame were sorted according to a set of stratification variables prior to sampling. For example, Austria employed implicit stratification by district and school size within each regional stratum. Regardless of the other stratification variables used, all countries used implicit stratification by a measure of size (MOS) of the school.

All countries used a systematic (random start, fixed interval) probability-proportional-to-size (PPS) sampling approach to sample schools. Note that when this method is combined with an implicit stratification procedure, the allocation of schools in the sample is proportional to the size of the implicit strata. Within the sampled schools, classes were sampled using a systematic random method in all countries except Morocco and Singapore, where classes were sampled with probability proportional to size, and students within classes sampled with equal probability.

The PIRLS 2006 sample designs were implemented in an acceptable manner by all participants.

9.2.4 Target Population Sizes

Exhibit 9.3 shows the number of schools and students in each participant's target population, based on the sampling frame used to select the PIRLS 2006 sample, as well as the number of sampled schools and students that participated in the study, and an estimate of the student population size based on the student sample. The sample figures were derived using sampling weights (see Section 9.3). The population size estimate did not take into account the portion of the population excluded within schools, and made no adjustment for changes in the population between the date when the information in the sampling frame was collected and the date of the PIRLS 2006 data collection—usually a 2-year interval. Nevertheless, a comparison of the two estimates of the population size can be seen as a check on the sampling procedure. In most cases, the estimated population size closely matched the population size from the sampling frame.

9.3 Calculating Sampling Weights

The method of estimation used to produce estimates of totals from PIRLS data was through a simple weighted sum of all the responding records for the variables of interest. Estimates of percentages or means then were taken as ratios of these estimated totals. The two-stage stratified cluster PPS design used in PIRLS generally results in differential probabilities of selection of the

Exhibit 9.3 PIRLS 2006 Population and Sample Sizes

Country	Population		Sample			Mean Age
	Schools	Students	Schools	Students	Est. Pop.	
Austria	3,256	96,535	158	5,067	83,170	10.3
Belgium Flemish	2,121	64,240	137	4,479	66,150	10.0
Belgium French	1,664	49,614	150	4,552	47,756	9.9
Bulgaria	2,303	76,056	143	3,863	63,372	10.9
Canada (Alberta)	1,060	40,148	150	4,243	36,657	9.9
Canada (British Columbia)	1,236	45,723	148	4,150	42,963	9.8
Canada (Nova Scotia)	278	10,317	201	4,436	9,672	10.0
Canada (Ontario)	3,736	155,325	180	3,988	139,838	9.8
Canada (Quebec)	1,855	91,895	185	3,748	78,281	10.1
Chinese Taipei	2,170	313,505	150	4,589	304,488	10.1
Denmark	1,896	67,144	145	4,001	63,232	10.9
England	15,114	621,949	148	4,036	551,208	10.3
France	30,731	727,452	169	4,404	739,793	10.0
Georgia	2,063	47,143	149	4,402	44,793	10.1
Germany	18,757	793 946	405	7,899	776,861	10.5
Hong Kong SAR	648	74,952	144	4,712	70,683	10.0
Hungary	2,809	109,750	149	4,068	104,649	10.7
Iceland	136	4,174	128	3,673	4,074	9.8
Indonesia	150,441	4,372,275	168	4,774	4,227,746	10.4
Iran, Islamic Rep. Of	47,562	1,248,474	236	5,411	1,158,946	10.2
Israel	1,742	105,856	149	3,908	85,633	10.1
Italy	7,474	536,285	150	3,581	512,460	9.7
Kuwait	209	27,416	149	3,958	27,420	9.8
Latvia	825	20,575	147	4,162	19,793	11.0
Lithuania	1,118	35,989	146	4,701	32,730	10.7
Luxembourg	171	5,438	178	5,101	5,169	11.4
Macedonia, Rep of	308	25,696	150	4,002	22,928	10.6
Moldova, Rep. Of	1,388	50,258	150	4,036	43,867	10.9
Morocco	15,616	637,009	159	3,249	566,973	10.8
Netherlands	6,831	182,716	139	4,156	176,681	10.3
New Zealand	1,852	58,137	243	6,256	56,576	10.0
Norway	2,413	61,167	135	3,837	61,641	9.8
Poland	13,005	427,500	148	4,854	395,209	9.9
Qatar	124	7,542	119	6,680	7,138	9.8
Romania	7,329	229,632	146	4,273	198,634	10.9
Russian Federation	39,779	1,293,420	232	4,720	1,225,219	10.8
Scotland	2,100	61,326	130	3,775	57,115	9.9
Singapore	178	49,731	178	6,390	49,200	10.4
Slovak Republic	2,068	59,541	167	5,380	52,451	10.4
Slovenia	440	18,050	145	5,337	17,612	9.9
South Africa	15,045	942,494	429	16,073	970,522	10.9
Spain	11,631	406,360	152	4,094	391,084	9.9
Sweden	3,693	117,069	147	4,394	101,809	10.9
Trinidad and Tobago	500	19,915	147	3,951	17,190	10.1
United States	57,917	3,672,510	183	5,190	3,351,959	10.1
Iceland (5)	136	4,174	35	1,379	4,092	10.8
Norway (5)	2,413	61,167	66	1,808	66,051	10.8

students, requiring a unique sampling weight for each participating classroom in the study. The PIRLS 2006 student sampling weight comprised a series of multiplicative components. A basic weight was formed from the inverse of the probability of selecting a student from the population. This basic weight was adjusted by multiplicative factors that account for non-responding schools, classes, and students.

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 regions. A school-level participation adjustment was then made in the school weight to compensate for any sampled schools that did not participate and were not replaced. That 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. This classroom weight was calculated independently for each participating school. If a sampled classroom in a school did not participate, or if the participation rate among students in a classroom fell below 50 percent, a classroom-level participation adjustment was made to the classroom weight. Classroom participation adjustment could occur only within "participating schools" (a school was considered as a "participating school" if and only if there was at least one sampled classroom with at least 50 percent of its students participating in the study). If one of two (or more) selected classrooms in a school did not participate, the classroom participation adjustment was computed at the explicit stratum level rather than at the school level to reduce the risk of bias.

The third and final step consisted of calculating a student weight. For most PIRLS participants, because intact classrooms were sampled, each student in the sampled classrooms was certain of selection, and so the student weight was 1.0. When students were further sampled within classrooms, as was the case in Morocco and Singapore, a student weight reflecting the probability of the sampled students being selected within the classroom was calculated. A non-participation adjustment was then made to adjust for sampled students who did not take part in the testing. This adjustment 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 these three weights including 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 2006 sample design required that school selection probabilities be proportional to the school size, generally defined as enrolment in the target grade. The basic first stage weight for the i^{th} 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 i^{th} school, and

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

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

For countries such as the Russian Federation and the United States that included a preliminary sampling stage, 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 preliminary stage weight and the first stage weight, as described earlier.

In order to avoid ending up with some basic first stage weights being less than unity, the size of large schools (schools with sizes larger than the sampling interval given by M/n), was set back to the sampling interval. As a result, these large schools were sampled with equal probability without having to use an explicit stratification approach as for previous PIRLS and TIMSS cycles.

In a similar way but for different reasons, the size of small schools (see Chapter 4) was set to a constant so that these small schools could be sampled with equal probability without having to use explicit stratification.

9.3.2 School Non-participation Adjustment

First stage weights were calculated for all sampled and replacement schools that participated (i.e., with at least one sampled classroom with at least half of its students participating in the study). A school-level participation adjustment was required to compensate for schools that were sampled but did not participate, and were not replaced. Sampled schools that were found to be ineligible were removed from the calculation of this adjustment.[2] The school-level participation adjustment was calculated separately for each explicit stratum, 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.

Because in Qatar and Iceland all schools were included in the sample (i.e., census of the school population), the following school-level adjustment was used:

$$A_{sc} = \frac{m_s + m_{nr}}{m_s}$$

where m_s was the number of originally sampled students from schools that participated, and m_{nr} the number of originally sampled students from schools that did not participate.

The final first stage weight for the i^{th} 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 probability of a classroom within a sampled school being selected. All but Morocco and Singapore sampled classrooms within schools with equal probability. In these two exceptions, where student sub-sampling was involved, classrooms were sampled using PPS

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

techniques. Procedures for calculating sampling weights are presented below for both approaches.

Equal Probability Weighting: For the i^{th} school, let C^i be the total number of classrooms and c^i the number of sampled classrooms in the study. Using equal probability sampling, the basic second stage weight assigned to all sampled classrooms in the ith school was:

$$BW_{cl1}^{i} = \frac{C^i}{c^i}$$

For most PIRLS participants, c^i took the values 1, 2 or 3. Some PIRLS participants sampled all classrooms in a selected school.

Probability Proportional to Size Weighting (Morocco and Singapore only): For the i^{th} school, let $k^{i,j}$ be the size of the j^{th} classroom. Using PPS sampling, the final second stage weight assigned to the j^{th} sampled classroom in the i^{th} school was:

$$BW_{cl2}^{i,j} = \frac{K^i}{c^i \cdot k^{i,j}}$$

where c^i was the number of sampled classrooms in the i^{th} school, as defined earlier, and

$$K^i = \sum_{j=1}^{c^i} k^{i,j}$$

Singapore sampled two classrooms $(c^i = 2)$ and Morocco sampled a single classroom $(c^i = 1)$.

9.3.4 Classroom Non-participation Adjustment

Second stage weights were calculated for all sampled classrooms in the sampled schools and replacement schools that participated. A classroom-level participation adjustment was applied to compensate for classrooms that did not participate or where student participation rate was below 50 percent. Sampled classrooms with student participation below 50 percent were given a weight of zero and considered to be non-participating. The classroom-level participation

adjustment was calculated separately for each explicit stratum rather than school to minimize the risk of bias.

The adjustment was calculated as follows:

$$A_{cl} = \frac{\sum_{i}^{s+r1+r2} 1/c^i}{\sum_{i}^{s+r1+r2} \delta_i / c^i}$$

where c^i was the number of sampled classrooms in the i^{th} school, as defined earlier, and δ_i takes on value 1 if the classroom participated and 0 otherwise.

When no sub-sampling of classrooms was involved, the final second stage weight assigned to all sampled classrooms in the i^{th} school became:

$$FW_{cl1}^{i} = A_{cl} \cdot BW_{cl1}^{i}$$

When classrooms were sub-sampled within schools, the final second stage weight assigned to the j^{th} sampled classroom in the i^{th} school became:

$$FW_{cl2}^{i,j} = A_{cl} \cdot BW_{cl2}^{i,j}$$

9.3.5 The Third Stage (Student) Weight

The third stage weight represented the inverse of the probability of a student in a sampled class being selected. When intact classrooms that included all students were sampled, as was the case for all but two PIRLS 2006 participants, this probability was unity. However, the probability of selection varied when students were sampled within classrooms. Procedures for calculating weights are presented below for both sampling approaches. The third stage weight is calculated independently for each sampled classroom.

Sampling Intact Classrooms: The basic third stage weight for the j^{th} classroom in the i^{th} school was simply:

$$BW_{st1}^{i,j} = 1.0$$

Subsampling Students: (Morocco and Singapore only) The basic third stage weight for the j^{th} classroom in the i^{th} school was:

$$BW_{st2}^{i,j} = \frac{k^{i,j}}{s^{i,j}}$$

where $k^{i,j}$ was the size of the j^{th} classroom in the i^{th} school, as defined earlier, and $s^{i,j}$ was the number of sampled students per sampled classroom.

9.3.6 Adjustment for Student Non-participation

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

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

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

The third and final stage weight for students the j^{th} classroom in the i^{th} school thus became:

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

where Δ equals 1 when there was no student sub-sampling and 2 when students were sub-sampled within classrooms.

9.3.7 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. For example, when no sub-sampling of classrooms was involved, this product is given by:

$$W^{i,j} = FW_{sc}^{i} \cdot FW_{cl1}^{i} \cdot FW_{st1}^{i,j}$$

or

$$W^{i,j} = A_{sc} \cdot BW^{i}_{sc} \cdot FW^{i}_{cl1} \cdot A^{i,j}_{st} BW^{i,j}_{st\Delta}$$

When classrooms were sub-sampled within schools, the overall sampling weight was:

$$W^{i,j} = FW^{i}_{sc} FW^{i,j}_{cl2} \cdot FW^{i,j}_{st\Delta}$$

or

$$W^{i,j} = A_{sc} \cdot BW^{i}_{sc} \cdot FW^{i,j}_{cl2} \cdot A^{i,j}_{st} BW^{i,j}_{st\Delta}$$

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

9.4 Calculating School and Student Participation Rates

Since non-participation by sampled schools, classrooms, or students can lead to bias in the study results, a variety of participation rates were computed to show the level of success each PIRLS participant achieved in securing participation from their sampled schools, classrooms, and students. To monitor school participation, two school participation rates were computed: one based on originally sampled schools only, and one based on sampled and both first and second replacement schools. Classroom and student participation rates also were computed, as were overall participation rates.

9.4.1 Unweighted School Participation Rates

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

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

R^{sc-r}_{unw} = 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. A school was labelled as a "participating school" if at least one of its sampled classrooms had at least a 50 percent student participation rate. The rates were calculated as follows:

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

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

9.4.2 Unweighted Classroom Participation Rates

The unweighted classroom participation rate was computed as follows:

$$R_{unw}^{cl} = \frac{\sum_{i}^{s+r1+r2} c_{\star}^{i}}{\sum_{i}^{s+r1+r2} c_i}$$

where c^i was the number of sampled classrooms in the i^{th} school, and $c_{\star}^{i}$ was the number of participating sampled classrooms in the i^{th} school. Both summations are over all participating schools.

9.4.3 Unweighted Student Participation Rates

The unweighted student participation rate was computed as follows where summations are done over all participating schools and over all classrooms with at least 50 percent of its students participating in the study:

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

9.4.4 Unweighted Overall Participation Rates

Two unweighted overall participation rates were computed for each PIRLS participant. They were as follows:

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

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

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

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

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

9.4.5 Weighted School Participation Rates

Two weighted school-level participation rates were computed for each PIRLS participant. They were as follows:

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

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

The weighted school participation rates were calculated as follows:

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

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

where both the numerator and denominator were summations over all responding students and the appropriate classroom-level and student-level sampling weights were used. Δ takes the value one when no sub-sampling was involved and two otherwise. 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 enrolment in the target population. The numerator, however, changes from one equation to the next. Only students from originally sampled schools and from classrooms with at least 50 percent of their students participating in the study 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.

9.4.6 Weighted Classroom Participation Rates

The weighted classroom participation rate was computed as follows:

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

where both the numerator and denominator were summations over all responding students from classrooms with at least 50 percent of their students participating in the study, and the appropriate student-level sampling weights were used. Note that the basic classroom-level weight appears in the numerator, whereas the final classroom-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, either sampled or replacement.

9.4.7 Weighted Student Participation Rates

The weighted student participation rate was computed as follows:

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

where both the numerator and denominator were summations over all responding students from participating schools. 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 classroom-level participation rate for all participating schools, either sampled or replacement.

9.4.8 Weighted Overall Participation Rates

Two 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-r} = 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, weighted classroom participation rate, and the weighted student participation rate. They were computed as follows:

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

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

Weighted school, classroom, student, and overall participation rates were computed for each PIRLS participant using these procedures.

9.5 Meeting PIRLS's Standards for Sampling Participation

PIRLS participants understood that the goal for sampling participation was 100 percent for all sampled schools, classrooms, and students. Guidelines for reporting achievement data for PIRLS participants securing less than full participation were modeled after IEA's TIMSS and PIRLS previous studies. As summarized in Exhibit 9.4, countries were assigned to one of three categories on the basis of their sampling participation. Countries in Category 1 were considered to have met the PIRLS 2006 sampling requirements, and to have an acceptable participation rate. Countries in Category 2 met the participation 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 2006 was to have as many countries as possible achieve Category 1 status.

Exhibits 9.5 through 9.8 present the school, classroom, student, and overall participation rates and achieved sample sizes for each of the PIRLS 2006 participants. Almost all participants had excellent participation rates and belong in Category 1. However, Belgium (Flemish), the Netherlands, Scotland, and the United States met the sampling requirements only after including replacement schools, and therefore belong in Category 2. Although Norway had overall participation rates after including replacement schools of just below 75 percent (71%), it was decided during the sampling adjudication that this rate did not warrant placement in Category 3. Instead, results for that country in the international report 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 international report exhibits, 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 nearest whole percent). Countries in this category were annotated in the international report exhibits, 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 exhibits, below the other countries, in the international report. These countries were presented in alphabetical order.

9.6 Trends in Student Populations

Because an important goal of the PIRLS 2006 assessment was to measure changes in fourth-grade students' reading achievement since 2001, it is important to track any changes in population composition and coverage since then that might be related to student achievement. Exhibit 9.9 presents, for each country, four attributes of the populations sampled in 2001 and 2006: number of years of formal schooling, average student age, the score on the UNDP's human development index, and the percentage of students in the national desired population excluded from the assessment. Most countries and provinces were very similar with regard to these attributes across the two years, although it is noteworthy than the Russian Federation and Slovenia underwent structural changes in the age at which children enter schools that are reflected in their samples. In 2001, the Russian sample contained third-grade students from some regions and fourth-grade students from others, whereas all students were in fourth grade in 2006. Slovenia is in transition towards having all children begin school at an earlier age so that they all will have four years of primary schooling instead of three years, as was the case in 2001. However, the transition was not complete in 2006.

Exhibit 9.5 PIRLS 2006 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
Austria	100%	100%	160	158	158	0	158
Belgium Flemish	69%	92%	150	149	102	35	137
Belgium French	85%	100%	150	150	129	21	150
Bulgaria	88%	97%	150	147	130	13	143
Canada (Alberta)	100%	100%	150	150	150	0	150
Canada (British Columbia)	98%	99%	150	150	147	1	148
Canada (Nova Scotia)	99%	100%	201	201	200	1	201
Canada (Ontario)	88%	90%	200	198	173	7	180
Canada (Quebec)	96%	96%	200	194	185	0	185
Chinese Taipei	98%	100%	150	150	147	3	150
Denmark	89%	99%	150	146	128	17	145
England	86%	99%	150	150	129	19	148
France	94%	97%	175	175	164	5	169
Georgia	94%	100%	152	149	139	10	149
Germany	97%	99%	410	407	397	8	405
Hong Kong SAR	91%	100%	150	144	130	14	144
Hungary	99%	100%	150	149	147	2	149
Iceland	99%	99%	136	131	128	0	128
Indonesia	99%	100%	170	168	166	2	168
Iran, Islamic Rep. Of	100%	100%	240	236	235	1	236
Israel	98%	100%	150	149	146	3	149
Italy	91%	100%	150	150	136	14	150
Kuwait	99%	99%	150	150	149	0	149
Latvia	97%	98%	150	150	145	2	147
Lithuania	99%	100%	150	146	144	2	146
Luxembourg	100%	100%	183	178	178	0	178
Macedonia, Rep of	100%	100%	150	150	149	1	150
Moldova, Rep. Of	98%	100%	150	150	148	2	150
Morocco	98%	99%	160	160	156	3	159
Netherlands	70%	93%	150	150	104	35	139
New Zealand	92%	99%	250	250	220	23	243
Norway	68%	82%	178	177	118	17	135
Poland	99%	100%	150	148	147	1	148
Qatar	100%	100%	123	119	119	0	119
Romania	99%	99%	150	147	146	0	146
Russian Federation	100%	100%	232	232	232	0	232
Scotland	69%	87%	150	150	101	29	130
Singapore	100%	100%	178	178	178	0	178
Slovak Republic	93%	98%	174	171	155	12	167
Slovenia	93%	97%	150	150	140	5	145
South Africa	96%	99%	441	438	422	7	429
Spain	99%	100%	152	152	149	3	152
Sweden	100%	100%	150	147	147	0	147
Trinidad and Tobago	99%	99%	150	149	147	0	147
United States	57%	86%	222	214	120	63	183
Iceland (5)	100%	100%	35	35	35	0	35
Norway(5)	51%	68%	105	105	56	10	66

Exhibit 9.6 PIRLS 2006 School 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
Austria	98%	5,431	24	208	5,199	132	5,067
Belgium Flemish	99%	4,608	10	47	4,551	72	4,479
Belgium French	95%	4,810	19	14	4,777	225	4,552
Bulgaria	97%	4,156	37	135	3,984	121	3,863
Canada (Alberta)	96%	4,773	79	250	4,444	201	4,243
Canada (British Columbia)	95%	4,663	68	244	4,351	201	4,150
Canada (Nova Scotia)	96%	4,884	79	189	4,616	180	4,436
Canada (Ontario)	97%	4,436	40	252	4,144	156	3,988
Canada (Quebec)	84%	4,639	50	99	4,490	742	3,748
Chinese Taipei	99%	4,746	62	55	4,629	40	4,589
Denmark	97%	4,349	51	154	4,144	143	4,001
England	93%	4,492	117	38	4,337	301	4,036
France	98%	4,558	55	16	4,487	83	4,404
Georgia	98%	4,837	120	209	4,508	106	4,402
Germany	94%	8,395	49	44	8,302	403	7,899
Hong Kong SAR	97%	4,917	25	34	4,858	146	4,712
Hungary	97%	4,265	17	46	4,202	134	4,068
Iceland	91%	4,200	47	102	4,051	378	3,673
Indonesia	98%	4,981	99	0	4,882	108	4,774
Iran, Islamic Rep. Of	99%	5,609	122	22	5,465	54	5,411
Israel	93%	4,378	5	179	4,194	286	3,908
Italy	97%	3,882	31	153	3,698	117	3,581
Kuwait	89%	4,467	0	0	4,467	509	3,958
Latvia	94%	4,469	14	17	4,438	276	4,162
Lithuania	92%	5,400	67	183	5,150	449	4,701
Luxembourg	99%	5,342	15	158	5,169	68	5,101
Macedonia, Rep of	96%	4,209	33	11	4,165	163	4,002
Moldova, Rep. Of	95%	4,281	32	0	4,249	213	4,036
Morocco	95%	3,444	43	0	3,401	152	3,249
Netherlands	97%	4,366	63	5	4,298	142	4,156
New Zealand	96%	6,872	130	196	6,546	290	6,256
Norway	87%	4,570	27	134	4,409	572	3,837
Poland	95%	5,410	21	232	5,157	303	4,854
Qatar	94%	7,490	305	47	7,138	458	6,680
Romania	98%	4,463	97	0	4,366	93	4,273
Russian Federation	97%	4,911	20	35	4,856	136	4,720
Scotland	94%	4,123	66	41	4,016	241	3,775
Singapore	95%	6,760	67	0	6,693	303	6,390
Slovak Republic	96%	5,741	34	105	5,602	222	5,380
Slovenia	96%	5,596	12	27	5,557	220	5,337
South Africa	91%	17,934	475	35	17,424	1,351	16,073
Spain	97%	4,391	12	143	4,236	142	4,094
Sweden	96%	4,653	33	33	4,587	193	4,394
Trinidad and Tobago	95%	4,237	77	0	4,160	209	3,951
United States	96%	5,761	160	159	5,442	252	5,190
Iceland (5)	88%	1,618	15	42	1,561	182	1,379
Norway (5)	84%	2,238	14	62	2,162	354	1,808

Exhibit 9.7 PIRLS 2006 Participation Rates (Unweighted)

Country	School Participation Before Replacement	School Participation After Replacement	Classes Participation	Student Participation	Overall Participation Before Replacement	Overall Participation After Replacement
Austria	100%	100%	99%	98%	97%	97%
Belgium Flemish	69%	92%	100%	98%	67%	91%
Belgium French	86%	100%	100%	95%	82%	95%
Bulgaria	88%	97%	100%	97%	85%	94%
Canada (Alberta)	100%	100%	100%	96%	96%	96%
Canada (British Columbia)	98%	99%	100%	95%	94%	94%
Canada (Nova Scotia)	100%	100%	100%	96%	96%	96%
Canada (Ontario)	87%	91%	100%	96%	84%	88%
Canada (Quebec)	95%	95%	100%	84%	80%	80%
Chinese Taipei	98%	100%	100%	99%	97%	99%
Denmark	88%	99%	100%	97%	85%	96%
England	86%	99%	100%	93%	80%	92%
France	94%	97%	100%	98%	92%	95%
Georgia	93%	100%	100%	98%	91%	98%
Germany	98%	100%	100%	95%	93%	95%
Hong Kong SAR	90%	100%	100%	97%	88%	97%
Hungary	99%	100%	100%	97%	96%	97%
Iceland	98%	98%	100%	91%	89%	89%
Indonesia	99%	100%	100%	98%	97%	98%
Iran, Islamic Rep. Of	100%	100%	100%	99%	99%	99%
Israel	98%	100%	100%	93%	91%	93%
Italy	91%	100%	100%	97%	88%	97%
Kuwait	99%	99%	99%	89%	88%	88%
Latvia	97%	98%	100%	94%	91%	92%
Lithuania	99%	100%	100%	91%	90%	91%
Luxembourg	100%	100%	100%	99%	99%	99%
Macedonia, Rep of	99%	100%	100%	96%	95%	96%
Moldova, Rep. Of	99%	100%	100%	95%	94%	95%
Morocco	98%	99%	100%	96%	93%	95%
Netherlands	69%	93%	100%	97%	67%	90%
New Zealand	88%	97%	100%	96%	84%	93%
Norway	67%	76%	100%	87%	58%	66%
Poland	99%	100%	100%	94%	94%	94%
Qatar	100%	100%	100%	94%	94%	94%
Romania	99%	99%	100%	98%	97%	97%
Russian Federation	100%	100%	100%	97%	97%	97%
Scotland	67%	87%	100%	94%	63%	82%
Singapore	100%	100%	100%	96%	96%	96%
Slovak Republic	91%	98%	100%	96%	87%	94%
Slovenia	93%	97%	100%	96%	90%	93%
South Africa	96%	98%	100%	92%	89%	90%
Spain	98%	100%	100%	97%	95%	97%
Sweden	100%	100%	100%	96%	96%	96%
Trinidad and Tobago	99%	99%	100%	95%	94%	94%
United States	56%	86%	100%	95%	53%	81%
Iceland (5)	100%	100%	100%	88%	88%	88%
Norway (5)	53%	63%	97%	84%	43%	51%

Exhibit 9.8 PIRLS 2006 Participation Rates (Weighted)

Countries	School Participation		Classroom Participation	Student Participation	Overall Participation	
	Before Replacement	After Replacement			Before Replacement	After Replacement
Austria	100%	100%	99%	98%	97%	97%
Belgium (Flemish)	69%	92%	100%	99%	68%	91%
Belgium (French)	85%	100%	100%	95%	81%	95%
Bulgaria	88%	97%	100%	97%	85%	94%
Canada, Alberta	100%	100%	100%	96%	96%	96%
Canada, British Columbia	98%	99%	100%	95%	93%	94%
Canada, Nova Scotia	99%	100%	100%	96%	96%	96%
Canada, Ontario	88%	90%	100%	97%	85%	87%
Canada, Quebec	96%	96%	100%	84%	81%	81%
Chinese Taipei	98%	100%	100%	99%	97%	99%
Denmark	89%	99%	100%	97%	86%	96%
England	86%	99%	100%	93%	80%	92%
France	94%	97%	100%	98%	92%	95%
Georgia	94%	100%	100%	98%	93%	98%
Germany	97%	99%	100%	94%	90%	92%
Hong Kong SAR	91%	100%	100%	97%	89%	97%
Hungary	99%	100%	100%	97%	96%	97%
Iceland	99%	99%	100%	91%	90%	90%
Indonesia	99%	100%	100%	98%	97%	98%
Iran, Islamic Rep. of	100%	100%	100%	99%	99%	99%
Israel	98%	100%	100%	93%	91%	93%
Italy	91%	100%	100%	97%	88%	97%
Kuwait	99%	99%	99%	89%	88%	88%
Latvia	97%	98%	100%	94%	91%	92%
Lithuania	99%	100%	100%	92%	90%	92%
Luxembourg	100%	100%	100%	99%	99%	99%
Macedonia, Rep. of	100%	100%	100%	96%	96%	96%
Moldova, Rep. of	98%	100%	100%	95%	93%	95%
Morocco	98%	99%	100%	95%	93%	94%
Netherlands	70%	93%	100%	97%	67%	90%
New Zealand	92%	99%	100%	96%	88%	95%
Norway	68%	82%	100%	87%	58%	71%
Poland	99%	100%	100%	95%	94%	95%
Qatar	100%	100%	100%	94%	94%	94%
Romania	99%	99%	100%	98%	97%	97%
Russian Federation	100%	100%	100%	97%	97%	97%
Scotland	69%	87%	100%	94%	65%	81%
Singapore	100%	100%	100%	95%	95%	95%
Slovak Republic	93%	98%	100%	96%	89%	94%
Slovenia	93%	97%	100%	96%	90%	93%
South Africa	94%	96%	100%	92%	86%	88%
Spain	99%	100%	100%	97%	95%	97%
Sweden	100%	100%	100%	96%	96%	96%
Trinidad and Tobago	99%	99%	100%	95%	94%	94%
United States	57%	86%	100%	96%	54%	82%

Exhibit 9.9 Trends in PIRLS Student Populations

Country	Years of Formal Schooling		Average Age		Human Development Index		Overall Exclusion Rate	
	2006	2001	2006	2001	2006[1]	2001[2]	2006	2001
Bulgaria	4	4	10.9	10.9	0.816	0.772	6.4%	2.7%
Canada, Ontario	4	4	9.8	9.9	0.950	0.936	8.3%	6.6%
Canada, Quebec	4	4	10.1	10.2	0.950	0.936	3.6%	3.3%
England	5	5	10.3	10.2	0.940	0.923	2.4%	5.7%
France	4	4	10.0	10.1	0.942	0.924	3.8%	5.3%
Germany	4	4	10.5	10.5	0.932	0.921	0.7%	1.8%
Hong Kong SAR	4	4	10.0	10.2	0.927	0.880	3.9%	2.8%
Hungary	4	4	10.7	10.7	0.869	0.829	3.7%	2.1%
Iceland	4	4	9.8	9.7	0.960	0.932	3.8%	3.1%
Iran	4	4	10.2	10.4	0.746	0.714	3.8%	0.5%
Israel	4	4	10.1	10.0	0.927	0.893	22.5%	22.4%
Italy	4	4	9.7	9.8	0.940	0.909	5.3%	2.9%
Kuwait	4	4	9.8	9.9	0.871	0.818	0.3%	0.0%
Latvia	4	4	11.0	11.0	0.845	0.791	4.7%	4.6%
Lithuania	4	4	10.7	10.9	0.857	0.803	5.1%	3.8%
Macedonia	4	4	10.6	10.7	0.796	0.766	4.9%	4.2%
Moldova	4	4	10.9	10.8	0.694	0.699	0.6%	0.5%
Morocco	4	4	10.8	11.2	0.640	0.596	1.1%	1.0%
Netherlands	4	4	10.3	10.3	0.947	0.931	3.6%	3.7%
New Zealand	5	5	10.0	10.1	0.936	0.913	5.3%	3.2%
Norway	4	4	9.8	10.0	0.965	0.939	3.8%	2.8%
Romania	4	4	10.9	11.1	0.805	0.772	2.4%	4.5%
Russian Federation	4	3 or 4	10.8	10.3	0.797	0.775	7.7%	6.6%
Scotland	5	5	9.9	9.8	0.940	0.923	2.3%	4.7%
Singapore	4	4	10.4	10.1	0.916	0.876	0.9%	1.4%
Slovak Republic	4	4	10.4	10.3	0.856	0.831	3.6%	2.0%
Slovenia	3 or 4	3	9.9	9.8	0.910	0.874	0.8%	0.3%
Sweden	4	4	10.9	10.8	0.951	0.936	3.9%	5.0%
United States	4	4	10.1	10.2	0.948	0.934	5.9%	5.3%

1 Taken from the United Nations Development Programme's *Human Development Report 2006*, p. 283-286

2 Taken from the United Nations Development Programme's *Human Development Report 2001*, p. 141-144

References

TIMSS & PIRLS International Study Center. (2004). *PIRLS 2006 school sampling manual*. Chestnut Hill, MA: Boston College.

IEA. (2005). WinW3S: Within-school sampling software for Windows [Computer software and manual]. Hamburg: IEA Data Processing and Research Center.

Chapter 10

Item Analysis and Review

Michael O. Martin, Ann M. Kennedy, and Kathleen L. Trong

10.1 Overview

An important stage in creating the PIRLS 2006 achievement scale was an extensive review of the item statistics prior to item response theory (IRT) scaling. This review was conducted by the TIMSS & PIRLS International Study Center and involved evaluating the psychometric characteristics of each item within and across the participating countries. The purpose of this review was to ensure the quality of PIRLS achievement data by screening items for unusual item characteristics that could be attributed to an error, identifying the source, and rectifying the problem. For example, an item with low discrimination in one country atypical of the item's discrimination power in general may indicate a translation or printing problem. Also, for the trend items, item statistics were compared between 2001 and 2006.

In the few cases where country-level problems were identified, the TIMSS & PIRLS International Study Center consulted translation verification materials, checked printed booklets, and contacted National Research Coordinators (NRCs) to determine the source of the problem. When necessary, the item was removed from the international database for that country. This chapter describes the review process and the basic item statistics that were employed, using examples from the assessment.

10.2 Statistics for Item Analysis

As a first step, the TIMSS & PIRLS International Study Center created data almanacs containing the basic statistics for each achievement item. Exhibits 10.1 and 10.2 show examples of these statistics for a multiple-choice and constructed-response item, respectively. As these exhibits show, statistics were computed for each country individually, as well as on average internationally. The five Canadian provinces, listed below the international average row, were not included in the calculation of the international average.

For each item, almanacs include the number of students who were administered the item, item difficulty, item discrimination, and the percentage of boys and girls who responded correctly. For multiple-choice items, the percentage of students who chose each option and the percentage of students who omitted or did not reach the item was computed. In addition, the point-biserial correlation between each option and the total score was calculated. For constructed-response items, the percentage of students at each score level, the difficulty and discrimination for each score level (items could have up to 3 points), the number of responses that were double-scored, and the reliability between the two scorers were calculated. More detailed descriptions of these statistics are provided below.

- N: The number of students who were administered the item. If a student did not reach the item, it was considered not administered during item analysis.[1]
- Diff: Item difficulty, calculated as the percentage of students providing a correct response to the item. For constructed-response items worth more than one point, this is the students' average score as a percentage of the maximum score points for the item. Items that were not reached by the students were treated as not administered when computing this statistic.
- Disc: Item discrimination, calculated as the correlation between a correct response to the item and the total score on all items in the test booklet.[2] For constructed-response items worth more than one point, the correlation between the number of score points and total score was used. Items exhibiting good measurement properties should have a moderately positive correlation.

1 For the purpose of item analysis and item parameter estimation for scaling, items not reached by the student were considered not administered. However, these items were treated as incorrect when estimating student proficiency.

2 For the purpose of computing the discrimination index, the total score was the percentage of items a student answered correctly.

Exhibit 10.1 International Item Statistics for a Multiple-choice Item

Progress in International Reading Literacy Study - PIRLS 2006 Assessment Results
International Item Statistics (Unweighted) - 4th Grade
For Internal Review Only: DO NOT CITE OR CIRCULATE

Acquire and Use Information: Focus on and Retrieve Explicitly Stated Information and Ideas (Searching for Food)
Label: How do other ants from nest find food too (R021S04M - S04)
Item Type = MC Key = C

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	Avg. Score Girls	Avg. Score Boys	Flags
Austria	1008	69.3	0.51	13.1	5.0	69.3	10.1	0.0	2.5	0.0	-0.22	-0.25	0.51	-0.25	0.00	-0.18	-0.40	65.8	72.6	____F_B
Belgium (Flemish)	859	65.0	0.42	19.7	2.1	65.0	12.3	0.0	0.9	0.0	-0.27	-0.10	0.42	-0.23	0.00	-0.05	-0.31	63.0	67.0	__H_F__
Belgium (French)	907	60.5	0.49	17.4	6.7	60.5	13.6	0.0	1.8	0.0	-0.25	-0.22	0.49	-0.20	0.00	-0.15	-0.70	57.9	63.1	____F__
Bulgaria	756	69.7	0.54	17.2	4.5	69.7	8.3	0.0	0.3	0.0	-0.29	-0.28	0.54	-0.29	0.00	-0.05	-0.44	68.9	70.5	____F__
Chinese Taipei	912	82.1	0.49	10.0	3.4	82.1	3.7	0.0	0.8	0.0	-0.34	-0.27	0.49	-0.13	0.00	-0.12	-1.32	81.1	83.1	__E_F__
Denmark	770	78.7	0.46	10.4	3.0	78.7	7.3	0.0	0.6	0.1	-0.26	-0.27	0.46	-0.24	0.00	-0.07	-0.88	80.7	76.5	____F__
England	808	69.4	0.53	15.6	4.1	69.4	9.8	0.0	1.1	0.1	-0.22	-0.29	0.53	-0.30	0.00	-0.15	-0.58	69.9	69.0	____F__
France	875	80.5	0.42	8.9	3.4	80.5	6.1	0.0	1.1	0.0	-0.20	-0.25	0.42	-0.22	0.00	-0.10	-1.50	83.4	77.6	__E_F_G
Georgia	845	43.1	0.53	21.7	17.0	43.1	16.6	0.0	1.7	0.1	-0.17	-0.26	0.53	-0.21	0.00	-0.12	-0.32	39.8	46.4	__H____
Germany	1572	78.1	0.51	8.5	4.2	78.1	8.0	0.0	1.3	0.1	-0.28	-0.25	0.51	-0.25	0.00	-0.14	-0.86	78.1	78.0	__E_F__
Hong Kong, SAR	947	78.8	0.44	14.7	2.4	78.8	4.0	0.0	0.1	0.0	-0.28	-0.31	0.44	-0.16	0.00	-0.05	-0.99	79.5	78.1	__E_F__
Hungary	791	67.6	0.54	11.9	7.6	67.6	11.3	0.0	1.6	0.0	-0.31	-0.20	0.54	-0.25	0.00	-0.15	-0.40	62.2	72.9	____F_B
Iceland	730	65.6	0.49	13.4	5.5	65.6	14.4	0.0	1.1	0.1	-0.25	-0.25	0.49	-0.24	0.00	-0.05	-0.73	64.9	66.3	____F__
Indonesia	915	38.3	0.41	21.6	14.4	38.3	24.3	0.0	1.4	0.1	-0.09	-0.22	0.41	-0.17	0.00	-0.09	-0.51	40.0	36.4	_______
Iran, Islamic Rep. of	1070	52.4	0.50	22.2	12.0	52.4	11.4	0.0	2.0	0.3	-0.17	-0.29	0.50	-0.21	0.00	-0.13	-1.01	51.7	53.0	__E____
Israel	676	65.7	0.59	14.6	10.1	65.7	8.4	0.0	1.2	0.3	-0.29	-0.30	0.59	-0.24	0.00	-0.17	-1.07	61.8	69.5	__E_F_B
Italy	720	67.9	0.49	14.6	6.5	67.9	10.6	0.0	0.4	0.0	-0.31	-0.26	0.49	-0.17	0.00	-0.06	-0.48	67.9	68.0	____F__
Kuwait	669	28.6	0.43	14.8	27.8	28.6	20.9	0.0	7.9	1.5	-0.08	-0.11	0.43	-0.15	0.00	-0.15	-0.63	32.8	23.1	_____G
Latvia	817	66.3	0.45	17.9	4.4	66.3	10.6	0.0	0.7	0.1	-0.22	-0.24	0.45	-0.22	0.00	-0.08	-0.25	67.5	65.2	__H_F__
Lithuania	946	58.8	0.48	20.1	5.5	58.8	13.6	0.0	2.0	0.0	-0.20	-0.19	0.48	-0.29	0.00	-0.10	-0.12	56.8	60.7	__H_F__
Luxembourg	991	71.2	0.52	13.9	4.5	71.2	9.3	0.0	1.0	0.0	-0.29	-0.26	0.52	-0.25	0.00	-0.09	-0.40	68.2	74.4	____F_B
Macedonia, Rep. of	787	42.7	0.51	27.4	13.6	42.7	12.3	0.0	3.9	0.1	-0.12	-0.30	0.51	-0.20	0.00	-0.16	-0.50	43.6	41.8	_______
Moldova, Rep. of	776	56.2	0.55	19.6	11.3	56.2	12.4	0.0	0.5	0.0	-0.27	-0.29	0.55	-0.22	0.00	-0.03	-0.40	55.4	57.0	_______
Morocco	613	32.6	0.43	17.0	27.4	32.6	16.5	0.0	6.5	0.5	-0.09	-0.17	0.43	-0.13	0.00	-0.16	-1.12	32.7	32.5	__E____
Netherlands	795	70.8	0.45	16.6	2.8	70.8	9.7	0.0	0.1	0.0	-0.31	-0.17	0.45	-0.20	0.00	-0.01	-0.56	69.9	71.8	____F__
New Zealand	1204	63.5	0.55	18.7	5.1	63.5	11.3	0.0	1.4	0.2	-0.25	-0.33	0.55	-0.24	0.00	-0.14	-0.40	65.0	62.1	____F__
Norway	744	58.1	0.45	21.2	3.0	58.1	15.6	0.0	2.2	0.3	-0.23	-0.21	0.45	-0.18	0.00	-0.18	-0.41	56.1	60.0	____F__
Poland	943	64.6	0.50	11.6	6.5	64.6	15.8	0.0	1.6	0.1	-0.22	-0.25	0.50	-0.23	0.00	-0.15	-0.68	63.6	65.5	____F__
Qatar	1287	31.9	0.50	17.1	25.2	31.9	24.0	0.0	1.9	0.4	-0.15	-0.22	0.50	-0.17	0.00	-0.04	-0.89	34.3	29.4	__E____
Romania	831	53.3	0.50	28.3	7.9	53.3	9.3	0.0	1.2	0.4	-0.24	-0.28	0.50	-0.18	0.00	-0.13	-0.42	51.6	55.0	____F__
Russian Federation	915	73.7	0.47	12.6	2.8	73.7	10.6	0.0	0.3	0.0	-0.23	-0.18	0.47	-0.30	0.00	-0.11	-0.36	72.7	74.7	____F__
Scotland	752	64.0	0.52	17.3	4.3	64.0	13.4	0.0	1.1	0.0	-0.29	-0.21	0.52	-0.25	0.00	-0.09	-0.49	67.1	60.8	____F__
Singapore	1260	73.7	0.57	17.1	3.5	73.7	5.4	0.0	0.3	0.1	-0.36	-0.27	0.57	-0.27	0.00	-0.05	-0.75	74.8	72.6	____F__
Slovak Republic	1073	68.2	0.52	12.5	8.5	68.2	10.3	0.0	0.6	0.1	-0.21	-0.25	0.52	-0.31	0.00	-0.08	-0.58	69.7	66.8	____F__
Slovenia	1044	59.4	0.43	24.0	5.3	59.4	10.2	0.0	1.1	0.1	-0.23	-0.20	0.43	-0.20	0.00	-0.06	-0.22	58.7	60.0	__H_F__
South Africa	2771	24.5	0.47	25.8	22.3	24.5	16.0	0.0	11.4	0.5	-0.07	-0.16	0.47	-0.07	0.00	-0.24	-0.63	24.2	24.9	_C_____
Spain	793	51.2	0.49	26.4	7.3	51.2	14.0	0.0	1.1	0.0	-0.25	-0.21	0.49	-0.21	0.00	-0.08	-0.05	46.6	55.2	__H_F_B
Sweden	854	62.3	0.43	19.3	4.7	62.3	11.7	0.0	2.0	0.0	-0.21	-0.17	0.43	-0.18	0.00	-0.22	0.01	58.8	65.8	__H_F_B
Trinidad and Tobago	780	50.1	0.55	25.5	8.3	50.1	14.5	0.0	1.5	0.3	-0.28	-0.25	0.55	-0.20	0.00	-0.10	-0.63	49.2	51.0	____F__
United States	1041	60.7	0.46	26.5	4.2	60.7	8.5	0.0	0.1	0.1	-0.29	-0.26	0.46	-0.16	0.00	-0.01	-0.20	60.4	61.0	__H_F__
International Avg.	946	60.5	0.49	17.7	8.2	60.5	11.9	0.0	1.8	0.1	-0.23	-0.24	0.49	-0.21	0.00	-0.11	-0.58	59.9	61.0	____F_B
Canada, Alberta	855	71.1	0.41	17.8	2.5	71.1	8.2	0.0	0.5	0.0	-0.26	-0.11	0.41	-0.23	0.00	-0.07	-0.40	70.1	72.1	____F__
Canada, British Colum	809	69.0	0.44	16.9	2.6	69.0	10.6	0.0	0.9	0.0	-0.25	-0.20	0.44	-0.23	0.00	-0.10	-0.31	69.1	68.9	____F__
Canada, Nova Scotia	863	63.0	0.48	19.7	2.8	63.0	13.3	0.0	1.2	0.0	-0.25	-0.26	0.48	-0.22	0.00	-0.12	-0.33	61.4	64.6	____F__
Canada, Ontario	781	65.7	0.44	19.3	4.1	65.7	10.4	0.0	0.5	0.0	-0.23	-0.24	0.44	-0.21	0.00	-0.06	-0.38	62.5	68.8	____F__
Canada, Quebec	762	62.9	0.48	19.4	4.2	62.9	12.5	0.0	1.0	0.0	-0.23	-0.22	0.48	-0.24	0.00	-0.12	-0.42	60.2	65.5	____F__

Keys: Diff= Percent Correct Score; Disc= Item Discrimination; Pct_A...D= Percent Choosing Each Option; Pct_In, OM, NR= Percent Invalid, Omitted, Not Reached; PB_A...D= Point Biserial for Each Option; PB_OM= Point Biserial for Omitted; RDIFF= Rasch Difficulty

Flags: A= Ability not ordered/Attractive distractor; B= Boys outperform girls; C= Difficulty less than chance; D= Negative/low discrimination; E= Easier than average; F= Distractor chosen by less than 10%; G= Girls outperform boys; H= Harder than average; V= Difficulty greater than 95%

Exhibit 10.2 International Item Statistics for a Constructed-response Item

Progress in International Reading Literacy Study - PIRLS 2006 Assessment Results 7
International Item Statistics (Unweighted) - 4th Grade
For Internal Review Only: DO NOT CITE OR CIRCULATE

Acquire and Use Information: Make Straightforward Inferences (Antarctica)
Label: Ways penguins keep warm (R011A07C - A07)
Item Type = CR Key = X

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	Reliability Cases	Reliability Score	Avg. Score Girls	Avg. Score Boys	Flags
Austria	1018	69.1	0.72	5.0	19.6	28.7	43.4	3.2	0.0	-0.38	-0.36	0.00	0.57	-0.31	0.47	238	95.0	73.3	64.9	__H_F_G
Belgium (Flemish)	887	77.4	0.69	5.4	10.7	28.5	54.8	0.6	0.0	-0.46	-0.33	-0.13	0.56	-0.15	0.49	198	95.5	78.3	76.5	__H_F__
Belgium (French)	905	59.3	0.76	13.3	17.6	27.5	35.1	6.5	0.1	-0.47	-0.22	0.17	0.55	-0.38	0.45	209	98.1	61.4	57.3	__H___
Bulgaria	772	79.7	0.81	4.7	9.3	18.3	64.4	3.4	0.3	-0.48	-0.39	-0.11	0.68	-0.37	0.21	174	95.4	80.7	78.7	___F__
Chinese Taipei	910	85.5	0.71	2.2	11.0	10.3	74.9	1.5	0.3	-0.31	-0.42	-0.13	0.62	-0.35	-0.33	217	97.7	87.2	83.9	__E_F__
Denmark	801	63.7	0.73	12.2	19.4	28.2	38.5	1.7	0.4	-0.51	-0.24	0.08	0.55	-0.25	0.93	185	95.7	69.2	57.9	__H___G
England	803	69.3	0.80	12.6	13.0	24.8	48.4	1.2	0.1	-0.61	-0.29	0.03	0.63	-0.23	0.37	196	98.5	75.7	63.5	_____G
France	884	70.8	0.74	10.2	13.2	12.6	58.0	6.0	0.1	-0.43	-0.24	-0.03	0.65	-0.41	0.39	240	95.4	73.8	68.2	_____
Georgia	872	58.0	0.77	10.3	17.9	31.8	30.8	9.2	0.3	-0.44	-0.21	0.17	0.56	-0.41	0.32	205	95.1	62.8	53.9	_____G
Germany	1574	70.0	0.70	5.0	17.2	37.2	39.5	1.2	0.1	-0.37	-0.40	0.03	0.52	-0.27	0.63	506	90.9	71.9	68.2	__H_F_G
Hong Kong, SAR	940	90.0	0.65	2.0	4.6	8.8	82.6	2.0	0.0	-0.36	-0.26	-0.15	0.55	-0.41	-0.02	209	96.2	90.8	89.2	__E_F__
Hungary	824	74.6	0.72	5.2	13.7	25.5	53.0	2.5	0.1	-0.36	-0.33	-0.10	0.59	-0.36	0.35	225	98.2	78.8	70.1	___F_G
Iceland	735	65.1	0.80	13.9	17.6	16.1	48.6	3.9	0.9	-0.53	-0.27	0.07	0.66	-0.32	0.11	200	97.0	68.7	61.2	__E__G
Indonesia	930	38.2	0.77	40.6	15.2	19.8	20.0	4.4	1.0	-0.63	0.03	0.29	0.56	-0.17	0.15	234	94.9	40.4	36.2	__E____
Iran, Islamic Rep. of	1059	43.0	0.78	26.4	20.6	19.1	23.4	10.5	1.0	-0.49	-0.05	0.26	0.58	-0.34	0.18	231	96.5	46.0	40.5	_____G
Israel	781	68.7	0.83	12.4	11.3	19.8	51.7	4.7	1.3	-0.58	-0.21	0.07	0.65	-0.39	0.07	259	84.2	70.7	66.7	__E____
Italy	709	73.3	0.75	7.6	15.8	18.8	55.6	2.3	0.0	-0.54	-0.26	-0.05	0.61	-0.29	0.60	187	94.7	75.1	71.6	__H_F__
Kuwait	708	25.0	0.79	28.5	6.2	10.3	16.1	38.8	6.3	-0.26	0.07	0.28	0.67	-0.42	0.00	165	84.2	29.5	19.8	__E_F_G
Latvia	828	84.0	0.69	3.7	8.9	17.0	69.7	0.6	0.0	-0.44	-0.38	-0.16	0.58	-0.21	-0.01	200	89.5	87.1	81.0	__E_F_G
Lithuania	935	74.8	0.68	5.0	15.1	28.3	50.9	0.6	0.0	-0.40	-0.40	-0.08	0.55	-0.15	0.29	199	94.5	79.3	70.9	___F_G
Luxembourg	1010	74.3	0.73	3.5	13.9	33.2	47.5	2.0	0.0	-0.35	-0.42	-0.08	0.58	-0.28	0.82	189	96.3	75.9	72.7	__H_F__
Macedonia, Rep. of	781	58.7	0.82	17.0	14.5	22.0	39.2	7.3	1.6	-0.51	-0.19	0.12	0.66	-0.36	0.09	177	91.0	63.2	54.4	__E__G
Moldova, Rep. of	800	62.9	0.76	11.0	18.4	32.6	35.0	3.0	0.5	-0.51	-0.19	0.04	0.59	-0.32	0.45	191	99.5	68.5	57.6	__H__G
Morocco	604	24.5	0.78	49.5	13.4	8.4	14.4	14.2	3.0	-0.43	0.05	0.26	0.68	-0.27	0.03	150	84.7	24.7	24.3	__E_F__
Netherlands	822	75.3	0.67	4.7	11.2	37.2	46.7	0.1	0.0	-0.39	-0.38	-0.13	0.54	-0.11	0.52	178	99.4	78.4	72.2	__H_F_G
New Zealand	1232	68.8	0.81	10.6	14.9	21.8	49.3	3.4	0.5	-0.51	-0.27	0.04	0.63	-0.39	0.47	237	94.1	72.7	65.0	__H__G
Norway	793	56.1	0.74	13.9	24.5	28.0	29.3	4.4	0.3	-0.45	-0.23	0.17	0.55	-0.33	0.52	207	86.5	60.4	51.9	__H__G
Poland	961	73.6	0.77	10.5	10.1	15.0	60.2	4.2	0.0	-0.52	-0.25	-0.05	0.66	-0.36	0.21	203	98.0	77.7	69.0	__H__G
Qatar	1303	33.1	0.79	41.1	12.9	13.7	19.6	12.7	1.1	-0.57	0.08	0.27	0.63	-0.26	-0.07	192	95.3	36.2	29.9	__E___G
Romania	849	64.2	0.80	14.0	15.3	22.6	44.1	4.0	0.6	-0.52	-0.21	0.06	0.63	-0.35	0.46	208	97.1	68.4	60.3	__H__G
Russian Federation	943	91.4	0.63	1.5	4.7	10.8	82.6	0.4	0.1	-0.37	-0.39	-0.20	0.53	-0.17	-0.46	211	98.6	93.3	89.2	__E_F_G
Scotland	752	72.1	0.77	10.0	12.0	23.9	52.1	2.0	0.0	-0.55	-0.28	-0.02	0.61	-0.28	0.36	176	98.3	77.6	66.5	___F_G
Singapore	1276	80.7	0.77	7.0	7.5	18.7	65.8	1.0	0.2	-0.61	-0.24	-0.14	0.63	-0.19	0.33	220	98.2	85.0	76.3	___F_G
Slovak Republic	1070	75.5	0.75	6.8	12.7	22.7	56.1	1.7	0.0	-0.50	-0.33	-0.07	0.61	-0.26	0.12	173	97.7	75.2	75.7	__E_F__
Slovenia	1063	71.1	0.75	8.9	14.0	24.0	50.4	2.6	0.1	-0.48	-0.34	0.00	0.60	-0.27	0.26	247	96.8	75.9	66.2	___F_G
South Africa	2678	19.2	0.81	56.8	10.6	7.9	10.4	14.3	5.5	-0.46	0.13	0.30	0.67	-0.20	0.15	209	77.0	21.6	16.7	__E_FRG
Spain	816	64.4	0.74	11.9	15.7	32.8	37.3	2.3	0.5	-0.51	-0.24	0.05	0.56	-0.24	0.40	208	80.8	62.1	66.8	_____
Sweden	882	76.8	0.73	5.3	11.2	28.8	53.9	0.8	0.2	-0.49	-0.34	-0.08	0.56	-0.17	0.38	236	91.9	78.3	75.3	___F__
Trinidad and Tobago	778	48.4	0.84	26.5	16.8	18.1	30.7	7.8	1.1	-0.59	-0.07	0.22	0.66	-0.34	0.37	201	94.5	53.2	43.7	_____G
United States	1033	71.6	0.79	9.5	14.8	24.0	50.6	1.1	0.1	-0.54	-0.35	-0.03	0.64	-0.21	0.38	238	92.9	75.4	67.6	___F_G
International Avg.	958	65.1	0.75	13.7	13.7	21.9	45.9	4.9	0.7	-0.47	-0.24	0.03	0.60	-0.29	0.29	213	93.9	68.1	62.0	_____G
Canada, Alberta	848	80.6	0.71	2.9	11.3	23.1	61.4	1.2	0.0	-0.32	-0.42	-0.17	0.60	-0.32	0.19	65	84.6	81.0	80.3	____F__
Canada, British Colum	831	80.0	0.74	4.2	10.5	22.1	61.7	1.4	0.2	-0.42	-0.41	-0.11	0.60	-0.26	0.31	60	93.3	81.5	78.6	____F__
Canada, Nova Scotia	883	74.8	0.77	7.8	12.3	22.0	56.1	1.8	0.3	-0.51	-0.30	-0.06	0.62	-0.31	0.33	98	93.9	79.7	69.7	____F_G
Canada, Ontario	785	73.2	0.79	8.9	13.5	22.9	53.4	1.3	0.4	-0.55	-0.34	-0.04	0.64	-0.20	0.41	155	94.2	74.8	71.6	____F__
Canada, Quebec	732	73.6	0.71	5.6	16.7	23.4	52.5	1.9	0.3	-0.39	-0.37	-0.01	0.56	-0.33	0.33	206	93.7	76.2	70.9	____F_G

Keys: Diff= Percent Correct Score; Disc= Item Discrimination; Pct_0...3= Percent Obtaining Score Level; Pct_OM, NR= Percent Omitted, Not Reached;
PB_0...2= Point Biserial for Score Level; PB_OM= Point Biserial for Omitted; RDIFF= Rasch Difficulty;
Reliability (Cases) = Responses Double Scored; Reliability (Score) = Percent Agreement on Score
Flags: A= Ability not ordered/Attractive distractor; B= Boys outperform girls; C= Difficulty less than chance; D= Negative/low discrimination; E= Easier than average;
F= Distractor chosen by less than 10%; G= Girls outperform boys; H= Harder than average; R= Scoring reliability < 80%; V= Difficulty greater than 95%

- Pct_A, Pct_B, Pct_C, Pct_D: Used for multiple-choice items only, each column indicates the percentage of students choosing the particular response option (A, B, C, or D). Students who did not reach the item were excluded from the denominator for these calculations.
- Pct_0, Pct_1, Pct_2, Pct_3: Used for constructed-response items only, each column indicates the percentage of students earning the particular number of score points (0, 1, 2, or 3) for that item. Students who did not reach the item were excluded from the denominator for these calculations.
- Pct_In: Used for multiple-choice items only, this column indicates the percentage of students who provided an invalid response to the item. A typical invalid response was a student selecting multiple response options for an item.
- Pct_OM: This column indicates the percentage of students who reached the item but did not provide a response. Students who did not reach the item were excluded from the denominator when calculating this statistic.
- Pct_NR: This column indicates the percentage of students who did not reach the item. An item was considered not reached if the student did not respond to any subsequent items in the test booklet, and the previous item was omitted.
- PB_A, PB_B, PB_C, PB_D: Used for multiple-choice items only, each column indicates the point-biserial correlation between the particular option (A, B, C, or D) and the total score. Items with good psychometric properties have near-zero or negative correlation coefficients for the incorrect options and a moderately positive correlation coefficient for the correct option.
- PB_0, PB_1, PB_2, PB_3: Used for constructed-response items only, each column indicates the correlation between the particular score level (0, 1, 2, or 3) and the total score. Items with good psychometric properties should increase with each score level.
- PB_In: Used for multiple-choice items only, this column indicates the correlation between an invalid response to the item (i.e., selecting multiple response options) and total score. For an item with good psychometric properties, this should be negative or near zero.

- PB_OM: This column indicates the correlation between a binary variable indicating an omitted response to the item and the total score. For an item with good psychometric properties, this 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 logits (with a positive logit indicating a difficult item) and was scaled so that the average Rasch item difficulty was zero within each country.
- Avg. Score Girls/Boys: These columns indicate the average difficulty for the item separately for boys and girls.
- Reliability Cases: To provide a measure of the scoring reliability of constructed-response items, those items in approximately one quarter of the test booklets were scored by two independent scorers in each country. This column indicates the number of responses that were double-scored for an item.
- Reliability Score: Used for constructed-response items only, this column indicates the percentage of exact agreement between two independent scorers.

As an aid during the review process, the almanacs also include a series of "flags" to highlight conditions that may warrant a closer look. While not all flags necessarily signify a faulty item, they draw attention to potential problems.

The following conditions are flagged:

- Difficulty levels for the item are significantly different for boys and girls;
- Item difficulty is less than chance (e.g., 25% for multiple-choice items);
- Item difficulty exceeds 95 percent;
- Item discrimination (i.e., the point-biserial for the correct option) is less than 0.2;
- Rasch difficulty estimate is below the average of all items;
- Rasch difficulty estimate is above the average of all items;
- For multiple choice items, a greater percentage of students chose the incorrect response than the percentage of students who chose the correct option, or the point-biserial correlation for one or more of the distracters exceeds zero;

- Less than 10 percent of students chose one or more of the distracters (for multiple-choice items) or earned one or more of the score levels (for constructed-response items);
- Scoring reliability is less than 80 percent; and
- Students with lower total scores are more likely to answer an item correctly than students with higher total scores.

In addition to item-level statistics, the TIMSS & PIRLS International Study Center also examined descriptive statistics for each test block as a whole to gain a sense of the overall performance of the items associated with each passage. These statistics included the number of students who were administered the block, the minimum, maximum, and average number of items students answered correctly, the standard deviation, the percent correct overall and, for boys and girls separately, the minimum, maximum, and average point-biserial correlation across the items, and the Cronbach's alpha reliability coefficient for the block. Each of these was calculated for individual countries and on average internationally.

Eleven countries and the five Canadian provinces administered the assessment in more than one language.[3] In these cases, an additional step was taken to examine item statistics for each language group. Due to the fact that some language groups make up a small proportion of the overall sample in a country, problems with a particular language within a country may not have been evident when the languages are combined, and may indicate a translation error for that language. The same process and statistics that were described above were used for this step.

10.3 Examining Item-by-Country Interactions

While it is reasonable to expect country performance to vary somewhat across items, countries with high average performance should generally perform well on each of the items, and countries with lower overall performance should do less well on individual items. When this pattern is not followed (i.e., a high-performing country does poorly on a particular item), this is called an item-by-country interaction. If this interaction is large, it could indicate a problem with the item that should be investigated and addressed.

To easily detect item-by-country interactions, the TIMSS & PIRLS International Study Center created plots that graphically display the Rasch difficulties for each item. Exhibit 10.3 displays the 95 percent confidence interval

3 See Chapter 5 for details about translation procedures.

for the national average Rasch difficulty estimate. The limits for the confidence intervals were computed as follows:

$$\text{Upper Limit} = \frac{e^{RDIFF_{ik} + SE_{RDIFF_{ik}} \text{x} Z_b}}{1 + e^{RDIFF_{ik} + SE_{RDIFF_{ik}} \text{x} Z_b}}$$

$$\text{Lower Limit} = \frac{e^{RDIFF_{ik} - SE_{RDIFF_{ik}} \text{x} Z_b}}{1 + e^{RDIFF_{ik} - SE_{RDIFF_{ik}} \text{x} Z_b}}$$

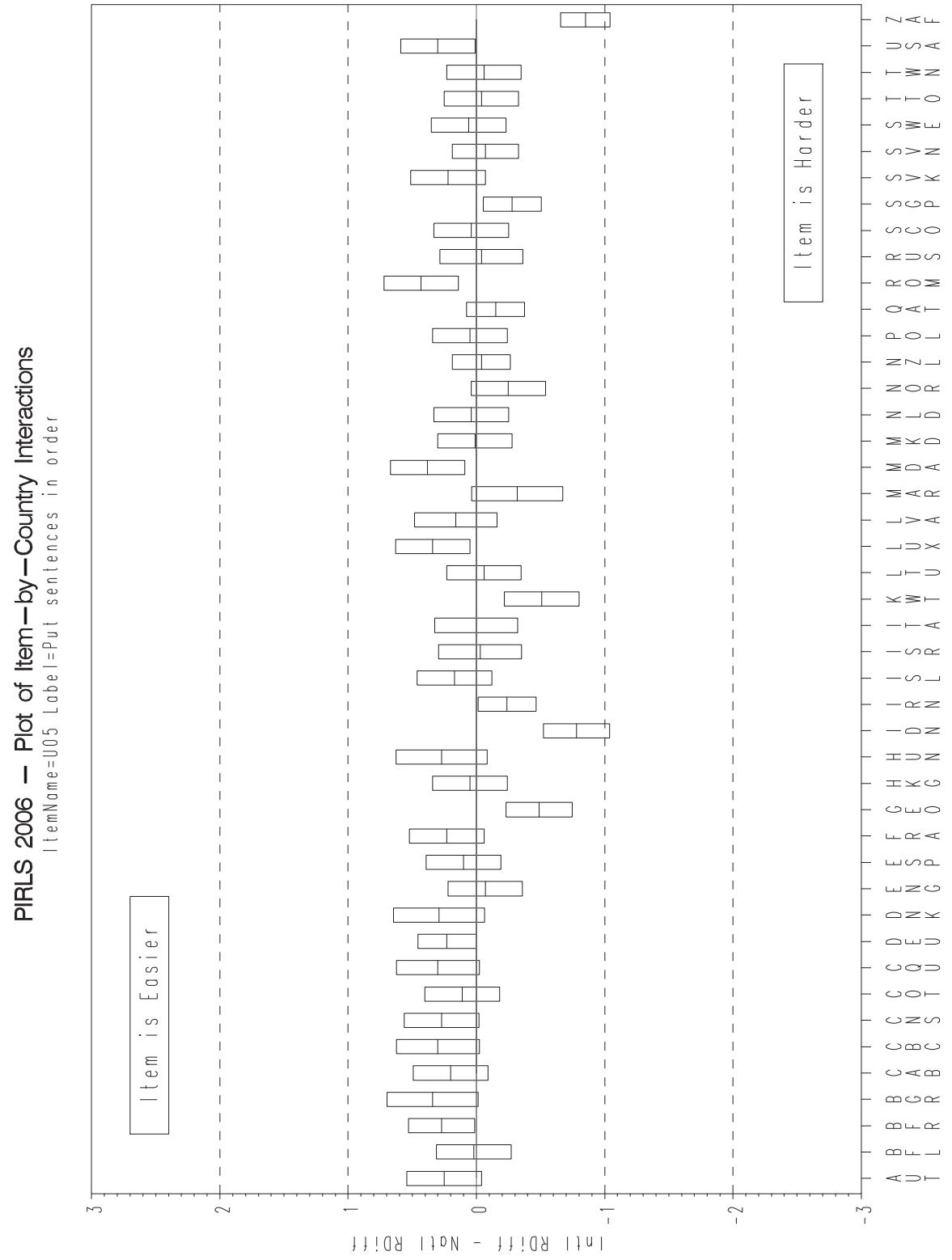
where $RDIFF_{ik}$ is the Rasch difficulty of item k within country i, $SE_{RDIFFik}$ is the standard error of the difficulty of item k in country i, and Z_b is the critical value from the Z distribution, corrected for multiple comparisons using the Bonferroni procedure. With the international average Rasch difficulty scaled to zero as a reference point, countries with a positive difference between the national and international Rasch difficulty found the item easier, and countries with a negative difference found the item more difficult.

10.4 Trend Item Analysis

Because an important part of the PIRLS 2006 assessment was the measuring trends across cycles, there was an additional stage of the review process to ensure that the trend items had similar characteristics in both cycles (i.e., an item that was relatively easy in 2001 should be relatively easy in 2006). The comparison between cycles was made in a number of ways. For each trend country, almanacs of item statistics displayed the percentage of students within each score category (or response option, for multiple-choice items) for each cycle, as well as the difficulty of the item and the percent correct by gender. While some changes were anticipated as countries' overall reading achievement may have improved or declined, items were noted if the percent correct changed by more than 15 percent for a particular country.

The TIMSS and PIRLS International Study Center used two different graphical displays to examine the differences between item difficulties in 2001 to 2006. The first of these, shown in Exhibit 10.4, displayed the difference in Rasch difficulty estimates (in logits) between the two assessment cycles. A positive difference indicates that the item was relatively easier in a country in 2006, and a negative difference indicates that an item was relatively more difficult. The second shows a country's performance on all trend items simultaneously.

Exhibit 10.3 Sample Plot of Item-by-Country Interactions for a PIRLS 2006 Item



Individually for each country, a scatterplot graphed the Rasch difficulty of each item in 2001 against the difficulty for that item in 2006. Where there are no differences between the difficulties in 2001 and 2006, the data points will align on or near the diagonal indicating a one-to-one correlation between cycles.

These graphs were used in conjunction with one another to detect items that performed differently in the two cycles. When such items were found, the source of the difference was investigated using booklets from both cycles, translation verifier's comments, national adaptation forms, and trend scoring reliability data.

10.5 Scoring Reliability for Constructed-response Items

Almost two thirds of the score points in the PIRLS 2006 assessment were from constructed-response items. In order to include these types of items in the assessment, it is essential that they are scored reliably within and across countries. In other words, a particular student response should receive the same score, regardless of the scorer. To ensure that this was the case, specific scoring guides were created for each constructed-response item that provided descriptions of each score level and sample student responses. Countries received extensive training in the application of each of these guides, using genuine student responses as examples and practice materials. Procedures for organizing and monitoring the scoring sessions were provided in the *PIRLS 2006 Survey Operations Procedures Unit 4* (TIMSS & PIRLS International Study Center, 2005). In addition to this training, countries were required to provide several types of scoring reliability data to document the consistency with which the scoring guides were applied. These are described in the following sections.

10.5.1 Within-country Scoring Reliability

This first type of scoring reliability data documented the extent to which items were scored consistently within each country. For each constructed-response item, 200 randomly selected student responses were independently marked by two scorers in the country. The percent agreement between these scorers was included in the item almanacs described above and examined as part of the review process. During this review, items were noted for further examination if agreement for a particular country fell below 70 percent. On average, the exact percent agreement across items was very high at 93 percent. All countries had an average percent of exact agreement above 81 percent. The average and range of these percentages is presented in Exhibit 10.6.

Exhibit 10.4 Sample Plot of Difference in Rasch Difficulties for a PIRLS 2006 Item

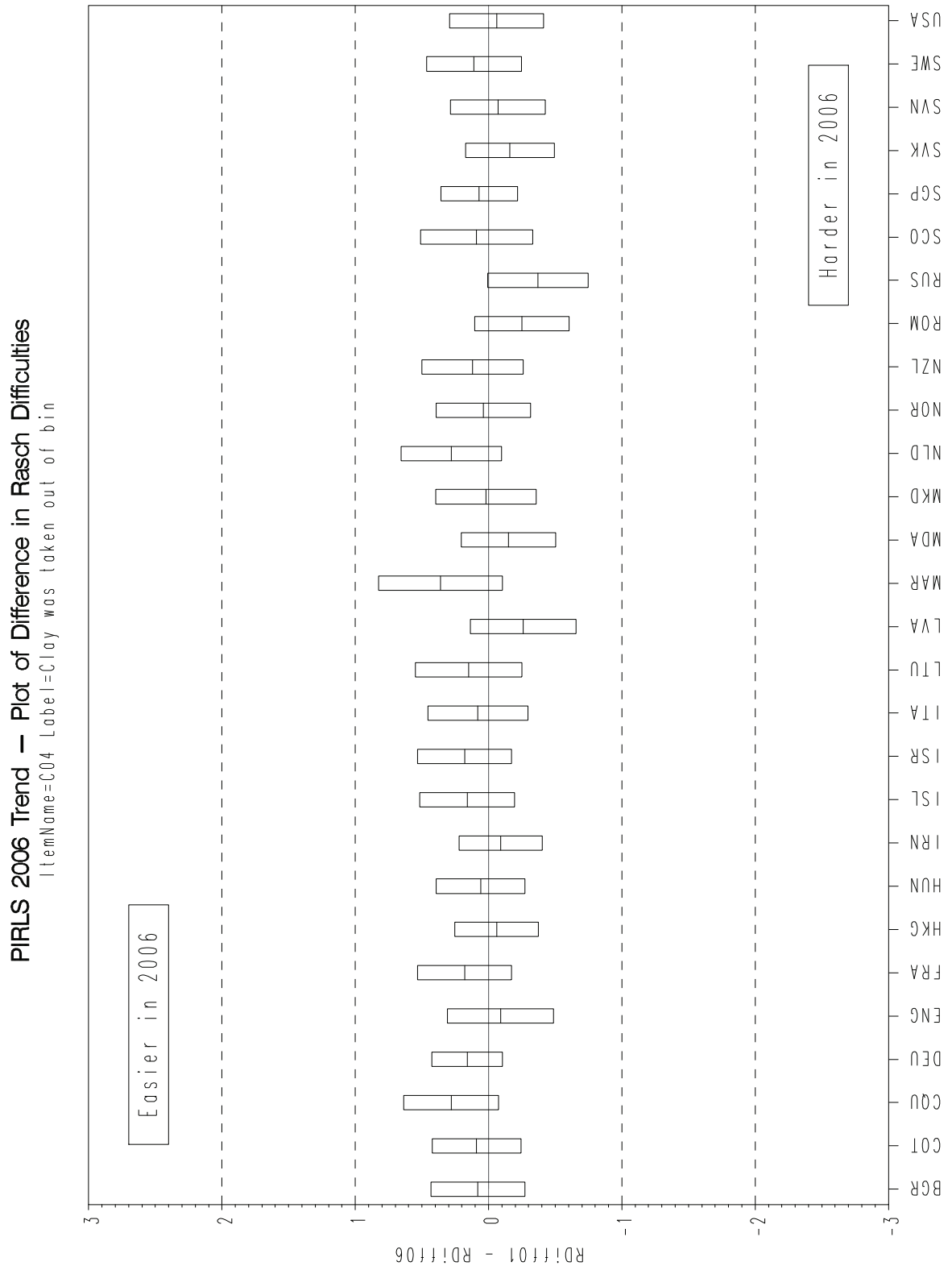


Exhibit 10.5 Sample Plot of Rasch Difficulties by Country for a PIRLS 2006 Item

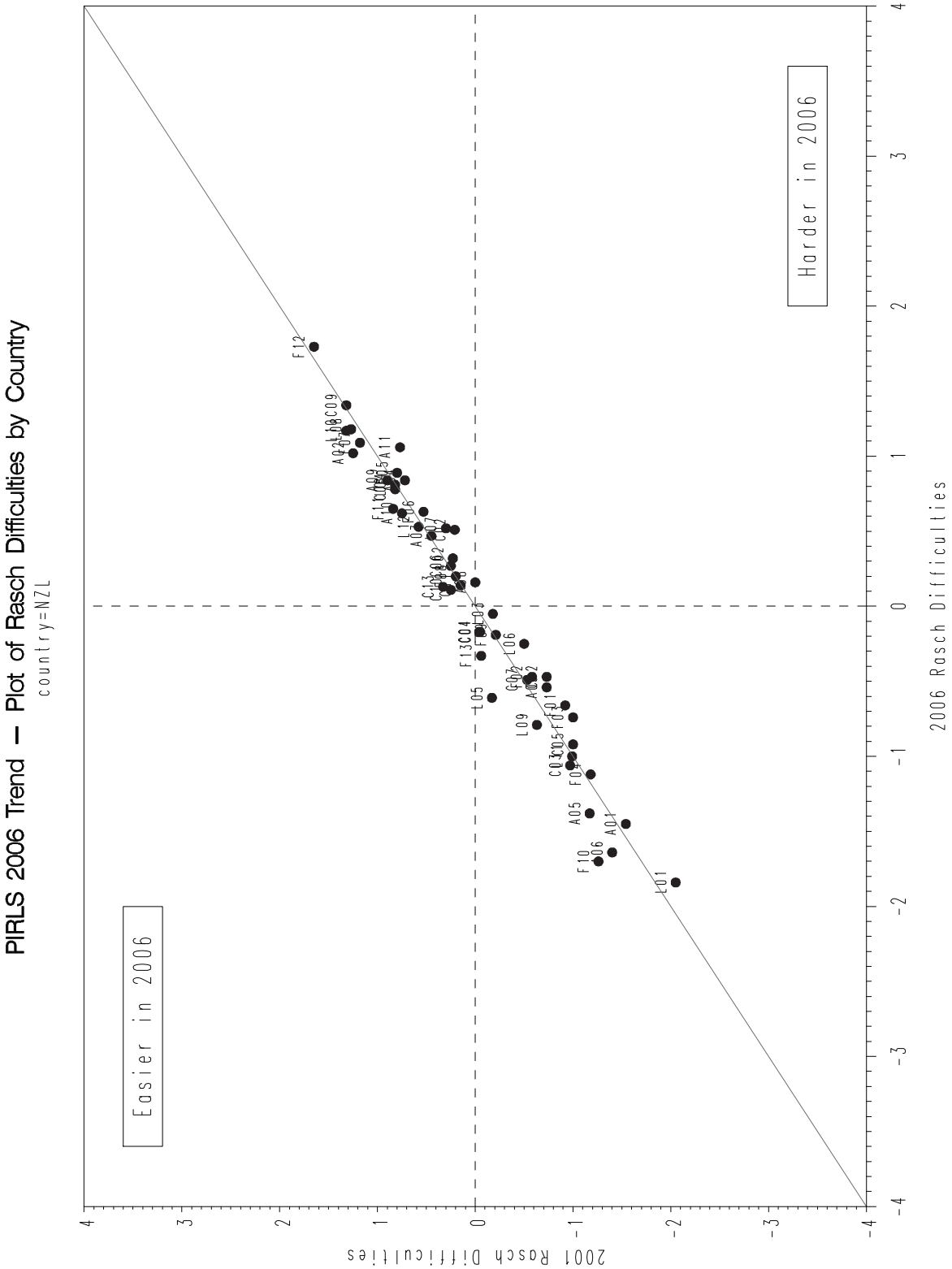


Exhibit 10.6 PIRLS 2006 Within-country Constructed-response Scoring Reliability Data

Countries	Average of Percent Exact Agreement Across Items	Range of Percent Exact Agreement	
		Minimum	Maximum
Austria	95	80	100
Belgium (Flemish)	90	73	99
Belgium (French)	97	90	100
Bulgaria	98	94	100
Canada, Alberta	91	67	100
Canada, British Columbia	92	70	100
Canada, Nova Scotia	93	84	100
Canada, Ontario	94	80	100
Canada, Quebec	95	87	100
Chinese Taipei	95	78	100
Denmark	97	90	100
England	98	93	100
France	89	69	100
Georgia	85	65	98
Germany	89	76	99
Hong Kong SAR	96	85	100
Hungary	98	89	100
Iceland	95	88	99
Indonesia	95	76	100
Iran, Islamic Rep. of	93	83	99
Israel	91	80	98
Italy	95	85	100
Kuwait	86	80	95
Latvia	90	78	100
Lithuania	97	91	100
Luxembourg	94	82	100
Macedonia, Rep. of	88	78	96
Moldova, Rep. of	99	97	100
Morocco	89	71	97
Netherlands	99	93	100
New Zealand	93	80	98
Norway	83	66	97
Poland	97	93	100
Qatar	97	93	99
Romania	99	96	100
Russian Federation	99	97	100
Scotland	97	89	100
Singapore	98	94	100
Slovak Republic	96	88	100
Slovenia	98	92	100
South Africa	82	63	92
Spain	81	61	96
Sweden	92	72	100
Trinidad and Tobago	93	71	100
United States	93	82	100
International Avg.	93	82	99

10.5.2 Cross-country Scoring Reliability

It also was important to document the consistency of scoring across countries. To accomplish this goal, the TIMSS & PIRLS International Study Center collected 200 student responses to each of the 23 constructed-response items from four assessment blocks, for a total of 4,600 student responses. Due to the wide range of languages used in PIRLS 2006 and the logistic issues this presents for cross-country scoring, responses were only collected from participants that administered the assessment in English (the Canadian province of Ontario, England, New Zealand, Scotland, Singapore, South Africa, and the United States). These responses were scanned by the IEA Data Processing and Research Center (DPC) and provided to each country in a software program that facilitated the scoring process.

Countries were asked to provide at least two scorers who were proficient in English to score this set of responses, following the main scoring activities. Each student response was scored by 62 scorers from across the countries, giving a total of 1,891 comparisons for each student response to each item, and 378,200 total comparisons across the set of 200 responses per item.[4] Exhibit 10.7 shows the percentage of exact agreement for each item. On average, there was 87 percent agreement, with variation across the individual items.

4 The number of comparisons varies across items because not all scorers scored all items.

Exhibit 10.7 PIRLS 2006 Cross-country Constructed-response Scoring Reliability

Purpose	Item Label	Total Valid Comparisons	Exact Percent Agreement
Literary Experience	Flowers F06C	377504	91%
	Flowers F07C	377957	80%
	Flowers F08C	375960	92%
	Flowers F09C	378078	93%
	Flowers F10C	376869	97%
	Flowers F12C	375684	63%
	Unbelievable Night U05C	377224	99%
	Unbelievable Night U06C	377385	93%
	Unbelievable Night U08C	378078	76%
	Unbelievable Night U10C	377453	96%
	Unbelievable Night U12C	377302	87%
Acquire and Use Information	Antartica A01C	378200	95%
	Antartica A03C	378139	98%
	Antartica A04C	377542	89%
	Antartica A07C	378139	88%
	Antartica A08C	377722	80%
	Antartica A09C	377370	83%
	Antartica A11C	377363	81%
	Day Hiking N02C	377897	91%
	Day Hiking N03C	378139	94%
	Day Hiking N08C	376927	92%
	Day Hiking N11C	377773	77%
	Day Hiking N12C	330146	76%
	Average Percent Agreement		87%

10.5.3 Trend Scoring Reliability

Extensive efforts were made to ensure that constructed-response items were scored consistently across testing cycles. In preparation for this, the IEA DPC scanned 200 student responses to each of the 26 trend constructed-response items from the PIRLS 2001 reliability booklet samples and provided the responses, along with the original scores from 2001, in a scoring software program.[5] As part of the scoring training activities, at least two scorers in each trend country were asked to score the set of student responses for each item from the four trend blocks, for a total of 5,200 student responses. After scoring half of these responses, scorers used the software to compare their scores to one

5 A number of participants were unable to complete the trend scoring reliability task because of software difficulties or because it was not possible to scan their 2001 students booklets.

another, as well as to the original score given in 2001. If agreement for any item fell below 85 percent, retraining was required and the responses for that item were rescored. Exhibit 10.8 shows the results of the trend scoring reliability task. Overall, the percent agreement across the trend constructed-response items was high—90 percent on average across countries.

Exhibit 10.8 PIRLS 2006 Trend Scoring Reliability (2001–2006) for the Constructed-response Items

Countries	Average Percent Exact Agreement Across Items
Canada, Ontario	–
Canada, Quebec	–
England	89
France	90
Germany	88
Hong Kong SAR	93
Hungary	91
Iceland	–
Iran, Islamic Rep. of	92
Israel	96
Italy	91
Latvia	84
Lithuania	92
Macedonia, Rep. of	81
Moldova, Rep. of	–
Morocco	–
Netherlands	93
New Zealand	90
Norway	90
Romania	–
Russian Federation	–
Scotland	88
Singapore	88
Slovak Republic	92
Slovenia	–
Sweden	89
United States	93
International Avg.	90

A dash (–) indicates data are not available.

10.6 Item Review Procedures

Using the range of techniques described in the previous sections, the TIMSS & PIRLS International Study Center reviewed the performance of each item within every participating country to ensure that the results were comparable internationally. In particular, items with the following problems were considered for possible deletion from the international database:

- An error was detected in the translation of the item that was not fixed before test administration;
- Data checking revealed a multiple-choice item with more or fewer options than the international version;
- The item analysis showed the item to have a negative biserial, or, for items with more than one score point, point biserials that did not increase with each score level;
- The item-by-country interaction results showed a very large interaction for a particular country;
- For constructed-response items, the within-country scoring reliability data showed an agreement of less than 70 percent; and
- For trend items, an item performed substantially differently in 2006 compared to 2001.

In cases where a potential problem was detected, test booklets (including PIRLS 2001 booklets, if necessary), and documentation from the translation verification and national adaptation process were reviewed. Additionally, the NRC was consulted for clarification or confirmation of translation, printing, or scoring problems. Of the 126 items used in the assessment, only one item was removed from the international database for all countries. In only three countries were one or two additional items problematic for international comparisons. The main cause of errors in items was translation or printing errors. A list of deleted and recoded items is provided in Appendix C.

References

TIMSS & PIRLS International Study Center. (2005). *Survey operations procedures unit 4: Scoring the PIRLS 2006 assessment*. Chestnut Hill, MA: Boston College.

Chapter 11

Scaling the PIRLS 2006 Reading Assessment Data

Pierre Foy, Joseph Galia, and Isaac Li

11.1 Overview

PIRLS 2006 had ambitious goals for broad coverage of the reading purposes and processes as described in its assessment framework[1] and for measuring trends across assessment cycles. To achieve these goals, the PIRLS 2006 assessment consisted of 10 reading passages and items arranged into 40-minute assessment blocks, four of which were retained from the 2001 assessment in order to serve as the foundation for measuring trends. PIRLS used a matrix-sampling design[2] to assign assessment blocks to student booklets—two blocks per student booklet—so that a comprehensive picture of the reading achievement of fourth-grade students in participating countries could be assembled from the booklets completed by individual students. PIRLS relied on Item Response Theory (IRT) scaling to combine the student responses and provide accurate estimates of reading achievement in the student population of each participating country, as well as measure trends in reading achievement among countries that also participated in the 2001 assessment. The PIRLS scaling methodology 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 conditioning and plausible values methodology used in scaling the PIRLS 2006 data, and then

1 The PIRLS 2006 assessment framework is described in Mullis, Kennedy, Martin, & Sainsbury (2006).

2 The PIRLS 2006 achievement test design is described in Chapter 2.

describes how this approach was applied to the PIRLS 2006 data and to the data from the previous PIRLS 2001 study in order to measure trends in achievement. The PIRLS scaling was carried out at the TIMSS & PIRLS International Study Center at Boston College, using software from Educational Testing Service.[3]

11.2 PIRLS 2006 Scaling Methodology[4]

The IRT 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 to measure trends in mathematics and science.

Three distinct IRT models, depending on item type and scoring procedure, were used in the analysis of the PIRLS 2006 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 student'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 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 response options.

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 student whose proficiency on a scale k is characterized by the unobservable variable θ_k will respond correctly to item i as:

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 and 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 multiple imputation 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).

$$\textbf{(1)} \quad P\left(x_i = 1 \mid \theta_k, a_i, b_i, c_i\right) = c_i + \frac{1 - c_i}{1 + \exp\left(-1.7 \cdot a_i(\theta_k - b_i)\right)} \equiv P_{i,1}\left(\theta_k\right)$$

where

x_i is the response to item *i*, 1 if correct and 0 if incorrect;

θ_k is the proficiency of a student on a scale *k* (note that a student 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 of item *i*, characterizing its difficulty;

c_i is the lower asymptote parameter of item *i*, reflecting the chances of students with very low proficiency selecting the correct answer.

The probability of an incorrect response to the item is defined as:

$$\textbf{(2)} \quad P_{i,0} = P\left(x_i = 0 \mid \theta_k, a_i, b_i, c_i\right) = 1 - P_{i,1}\left(\theta_k\right)$$

The two-parameter (2PL) model was used for the short constructed-response items that were scored as either correct or incorrect. The form of the 2PL model is the same as Equations (1) and (2) with the c_i parameter fixed at zero.

11.2.2 IRT Model for Polytomous Items

In PIRLS 2006, as 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 student 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 as:

$$\textbf{(3)} \quad P\left(x_i = l \mid \theta_k, a_i, b_i, d_{i,1}, \cdots, d_{i,m_i-1}\right) = \frac{\exp\left(\sum_{v=0}^{l} 1.7 \cdot a_i\left(\theta_k - b_i + d_{i,v}\right)\right)}{\sum_{g=0}^{m_i-1} \exp\left(\sum_{v=0}^{g} 1.7 \cdot a_i\left(\theta_k - b_i + d_{i,v}\right)\right)} \equiv P_{i,l}\left(\theta_k\right)$$

where

m_i is the number of response categories for item *i*, either 3 or 4;

x_i is the response to item *i*, ranging between 0 and $m_i - 1$;

θ_k is the proficiency of a student on a scale *k*;

a_i is the slope parameter of item *i*;

b_i is its location parameter, characterizing its difficulty;

$d_{i,l}$ is the category *l* threshold parameter.

The indeterminacy of model parameters in the polytomous model is resolved by setting $d_{i,0} = 0$ and $\sum_{j=1}^{m_i - 1} d_{i,j} = 0$.

For all of the IRT models there is a linear indeterminacy between 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, as was done for PIRLS in 2001. 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 a student's proficiency) and the specified parameters of the item, and are unaffected by the demographic characteristics or unique experiences of the students, 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:

(4) $$P\left(x \mid \theta_k, \textit{item parameters}\right) = \prod_{i=1}^{n} \prod_{l=0}^{m_i - 1} P_{i,l}\left(\theta_k\right)^{u_{i,l}}$$

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 dichotomously scored items, and $u_{i,l}$ is an indicator variable defined as:

(5)
$$u_{i,l} = \begin{cases} 1 & \textit{if response } x_i \textit{ is in category } l; \\ 0 & \textit{otherwise}. \end{cases}$$

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. In PIRLS 2006, the item parameters for each scale were estimated independently of the parameters of other scales. 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 student.

11.2.3 Proficiency Estimation Using Plausible Values

Most cognitive skills testing is concerned with accurately assessing the performance of individual students 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 each individual's estimated θ.

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. This design solicits relatively few responses from each sampled student while maintaining a wide range of content representation when responses are aggregated across all students. With this approach, however, the advantage of estimating population characteristics more efficiently is 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 that can be used in analyses with standard statistical software. A detailed review of the plausible values methodology is given in Mislevy (1991).[6]

The following 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 $t(\theta, y)$, 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 students. The 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:

(6)

$$\begin{aligned} t^*(x, y) &= E\left| t(\underline{\theta}, \underline{y}) \,|\, \underline{x}, \underline{y} \right| \\ &= \int t(\underline{\theta}, \underline{y})\, p\left(\underline{\theta} \,\middle|\, \underline{x}, \underline{y}\right)\, d\underline{\theta} \end{aligned}$$

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 items. These values are referred to as imputations in the sampling literature, and as plausible values in large-scale surveys such as PIRLS, TIMSS, NAEP, NALS, and IALS. The value of θ for any student 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 uncertainly 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 the above equation; the variance among them reflects the uncertainty due to not observing $\underline{\theta}$. It should be noted that this variance does not include the variability of sampling from the population. That variability is estimated separately by jackknife variance estimation procedures, which are discussed in Chapter 12.

6 Along with theoretical justifications, Mislevy presents comparisons with standard procedures; discusses biases that arise in some secondary analyses; and offers numerical examples.

Note that 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 student j are drawn from the conditional distribution $P\left(\theta_j \middle| x_j, y_j, \Gamma, \Sigma\right)$, where Γ is a matrix of regression coefficients for the background variables, and Σ is a common variance matrix of residuals. Using standard rules of probability, the conditional probability of proficiency can be represented as:.

(7) $$P\left(\theta_j \middle| x_j, y_j, \Gamma, \Sigma\right) \propto P\left(x_j \middle| \theta_j, y_j, \Gamma, \Sigma\right) P\left(\theta_j \middle| y_j, \Gamma, \Sigma\right) = P\left(x_j \middle| \theta_j\right) P\left(\theta_j \middle| y_j, \Gamma, \Sigma\right)$$

where θ_j is a vector of scale values, $P\left(x_j \middle| \theta_j\right)$ is the product over the scales of the independent likelihoods induced by responses to items within each scale, and $P\left(\theta_j \middle| y_j, \Gamma, \Sigma\right)$ is the multivariate joint density of proficiencies for the scales, conditional on the observed values 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 section.

11.2.4 Conditioning

A multivariate normal distribution was assumed for $P\left(\theta_j \middle| y_j, \Gamma, \Sigma\right)$, 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 of variables to be used in Γ. Typically, components accounting for 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:

(8) $$\theta = \Gamma' y^c + \varepsilon$$

7 For further discussion, see Mislevy, Beaton, Kaplan, and Sheehan (1992).

where 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.

Note that in order to be strictly correct for all functions Γ of θ, it is necessary that $P(\theta \mid y)$ 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 is discussed in detail in Mislevy (1991). In PIRLS 2006, 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 (7).

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 $P(\Gamma, \Sigma \mid x_j, y_j)$ 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_j and variance Σ_j^p of the posterior distribution in equation (7), where *p* is the number of scales, 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_j and variance Σ_j^p. These three steps are repeated five times, producing five imputations of θ_j for each sampled student.

For students with an insufficient number of responses, the Γ's and Σ's described in the previous paragraph are fixed. Hence, all students—regardless of the number of items attempted—are assigned a set of plausible values.

The plausible values can then be employed to evaluate equation (6) for an arbitrary function T as follows:

- Using the first vector of plausible values for each student, evaluate T as if the plausible values were the true values of θ. Denote the result as T_1.
- Evaluate the sampling variance of T_1, or Var_1, with respect to students' first vector of plausible values.
- Carry out steps 1 and 2 for the second through fifth vectors of plausible values, thus obtaining T_u and Var_u for $u = 2, \ldots, 5$.
- 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:

$$\hat{T} = \frac{\sum_u T_u}{5}$$

- An estimate of the variance of $\hat{T}$ is the sum of two components: an estimate of Var_u obtained by averaging as in the previous step, and the variance among the T_u's.

Let $\bar{U} = \frac{\sum_u Var_u}{M}$, and let $B_M = \frac{\sum_u \left(T_u - \hat{T}\right)^2}{M-1}$ be the variance among the M plausible values. Then the estimate of the total variance of $\hat{T}$ is:

(9)
$$Var\left(\hat{T}\right) = \bar{U} + \left(1 + M^{-1}\right) B_M$$

The first component in $Var\left(\hat{T}\right)$ reflects the uncertainty due to sampling students from the population; the second reflects the uncertainty due to the fact that sampled students' θ's are not known precisely, but only indirectly through x and y.

11.2.6 Working with Plausible Values

The plausible values methodology was used in PIRLS 2006 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 student reflects the uncertainty associated with proficiency estimates for individual students. However, retaining this component of uncertainty requires that additional analytical procedures be used to estimate students' proficiencies.

If the θ values were observed for all sampled students, the statistic $(t-T)/U^{1/2}$ would follow a *t*-distribution with *d* degrees of freedom. Then the incomplete-data statistic $\left(T-\hat{T}\right) / \left[Var\left(\hat{T}\right)\right]^{1/2}$ is approximately *t*-distributed, with degrees of freedom (Johnson & Rust, 1993) given by:

(10)

$$\nu = \frac{1}{\frac{f_M^{\,2}}{M-1} + \frac{\left(1-f_M\right)^2}{d}}$$

where *d* is the degrees of freedom for the complete-data statistic, and f_M is the proportion of total variance due to not observing the values:

(11)

$$f_M = \frac{\left(1+M^{-1}\right) B_M}{Var\left(\hat{T}\right)}$$

When B_M is small relative to $\bar{U}$, the reference distribution for the incomplete-data statistic differs little from the reference distribution for the corresponding complete-data statistic. If, in addition, d is large, the normal approximation can be used instead of the *t*-distribution.

For a *k*-dimensional function *T*, such as the *k* coefficients in a multiple regression analysis, each U and $\bar{U}$ is a covariance matrix, and B_M is an average of squares and cross-products rather than simply an average of squares. In this case, the quantity $\left(\underline{T}-\underline{\hat{T}}\right) Var^{-1}\left(\underline{\hat{T}}\right)\left(\underline{T}-\underline{\hat{T}}\right)'$ 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:

(12) $$f_M = \left(1+M^{-1}\right) Trace\left[B_M Var^{-1}\left(\hat{T}\right)\right] \Big/ k$$

For the same reason that the normal distribution can approximate the *t*-distribution, a chi-square distribution with *k* degrees of freedom can be used in place of the *F*-distribution for evaluating the significance of the above quantity $\left(\underline{T}-\underline{\hat{T}}\right) Var^{-1}\left(\underline{\hat{T}}\right)\left(\underline{T}-\underline{\hat{T}}\right)'$.

Statistics $\hat{T}$, the estimates of proficiency 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 2006 analyses included nearly all background variables.

11.3 Implementing the Scaling Procedures for the PIRLS 2006 Assessment Data

The application of IRT scaling and plausible value methodology to the PIRLS 2006 assessment data involved four major tasks: calibrating the achievement test items (estimating model parameters for each item), creating principal components from the student and home questionnaire data for use in conditioning; generating IRT scale scores (proficiency scores) for overall reading, the two purposes of reading (reading for literary experience and reading to acquire and use information) and the two processes of reading (processes of retrieving and straightforward inferencing and processes of interpreting, integrating, and evaluating); and placing the proficiency scores on the metric used to report the results from 2001. The PIRLS reporting metric was established by setting the average of the mean scores of the countries that participated in PIRLS 2001 to 500 and the standard deviation to 100. To enable comparisons between 2006 and 2001, the PIRLS 2006 data also were placed on this metric.

11.3.1 Calibrating the PIRLS 2006 Test Items

In striving to measure trends in a changing world, PIRLS releases a number of assessment blocks after each assessment year and replaces them with newly developed blocks that incorporate current thinking of reading literacy and approaches to reading instruction. A number of assessment blocks also are kept secure to be used again in future assessments. The PIRLS 2006 item calibration is based on all items from 2006 and 2001 and all countries that participated in both assessments. This is known as concurrent calibration. The common items are used to ensure that there is sufficient overlap between the current assessment and the previous one, however, the 2001 items that were ultimately released and the items that were developed for 2006 also contribute to setting the PIRLS 2006 scales. Exhibit 11.1 shows the distribution of items included in the PIRLS 2006 calibrations for all five PIRLS scales. The 174 items included in the overall scale were divided between those measuring reading for literary experience (89 items) and for information (85 items) for calibrating the two reading purposes scales, and between those measuring retrieving and straightforward inferencing (96 items) and those measuring interpreting, integrating, and evaluating (78 items) for calibrating the two comprehension processes scales. Exhibit 11.2 lists the countries included in the item calibrations and their sample sizes for both assessment years. A total of 225,542 students from 26 countries contributed to the item calibrations.

Exhibit 11.1 Items Included in the PIRLS 2006 Item Calibrations

PIRLS Scales		Items Unique to PIRLS 2001		Items Unique to PIRLS 2006 [1]		Items in Both Assessment Cycles		Total	
		Number	Points	Number	Points	Number	Points	Number	Points
Overall Reading		49	67	76	99	49	66	174	232
Purposes of Reading	Literary Experience	25	33	38	51	26	33	89	117
	Acquire and Use Information	24	34	38	48	23	33	85	115
Processes of Reading	Retrieving and Straightforward Inferencing	22	24	44	47	30	36	96	107
	Interpreting, Integrating, and Evaluating	27	43	32	52	19	30	78	125

1 Item R021S08M was removed from all item calibrations because of poor psychometric properties.

Exhibit 11.2 Samples Included in the PIRLS 2006 Item Calibrations

Countries	Sample Sizes	
	PIRLS 2006	PIRLS 2001
Bulgaria	3,863	3,460
England	4,036	3,156
France	4,404	3,538
Germany	7,899	7,633
Hong Kong SAR	4,712	5,050
Hungary	4,068	4,666
Iceland	3,673	3,676
Iran, Islamic Rep. of	5,411	7,430
Israel	3,908	3,973
Italy	3,581	3,502
Latvia	4,162	3,019
Lithuania	4,701	2,567
Macedonia, Rep. of	4,002	3,711
Moldova, Rep. of	4,036	3,533
Morocco	3,249	3,153
Netherlands	4,156	4,112
New Zealand	6,256	2,488
Norway	3,837	3,459
Romania	4,273	3,625
Russian Federation	4,720	4,093
Scotland	3,775	2,717
Singapore	6,390	7,002
Slovak Republic	5,380	3,807
Slovenia	5,337	2,952
Sweden	4,394	6,044
United States	5,190	3,763
Total	119,413	106,129

In line with the PIRLS assessment framework, IRT scales were constructed for reporting student achievement in overall reading, as well as for reporting separately for each of the two purposes of reading and the two processes of reading. The first step in constructing these scales was to estimate the IRT model item parameters for each item on each of the five PIRLS scales. This item calibration was conducted using the commercially-available PARSCALE software (Muraki & Bock, 1991; version 4.1). Item calibration included data from PIRLS 2006 and PIRLS 2001 for countries that participated in both assessment years in order to measure trends from 2001 to 2006. The assessment data were weighted to ensure that the data from each country and each assessment year contributed equally to the item calibration.

Five separate item calibrations were run: one for the overall reading scale; one for each of the two purposes of reading—literary experience and acquire and use information; and one for each of the two processes of reading—retrieving and straightforward inferencing and interpreting, integrating, and evaluating. Exhibits D.1 through D.5 in Appendix D display the item parameters estimated from the five calibration runs. All items and all students involved in the calibration process were included in the calibration of the overall reading scale. Interim reading scores[8] were produced as a by-product of this first calibration for use in generating conditioning variables. For the calibration of the literary experience scale, only items from literary assessment blocks and only those students completing a booklet with a literary block (183,431) were included. Similarly, only items from information assessment blocks and only those students completing a booklet with an information block (183,253) were included in the calibration of the acquire and use information scale. The situation was somewhat different for the two processes of reading since all assessment blocks, regardless of their purpose of reading, had a mix of items classified in the two processes of reading. Thus, only items classified in the retrieving and straightforward inferencing process and nearly all students[9] (225,539) were included in the calibration of the retrieving and straightforward inferencing scale, and only items classified in the interpreting, integrating, and evaluating process and nearly all students (225,435) were included in the calibration of the interpreting, integrating, and evaluating scale.

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.

In PIRLS 2006, as in 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.

8 Because each student responded to only a subset of the assessment item pool, these interim scores, known as expected a priori—or EAP—scores, were not sufficiently reliable for reporting PIRLS results. The plausible value proficiency scores were used for this purpose.

9 Three students did not respond to any items classified in the "retrieving and straightforward inferencing" process and 107 students did not respond to any items classified in the "interpreting, integrating, and evaluating" process.

This approach was considered optimal for item parameter estimation. However, 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 2006 Data

After the calibrations 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 2006 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 function curves calculated from the posterior distributions of the θ's for each student that responded to the item. Graphical plots of the theoretical and empirical item response function curves are called item characteristic curves (ICC).

Exhibit 11.3 shows an ICC 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. The theoretical curve based on the estimated item parameters is shown as a solid line. Empirical results are represented by circles. The empirical results were obtained by first dividing the proficiency scale into intervals of equal size and then counting the number of students responding to the item whose EAP scores from PARSCALE fell in each interval. Then the proportion of students in each interval that responded correctly to the item was calculated.[10] In the exhibit, the center of each circle represents this empirical proportion of correct responses. The size of each circle is proportional to the number of students contributing to the estimation of its empirical proportion correct.

10 These calculations were performed using the SENWGT.

Exhibit 11.3 PIRLS 2006 Reading Assessment Example Item Response Function for a Dichotomous Item

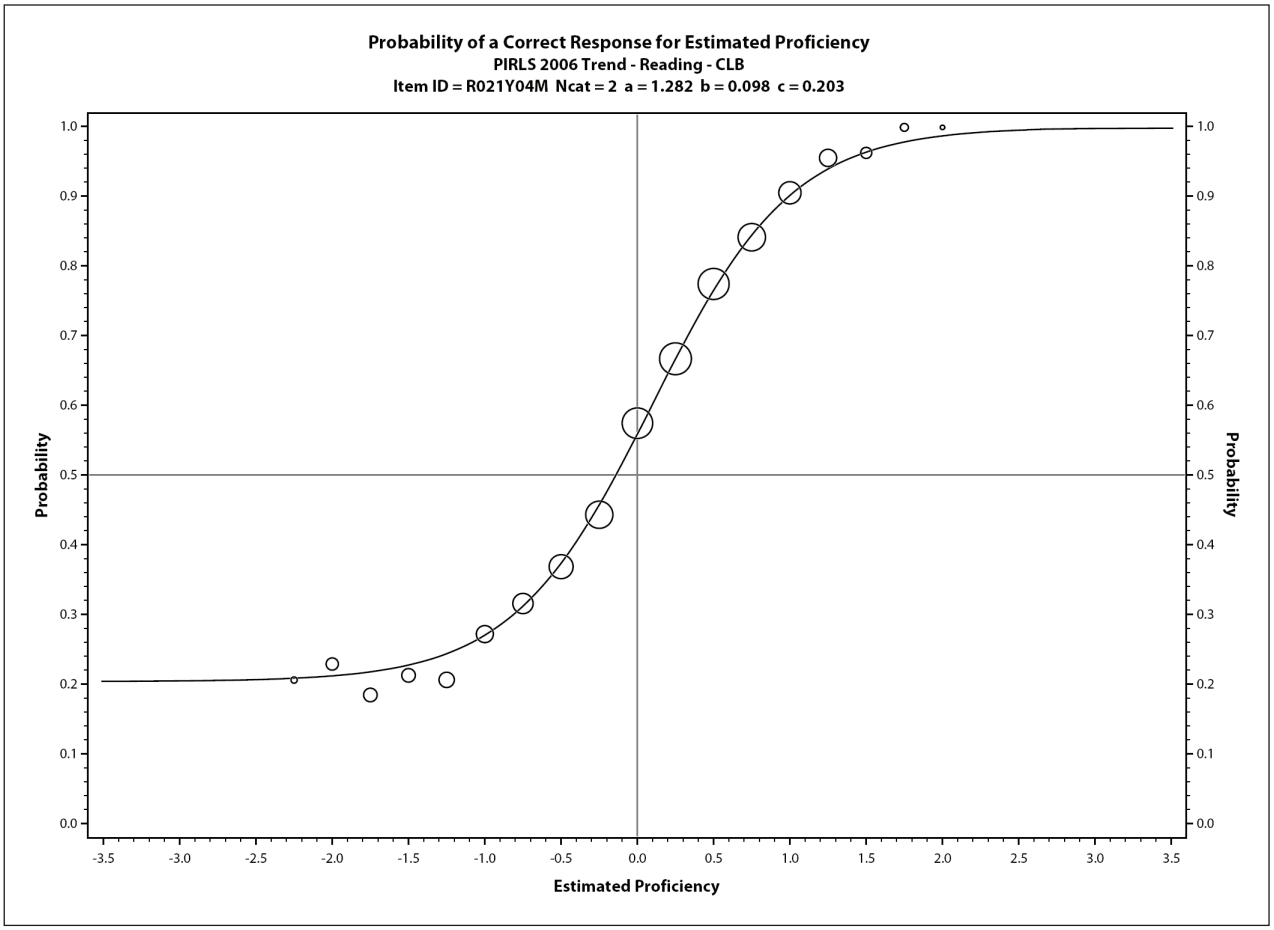
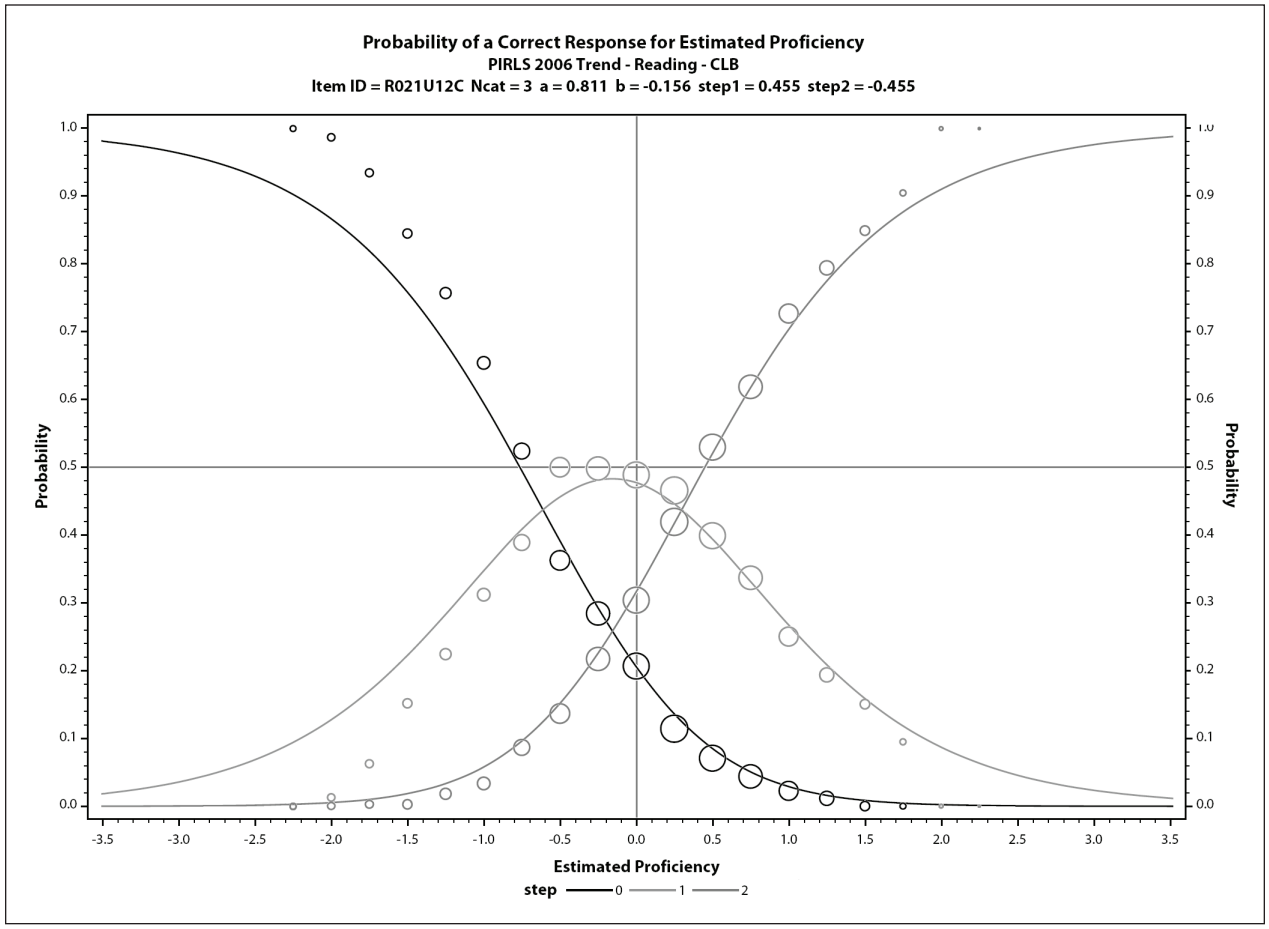


Exhibit 11.4 contains an ICC plot of the empirical and theoretical item response functions for a polytomous item. As for the dichotomous item plot, the horizontal axis represents the proficiency scale, but the vertical axis represents the probability of having a response in a given response category. The theoretical curves based on the estimated item parameters are shown as solid lines. Empirical results are represented by circles. The interpretation of the circles is the same as in Exhibit 11.3. For items where the IRT model fits the data well, the empirical results fall near the theoretical curves.

Exhibit 11.4 PIRLS 2006 Reading Assessment Example Item Response Function for a Polytomous Item



11.3.4 Variables for Conditioning the PIRLS 2006 Data

PIRLS 2006 used all background variables from the student background questionnaire and the Learning to Read Survey questionnaire. 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 reduce the number of variables while explaining most of their common variance. Principal components for the PIRLS 2006 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 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.[11] 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. Because the principal components analysis was performed separately for each country, different numbers of principal components were required to account for 90% of the common variance in each country's background variables.

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. These additional variables are characterized as primary conditioning variables. Exhibit 11.5 shows the total number of variables that were used for conditioning.

11 The process of generating criterion-scaled variables is described in Beaton (1969).

Exhibit 11.5 Number of Variables Used for Conditioning in PIRLS 2006

Countries	Sample Sizes	Number of Background Variables Available	Conditioning Variables	
			Principal Components	Primary Conditioning Variables
Austria	5,067	526	295	2
Belgium (Flemish)	4,479	520	286	2
Belgium (French)	4,552	514	291	2
Bulgaria	3,863	528	282	2
Canada, Alberta	4,243	495	274	3
Canada, British Columbia	4,150	495	274	3
Canada, Nova Scotia	4,436	495	279	3
Canada, Ontario	3,988	495	272	3
Canada, Quebec	3,748	495	275	3
Chinese Taipei	4,589	519	295	2
Denmark	4,001	528	289	2
England	4,036	528	280	2
France	4,404	516	293	2
Georgia	4,402	518	297	2
Germany	7,899	520	291	2
Hong Kong SAR	4,712	530	299	2
Hungary	4,068	497	278	2
Iceland	3,673	506	284	2
Indonesia	4,774	492	291	2
Iran, Islamic Rep. of	5,411	530	297	2
Israel	3,908	530	296	3
Italy	3,581	530	292	2
Kuwait	3,958	509	299	2
Latvia	4,162	525	292	3
Lithuania	4,701	511	290	2
Luxembourg	5,101	522	292	2
Macedonia, Rep. of	4,002	528	303	3
Moldova, Rep. of	4,036	530	294	3
Morocco	3,249	506	286	2
Netherlands	4,156	520	281	2
New Zealand	6,256	520	287	8
Norway	3,837	522	283	3
Poland	4,854	501	284	2
Qatar	6,680	526	310	2
Romania	4,273	530	289	3
Russian Federation	4,720	500	282	2
Scotland	3,775	528	277	2
Singapore	6,390	526	296	2
Slovak Republic	5,380	524	293	3
Slovenia	5,337	518	290	2
South Africa	14,657	503	312	12
Spain	4,094	528	285	6
Sweden	4,394	528	289	2
Trinidad and Tobago	3,951	491	281	2
United States[1]	5,190	285	166	7

1 The United States did not administer the "Learning to Read Survey" questionnaire, thus reducing the number of background variables available for conditioning.

11.3.5 Generating IRT Proficiency Scores for the PIRLS 2006 Data

The MGROUP program (Sheehan, 1985; version 3.2)[12] 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, and generates as output the plausible values that represent student proficiency. For each of the 45 PIRLS participants listed in Exhibit 11.5, it was necessary to run MGROUP three times to produce the PIRLS 2006 assessment scales: one unidimensional run for the overall reading scale, one multidimensional run for the reading purposes scales, and one multidimensional run for the comprehension processes scales. Thus a total of 135 (45x3) MGROUP runs were required to obtain proficiency scores for PIRLS 2006.

In addition to generating plausible values for the PIRLS 2006 data, the parameters estimated at the calibration stage also were used to generate plausible values on all five PIRLS scales using the 2001 data for the 26 trend countries that participated in both assessment years. These plausible values for the trend countries are called "link scores." Link scores were also produced for the Canadian provinces of Ontario and Quebec for evaluation purposes. Producing the link scores required 84 additional MGROUP runs.

Plausible values generated by the conditioning 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 an expected mean of zero across all countries.

11.3.6 Transforming the Proficiency Scores to Measure Trends between 2001 and 2006

To provide results for PIRLS 2006 comparable to the results from the PIRLS 2001 assessment, the 2006 proficiency scores (plausible values) had to be transformed to the metric used in 2001. To accomplish this, the means and standard deviations of the link scores for all five PIRLS scales were made to match the means and standard deviations of the scores reported in the 2001 assessment by applying the appropriate linear transformations. These linear transformations are given by:

(13)

$$PV_{k,i}^{*} = A_{k,i} + B_{k,i} \cdot PV_{k,i}$$

12 The MGROUP program was provided by ETS under contract to the TIMSS and PIRLS International Study Center at Boston College.

where

$PV_{k,i}$ is the plausible value *i* of scale *k* prior to transformation;

$PV_{k,i}^{*}$ is the plausible value *i* of scale *k* after transformation;

and $A_{k,i}$ and $B_{k,i}$ are the linear transformation constants.

The linear transformation constants were obtained by first computing, using the senate weight, the international means and standard deviations of the proficiency scores for all five PIRLS scales using the plausible values generated in 2001 for the 26 trend countries. Next, the same calculations were done using the 2006 link scores of the 26 trend countries. The linear transformation constants are defined as:

(14)

$$\begin{aligned} B_{k,i} &= \sigma_{k,i} / \sigma_{k,i}^{*} \\ A_{k,i} &= \mu_{k,i} - B_{k,i}\ \mu_{k,i}^{*} \end{aligned}$$

where

$\mu_{k,i}$ is the international mean of scale *k* based on plausible value *i* released in 2001;

$\mu_{k,i}^{*}$ is the international mean of scale *k* based on plausible value *i* of the 2006 link scores;

$\sigma_{k,i}$ is the international standard deviation of scale *k* based on plausible value *i* released in 2001;

$\sigma_{k,i}^{*}$ is the international standard deviation of scale *k* based on plausible value *i* of the 2006 link scores.

Exhibit 11.6 shows the linear transformation constants that were computed.

Once the linear transformation constants were established, all of the proficiency scores from the 2006 assessment were transformed by applying the

Exhibit 11.6 Linear Transformation Constants Used for the PIRLS 2006 Data

Scale		Plausible Values	PIRLS 2001 Scores		2006 "Link Scores"		$A_{k,i}$	$B_{k,i}$
			Mean	Standard Deviation	Mean	Standard Deviation		
Overall Reading		PV1	514.9855	91.9947	-0.0435	0.9018	102.0152	519.4237
		PV2	514.8861	92.1770	-0.0399	0.8999	102.4341	518.9761
		PV3	514.8006	92.3735	-0.0400	0.9006	102.5698	518.8983
		PV4	514.8252	92.2470	-0.0390	0.8991	102.5944	518.8265
		PV5	514.7781	92.2987	-0.0414	0.9012	102.4191	519.0163
Purposes of Reading	Literary Experience	PV1	514.6110	92.5091	0.1699	0.9962	92.8640	498.8316
		PV2	514.5735	92.5937	0.1694	0.9947	93.0840	498.8075
		PV3	514.4664	92.4649	0.1722	0.9979	92.6575	498.5120
		PV4	514.6021	92.6655	0.1711	0.9965	92.9868	498.6943
		PV5	514.4937	92.7265	0.1723	0.9988	92.8355	498.5015
	Acquire and Use Information	PV1	514.5481	92.2754	0.0737	0.9664	95.4885	507.5074
		PV2	514.3908	92.2856	0.0735	0.9672	95.4108	507.3758
		PV3	514.6731	92.0767	0.0708	0.9656	95.3604	507.9254
		PV4	514.5654	92.0253	0.0709	0.9671	95.1549	507.8201
		PV5	514.5440	92.1947	0.0681	0.9663	95.4119	508.0446
Processes of Reading	Retrieving and Straightforward Inferencing	PV1	514.3950	93.7040	0.0233	0.9984	93.8557	512.2098
		PV2	514.6367	93.6133	0.0216	0.9995	93.6559	512.6105
		PV3	514.4507	93.7383	0.0206	0.9971	94.0063	512.5182
		PV4	514.3605	93.5307	0.0215	1.0014	93.4021	512.3508
		PV5	514.3732	93.7112	0.0220	1.0000	93.7112	512.3132
	Interpreting, Integrating, and Evaluating	PV1	515.2249	90.9159	-0.1075	0.9767	93.0841	525.2357
		PV2	515.0767	91.1286	-0.1112	0.9806	92.9353	525.4140
		PV3	515.0542	91.1681	-0.1101	0.9814	92.8949	525.2859
		PV4	515.0299	90.9888	-0.1118	0.9845	92.4232	525.3639
		PV5	515.0590	91.1741	-0.1133	0.9826	92.7873	525.5682

same linear transformations for all countries. This provided achievement scores for the PIRLS 2006 assessment that were directly comparable to the scores from the 2001 assessment.

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,* 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, *Statistical theories of mental test scores*, (pp.397–479). Reading, MA: Addison-Wesley Publishing.

Johnson, E.G., & Rust, K.F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics, 17,* 175–190.

Lord, F.M.,& Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdales, NJ: Lawrence Erlbaum Associates.

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*, 133–161.

Mislevy, R.J., Johnson, E.G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, *17*, 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). (no. 15-TR-20) Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Mullis, I.V.S., Kennedy, A.M., Martin, M.O., & Sainsbury, M. (2006). *PIRLS 2006 assessment framework and specifications* (2nd ed.). Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., & Gonzalez, E.J., (2004), *International achievement in the processes of reading comprehension: Results from PIRLS 2001 in 35 countries*, Chestnut Hill, MA: Boston College.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16(2)*, 159-176.

References *(continued)*

Muraki, E., & Bock, R.D. (1991). PARSCALE: Parameter scaling of rating data [Computer software and manual], Chicago, IL: Scientific Software, Inc.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Sheehan, K. M. (1985). M-GROUP: Estimation of group effects in multivariate models [Computer program]. Princeton, NJ: Educational Testing Service.

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) (No. 15-TR-20). 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.

Chapter 12

Reporting Student Achievement in Reading

Ann M. Kennedy and Kathleen L. Trong

12.1 Overview

The *PIRLS 2006 International Report* (Mullis, Martin, Kennedy, & Foy, 2007) presents a summary of reading achievement at the fourth grade in the 45 participating countries and provinces, as well as trends for those countries that also participated in PIRLS 2001. This chapter explains how the PIRLS International Benchmarks were established and the scale-anchoring process used to describe student achievement at each of these benchmarks. Additionally, the statistical procedures used to estimate the sampling and imputation variance that result from the PIRLS sampling and assessment design are described, as well as the methods used to calculate key statistics across countries.

12.2 PIRLS 2006 International Benchmarks of Student Achievement

As described in the previous chapter, substantial effort was put into creating the PIRLS reading achievement scale. To make full use of this information, it is essential that readers understand what scores on the scale mean. In other words, what skills did a student who scored 500 demonstrate? To facilitate this, the PIRLS International Benchmarks were created, and scale anchoring was used to describe student achievement at these points along the scale. The associated scale score for each benchmark is shown in Exhibit 12.1.

Exhibit 12.1 PIRLS 2006 International Benchmarks

Scale Score	International Benchmark
625	Advanced International Benchmark
550	High International Benchmark
475	Intermediate International Benchmark
400	Low International Benchmark

The PIRLS International Benchmarks are a set of unchanging points along the achievement scale that can be used to measure student achievement across countries and over time. It should be noted that the PIRLS 2006 International Benchmarks were established using procedures different from those in 2001. In PIRLS 2001, percentiles were used to determine benchmarks. That is, the points used to describe achievement were the Top 10 Percent (90^{th} percentile), Upper Quarter (75^{th} percentile), Median (50^{th} percentile), and Lower Quarter (25^{th} percentile). However, because benchmarks based on percentiles necessarily would be recalculated in each cycle according to the countries participating in that cycle, they would fluctuate as a greater range of countries participate in the future. To enable using the benchmarks to make comparisons across assessment cycles, the points need to be kept the same from cycle to cycle. Therefore, beginning in TIMSS 2003, permanent benchmarks were chosen for use with both IEA's TIMSS and PIRLS studies that were similar to those anchored in TIMSS 1999 for both mathematics and science (Gonzalez, Galia, Arora, Erberber, & Diaconu, 2004).

For reporting purposes, the 2006 benchmarks were applied to the 2001 data to allow for comparison across cycles. The permanent benchmarks are evenly distributed along the scale and are more dispersed than those in PIRLS 2001, with the 2006 benchmarks ranging from 400 (Low) to 625 (Advanced), whereas the 2001 benchmarks ranged from 435 (Lower Quarter) to 615 (Top 10 Percent). This greater breadth will be better able to capture the variance of achievement as more diverse countries participate in future assessments.

12.2.1 Identifying Students Achieving at Each Benchmark

Criteria were established for identifying students who scored at each of these International Benchmarks. As has been done in previous IEA studies, across all the PIRLS 2006 participants, all students scoring within +/- 5 score points of the benchmark were included in scale-anchoring analyses. This is done to create student groups that are large enough for analysis purposes, but small enough

that each benchmark remains clearly distinguished from the others. These ranges and the number of students scoring within each range in PIRLS 2006 are displayed in Exhibit 12.2.

Exhibit 12.2 Range Around Each International Benchmark and Number of Students Within Range

	Low International Benchmark 400	Intermediate International Benchmark 475	High International Benchmark 550	Advanced International Benchmark 625
Range of Scale Scores	395-405	470-480	545-555	620-630
Number of Students	2,681	6,484	10,360	4,844

12.2.2 Identifying Items Characterizing Achievement at Each Benchmark

Once the students achieving at each benchmark were identified, criteria were established to determine the items that these students were likely to answer correctly and that discriminate between the benchmarks (e.g., between the High and Advanced International Benchmarks). This allows for the development of descriptions of skills that students at each benchmark demonstrated through scale anchoring. To determine which items students at each anchor level were likely to answer correctly, the percent correct for those students was calculated for each item at each benchmark. For this analysis, students across the PIRLS 2006 participants were weighted so that students in each country contributed proportional to the size of the student population in that country.

For dichotomously scored items, the percent of students at each anchor point who answered each item correctly was computed. For constructed-response items with multiple score points (i.e., 2 or 3), each score level was treated separately because the different score levels may demonstrate different reading skills. For example, for a 2-point item, the percent of students at each anchor point earning only partial credit (1 point) was computed. In addition, the percent of students at each anchor point earning at least partial credit (1 or 2 points) was computed. This allowed the different score levels of an item to potentially anchor at different benchmarks.

Except at the Low International Benchmark, establishing criteria to identify items that were answered correctly by most students at the benchmark, but by fewer students at the next lower point, required considering achievement at adjacent benchmarks. For multiple-choice items, the criterion of 65 percent was used, 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 this means that students were more likely to answer the item incorrectly than correctly. For example, if 65 percent of students scoring at the High International Benchmark answered a particular multiple-choice item correctly, but less than 50 percent of students at the Intermediate International Benchmark did so, this would be an anchor item for the High International Benchmark. For constructed-response items, a criterion of 50 percent was used, since there is no possibility of guessing to take into account, with no criterion for lower points.

Anchored Items

The criteria used to identify items that "anchored" at each of the four PIRLS 2006 International Benchmarks are outlined below.

An item anchored at the Low International Benchmark if:

- For a constructed-response item, at least 50 percent of students received either partial credit (e.g., at least 1 or at least 2 points, depending upon the maximum number of score points) or the full-credit score value (1, 2, or 3);
- For a multiple-choice item, at least 65 percent of students answered the item correctly. At the lowest level, only the 65 percent criterion is necessary, as there is no lower level from which to discriminate.

An item anchored at the Intermediate International Benchmark if:

- For a constructed-response item, at least 50 percent of students received at least partial or full credit;
- For a multiple-choice item at least 65 percent of students at the Intermediate International Benchmark, and less than 50 percent of students at the Low International Benchmark, answered the item correctly.

An item anchored at the High International Benchmark if:

- For a constructed-response item, at least 50 percent of students received at least partial or full credit;
- For a multiple-choice item, at least 65 percent of students at the High International Benchmark, and less than 50 percent of students at the Intermediate International Benchmark, answered the item correctly.

An item anchored at the Advanced International Benchmark if:

- For a constructed-response item, at least 50 percent of students received at least partial or full credit;
- For a multiple-choice item, at least 65 percent of students, and less than 50 percent of students at the High International Benchmark, answered the item correctly.

Almost Anchored Items

Not all items were assumed to be able to meet the anchoring criteria. Some items nearly met the 65 percent criterion, but did not discriminate between the anchor levels. Others discriminated well between anchor levels, but did not quite meet the 65 percent criterion.

The following criteria were established for those items nearly satisfying the anchoring criteria.

- An item "almost anchored" if more than 60 percent of students at a level answered an item correctly, and less than 50 percent of the students at the next lowest level answered correctly (the discrimination criterion is met).
- An item "anchored (only 60-65)" if more than 60 percent of students at a level answered an item correctly, but 50 percent or more students at the next lowest level answered correctly (the discrimination criterion is not met).

It is important to note that since there is no discrimination criterion for constructed-response items, the descriptions of the criteria for nearly meeting the anchoring requirements are for multiple-choice items only.

Items Too Difficult to Anchor

An item was too difficult to anchor if, for constructed-response items, less than 50 percent of students at the Advanced International Benchmark received at least partial or full credit, depending on the maximum score level for the item. For a multiple-choice item to be considered too difficult to anchor, less than 60 percent of students at the Advanced International Benchmark were able to answer correctly.

The results of the PIRLS 2006 scale anchoring of reading achievement are presented below in Exhibit 12.3. As this exhibit shows, considering items

that met the less stringent anchoring criteria added a substantial amount of information that could be used to describe student performance beyond what would have been available using only items that anchored.

Exhibit 12.3 Number of Items Anchoring at Each Benchmark

	Anchored	Almost Anchored	Met 60-65% Criterion	Total
Low (400)	9	4	0	13
Intermediate (475)	28	6	7	41
High (550)	53	5	9	67
Advanced (625)	28	0	6	34
			Too Difficult to Anchor	10
			Total	165

12.2.1 Expert Review of Anchor Items by Content Area

Once the empirical analysis identifying the items that anchored at each International Benchmark was completed, the items were reviewed by the PIRLS 2006 Reading Development Group (RDG), with the goal of developing descriptions of student performance. Members of the RDG were provided binders for each of the reading purposes, literary and informational, with their respective items grouped by benchmark and sorted by anchoring criteria. In other words, within the literary binder, there was a section for items that anchored at each benchmark, and in each section, the items that anchored appeared first, followed by those that almost anchored and those that met only the 60 to 65 percent criteria. For each item, the following information was displayed: item stem, answer key (for multiple-choice items), scoring guide for (constructed-response items), reading purpose, reading process, percent correct at each anchor point, overall international percent correct, and whether or not the item was released.

Using these materials, the descriptive portion of the scale anchoring analysis was conducted in Copenhagen, Denmark in April 2007. The task included developing a short description of the knowledge, understanding, or skills demonstrated by at least a partial-credit response for some constructed-response items, or by a full-credit response for a multiple-choice item or the maximum score level of a constructed-response item. Then, the item level descriptions for each International Benchmark were used to generalize and

draft a summary of the level of comprehension shown by students at each of the benchmarks. Following the meeting, the drafts were edited and presented in the international report. Additionally, example items that were selected to illustrate the benchmark descriptions were included in the international report.

Exhibit 12.4 presents the number of items (or point values, for multiple-point constructed-response items) that met one of the anchoring criteria for each benchmark, presented by reading purpose, as well as the number of items that were too difficult to anchor.

Exhibit 12.4 Number of Items Anchoring at Each Benchmark

	Low Benchmark	Intermediate Benchmark	High Benchmark	Advanced Benchmark	Too Difficult to Anchor	Total
Reading for Literary Purposes	5	24	37	15	3	84
Reading for Information	8	17	30	19	7	81

12.3 Capturing the Uncertainty in the PIRLS Student Achievement Measures

As discussed in previous chapters on sampling and scaling, PIRLS made extensive use of probability sampling techniques to sample students, and applied matrix sampling methods to administer a subset of the PIRLS 2006 assessment materials to each individual student. While this approach minimized the response burden to students, there is some variance or uncertainty in the statistics as a consequence. This uncertainty is measured and reported by providing an estimate of its standard error together with each statistic in the *PIRLS 2006 International Report* (Mullis, Martin, Kennedy, & Foy, 2007). For the achievement results, these standard errors reflect the uncertainty of the proficiency estimates due to two variance components—sampling variance and imputation variance.

12.3.1 Estimating Sampling Variance

There are several options for estimating sampling errors that take into account a complex sampling design, such as the stratified multistage cluster sampling applied in PIRLS 2006 (Brick, Morganstein, & Valliant, 2000). PIRLS uses a variation of the jackknife repeated replication (JRR) technique (Johnson & Rust, 1992) because it is computationally straightforward and provides approximately

unbiased estimates of the sampling errors of means, totals, and percentages. This technique assumes that the primary sampling units (PSUs) can be paired in a manner consistent with the sample design, with each pair belonging to a pseudo-stratum for variance estimation purposes. The JRR technique appropriately accounts for the combined effect of the between- and within-PSU contributions to the sampling variance.

The application of JRR involves pairing schools to sampling zones, and randomly selecting one of these schools to double its contribution and set the contribution of its paired school to zero, constructing a number of "pseudo-replicates" of the original sample. The statistic of interest is computed once for the original sample, and once again for each pseudo-replicate sample, with the variation between the estimates for each of the replicate samples and the original sample estimate being the jackknife estimate of the sampling error of the statistic.

12.3.2 Constructing Sampling Zones for Sampling Variance Estimation

Statistics Canada worked through the list of sampled schools for each PIRLS participating country and Canadian province to apply the JRR technique. Sampled schools were paired and assigned to a series of groups known as "sampling zones". Organized according to the order in which they were selected, the first and second sampled schools were assigned to the first sampling zone, the third and fourth schools to the second zone, and continuing through the list. 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 purpose of calculating the jackknife standard error. Each zone then consisted of a pair of schools or "quasi" schools. Exhibit 12.5 shows the range of sampling zones used in each country.

Exhibit 12.5 Number of Sampling Zones Used in PIRLS 2006 and PIRLS 2001

Countries	PIRLS 2006 Sampling Zones	PIRLS 2001 Sampling Zones
Austria	75	◊
Belgium (Flemish)	70	◊
Belgium (French)	75	◊
Bulgaria	74	75
Canada, Alberta	75	◊
Canada, British Columbia	74	◊
Canada, Nova Scotia	75	◊
Canada, Ontario	75	◊
Canada, Quebec	75	◊
Chinese Taipei	75	◊
Denmark	73	◊
England	75	66
France	75	73
Georgia	75	◊
Germany	75	75
Hong Kong SAR	74	74
Hungary	75	75
Iceland	75	75
Indonesia	75	◊
Iran, Islamic Rep. of	75	75
Israel	75	74
Italy	75	75
Kuwait	75	◊
Latvia	74	71
Lithuania	75	73
Luxembourg	75	◊
Macedonia, Rep. of	75	73
Moldova, Rep. of	75	75
Morocco	75	59
Netherlands	71	67
New Zealand	75	75
Norway	75	69
Poland	74	◊
Qatar	75	◊
Romania	75	73
Russian Federation	74	61
Scotland	66	59
Singapore	75	75
Slovak Republic	74	75
Slovenia	73	75
South Africa	75	◊
Spain	75	◊
Sweden	74	75
Trinidad and Tobago	75	◊
United States	47	52

A diamond (◊) indicates the country did not participate in the 2001 assessment.

12.3.3 Computing Sampling Variance Using the JRR Method

The JRR algorithm assumes that there are *H* sampling zones within each country, each containing two sampled schools selected independently. The equation to compute the JRR variance estimate of a statistic *t* from the sample for a country is as follows:

$$Var_{jrr}(t) = \sum_{h=1}^{H} \left[t(J_h) - t(S) \right]^2$$

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 h^{th} jackknife replicate. This is computed using all cases except those in the h^{th} zone of the sample. For those in the h^{th} 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 process is 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 h^{th} pair.

Therefore, in PIRLS 2006, the computation of the JRR variance estimate for any statistic required the computation of the statistic up to 76 times for any given country: once to obtain the statistic for the whole sample, and as many as 75 times to obtain the statistics for each of the jackknife replicates (J_h). The number of jackknife replicates for a given country depended on the number of implicit strata or sampling zones defined for that country.

Replicate weights used in calculations of statistics were created by doubling and zeroing the weights of the selected units within the sampling zones. Within a zone, one of the schools was randomly assigned an indicator (u_i), code of 1 or 0 so that one member of the pair was assigned 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 weight $W_h^{g,i,j}$ 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.

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 $W_0^{g,i,j}$ is the overall sampling weight, which is the product of the final school weight, classroom weight, and student weight.

The replicate weights for a single case were then computed as

$$W_h^{g,i,j} = W_0^{g,i,j} \cdot k_{hi}$$

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.

The replicate weights were not included as permanent variables in the PIRLS 2006 international database. Instead, they were created temporarily for each analysis by the sampling variance estimation program. For each country, PIRLS computed 75 replicate weights 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 total number of zones, were equal to the overall sampling weight. While computing 75 replicate weights for each country had no effect on the size of the error variance computed using the jackknife formula, the process facilitated the computation of standard errors for a number of countries simultaneously.

12.3.4 Estimating Imputation Variance

As described in Chapter 2, a matrix-sampling test design was used such that an individual student was administered a single test booklet containing only a portion of the PIRLS 2006 assessment. Using the scaling techniques described in Chapter 11, the results were aggregated across all booklets to provide results for the entire assessment, and plausible values were generated as estimates of student performance on the assessment as a whole. The variability among these estimates, or imputation error, for each variable was combined with the sampling error for that variable, providing an appropriate standard error that incorporates both error components.

To compute the imputation variance for any estimable statistic, t_m (e.g., mean, difference between means, or percentiles), the statistic must first be calculated for each set of *M* plausible values, where $m = 1, 2, \ldots, 5$.[1]

Once the statistics are computed, the imputation variance is computed as:

$$Var_{imp} = \left(1 + 1/M\right) Var\left(t_1, \ldots, t_M\right)$$

where *M* is the number of plausible values used in the calculation, and $Var\left(t_1, \ldots, t_M\right)$ is the variance of the *M* estimates computed using each plausible value.

12.3.5 Combining Sampling and Imputation Variance

In reporting reading proficiency statistics, PIRLS presented all calculated statistics with their standard errors, which incorporate both sampling and imputation variance components. The standard errors were computed using the following formula:[2]

$$Var\left(t_{pv}\right) = Var_{jrr}\left(t_1\right) + Var_{imp}$$

where $Var_{jrr}\left(t_1\right)$ is the sampling variance for the first plausible value and *Varimp* is the imputation variance. The *PIRLS 2006 User Guide for the International Database* (Foy & Kennedy, 2008) includes programs, for both SAS and SPSS statistical packages, that compute each of these variance components for the PIRLS 2006 data.

12.4 Calculating National and International Statistics for Student Achievement

This section describes the procedures for computing the statistics used to summarize reading achievement in the *PIRLS 2006 International Report*, including mean achievement scale scores based on plausible values, gender differences in average achievement, and performance on example items.

1 The general procedure for estimating the imputation variance using plausible values is described in Mislevy, R.J., Beaton, A.E., Kaplan, B., & Sheehan, K.M. (1992).

2 With unlimited computing resources, computing the imputation variance for the plausible values and the JRR sampling variance for each of the plausible values (pv) (i.e., computing the same statistic as many as 380 times: once for each pv using the overall sampling weight, and then 75 times for each pv using the complete set of replicate weights) is ideal. An acceptable shortcut, however, is to compute the JRR variance component using one pv, and then the imputation variance using the five pv. Using this approach, a statistic would be computed only 80 times.

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:

$$\bar{X}_{pvl} = \frac{\sum_{j=1}^{N} W^{i,j} . pv_{lj}}{\sum_{j=1}^{N} W^{i,j}}$$

where

- $\bar{X}_{pvl}$ is the country mean for plausible value l
- pv_{lj} 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, and
- N is the number of students in the country's sample.

Exhibits 12.6 through 12.10 provide basic summary statistics for reading achievement overall, as well as by purposes and processes. Each exhibit presents the student sample size, the mean achievement scale score 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.

12.4.1 Comparing Achievement Differences Across Countries

In reporting student achievement in the international report, PIRLS compares average performance of a participant with that of the other participants. Differences in mean achievement between countries are considered statistically significant if the absolute difference between them, divided by the standard error of the difference, is greater than the critical value. For differences between countries, which can be considered as independent samples, the standard error of the difference between means is 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. The means and standard errors used in the calculation of statistical significance for

reading achievement overall and by purposes and processes are presented in Exhibits 12.6-12.9.

The significance tests presented were not adjusted for multiple comparisons among countries. 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, the results vary depending on the number of countries included in the adjustment, leading to apparently conflicting results from comparisons using different combinations of countries.

12.4.2 Comparing National Average Achievement to the PIRLS Scale Average

Several exhibits in the international report compare the mean achievement for a country with the PIRLS scale average (500, with no standard error), together with a test of the statistical significance of the difference. The standard error of the difference is equal to the standard error of the mean achievement score for the country.

12.4.3 Reporting Gender Differences Within Countries

Gender differences were reported in overall student achievement in reading, as well as in the reading purposes and processes scales. Gender differences were presented in an exhibit showing mean achievement for girls and boys and their differences, with an accompanying graph indicating whether the difference 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 samples for the purpose of statistical tests. Accordingly, PIRLS applied a jackknife procedure for correlated samples to estimate the standard errors of the differences. This procedure involved computing the average difference between boys and girls in each country once for each of the 75 replicate samples, and five more times, once for each plausible value, as described in the earlier section on estimating imputation variance.

Exhibit 12.6 Summary Statistics and Standard Errors in Overall Reading Achievement

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Austria	5,067	538.296	63.654	2.105	2.200
Belgium (Flemish)	4,479	547.044	55.622	1.866	1.964
Belgium (French)	4,552	499.666	68.585	2.590	2.640
Bulgaria	3,863	547.032	82.682	4.341	4.366
Chinese Taipei	4,589	535.371	64.143	1.928	2.040
Denmark	4,001	546.346	69.712	2.257	2.266
England	4,036	539.483	86.845	2.464	2.560
France	4,404	521.593	66.584	2.061	2.066
Georgia	4,402	470.836	74.877	3.075	3.138
Germany	7,899	547.591	66.977	2.094	2.175
Hong Kong SAR	4,712	563.911	59.327	2.337	2.354
Hungary	4,068	550.889	70.238	2.931	2.976
Iceland	3,673	510.597	68.107	1.125	1.289
Indonesia	4,774	404.737	78.616	4.039	4.074
Iran, Islamic Rep. of	5,411	420.933	94.685	3.044	3.088
Israel	3,908	512.462	98.825	3.345	3.348
Italy	3,581	551.468	67.854	2.882	2.932
Kuwait	3,958	330.300	110.751	3.632	4.216
Latvia	4,162	540.912	62.635	2.210	2.335
Lithuania	4,701	537.033	56.895	1.610	1.640
Luxembourg	5,101	557.195	66.405	0.873	1.084
Macedonia, Rep. of	4,002	442.395	101.330	3.940	4.089
Moldova, Rep. of	4,036	499.884	69.038	3.025	3.037
Morocco	3,249	322.580	109.139	5.797	5.938
Netherlands	4,156	547.152	53.026	1.458	1.520
New Zealand	6,256	531.715	86.948	1.974	2.016
Norway	3,837	498.008	66.601	2.442	2.553
Poland	4,854	519.389	75.250	2.205	2.356
Qatar	6,680	353.436	95.575	1.070	1.090
Romania	4,273	489.473	91.463	4.998	5.012
Russian Federation	4,720	564.744	68.744	3.301	3.355
Scotland	3,775	527.355	79.862	2.755	2.791
Singapore	6,390	558.273	76.658	2.835	2.883
Slovak Republic	5,380	530.815	74.164	2.732	2.755
Slovenia	5,337	521.531	70.721	2.072	2.087
South Africa	14,657	301.613	136.181	5.467	5.555
Spain	4,094	512.504	70.965	2.394	2.482
Sweden	4,394	549.282	63.642	2.168	2.280
Trinidad and Tobago	3,951	435.588	103.316	4.863	4.885
United States	5,190	539.925	74.063	3.541	3.549

Exhibit 12.7 Summary Statistics and Standard Errors in Reading Achievement for Literary Purposes

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Austria	5,067	537.074	62.275	1.999	2.112
Belgium (Flemish)	4,479	543.807	57.634	1.878	1.908
Belgium (French)	4,552	499.482	67.463	2.401	2.419
Bulgaria	3,863	542.150	83.832	4.483	4.513
Chinese Taipei	4,589	530.438	69.442	1.946	1.994
Denmark	4,001	547.387	68.435	2.212	2.626
England	4,036	538.707	89.363	2.493	2.605
France	4,404	516.297	65.632	2.000	2.405
Georgia	4,402	476.456	75.489	3.130	3.238
Germany	7,899	548.768	66.452	1.992	2.161
Hong Kong SAR	4,712	556.926	64.015	2.538	2.607
Hungary	4,068	556.761	70.087	2.861	2.928
Iceland	3,673	514.476	65.901	1.026	1.660
Indonesia	4,774	397.186	78.412	3.889	3.922
Iran, Islamic Rep. of	5,411	426.209	96.459	3.076	3.147
Israel	3,908	516.439	97.702	3.203	3.429
Italy	3,581	551.490	73.744	3.147	3.269
Kuwait	3,958	340.428	108.051	3.509	3.659
Latvia	4,162	539.283	63.419	2.085	2.386
Lithuania	4,701	541.633	58.441	1.771	1.933
Luxembourg	5,101	554.897	68.090	0.802	0.954
Macedonia, Rep. of	4,002	438.603	97.225	3.574	3.735
Moldova, Rep. of	4,036	492.228	68.133	2.621	2.814
Morocco	3,249	317.357	116.430	6.240	6.452
Netherlands	4,156	544.552	56.522	1.636	1.837
New Zealand	6,256	527.324	86.488	2.017	2.059
Norway	3,837	501.131	66.508	2.464	2.508
Poland	4,854	523.138	77.809	2.263	2.482
Qatar	6,680	358.373	96.300	1.026	1.255
Romania	4,273	493.009	91.085	4.806	4.840
Russian Federation	4,720	561.032	69.422	3.192	3.297
Scotland	3,775	526.900	81.191	2.464	2.575
Singapore	6,390	551.518	80.283	2.904	2.915
Slovak Republic	5,380	533.326	74.230	2.773	2.864
Slovenia	5,337	519.435	68.958	1.977	2.032
South Africa	14,657	299.431	134.651	5.150	5.249
Spain	4,094	516.423	75.241	2.632	2.694
Sweden	4,394	546.026	61.406	2.169	2.256
Trinidad and Tobago	3,951	434.137	103.793	4.586	4.631
United States	5,190	540.658	77.645	3.434	3.571

Exhibit 12.8 Summary Statistics and Standard Errors in Reading Achievement for Informational Purposes

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Austria	5,067	536.131	64.668	2.216	2.309
Belgium (Flemish)	4,479	547.126	53.133	1.763	2.036
Belgium (French)	4,552	497.958	67.894	2.617	2.785
Bulgaria	3,863	549.828	82.797	4.272	4.355
Chinese Taipei	4,589	538.261	58.521	1.693	1.815
Denmark	4,001	541.709	71.717	2.298	2.407
England	4,036	537.069	84.067	2.383	2.530
France	4,404	526.076	66.505	1.985	2.110
Georgia	4,402	465.178	77.053	3.324	3.552
Germany	7,899	544.445	66.448	2.142	2.265
Hong Kong SAR	4,712	568.232	55.924	2.215	2.250
Hungary	4,068	541.154	70.292	2.953	3.081
Iceland	3,673	505.181	71.493	1.194	1.383
Indonesia	4,774	417.685	82.163	4.151	4.165
Iran, Islamic Rep. of	5,411	419.796	90.683	3.023	3.122
Israel	3,908	507.409	98.601	3.437	3.619
Italy	3,581	548.937	64.080	2.727	2.934
Kuwait	3,958	326.510	117.862	4.044	4.296
Latvia	4,162	539.895	62.530	2.247	2.390
Lithuania	4,701	529.879	54.480	1.597	1.628
Luxembourg	5,101	556.644	63.982	0.794	0.971
Macedonia, Rep. of	4,002	449.857	102.559	3.996	4.174
Moldova, Rep. of	4,036	508.045	70.362	3.022	3.042
Morocco	3,249	334.506	104.921	5.869	6.020
Netherlands	4,156	547.557	49.555	1.323	1.594
New Zealand	6,256	533.516	83.737	2.083	2.234
Norway	3,837	494.263	68.335	2.642	2.754
Poland	4,854	515.055	72.322	1.992	2.191
Qatar	6,680	356.046	93.687	0.968	1.621
Romania	4,273	487.202	88.408	4.929	4.943
Russian Federation	4,720	563.774	65.985	3.270	3.346
Scotland	3,775	526.952	77.890	2.448	2.556
Singapore	6,390	563.166	70.399	2.665	2.832
Slovak Republic	5,380	526.803	72.807	2.513	2.644
Slovenia	5,337	522.956	70.774	2.175	2.390
South Africa	14,657	315.626	131.904	5.083	5.150
Spain	4,094	508.187	67.542	2.415	2.889
Sweden	4,394	548.617	67.171	2.229	2.351
Trinidad and Tobago	3,951	440.119	99.288	4.341	4.586
United States	5,190	537.164	69.909	3.298	3.440

Exhibit 12.9 Summary Statistics and Standard Errors in Reading Achievement for Retrieving and Straightforward Inferencing Processes

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Austria	5,067	544.012	65.937	2.002	2.087
Belgium (Flemish)	4,479	544.562	59.380	1.671	1.920
Belgium (French)	4,552	501.166	70.840	2.488	2.637
Bulgaria	3,863	537.648	78.864	4.144	4.233
Chinese Taipei	4,589	540.923	67.730	1.915	1.961
Denmark	4,001	550.980	78.411	2.532	2.691
England	4,036	533.309	90.704	2.477	2.841
France	4,404	523.467	67.235	1.964	2.098
Georgia	4,402	477.963	73.262	3.267	3.320
Germany	7,899	554.563	71.815	2.110	2.624
Hong Kong SAR	4,712	557.528	59.249	2.432	2.515
Hungary	4,068	543.514	69.262	2.623	2.781
Iceland	3,673	516.355	72.863	1.125	1.227
Indonesia	4,774	409.457	77.640	3.854	3.927
Iran, Islamic Rep. of	5,411	427.870	96.146	3.204	3.294
Israel	3,908	507.349	94.658	3.000	3.216
Italy	3,581	544.103	69.523	2.768	2.816
Kuwait	3,958	336.978	106.949	3.303	3.865
Latvia	4,162	534.034	64.956	2.317	2.462
Lithuania	4,701	531.073	60.179	1.715	1.899
Luxembourg	5,101	565.086	72.780	0.849	1.205
Macedonia, Rep. of	4,002	445.981	97.680	3.787	3.830
Moldova, Rep. of	4,036	485.985	68.782	2.820	2.870
Morocco	3,249	336.209	103.833	6.014	6.170
Netherlands	4,156	551.212	60.907	1.619	2.036
New Zealand	6,256	523.595	86.261	2.148	2.269
Norway	3,837	501.977	71.976	2.246	2.291
Poland	4,854	515.977	75.800	2.198	2.356
Qatar	6,680	360.581	94.470	0.963	1.202
Romania	4,273	488.843	88.819	5.114	5.203
Russian Federation	4,720	562.323	70.091	3.223	3.438
Scotland	3,775	524.682	81.700	2.569	2.810
Singapore	6,390	560.224	84.587	3.227	3.293
Slovak Republic	5,380	529.011	74.858	2.697	2.754
Slovenia	5,337	518.658	72.106	1.972	2.063
South Africa	14,657	306.569	130.940	5.163	5.322
Spain	4,094	508.235	69.074	2.484	2.515
Sweden	4,394	550.238	68.608	2.226	2.360
Trinidad and Tobago	3,951	438.496	102.661	4.596	4.708
United States	5,190	532.155	78.000	3.312	3.339

Exhibit 12.10 Summary Statistics and Standard Errors in Reading Achievement for Interpreting, Integrating, and Evaluating Processes

Country	Sample Size	Mean Proficiency	Standard Deviation	Jackknife Sampling Error	Overall Standard Error
Austria	5,067	530.060	64.056	2.115	2.213
Belgium (Flemish)	4,479	547.098	53.047	1.770	1.819
Belgium (French)	4,552	496.888	66.771	2.445	2.460
Bulgaria	3,863	552.640	83.558	4.409	4.428
Chinese Taipei	4,589	529.729	62.136	1.775	1.858
Denmark	4,001	542.249	62.359	2.004	2.326
England	4,036	543.082	81.437	2.248	2.450
France	4,404	517.834	66.101	2.135	2.291
Georgia	4,402	461.308	80.216	3.403	3.539
Germany	7,899	540.149	65.042	2.096	2.162
Hong Kong SAR	4,712	565.539	58.882	2.308	2.436
Hungary	4,068	553.827	67.512	2.718	2.990
Iceland	3,673	503.032	65.802	1.062	1.267
Indonesia	4,774	404.170	80.212	3.956	4.133
Iran, Islamic Rep. of	5,411	417.701	92.501	3.200	3.280
Israel	3,908	516.149	97.497	3.357	3.569
Italy	3,581	555.668	64.610	2.775	2.852
Kuwait	3,958	329.859	113.084	3.761	3.953
Latvia	4,162	545.200	58.113	1.863	1.882
Lithuania	4,701	540.190	53.081	1.590	1.635
Luxembourg	5,101	548.282	62.905	0.754	0.888
Macedonia, Rep. of	4,002	439.069	104.679	3.914	4.028
Moldova, Rep. of	4,036	515.334	67.212	2.789	2.919
Morocco	3,249	312.993	116.086	6.430	6.552
Netherlands	4,156	542.283	50.528	1.328	1.496
New Zealand	6,256	537.930	81.422	2.072	2.182
Norway	3,837	494.934	65.567	2.165	2.415
Poland	4,854	521.798	72.491	2.150	2.290
Qatar	6,680	355.309	92.402	0.908	1.553
Romania	4,273	490.000	90.988	5.228	5.321
Russian Federation	4,720	562.554	66.192	3.205	3.248
Scotland	3,775	528.473	76.794	2.455	2.561
Singapore	6,390	555.562	69.400	2.672	2.705
Slovak Republic	5,380	531.238	71.335	2.755	2.791
Slovenia	5,337	523.322	66.245	1.916	1.959
South Africa	14,657	313.039	130.433	5.143	5.284
Spain	4,094	515.320	71.554	2.571	2.615
Sweden	4,394	546.476	62.141	2.040	2.187
Trinidad and Tobago	3,951	436.522	100.312	4.720	5.032
United States	5,190	545.830	67.134	3.204	3.331

12.4.4 Reporting Student Performance on Individual Items

To describe the PIRLS International Benchmarks, PIRLS provides several examples of achievement items from the assessment together with the percentages of students in each country responding correctly to or earning partial or full credit on the items. The basis for calculating these percentages was the total number of students that were administered the 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, it was the weighted percentage of students that achieved at least partial credit or full credit on the item. Omitted and not-reached items were treated as incorrect.

When the percent correct for example items was computed, student responses were classified in the following way.

For multiple-choice items, the responses to item *j* were classified as:

- Correct (C_j) when the correct option for an item was selected,
- Incorrect (W_j) when the incorrect option or no option at all was selected,
- Invalid (I_j) when two or more choices were made on the same question,
- 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 constructed-response items, student responses to item *j* were classified as:

- Correct (C_j) when the maximum number of points was obtained on the question,
- Incorrect (W_j) when the wrong answer or an answer not worth all the points in the question was given,
- Invalid (N_j) when the student's response was not legible or interpretable, or 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 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.

References

Brick, J.M., Morganstein, D., & Valliant, R. (2000). *Analysis of complex sample data using replication.* Rockville, MD: Westat.

Foy, P. & Kennedy, A.M. (Eds). (2008). *PIRLS 2006 user guide for the international database.* Chestnut Hill, MA: Boston College.

Gonzalez, E.J., Galia, J., Arora, A, Erberber, E., & Diaconu, D. (2004). Reporting student achievement in mathematics and science. In M.O. Martin, I.V.S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 technical report* (pp. 275-307). 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.

Mislevy, R.J., Beaton, A.E., Kaplan, B., & Sheehan, K.M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29,* 133-161.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., & Foy, P. (2007). *PIRLS 2006 international report: IEA's progress in international reading literacy study in primary schools in 40 countries*. Chestnut Hill, MA: Boston College.

Chapter 13

Reporting PIRLS 2006 Questionnaire Data

Kathleen L. Trong and Ann M. Kennedy

13.1 Overview

Through the *PIRLS 2006 International Report* (Mullis, Martin, Kennedy, & Foy, 2007), PIRLS strove to present factors related to teaching and learning reading helpful in understanding the reading achievement results. To describe the educational context for reading achievement, data on hundreds of background variables were collected from students, teachers, schools, parents, and ministries of education. This information was summarized in a concise manner to make it as accessible and useful as possible for policymakers, researchers, and educators. This chapter describes the procedures used to analyze these background data and create the indices reported in Chapters 3 through 7 of the *PIRLS 2006 International Report*. The description includes an explanation of initial exploratory analyses, reporting methods for individual and derived variables, and the review process for exhibits.

PIRLS background data were collected through the five background questionnaires used to gather information at various levels of the education system, as described in Chapter 3. These include:

- The *Student Questionnaire*, which collected information about students' literacy-related activities and resources in and outside of school.
- The *Learning to Read Survey* (home questionnaire), which collected information from parents about literacy-related activities and resources

at home, their attitudes toward reading, and their perceptions of their child's school.

- The *Teacher Questionnaire*, which collected information from teachers about the reading instruction in the classroom and the school as a whole, as well as information about teachers' background and training.
- The *School Questionnaire*, which collected information from school principals about schools' reading curriculum and policies, in addition to the schools' demographics and resources.
- The *Curriculum Questionnaire*, which collected information from National Research Coordinators (NRCs) about the nationally (or regionally) defined reading curriculum in primary schools.

Based on responses to the questions in these questionnaires and in line with the conceptual framework for contexts for learning to read described in the PIRLS *2006 Assessment Framework and Specifications* (Mullis, Kennedy, Martin, & Sainsbury, 2006), a subset of variables was selected for analysis and reporting. Often, several variables were intended to measure a single construct. For reporting, these variables were combined to form a single index variable.

13.2 Exploratory Analyses

Planning for reporting the questionnaire data began with a review of the questionnaires administered in PIRLS 2006 and the previous cycle, PIRLS 2001. Staff at the TIMSS & PIRLS International Study Center identified variables that had been used in 2001 to determine if trends could be measured, and if improvements in construct measurement could be made to indices developed in 2001 by adding new items from the PIRLS 2006 questionnaires. Newly developed variables were reviewed in the context of the PIRLS 2006 framework to identify variables for creating new indices.

Following this preliminary step, data almanacs consisting of statistical summaries of all background variables for the student, teacher, school, and home questionnaires were reviewed. These almanacs presented descriptive statistics such as, for categorical variables, the percentage of students in each category and mean reading achievement, and for continuous variables, the minimum, maximum, mean, mode, and percentile scores. For every variable, these statistics were presented separately for each country and averaged internationally. The data review allowed the TIMSS & PIRLS International Study Center to

examine the quality of the data, as well as identify data patterns for reporting purposes. If anything unusual was noted in the data, national versions of the questionnaires and national adaptations were revisited. If further clarification was needed, the National Research Coordinator (NRC) was contacted.

Because the Southern Hemisphere countries (New Zealand, Singapore, and South Africa) administered the assessment at the end of 2005, their data were available for use in exploratory analyses before the data from the Northern Hemisphere countries became available. These analyses had three primary purposes: identifying new indices that could be created from variables added in the 2006 cycle, ensuring that indices used in 2001 still performed similarly in 2006, and exploring the impact of improving indices created in PIRLS 2001 by adding an extra component variable. The exploratory analyses included principal components analyses to examine the dimensionality of proposed indices using different combinations of variables. In determining whether to add a variable to an index, analyses were conducted to ensure that any new variable was highly correlated with the other items used in the index, the variables of the modified index formed a single factor, and inclusion of a new variable would not cause any major fluctuations in the index distribution. If a variable did not meet these criteria, the index was left unchanged. At this stage, all analyses were conducted using SPSS 14.0 for Windows (SPSS Inc., 2005).

13.3 Reporting Background Data

The most straightforward way that PIRLS background data were presented was by simply reporting the percentage of students responding to each category of a variable, often accompanied by the mean reading achievement of the students in each category. This presented readers with a descriptive summary of how the responses were distributed within and across countries in a manner that is easy to understand and interpret. In some cases, response categories were collapsed.

13.4 Computing Derived Variables

In the PIRLS questionnaires, there were often several questions asked about various aspects of a single construct. In these cases, responses to the individual items were combined to create a derived variable that provided a more comprehensive picture of the construct of interest than the individual variables could on their own. In addition, these derived variables could be expected to

be more reliable, since random errors from individual variables tend to cancel each other out (DeVellis, 1991; Spector, 1992).

Student records were included in the derived variable calculation only if there were data available for two thirds of the variables involved. For example, if a derived variable was based on six component variables, students who were missing responses to more than two of these were counted as missing on the derived variable. Supplement 3 of the *PIRLS 2006 User Guide for the International Database* (Foy & Kennedy, 2008) provides a description of all derived variables included in the international database.

In the PIRLS reports, an index is a special type of derived variable that assigns students to one of three levels—high, medium, and low—on the basis of their responses to the component variables. The high category of an index represents the responses that are expected to characterize aspects of a positive literacy environment, and the low category those responses that are least supportive of literacy. The PIRLS indices are intended to describe factors related to the fostering of reading achievement in terms of the questions that were actually asked, thereby preserving a high degree of interpretability. For example, students at the high level of the PIRLS 2006 Index of Early Home Literacy Activities (described later in this chapter) had parents who reported often engaging with the student in each of six early literacy activities (read books, tell stories, sing songs, play with alphabet toys, play word games, and read aloud signs and labels) before the student began primary schooling. In contrast, students at the low level of this index had parents reporting never or almost never engaging the student in such activities.

In constructing an index, it was important that the component variables were intercorrelated so that together they formed a reliable scale, and also that they were correlated with student reading achievement. The process of identifying the response combinations that defined the high, medium, and low levels of the index also was informed by the relationship with achievement, but often these combinations were chosen based on a judgment of which responses could be expected to most effectively capture the construct's support for literacy or good practices.

13.5 Display of Background Data

PIRLS 2006 presented the background questionnaire data in Chapters 3–7 of the *PIRLS 2006 International Report*, with the first two chapters focusing on the student and parent questionnaires and the last three chapters utilizing data primarily from the teacher, school, and curriculum questionnaires.

In all of the exhibits, except those derived from the *Curriculum Questionnaire*, the student was the unit of analysis. In other words, data always were presented as the percentage of students possessing a particular characteristic, even if the information had been supplied by parents, teachers, or principals. This approach presents the data from the perspective of students' educational experiences and is consistent with the PIRLS sampling and assessment design. In many exhibits, the average reading achievement associated with the students in each category was also presented.

Exhibits generally were organized alphabetically by country, with an additional row showing the average internationally. However, in reporting some variables, including indices, countries were organized according to the percentage of students in the high category in descending order to help readers see the variation across countries more easily.

Since one of the major benefits of PIRLS is the ability to measure trends over time, data from PIRLS 2001 background questionnaires were included whenever possible. In these exhibits, the change from 2001 in percentage of students in variable categories was displayed for countries that participated in the 2001 assessment, with an arrow indicating if the percent in 2006 was significantly higher or lower. Several exhibits also focused on the differences between boys and girls, with arrows designating significant differences between the genders.

While most countries had very high response rates for the background questionnaires, in some cases the response rates were lower than acceptable. Because all of the data were presented with the student as the unit of analysis, this also was the way that response rates were calculated. The following special notations were used to convey information about response rates in the exhibits of the international report:

- An "r" next to the data indicates that responses were available for 70–84 percent of the students;

- An "s" next to the data indicates that responses were available for 50–69 percent of the students; and
- An "x" in place of the data indicates that responses were available for less than 50 percent of the students.

There also were other situations in which data was not shown. These were denoted in the following ways:

- When the percentage of students in a particular category was less than two percent, achievement data were replaced with by a tilde (~);
- When a country did not participate in the 2001 assessment, a diamond (◊) was shown in trend data columns; and
- When comparable data were not available for a particular country, a dash (-) was shown in the affected columns.

The absence of comparable data was usually because the country did not ask a particular question in one of the questionnaires. Most notably, there were no data available from parents in the United States, because that country did not administer the *Learning to Read Survey* and no data available from the *School Questionnaire* in Luxembourg because primary schools in Luxembourg do not have principals.

13.6 Summary of Background Indices

In the following section, the PIRLS 2006 indices presented in each background chapter of the international report are described. The composition of each index is briefly described, together with information about the reliability of the underlying scale (Cronbach's alpha) and its relationship to student reading achievement (the multiple correlation between the component variables of the index and achievement and the percent of variance in achievement accounted for by the component variables). While the creation of the indices relied heavily on judgments about desirable literacy environments, these statistics provide a sense of how well the component variables are related to one another and to reading achievement. When reviewing these exhibits, it is important to keep in mind that these indices are intended to act as international indicators. While within-country relationships were considered during development, index performance may vary due to the culturally embedded nature of these variables.

Chapter 3 of the international report focused on literacy-related activities in the home, including information on parents' background and attitudes, home resources, and activities parents have done with their child.

The Index of Early Home Literacy Activities (EHLA) attempts to categorize students according to their parents' reports about engaging in early literacy activities with the students before they began primary school. The index is presented in Exhibit 3.1 of the international report and also was reported in 2001. It is based on parents' reports of the frequency with which they engage with their child in the following activities 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, and read aloud signs and labels. An average was computed across the six items based on a 3-point scale: *never or almost never* = 1, *sometimes* = 2, and *often* = 3. A high level indicates an average score of greater than 2.33 through 3. A medium level indicates an average score of 1.67 through 2.33. A low level indicates an average score of 1 to less than 1.67.

As shown in Exhibit 13.1, the six activities form a fairly reliable scale, with a median reliability coefficient (Cronbach's alpha) across countries of 0.68. The median multiple correlation between the six activities and student achievement was 0.28, corresponding to an R-square of 0.08.

The Index of Home Educational Resources (HER) is intended to summarize the students' and parents' reports about aspects of the home environment and the extent to which it supports literacy. Presented in Exhibit 3.2 of the international report, this index, also reported in 2001, is 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: the number of children's books in the home and parents' education. A high level indicates more than 100 books in the home, more than 25 children's books, at least 3 of 4 educational aids, and at least one parent finished university. A low level indicates 25 or fewer books in the home, 25 or fewer children's books, no more than 2 educational aids, and parents that have not completed secondary education. A medium level includes all other combinations of responses.

Exhibit 13.2 shows that the component variables form a reliable scale, with a median reliability coefficient (Cronbach's alpha) across countries of 0.61. The median multiple correlation between the component variables and student achievement was 0.44, corresponding to an R-square of 0.19.

Exhibit 13.1 Index of Early Home Literacy Activities (EHLA) Statistics

Country	Cronbach's Alpha Between the Component Variables	Multiple R Between Student Reading Achievement and Component Variables	Percent of Variance in Student Reading Achievement Accounted for by the Component Variables
Austria	0.64	0.35	0.12
Belgium (Flemish)	0.67	0.30	0.09
Belgium (French)	0.64	0.35	0.12
Bulgaria	0.79	0.26	0.07
Canada, Alberta	0.74	0.26	0.07
Canada, British Columbia	0.75	0.24	0.06
Canada, Nova Scotia	0.73	0.26	0.07
Canada, Ontario	0.73	0.22	0.05
Canada, Quebec	0.68	0.30	0.09
Chinese Taipei	0.74	0.35	0.12
Denmark	0.66	0.28	0.08
England	0.72	0.33	0.11
France	0.63	0.33	0.11
Georgia	0.70	0.20	0.04
Germany	0.61	0.33	0.11
Hong Kong SAR	0.73	0.20	0.04
Hungary	0.63	0.26	0.07
Iceland	0.69	0.28	0.08
Indonesia	0.73	0.20	0.04
Iran, Islamic Rep. of	0.74	0.35	0.12
Israel	0.70	0.10	0.01
Italy	0.60	0.24	0.06
Kuwait	0.66	0.20	0.04
Latvia	0.62	0.22	0.05
Lithuania	0.64	0.26	0.07
Luxembourg	0.69	0.35	0.12
Macedonia, Rep. of	0.69	0.24	0.06
Moldova, Rep. of	0.69	0.26	0.07
Morocco	0.73	0.22	0.05
Netherlands	0.66	0.28	0.08
New Zealand	0.77	0.32	0.10
Norway	0.65	0.26	0.07
Poland	0.63	0.32	0.10
Qatar	0.62	0.17	0.03
Romania	0.78	0.44	0.19
Russian Federation	0.68	0.28	0.08
Scotland	0.72	0.26	0.07
Singapore	0.79	0.32	0.10
Slovak Republic	0.61	0.36	0.13
Slovenia	0.67	0.28	0.08
South Africa	0.59	0.22	0.05
Spain	0.64	0.32	0.10
Sweden	0.69	0.26	0.07
Trinidad and Tobago	0.73	0.35	0.12
United States	-	-	-
International Median	0.68	0.28	0.08

A dash (–) indicates data are not available.

Exhibit 13.2 Index of Home Educational Resources (HER) Statistics

Country	Cronbach's Alpha Between the Component Variables	Multiple R Between Student Reading Achievement and Component Variables	Percent of Variance in Student Reading Achievement Accounted for by the Component Variables
Austria	0.6	0.49	0.24
Belgium (Flemish)	0.57	0.45	0.20
Belgium (French)	0.61	0.46	0.21
Bulgaria	0.78	0.41	0.17
Canada, Alberta	0.48	0.35	0.12
Canada, British Columbia	0.51	0.33	0.11
Canada, Nova Scotia	0.5	0.40	0.16
Canada, Ontario	0.5	0.36	0.13
Canada, Quebec	0.55	0.37	0.14
Chinese Taipei	0.66	0.46	0.21
Denmark	0.55	0.33	0.11
England	-	-	-
France	0.65	0.46	0.21
Georgia	0.67	0.32	0.10
Germany	0.65	0.51	0.26
Hong Kong SAR	0.68	0.26	0.07
Hungary	0.72	0.56	0.31
Iceland	0.46	0.37	0.14
Indonesia	0.45	0.36	0.13
Iran, Islamic Rep. of	0.79	0.51	0.26
Israel	0.57	0.48	0.23
Italy	0.62	0.32	0.10
Kuwait	0.37	0.33	0.11
Latvia	0.57	0.39	0.15
Lithuania	0.67	0.42	0.18
Luxembourg	0.68	0.51	0.26
Macedonia, Rep. of	0.63	0.48	0.23
Moldova, Rep. of	0.6	0.32	0.10
Morocco	0.66	0.32	0.10
Netherlands	0.6	0.40	0.16
New Zealand	0.53	0.45	0.20
Norway	0.53	0.39	0.15
Poland	0.65	0.47	0.22
Qatar	0.38	0.28	0.08
Romania	0.74	0.51	0.26
Russian Federation	0.61	0.40	0.16
Scotland	0.6	0.44	0.19
Singapore	0.62	0.52	0.27
Slovak Republic	0.7	0.53	0.28
Slovenia	0.61	0.44	0.19
South Africa	0.57	0.57	0.33
Spain	0.62	0.41	0.17
Sweden	0.54	0.41	0.17
Trinidad and Tobago	0.53	0.42	0.18
United States	-	-	-
International Median	0.61	0.44	0.19

A dash (–) indicates data are not available.

The Index of Parents' Attitudes Toward Reading (PATR) groups students according to their parents' reports of their own preferences for reading. This index, developed originally for PIRLS 2001, is presented in Exhibit 3.10 of the international report. The index is based on parents' agreement with the following statements: 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. An average was computed across the five 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. A high level indicates an average of greater than 3 through 4. A medium level indicates an average of 2 through 3. A low level indicates an average of 1 to less than 2.

As shown in Exhibit 13.3, the five statements form a reliable scale, with a median reliability coefficient (Cronbach's alpha) across countries of 0.78. The median multiple correlation between the five statements and student achievement was 0.24, corresponding to an R-square of 0.06.

Chapter 4 of the international report presented students' reports on reading attitudes, self-concept, and out-of-school activities.

The Index of Students' Attitudes Toward Reading (SATR) categorizes students according to their own reading preferences. The index was developed in 2001 and is presented in Exhibit 4.1 of the international report. This index is based on students' agreement with the following statements: 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. An average was 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 statements were reverse-coded. A high level indicates an average of greater than 3 through 4. A medium level indicates an average of 2 through 3. A low level indicates an average of 1 to less than 2.

As shown in Exhibit 13.4, the component variables form a reliable scale, with a median reliability coefficient (Cronbach's alpha) across countries of 0.68. The median multiple correlation between the component variables and student achievement was 0.39, corresponding to an R-square of 0.15.

Exhibit 13.3 Index of Parents' Attitudes Toward Reading (PATR) Statistics

Country	Cronbach's Alpha Between the Component Variables	Multiple R Between Student Reading Achievement and Component Variables	Percent of Variance in Student Reading Achievement Accounted for by the Component Variables
Austria	0.86	0.30	0.09
Belgium (Flemish)	0.87	0.28	0.08
Belgium (French)	0.84	0.28	0.08
Bulgaria	0.84	0.28	0.08
Canada, Alberta	0.85	0.22	0.05
Canada, British Columbia	0.82	0.22	0.05
Canada, Nova Scotia	0.86	0.24	0.06
Canada, Ontario	0.82	0.20	0.04
Canada, Quebec	0.87	0.24	0.06
Chinese Taipei	0.72	0.17	0.03
Denmark	0.86	0.24	0.06
England	0.83	0.26	0.07
France	0.79	0.28	0.08
Georgia	0.64	0.22	0.05
Germany	0.82	0.32	0.10
Hong Kong SAR	0.65	0.10	0.01
Hungary	0.78	0.32	0.10
Iceland	0.83	0.20	0.04
Indonesia	0.58	0.17	0.03
Iran, Islamic Rep. of	0.63	0.22	0.05
Israel	0.72	0.35	0.12
Italy	0.82	0.28	0.08
Kuwait	0.71	0.14	0.02
Latvia	0.75	0.17	0.03
Lithuania	0.75	0.24	0.06
Luxembourg	0.81	0.32	0.10
Macedonia, Rep. of	0.67	0.40	0.16
Moldova, Rep. of	0.62	0.17	0.03
Morocco	0.59	0.14	0.02
Netherlands	0.84	0.30	0.09
New Zealand	0.83	0.28	0.08
Norway	0.84	0.22	0.05
Poland	0.78	0.26	0.07
Qatar	0.71	0.17	0.03
Romania	0.77	0.39	0.15
Russian Federation	0.75	0.20	0.04
Scotland	0.85	0.22	0.05
Singapore	0.72	0.22	0.05
Slovak Republic	0.80	0.36	0.13
Slovenia	0.79	0.28	0.08
South Africa	0.51	0.39	0.15
Spain	0.80	0.24	0.06
Sweden	0.84	0.24	0.06
Trinidad and Tobago	0.72	0.24	0.06
United States	-	-	-
International Median	0.78	0.24	0.06

A dash (–) indicates data are not available.

Exhibit 13.4 Index of Students' Attitudes Toward Reading (SATR) Statistics

Country	Cronbach's Alpha Between the Component Variables	Multiple R Between Student Reading Achievement and Component Variables	Percent of Variance in Student Reading Achievement Accounted for by the Component Variables
Austria	0.76	0.40	0.16
Belgium (Flemish)	0.76	0.36	0.13
Belgium (French)	0.66	0.42	0.18
Bulgaria	0.70	0.35	0.12
Canada, Alberta	0.76	0.42	0.18
Canada, British Columbia	0.75	0.41	0.17
Canada, Nova Scotia	0.78	0.40	0.16
Canada, Ontario	0.76	0.36	0.13
Canada, Quebec	0.72	0.42	0.18
Chinese Taipei	0.63	0.32	0.10
Denmark	0.76	0.33	0.11
England	0.78	0.40	0.16
France	0.66	0.42	0.18
Georgia	0.41	0.32	0.10
Germany	0.76	0.40	0.16
Hong Kong SAR	0.60	0.35	0.12
Hungary	0.77	0.39	0.15
Iceland	0.64	0.40	0.16
Indonesia	0.28	0.41	0.17
Iran, Islamic Rep. of	0.66	0.36	0.13
Israel	0.61	0.40	0.16
Italy	0.69	0.32	0.10
Kuwait	0.38	0.37	0.14
Latvia	0.72	0.37	0.14
Lithuania	0.69	0.39	0.15
Luxembourg	0.79	0.36	0.13
Macedonia, Rep. of	0.47	0.51	0.26
Moldova, Rep. of	0.56	0.26	0.07
Morocco	0.40	0.35	0.12
Netherlands	0.78	0.37	0.14
New Zealand	0.67	0.49	0.24
Norway	0.71	0.40	0.16
Poland	0.72	0.41	0.17
Qatar	0.43	0.39	0.15
Romania	0.63	0.35	0.12
Russian Federation	0.63	0.39	0.15
Scotland	0.75	0.42	0.18
Singapore	0.71	0.41	0.17
Slovak Republic	0.72	0.37	0.14
Slovenia	0.73	0.42	0.18
South Africa	0.34	0.37	0.14
Spain	0.64	0.32	0.10
Sweden	0.79	0.41	0.17
Trinidad and Tobago	0.56	0.39	0.15
United States	0.73	0.39	0.15
International Median	0.68	0.39	0.15

The Index of Students' Reading Self-Concept (SRSC) groups students by their perceptions of their own reading competencies. This index, reported in Exhibit 4.2 of the international report, was a slightly modified version of the index developed in 2001. The index is based on students' responses to the following statements: reading is very easy for me, I do not read as well as other students in my class, when I am reading by myself I understand almost everything I read, and I read slower than other students in my class. An average was 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 statements were reverse-coded. A high level indicates an average of greater than 3 through 4. A medium level indicates an average of 2 through 3. A low level indicates an average of 1 to less than 2. The statement "I read slower than other students in my class" is a new variable added to the index in PIRLS 2006, and was not a part of the PIRLS 2001 index calculations.

Exhibit 13.5 presents the statistics for the four component variables, which form a reliable scale, with a median reliability coefficient (Cronbach's alpha) across countries of 0.60. The median multiple correlation between the four statements and student achievement was 0.40, corresponding to an R-square of 0.16.

Chapter 5 of the international report describes the school curriculum for reading and organization for teaching reading. This includes reports of instructional time, class size, and the availability of specialists. This chapter did not include any indices. Chapter 6 focused on teachers and reading instruction, and presented information about teachers' backgrounds and use of various instructional techniques, resources, and assessment.

The Index of Reading for Homework (RFH) is a unique index for two reasons. First, it is comprised of only two variables. Second, the categories for grouping students are sensitive to differences across countries in the role of homework in reading instruction. The index is presented in Exhibit 6.23 of the international report, and was developed in 2001. Students were categorized according to 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? A high level indicates students are expected to spend more than 30 minutes at least 1–2 times a week. A low level indicates students are never assigned homework or are expected to spend no more than 30 minutes less than once a week. A medium level indicates all other combinations of the frequencies.

Exhibit 13.5 Index of Students' Reading Self-Concept (SRSC) Statistics

Country	Cronbach's Alpha Between the Component Variables	Multiple R Between Student Reading Achievement and Component Variables	Percent of Variance in Student Reading Achievement Accounted for by the Component Variables
Austria	0.64	0.35	0.12
Belgium (Flemish)	0.71	0.39	0.15
Belgium (French)	0.56	0.36	0.13
Bulgaria	0.68	0.37	0.14
Canada, Alberta	0.67	0.41	0.17
Canada, British Columbia	0.67	0.42	0.18
Canada, Nova Scotia	0.67	0.45	0.20
Canada, Ontario	0.66	0.42	0.18
Canada, Quebec	0.67	0.48	0.23
Chinese Taipei	0.56	0.40	0.16
Denmark	0.74	0.55	0.30
England	0.70	0.46	0.21
France	0.60	0.40	0.16
Georgia	0.56	0.41	0.17
Germany	0.62	0.42	0.18
Hong Kong SAR	0.57	0.41	0.17
Hungary	0.65	0.44	0.19
Iceland	0.66	0.46	0.21
Indonesia	0.28	0.36	0.13
Iran, Islamic Rep. of	0.45	0.46	0.21
Israel	0.50	0.39	0.15
Italy	0.53	0.35	0.12
Kuwait	0.31	0.35	0.12
Latvia	0.59	0.42	0.18
Lithuania	0.61	0.40	0.16
Luxembourg	0.68	0.47	0.22
Macedonia, Rep. of	0.49	0.49	0.24
Moldova, Rep. of	0.43	0.28	0.08
Morocco	0.30	0.26	0.07
Netherlands	0.73	0.36	0.13
New Zealand	0.59	0.45	0.20
Norway	0.67	0.42	0.18
Poland	0.70	0.50	0.25
Qatar	0.51	0.53	0.28
Romania	0.63	0.41	0.17
Russian Federation	0.55	0.33	0.11
Scotland	0.66	0.40	0.16
Singapore	0.60	0.39	0.15
Slovak Republic	0.65	0.45	0.20
Slovenia	0.66	0.47	0.22
South Africa	0.35	0.37	0.14
Spain	0.39	0.39	0.15
Sweden	0.71	0.46	0.21
Trinidad and Tobago	0.57	0.49	0.24
United States	0.65	0.39	0.15
International Median	0.60	0.40	0.16

As shown in Exhibit 13.6, the variables comprising this index have relatively lower reliability (an international median Cronbach's alpha of 0.28) and a weaker relationship to achievement (an international median multiple R and R-square of 0.0), as compared to other indices. These statistics suggest that while homework is an important part of instruction in many countries, often students receiving the greatest amounts of homework or spending most time on it may be those who do not read as well as other students.

Chapter 7 focused on school contexts such as schools' locations and resources and measures of school climate and safety.

The Index of Availability of School Resources (ASR) categorized students according to their principals' reports of the extent to which their schools were impacted by a lack of resources. The index, modified from the 2001 index, is presented in Exhibit 7.5 of the international report. This index is based on principals' reports of how much the school's capacity to provide instruction is affected by a shortage or inadequacy of the following: qualified teaching staff, teachers with a specialization in reading, second language teachers, 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 audio-visual resources. An average was computed based on a 4-point scale: *a lot* = 1, *some* = 2, *a little* = 3, and *not at all* = 4. Responses for each activity were averaged across each principal. A high level indicates an average of greater than 3 through 4. A medium level indicates an average of 2 through 3. A low level indicates an average of 1 to less than 2. "Second language teachers" was added to the PIRLS 2006 index, and is not included in the 2001 index calculations. "Teachers with a specialization in reading" was worded as "teachers qualified to teach reading" in 2001.

As shown in Exhibit 13.7, the component variables form a reliable scale, with a median reliability coefficient (Cronbach's alpha) across countries of 0.85. The median multiple correlation between the component variables and student achievement was 0.17, corresponding to an R-square of 0.03.

Exhibit 13.6 Index of Reading for Homework (RFH) Statistics

Country	Cronbach's Alpha Between the Component Variables	Multiple R Between Student Reading Achievement and Component Variables	Percent of Variance in Student Reading Achievement Accounted for by the Component Variables
Austria	-0.37	0.00	0.00
Belgium (Flemish)	0.18	0.10	0.01
Belgium (French)	0.22	0.10	0.01
Bulgaria	0.35	0.14	0.02
Canada, Alberta	0.45	0.00	0.00
Canada, British Columbia	0.44	0.00	0.00
Canada, Nova Scotia	0.38	0.10	0.01
Canada, Ontario	0.47	0.00	0.00
Canada, Quebec	0.04	0.00	0.00
Chinese Taipei	-0.07	0.00	0.00
Denmark	0.18	0.00	0.00
England	0.36	0.00	0.00
France	0.12	0.00	0.00
Georgia	0.48	0.10	0.01
Germany	0.05	0.00	0.00
Hong Kong SAR	0.35	0.14	0.02
Hungary	0.28	0.00	0.00
Iceland	0.33	0.00	0.00
Indonesia	0.22	0.00	0.00
Iran, Islamic Rep. of	0.44	0.00	0.00
Israel	0.17	0.10	0.01
Italy	0.14	0.00	0.00
Kuwait	-	-	-
Latvia	0.19	0.00	0.00
Lithuania	0.11	0.00	0.00
Luxembourg	0.43	0.00	0.00
Macedonia, Rep. of	0.56	0.10	0.01
Moldova, Rep. of	-0.17	0.00	0.00
Morocco	0.07	0.17	0.03
Netherlands	0.53	0.00	0.00
New Zealand	0.25	0.10	0.01
Norway	0.23	0.00	0.00
Poland	0.43	0.00	0.00
Qatar	0.34	0.10	0.01
Romania	0.41	0.00	0.00
Russian Federation	-0.16	0.00	0.00
Scotland	0.11	0.00	0.00
Singapore	0.63	0.10	0.01
Slovak Republic	0.38	0.00	0.00
Slovenia	0.32	0.00	0.00
South Africa	0.34	0.17	0.03
Spain	0.43	0.10	0.01
Sweden	0.25	0.00	0.00
Trinidad and Tobago	0.42	0.10	0.01
United States	0.54	0.10	0.01
International Median	0.28	0.00	0.00

A dash (–) indicates data are not available.

Exhibit 13.7 Index of Availability of School Resources (ASR) Statistics

Country	Cronbach's Alpha Between the Component Variables	Multiple R Between Student Reading Achievement and Component Variables	Percent of Variance in Student Reading Achievement Accounted for by the Component Variables
Austria	0.80	0.17	0.03
Belgium (Flemish)	0.88	0.14	0.02
Belgium (French)	0.80	0.14	0.02
Bulgaria	0.89	0.28	0.08
Canada, Alberta	0.87	0.14	0.02
Canada, British Columbia	0.84	0.17	0.03
Canada, Nova Scotia	0.90	0.14	0.02
Canada, Ontario	0.91	0.17	0.03
Canada, Quebec	0.87	0.17	0.03
Chinese Taipei	0.95	0.10	0.01
Denmark	0.79	0.14	0.02
England	0.85	0.14	0.02
France	0.73	0.17	0.03
Georgia	0.83	0.17	0.03
Germany	0.81	0.24	0.06
Hong Kong SAR	0.89	0.17	0.03
Hungary	-	-	-
Iceland	0.82	0.14	0.02
Indonesia	0.85	0.32	0.10
Iran, Islamic Rep. of	0.87	0.32	0.10
Israel	0.91	0.35	0.12
Italy	0.84	0.14	0.02
Kuwait	0.85	0.17	0.03
Latvia	0.92	0.20	0.04
Lithuania	0.89	0.17	0.03
Luxembourg	-	-	-
Macedonia, Rep. of	0.83	0.35	0.12
Moldova, Rep. of	0.78	0.17	0.03
Morocco	0.92	0.24	0.06
Netherlands	0.81	0.22	0.05
New Zealand	0.88	0.10	0.01
Norway	0.78	0.10	0.01
Poland	0.85	0.14	0.02
Qatar	0.92	0.22	0.05
Romania	0.89	0.28	0.08
Russian Federation	0.92	0.17	0.03
Scotland	0.82	0.10	0.01
Singapore	0.95	0.14	0.02
Slovak Republic	0.77	0.20	0.04
Slovenia	0.86	0.10	0.01
South Africa	0.82	0.46	0.21
Spain	0.92	0.20	0.04
Sweden	0.88	0.10	0.01
Trinidad and Tobago	0.80	0.22	0.05
United States	0.90	0.20	0.04
International Median	0.85	0.17	0.03

A dash (–) indicates data are not available.

The Index of Home-School Involvement (HSI) groups students according to principals' reports of the activities offered by their schools and parents' involvement in school activities. The exhibit, developed in 2001, is presented in Exhibit 7.8 of the international report. This index is based on principals' responses to questions about how often they hold parent-teacher conferences and communicate with parents about students' progress, and on parents' responses to questions about how often they attend meetings and events organized by the school. A high level indicates that four or more times a year, schools hold teacher-parent conferences and events at school that are attended by more than half of the parents, send home letters, calendars, newsletters, etc., with information about the school seven or more times a year, and send written reports (report cards) of child's performance four or more times a year. A low level indicates schools never hold teacher-parent conferences, or if they do, only between 0–25% of parents attend; schools hold events to which parents are invited once a year or less, to which 0–25% of parents attend; letters, calendars, newsletters, etc., are sent home three times a year or less; and written reports of children's performance are sent home once a year or less. A medium level indicates all other combinations.

Exhibit 13.8 presents the statistics for the component variables. The median reliability coefficient (Cronbach's alpha) across countries was 0.50. The median multiple correlation between the component variables and student achievement was 0.14, corresponding to an R-square of 0.02.

Exhibit 13.8 Index of Home-School Involvement (HSI) Statistics

Country	Cronbach's Alpha Between the Component Variables	Multiple R Between Student Reading Achievement and Component Variables	Percent of Variance in Student Reading Achievement Accounted for by the Component Variables
Austria	0.33	0.14	0.02
Belgium (Flemish)	0.38	0.10	0.01
Belgium (French)	0.31	0.14	0.02
Bulgaria	0.48	0.17	0.03
Canada, Alberta	0.26	0.14	0.02
Canada, British Columbia	0.43	0.14	0.02
Canada, Nova Scotia	0.19	0.10	0.01
Canada, Ontario	0.08	0.10	0.01
Canada, Quebec	0.44	0.17	0.03
Chinese Taipei	0.66	0.10	0.01
Denmark	0.25	0.10	0.01
England	0.34	0.17	0.03
France	0.39	0.17	0.03
Georgia	0.58	0.17	0.03
Germany	0.51	0.20	0.04
Hong Kong SAR	0.48	0.10	0.01
Hungary	0.53	0.22	0.05
Iceland	0.40	0.00	0.00
Indonesia	0.67	0.20	0.04
Iran, Islamic Rep. of	0.68	0.28	0.08
Israel	0.63	0.53	0.28
Italy	0.41	0.14	0.02
Kuwait	0.48	0.10	0.01
Latvia	0.52	0.14	0.02
Lithuania	0.52	0.14	0.02
Luxembourg	-	-	-
Macedonia, Rep. of	0.68	0.32	0.10
Moldova, Rep. of	0.55	0.10	0.01
Morocco	0.76	0.22	0.05
Netherlands	0.33	0.10	0.01
New Zealand	0.45	0.14	0.02
Norway	0.41	0.10	0.01
Poland	0.53	0.14	0.02
Qatar	0.76	0.14	0.02
Romania	0.59	0.22	0.05
Russian Federation	0.50	0.14	0.02
Scotland	0.43	0.10	0.01
Singapore	0.55	0.10	0.01
Slovak Republic	0.61	0.26	0.07
Slovenia	0.41	0.10	0.01
South Africa	0.66	0.45	0.20
Spain	0.50	0.20	0.04
Sweden	0.37	0.10	0.01
Trinidad and Tobago	0.54	0.28	0.08
United States	0.42	0.17	0.03
International Median	0.50	0.14	0.02

A dash (–) indicates data are not available.

The Index of Principals' Perceptions of School Climate (PPSC) categorizes students according to principals' perceptions of various factors related to the social climate of the school. The exhibit was modified from that in 2001, and is presented in Exhibit 7.12 of the international report. The index is based on principals' characterization of the following: teachers' job satisfaction, teachers' expectations for student achievement, parental support for student achievement, students' regard for school property, students' desire to do well in school, and students' regard for each other's welfare. An average was computed on a 5-point scale: *very low* = 1, *low* = 2, *medium* = 3, *high* = 4, and *very high* = 5. Responses for each activity were averaged across each principal. A high level indicates an average of greater than 3.67 through 5. A medium level indicates an average of 2.33 through 3.67. A low level indicates an average of 1 to less than 2.33. "Students' regard for each other's welfare" was added to the index in PIRLS 2006 and is not included in the 2001 index calculations.

As shown in Exhibit 13.9, the six variables form a reliable scale, with a median reliability coefficient (Cronbach's alpha) across countries of 0.79. The median multiple correlation between the six variables and student achievement was 0.20, corresponding to an R-square of 0.04.

The Index of Teacher Career Satisfaction (TCS) attempts to group students according to their teachers' reports of satisfaction with their current position and career choice as a whole. Developed in 2006, the index is presented in Exhibit 7.13 of the international report. The index is based on teachers' agreement with the following statements: I am content with my profession as a teacher, I am satisfied with being a teacher at this school, I would describe the teachers at this school as a satisfied group, I had more enthusiasm when I began teaching than I have now, and I do important work as a teacher. An average was computed across the five items based on a 4-point scale: *disagree a lot* = 1, *disagree a little* = 2, *agree a little* = 3, *agree a lot* = 4. Responses for negative statements were reverse-coded. A high level indicates an average of 3 through 4. A medium level indicates an average of 2 to less than 3. A low level indicates an average of 1 to less than 2.

As shown in Exhibit 13.10, the six statements form a fairly reliable scale, with a median reliability coefficient (Cronbach's alpha) across countries of 0.60. The median multiple correlation between the six statements and student achievement was 0.10, corresponding to an R-square of 0.01.

Exhibit 13.9 Index of Principals' Perceptions of School Climate (PPSC) Statistics

Country	Cronbach's Alpha Between the Component Variables	Multiple R Between Student Reading Achievement and Component Variables	Percent of Variance in Student Reading Achievement Accounted for by the Component Variables
Austria	0.72	0.22	0.05
Belgium (Flemish)	0.68	0.14	0.02
Belgium (French)	0.92	0.17	0.03
Bulgaria	0.80	0.24	0.06
Canada, Alberta	0.85	0.24	0.06
Canada, British Columbia	0.85	0.22	0.05
Canada, Nova Scotia	0.86	0.14	0.02
Canada, Ontario	0.86	0.14	0.02
Canada, Quebec	0.83	0.20	0.04
Chinese Taipei	0.79	0.10	0.01
Denmark	0.80	0.14	0.02
England	0.85	0.26	0.07
France	0.80	0.22	0.05
Georgia	0.79	0.14	0.02
Germany	0.74	0.30	0.09
Hong Kong SAR	0.85	0.10	0.01
Hungary	0.81	0.28	0.08
Iceland	0.92	0.00	0.00
Indonesia	0.73	0.10	0.01
Iran, Islamic Rep. of	0.74	0.28	0.08
Israel	0.79	0.30	0.09
Italy	0.76	0.14	0.02
Kuwait	0.74	0.17	0.03
Latvia	0.76	0.14	0.02
Lithuania	0.68	0.17	0.03
Luxembourg	-	-	-
Macedonia, Rep. of	0.74	0.35	0.12
Moldova, Rep. of	0.66	0.14	0.02
Morocco	0.87	0.26	0.07
Netherlands	0.72	0.22	0.05
New Zealand	0.88	0.26	0.07
Norway	0.73	0.10	0.01
Poland	0.77	0.10	0.01
Qatar	0.81	0.22	0.05
Romania	0.82	0.32	0.10
Russian Federation	0.74	0.20	0.04
Scotland	0.85	0.17	0.03
Singapore	0.80	0.20	0.04
Slovak Republic	0.77	0.26	0.07
Slovenia	0.76	0.10	0.01
South Africa	0.84	0.22	0.05
Spain	0.85	0.24	0.06
Sweden	0.71	0.17	0.03
Trinidad and Tobago	0.80	0.36	0.13
United States	0.85	0.24	0.06
International Median	0.79	0.20	0.04

A dash (–) indicates data are not available.

Exhibit 13.10 Index of Teacher Career Satisfaction (TCS) Statistics

Country	Cronbach's Alpha Between the Component Variables	Multiple R Between Student Reading Achievement and Component Variables	Percent of Variance in Student Reading Achievement Accounted for by the Component Variables
Austria	0.59	0.10	0.01
Belgium (Flemish)	0.69	0.10	0.01
Belgium (French)	0.59	0.10	0.01
Bulgaria	0.62	0.17	0.03
Canada, Alberta	0.71	0.10	0.01
Canada, British Columbia	0.68	0.10	0.01
Canada, Nova Scotia	0.62	0.10	0.01
Canada, Ontario	0.57	0.10	0.01
Canada, Quebec	0.70	0.10	0.01
Chinese Taipei	0.64	0.00	0.00
Denmark	0.67	0.10	0.01
England	0.75	0.17	0.03
France	0.69	0.10	0.01
Georgia	0.41	0.10	0.01
Germany	0.64	0.10	0.01
Hong Kong SAR	0.62	0.14	0.02
Hungary	0.65	0.14	0.02
Iceland	0.57	0.10	0.01
Indonesia	0.36	0.14	0.02
Iran, Islamic Rep. of	0.44	0.22	0.05
Israel	0.53	0.22	0.05
Italy	0.69	0.10	0.01
Kuwait	0.40	0.00	0.00
Latvia	0.61	0.00	0.00
Lithuania	0.52	0.00	0.00
Luxembourg	0.68	0.00	0.00
Macedonia, Rep. of	0.54	0.30	0.09
Moldova, Rep. of	0.44	0.00	0.00
Morocco	0.66	0.17	0.03
Netherlands	0.66	0.10	0.01
New Zealand	0.65	0.00	0.00
Norway	0.54	0.14	0.02
Poland	0.55	0.00	0.00
Qatar	0.60	0.10	0.01
Romania	0.48	0.14	0.02
Russian Federation	0.57	0.10	0.01
Scotland	0.69	0.00	0.00
Singapore	0.64	0.00	0.00
Slovak Republic	0.59	0.00	0.00
Slovenia	0.59	0.00	0.00
South Africa	0.64	0.33	0.11
Spain	0.51	0.10	0.01
Sweden	0.69	0.00	0.00
Trinidad and Tobago	0.59	0.14	0.02
United States	0.59	0.00	0.00
International Median	0.60	0.10	0.01

The Index of Parents' Perceptions of School Environment (PPSE) attempts to categorize students according to their parents' perceptions of the schools' efforts to provide a supportive learning environment. Newly developed in 2006, the index is presented in Exhibit 7.14 of the international report. The index is based on parents' agreement with the following statements: my child's school includes me in my child's education, my child's school should make a greater effort to include me in my child's education, my child's school cares about my child's progress in school, and my child's school does a good job in helping my child become better in reading. An average was computed across the four items based on a 4-point scale: *disagree a lot* = 1, *disagree a little* = 2, *agree a little* = 3, *agree a lot* = 4. Responses for negative statements were reverse-coded. A high level indicates an average of greater than 3 through 4. A medium level indicates an average of 2 through 3. A low level indicates an average of 1 to less than 2.

As shown in Exhibit 13.11, the reliability of this index, although quite high in many countries (Cronbach's alpha is above 0.75 in 12 countries), is low in some countries also. This suggests that some component variables may have different connotations in different contexts. For instance, parents may expect to be involved in their child's school to varying degrees in different countries. Therefore, their responses to that item may not coincide with other responses in the index, decreasing the overall reliability in some countries. The median multiple correlation between the component variables and student achievement was 0.17, corresponding to an R-square of 0.03.

The Index of Student Safety in School (SSS) groups students according to their perception of safety at school and their reports of incidents affecting safety. The index was developed for PIRLS 2006 and is presented in Exhibit 7.15 of the international report. This index is based on students' agreement with the statement "I feel safe when I am at school" and reports of stealing, bullying and injury happening to the students themselves or someone in their class in the last month. A high level indicates students agree a little or a lot with feeling safe at school, had one or fewer incidents happen to them, and had one or fewer incidents happen to someone in their class in the last month. A low level indicates that students disagree a little or a lot with feeling safe at school, had two or more incidents happen to them, and had two or more incidents happen to someone in their class in the last month. A medium level includes all other combinations of responses.

As shown in Exhibit 13.12, the component variables form a reliable scale, with a median reliability coefficient (Cronbach's alpha) across countries of 0.68. The median multiple correlation between the component variables and student achievement was 0.20, corresponding to an R-square of 0.04.

Exhibit 13.11 Index of Parents' Perceptions of School Environment (PPSE) Statistics

Country	Cronbach's Alpha Between the Component Variables	Multiple R Between Student Reading Achievement and Component Variables	Percent of Variance in Student Reading Achievement Accounted for by the Component Variables
Austria	0.71	0.22	0.05
Belgium (Flemish)	0.71	0.22	0.05
Belgium (French)	0.64	0.20	0.04
Bulgaria	0.27	0.14	0.02
Canada, Alberta	0.68	0.14	0.02
Canada, British Columbia	0.66	0.17	0.03
Canada, Nova Scotia	0.70	0.10	0.01
Canada, Ontario	0.66	0.10	0.01
Canada, Quebec	0.67	0.17	0.03
Chinese Taipei	0.21	0.10	0.01
Denmark	0.75	0.14	0.02
England	0.71	0.17	0.03
France	0.67	0.17	0.03
Georgia	0.35	0.14	0.02
Germany	0.73	0.20	0.04
Hong Kong SAR	0.36	0.10	0.01
Hungary	0.69	0.22	0.05
Iceland	0.72	0.10	0.01
Indonesia	-0.22	0.10	0.01
Iran, Islamic Rep. of	0.15	0.10	0.01
Israel	0.59	0.22	0.05
Italy	0.55	0.17	0.03
Kuwait	0.49	0.10	0.01
Latvia	0.53	0.20	0.04
Lithuania	0.54	0.14	0.02
Luxembourg	0.62	0.20	0.04
Macedonia, Rep. of	0.31	0.24	0.06
Moldova, Rep. of	0.32	0.20	0.04
Morocco	0.39	0.10	0.01
Netherlands	0.75	0.10	0.01
New Zealand	0.70	0.20	0.04
Norway	0.70	0.10	0.01
Poland	0.53	0.17	0.03
Qatar	0.55	0.17	0.03
Romania	0.33	0.14	0.02
Russian Federation	0.36	0.22	0.05
Scotland	0.71	0.10	0.01
Singapore	0.45	0.00	0.00
Slovak Republic	0.48	0.20	0.04
Slovenia	0.61	0.20	0.04
South Africa	0.14	0.24	0.06
Spain	0.67	0.20	0.04
Sweden	0.72	0.14	0.02
Trinidad and Tobago	0.54	0.22	0.05
United States	-	-	-
International Median	0.55	0.17	0.03

A dash (–) indicates data are not available.

Exhibit 13.12 Index of Student Safety in School (SSS) Statistics

Country	Cronbach's Alpha Between the Component Variables	Multiple R Between Student Reading Achievement and Component Variables	Percent of Variance in Student Reading Achievement Accounted for by the Component Variables
Austria	0.73	0.22	0.05
Belgium (Flemish)	0.70	0.26	0.07
Belgium (French)	0.70	0.17	0.03
Bulgaria	0.72	0.22	0.05
Canada, Alberta	0.70	0.26	0.07
Canada, British Columbia	0.69	0.17	0.03
Canada, Nova Scotia	0.70	0.22	0.05
Canada, Ontario	0.70	0.20	0.04
Canada, Quebec	0.72	0.22	0.05
Chinese Taipei	0.76	0.22	0.05
Denmark	0.60	0.17	0.03
England	0.69	0.26	0.07
France	0.72	0.20	0.04
Georgia	0.69	0.14	0.02
Germany	0.69	0.28	0.08
Hong Kong SAR	0.73	0.20	0.04
Hungary	0.68	0.22	0.05
Iceland	0.74	0.17	0.03
Indonesia	0.65	0.17	0.03
Iran, Islamic Rep. of	0.63	0.17	0.03
Israel	0.66	0.30	0.09
Italy	0.68	0.14	0.02
Kuwait	0.67	0.10	0.01
Latvia	0.65	0.20	0.04
Lithuania	0.68	0.20	0.04
Luxembourg	0.65	0.17	0.03
Macedonia, Rep. of	0.74	0.32	0.10
Moldova, Rep. of	0.67	0.10	0.01
Morocco	0.55	0.14	0.02
Netherlands	0.69	0.24	0.06
New Zealand	0.70	0.24	0.06
Norway	0.67	0.17	0.03
Poland	0.70	0.17	0.03
Qatar	0.60	0.24	0.06
Romania	0.70	0.22	0.05
Russian Federation	0.62	0.17	0.03
Scotland	0.67	0.20	0.04
Singapore	0.67	0.17	0.03
Slovak Republic	0.70	0.17	0.03
Slovenia	0.67	0.20	0.04
South Africa	0.54	0.20	0.04
Spain	0.70	0.14	0.02
Sweden	0.70	0.22	0.05
Trinidad and Tobago	0.60	0.17	0.03
United States	0.66	0.22	0.05
International Median	0.68	0.20	0.04

The Index of Principals' Perception of School Safety (PPSS) categorizes students according to their principals' perceptions of the degree to which various problems occur in their schools. The index, developed in 2001, is presented in Exhibit 7.16 of the international report. This index is based on principals' reports about the degree to which each of the following was a problem: classroom disturbances, cheating, profanity, vandalism, theft, intimidation or verbal abuse of other students, and physical conflicts among students. An average was computed on a 4-point scale: *serious problem* = 1, *moderate problem* = 2, *minor problem* = 3, *not a problem* = 4. A high level indicates an average of greater than 3 through 4. A medium level indicates an average of 2 through 3. A low level indicates an average of 1 to less than 2.

As shown in Exhibit 13.13, the component variables form a very reliable scale, with a median reliability coefficient (Cronbach's alpha) across countries of 0.87. The median multiple correlation between the component variables and student achievement was 0.14, corresponding to an R-square of 0.02.

13.7 Reviewing Questionnaire Exhibits

Based on preliminary analyses, analysis specifications were created for all derived variables, including indices. This documentation included the variables to be used and their sources, the way variables would be recoded and combined, and how the derived variable would be presented in the international report. The analysis specifications guided the programmers and TIMSS & PIRLS International Study Center production staff who implemented these analyses and created exhibits, and were made available to NRCs to aid their reviews of the exhibits. The final exhibits were produced using custom-designed SAS programs that calculated student reading achievement averages using all five imputed scores (plausible values) for each student, including standard errors calculated using the jackknife procedure (see Chapter 12).

Exhibit 13.13 Index of Principals' Perception of School Safety (PPSS) Statistics

Country	Cronbach's Alpha Between the Component Variables	Multiple R Between Student Reading Achievement and Component Variables	Percent of Variance in Student Reading Achievement Accounted for by the Component Variables
Austria	0.86	0.10	0.01
Belgium (Flemish)	0.85	0.10	0.01
Belgium (French)	0.88	0.20	0.04
Bulgaria	0.86	0.14	0.02
Canada, Alberta	0.85	0.14	0.02
Canada, British Columbia	0.85	0.14	0.02
Canada, Nova Scotia	0.82	0.14	0.02
Canada, Ontario	0.86	0.14	0.02
Canada, Quebec	0.86	0.17	0.03
Chinese Taipei	0.87	0.00	0.00
Denmark	0.84	0.14	0.02
England	0.87	0.24	0.06
France	0.85	0.24	0.06
Georgia	0.87	0.20	0.04
Germany	0.85	0.22	0.05
Hong Kong SAR	0.90	0.10	0.01
Hungary	0.84	0.22	0.05
Iceland	0.78	0.00	0.00
Indonesia	0.93	0.14	0.02
Iran, Islamic Rep. of	0.86	0.17	0.03
Israel	0.88	0.22	0.05
Italy	0.92	0.17	0.03
Kuwait	0.93	0.24	0.06
Latvia	0.88	0.20	0.04
Lithuania	0.84	0.10	0.01
Luxembourg	-	-	-
Macedonia, Rep. of	0.90	0.17	0.03
Moldova, Rep. of	0.95	0.10	0.01
Morocco	0.93	0.14	0.02
Netherlands	0.77	0.14	0.02
New Zealand	0.88	0.24	0.06
Norway	0.83	0.10	0.01
Poland	0.80	0.00	0.00
Qatar	0.94	0.17	0.03
Romania	0.94	0.10	0.01
Russian Federation	0.74	0.10	0.01
Scotland	0.79	0.10	0.01
Singapore	0.85	0.10	0.01
Slovak Republic	0.89	0.10	0.01
Slovenia	0.84	0.10	0.01
South Africa	0.88	0.33	0.11
Spain	0.91	0.14	0.02
Sweden	0.84	0.14	0.02
Trinidad and Tobago	0.86	0.17	0.03
United States	0.87	0.20	0.04
International Median	0.87	0.14	0.02

A dash (–) indicates data are not available.

Representatives from participating countries had several opportunities to review exhibits and make suggestions for additions and modifications. The draft exhibits first were reviewed, in conjunction with the *PIRLS 2006 International Report* outline, background data almanacs, and analysis notes, at the seventh NRC meeting in Queenstown, New Zealand in November 2006. At that time, data had been received and processed by the IEA Data Processing and Research Center for all but two participating countries, allowing NRCs to view their questionnaire results as they would be displayed in the report. Based on NRCs' comments, the exhibits and data were further refined for a second review at the eighth NRC meeting in Quebec City, Canada in June 2007. At this meeting, NRCs were provided with a draft of the *PIRLS 2006 International Report* containing complete versions of the report exhibits. NRCs approved these final exhibits and text with some suggested revisions, which were implemented by the TIMSS & PIRLS International Study Center staff for the report.

References

DeVellis, R. (1991). *Scale development: Theory and applications*. Newbury Park, CA: Sage Publications.

Foy, P., & Kennedy, A.M. (Eds.). (2008). *PIRLS 2006 user guide for the international database.* Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Kennedy, A.M., Martin, M.O., & Sainsbury, M. (2006). *PIRLS 2006 assessment framework and specifications* (2nd ed.). Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., & Foy, P. (2007). *PIRLS 2006 international report: IEA's progress in international reading literacy study in primary schools in 40 countries*. Chestnut Hill, MA: Boston College.

SAS Institute (2002). SAS system for Windows (version 9.1). Cary, NC: SAS Institute.

Spector, P. (1992). *Summated rating scale construction, an introduction* (Sage University Papers Series on Quantitative Applications in the Social Sciences, series no. 07-082). Beverly Hills, CA: Sage

SPSS Inc. (2005). SPSS for Windows (version 14.0). Chicago, IL: SPSS Inc.

Appendix A

Acknowledgements

Introduction

PIRLS 2006 was a collaborative effort involving hundreds of individuals around the world. This appendix recognizes the individuals and organizations for their contributions. Given that work on PIRLS 2006 has spanned approximately 5 years and has involved so many people and organizations, this list may not include all who contributed. Any omission is inadvertent.

Of the first importance, PIRLS 2006 is deeply indebted to the students, parents, teachers, and school principals who contributed their time and effort to the study.

Management and Coordination

PIRLS is a major undertaking of IEA, and together with the Trends in International Mathematics and Science Study (TIMSS), comprises the core of IEA's regular cycles of studies. The PIRLS assessment at the fourth grade complements TIMSS, which regularly assesses mathematics and science achievement at fourth and eighth grades.

The TIMSS & PIRLS International Study Center at Boston College has responsibility for the overall direction and management of the TIMSS and PIRLS projects. Headed by Drs. Ina V.S. Mullis and Michael O. Martin, the study center is located in the Lynch School of Education. Dr. Ann M. Kennedy is the PIRLS Project Coordinator. In carrying out the project, the TIMSS & PIRLS International Study Center worked closely with the IEA Secretariat in Amsterdam, which provided guidance overall and was responsible for

verification of all translations produced by the participating countries. The IEA Data Processing and Research Center in Hamburg was responsible for processing and verifying the data submitted by the participants; Statistics Canada in Ottawa was responsible for school and student sampling activities; and Educational Testing Service (ETS) in Princeton, New Jersey consulted on psychometric methodology and provided software for scaling the achievement data.

The Project Management Team, comprised of study directors and representatives from the TIMSS & PIRLS International Study Center, the IEA Secretariat, the IEA Data Processing and Research Center, Statistics Canada, and ETS met twice a year throughout the study to discuss the study's progress, procedures, and schedule. In addition, the study directors met with members of IEA's Technical Executive Group twice yearly to review technical issues.

Dr. Marian Sainsbury from the National Foundation for Educational Research in England (NFER) was the PIRLS 2006 Reading Coordinator and Dr. Patricia Donahue from ETS was a special reading assessment consultant. Together with the Reading Development Group, a panel of internationally recognized experts in reading research, instruction, and assessment, they provided excellent guidance throughout PIRLS 2006.

To work with the international team and coordinate within-country activities, each participating country designated an individual to be the PIRLS National Research Coordinator (NRC). The NRCs have the complicated and challenging task of implementing the PIRLS study in their countries in accordance with the PIRLS guidelines and procedures. The quality of the PIRLS assessment and data depends on the work of the NRCs and their colleagues in carrying out the very complex sampling, data collection, and scoring tasks involved. In addition, the Questionnaire Development Group, comprised of NRCs, provided advice on questionnaire development.

Continuing the tradition of truly exemplary work established in PIRLS 2001, the PIRLS 2006 NRCs (often the same NRCs as in 2001), performed their many tasks with dedication, competence, energy, and goodwill, and have been commended by the IEA Secretariat, the TIMSS & PIRLS International Study Center, the IEA Data Processing and Research Center, and Statistics Canada for their commitment to the project and the high quality of their work.

Funding

A project of this magnitude requires considerable financial support. IEA's major funding partners for PIRLS included the World Bank, the U.S. Department of Education through the National Center for Education Statistics, and those countries that contributed by way of fees. The financial support provided by Boston College and NFER also is gratefully acknowledged.

IEA Secretariat

Seamus Hegarty, IEA Chair
Hans Wagemaker, Executive Director
Barbara Malak, Manager, Membership Relations
Juriaan Hartenburg, Financial Manager
Suzanne Morony, Senior Manager Assistant

TIMSS & PIRLS International Study Center at Boston College

Ina V.S. Mullis, Co-Director
Michael O. Martin, Co-Director
Pierre Foy, Director of Sampling and Data Analysis
Ann M. Kennedy, Coordinator of Project Development and Operations, PIRLS Coordinator
Alka Arora, TIMSS Advanced 2008 Project Coordinator
Debra Berger, Production Editor
Marcie Bligh, Manager of Office Administration
Joann Cusack, Administrative Coordinator
Ebru Erberber, TIMSS Science Research Associate
Susan Farrell, Co-Manager of Publications
Joseph Galia, Senior Statistician/Programmer
Christine Hoage, Manager of Finance
Ieva Johansone, Survey Operations Coordinator
Isaac Li, Statistician/Programmer
Dana Milne, TIMSS Graduate Assistant
Jennifer Moher, Data Graphics Specialist
Mario Pita, Co-Manager of Publications
Corinna Preuschoff, TIMSS Mathematics Research Associate
Ruthanne Ryan, Data Graphics Specialist
Gabrielle Stanco, TIMSS Graduate Assistant
Feng Tian, TIMSS Graduate Assistant
Kathleen L. Trong, PIRLS Research Associate

IEA Data Processing and Research Center

Dirk Hastedt, Co-Director
Juliane Barth, Co-Manager, TIMSS & PIRLS Data Processing
Oliver Neuschmidt, Co-Manager, TIMSS & PIRLS Data Processing
Yasin Afana, Researcher
Milena Taneva, Researcher
Marta Kostek-Drosihn, Researcher
Sabine Meinck, Researcher
Olaf Zuehlke, Researcher
Christine Busch, Researcher
Alena Becker, Researcher
Simone Uecker, Researcher
Michael Jung, Researcher
Tim Daniel, Researcher
Dirk Oehler, Researcher
Stephan Petzchen, Senior Programmer
Ralph Carstens, Programmer
Hauke Heyen, Programmer
Harpreet Singh Choudry, Programmer

Statistics Canada

Marc Joncas, Senior Methodologist

Educational Testing Service

Mathias von Davier, Senior Research Scientist

Sampling Referee

Keith Rust, Vice President and Associate Director of the Statistical Group, Westat, Inc.

PIRLS 2006 Assessment Development

Reading Coordinator

Marian Sainsbury, NFER

Reading Assessment Consultant

Patricia Donahue, ETS

PIRLS 2006 Reading Development Group (RDG)

Dominique Lafontaine, *Service de Pedagogie Experimentale, Belgium*
Jan Mejding, *Danish University of Education, Denmark*
Sue Horner, *Qualifications and Curriculum Authority, England*
Renate Valtin, *Abteilung Grundschulpädagogik, Humboldt Universität, Germany*

Galina Zuckerman, *Psychological Institute, Russian Academy of Education, Russian Federation*
Elizabeth Pang and Selene Tan, *Ministry of Education HQ, Singapore*
Karen Wixson, *University of Michigan, United States*

PIRLS Questionnaire Development Group (QDG)

Meng Hong Wei, *The China National Institute of Education, China*
Marc Colmant, *Ministere de l'Education Nationale, France*
Knut Schwippert, *University of Hamburg, Institute for Comparative & Multicultural Education, Germany*
Gabriella Pavan de Gregorio, *Instituto Nazionale per la Valuatazione del Sistema Dell'Istruzione, Italy*
Bojana Naceva, *Bureau for Development of Education, Republic of Macedonia*
Mieke van Diepen, *Expertisecentrum Nederlands, Netherlands*
Ragnar Gees Solheim, *National Center for Reading Education and Research, Norway*
Larry Ogle, *National Center for Education Statistics, United States*

PIRLS 2006 National Research Coordinators (NRCs)

Austria

Günter Haider
Birgit Suchań
Austrian IEA Research Centre, Universität Salzburg

Belgium

Flemish

Jan Van Damme
Katholieke Universiteit Leuven

French

Annette Lafontaine
Université de Liège

Bulgaria

Tatyana Angelova
Felianka Kaftandjieva (through 2004)
University of Sofia

Canada

Alberta

Ping Yang
Learner Assessment Branch, Alberta Education

British Columbia

Diane Lalancette
Exams & Assessment Policy

Nova Scotia

Marthe Craig
Evaluation Coordinator, Evaluation Services

Ontario

Michael Kozlow
Francine Jaques (through 2004)
Education Quality and Accountability Office

Québec

Serge Baillargeon
Ministère de l'Éducation

Chinese Taipei

Hwawei Ko
Graduate Institute of Learning and Instruction
National Central University

Denmark
Jan Mejding
The Danish University of Education

England
Liz Twist
National Foundation for Educational Research in England and Wales

France
Marc Colmant
Ministère de l'Éducation Nationale

Georgia
Maia Miminoshvili
National Assessment and Examinations Center

Germany
Wilfried Bos
Sabine Hornberg
Institut fuer Schulentwicklungsfurschung
University of Dortmund

Hong Kong
Tse Shek-Kam
The University of Hong Kong

Hungary
Ildiko Balazsi
Péter Balkányi
Annamária Szazó Rábai (through 2004)
National Institute of Public Education
Centre for Evaluation Studies

Iceland
Brynhildur Scheving Thorsteinsson
Institute for Educational Research

Indonesia
Burhanuddin Tola
Center for Educational Assessment
Bahrul Hayat (through 2004)
Ministry of National Education

Iran, Islamic Republic of
Abdol'azim Karimi
Institute for Educational Research

Israel
Elite Olshtain
Hebrew University
Ruth Zuzovsky
Tel Aviv University

Italy
Silvana Serra
Gabriella Pavan de Gregorio (through 2005)
Instituto Nazionale per la Valuatazione del Sistema Dell'Istruzione

Kuwait
Abdul Ghani Al-Bazzaz
Ministry of Education

Latvia
Antra Ozola
University of Latvia

Lithuania
Aiste Elijio
Ministry of Education and Science

Luxembourg
Pierre Reding
Martin Frieberg
Ministère de l'Éducation Nationale

Macedonia, Republic of
Tanja Andonova
Pedagogical Institute of Macedonia
Bojana Naceva (through 2006)
Bureau for Development of Education

Moldova, Republic of
Ilie Nasu
Ministry of Education and Science

Morocco

Mohammed Sassi
Departement de l'Évaluation Nationale

Netherlands

Andrea Netten
Mieke van Diepen (through 2004)
Expertisecentrum Nederlands

New Zealand

Megan Chamberlain
Ministry of Education

Norway

Ragnar Gees Solheim
Victor van Daal
Finn-Egil Toennessen (through 2005)
National Centre for Reading,
Education and Reading Research
University of Stavanger

Poland

Krzysztof Konarzewski
Institute of Psychology
Polish Academy of Science

Qatar

Abdessalem Buslama
Marcus Broer (through 2006)
Evaluation Institute
Supreme Education Council
Office of Student Assessment

Romania

Gabriela Noveanu
Institute for Educational Sciences
Evaluation and Forecasting Division

Russian Federation

Galina Kovalyova
The Russian Academy of Education

Scotland

Fiona Fraser
Jo MacDonald (through 2003)
Scottish Office, Education Department

Singapore

Wong Look Kwang
New Yi Cheen (through 2005)
Research and Evaluation
Ministry of Education

Slovak Republic

Eva Obrancova
Zuzana Lukackova (through 2004)
SPU—National Institute for Education

Slovenia

Marjeta Doupona-Horvat
Educational Research Institute

South Africa

Sarah Howie
Elsie Venter
University of Pretoria

Spain

Mar Gonzalez Garcia
Flora Gil Traver (through 2005)
Instituto Nacional de Calidad y Evaluacion
del Sistema Educativo

Sweden

Bo Palaszewski
National Agency for Education
Caroline Liberg
Uppsala University

Trinidad and Tobago

Harrilal Seecharan
Mervyn Sambucharan
Division of Educational Research
and Evaluation

United States

Laurence Ogle
National Center for Education Statistics
U.S. Department of Education

Appendix B

Characteristics of National Samples

Introduction

For each country participating in PIRLS 2006, 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.1 Austria

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<3), and special education schools
- Within-school exclusions consisted of intellectually and functionally disabled students, and non-native language speakers

Sample Design

- Explicit stratification by region for a total of 9 explicit strata
- Implicit stratification by district (the number of districts varies by region) for a total of 121 implicit strata
- Sampled two classrooms per school whenever possible
- Small schools sampled with equal probabilities

Exhibit B.1 Allocation of School Sample in Austria

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Burgenland	5	0	5	0	0	0
Kärnten	11	1	10	0	0	0
Niederösterreich	31	0	31	0	0	0
Oberösterreich	30	0	30	0	0	0
Salzburg	11	0	11	0	0	0
Steiermark	23	0	23	0	0	0
Tirol	15	1	14	0	0	0
Vorarlberg	8	0	8	0	0	0
Wien	26	0	26	0	0	0
Total	160	2	158	0	0	0

B.2 Belgium (Flemish)

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<5), and special schools

Sample Design

- Explicit stratification by school type (Flemish community, public, private) for a total of 3 explicit strata
- Implicit stratification by province (Antwerpen, Limburg, Oost-Vlaanderen, Vlaams-Brabant, West-Vlaanderen) for a total of 15 implicit strata
- Sampled two classrooms per school whenever possible
- Small schools sampled with equal probabilities

Exhibit B.2 Allocation of School Sample in Belgium (Flemish)

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Flemish Community schools	20	0	15	4	0	1
Public schools	34	1	21	5	3	4
Private schools	96	0	66	16	7	7
Total	150	1	102	25	10	12

B.3 Belgium (French)

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<5), schools for disabled children, and hospital schools
- Within-school exclusions consisted of children with less than one year of instruction in French

Sample Design

- Explicit stratification by school type (state, communal, religious) for a total of 3 explicit strata
- Implicit stratification by region (Brabant Wallon, Bruxelles-Capitale, Hainault, Liège, Namur, Luxembourg) for a total of 18 implicit strata
- Sampled two classrooms per school whenever possible
- Small schools sampled with equal probabilities

Exhibit B.3 Allocation of School Sample in Belgium (French)

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
State	14	0	11	3	0	0
Communal	72	0	62	9	1	0
Religious	64	0	56	8	0	0
Total	150	0	129	20	1	0

B.4 Bulgaria

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<6), and special education schools

Sample Design

- Explicit stratification by region for a total of 9 explicit strata
- Implicit stratification by urbanization (urban, rural) for a total of 18 implicit strata

- Sampled two classrooms per school having at least 60 students (MOS≥60) and one classroom otherwise
- Small schools sampled with equal probabilities

Exhibit B.4 Allocation of School Sample in Bulgaria

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Burgas	18	0	15	3	0	0
Hashkovo	16	1	12	2	0	1
Lovech	16	0	14	1	0	1
Montana	10	0	9	1	0	0
Plovdiv	22	1	20	1	0	0
Ruse	14	1	12	1	0	0
Sofia City	19	0	17	2	0	0
Sofia Region	17	0	15	1	0	1
Varna	18	0	16	1	0	1
Total	150	3	130	13	0	4

B.5 Canada, Alberta

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<6), and online/correspondence students

Sample Design

- No explicit stratification
- Implicit stratification by school type (charter, Francophone, private, public, separate) for a total of 5 implicit strata
- Sampled two classrooms per school having at least 60 students (MOS≥60) and one classroom otherwise
- Small schools sampled with equal probabilities (MOS<16)

Exhibit B.5 Allocation of School Sample in Canada, Alberta

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Canada (Alberta)	150	0	150	0	0	0
Total	150	0	150	0	0	0

B.6 Canada, British Columbia

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<6), alternate and distance education schools, district distance education schools, and long term Provincial Resource Program (PRP) schools

Sample Design

- No explicit stratification
- Implicit stratification by school type (public, independent) for a total of 2 implicit strata
- Sampled two classrooms per school having at least 52 students (MOS≥52) and one classroom otherwise
- Small schools sampled with equal probabilities (MOS<14)

Exhibit B.6 Allocation of School Sample in Canada, British Columbia

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Canada (British Columbia)	150	0	147	1	0	2
Total	150	0	147	1	0	2

B.7 Canada, Nova Scotia

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<5)

Sample Design

- Explicit stratification by regional council for a total of 8 explicit strata
- No implicit stratification
- Sampled two classrooms per school having at least 77 students (MOS≥77) and one classroom otherwise
- Very large and small schools sampled with equal probabilities
- Census of schools in the four smallest regional councils (Strait Regional, Acadian Provincial, South Shore Regional, and Tri-County Regional)

Exhibit B.7 Allocation of School Sample in Canada, Nova Scotia

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Cape Breton-Victoria Regional	25	0	25	0	0	0
Strait Regional	16	0	16	0	0	0
Chignecto Central Regional	26	0	25	1	0	0
Halifax Regional	58	0	58	0	0	0
Annapolis Valey Regional	25	0	25	0	0	0
Acadien Provincial	16	0	16	0	0	0
South Shore Regional	17	0	17	0	0	0
Tri-County Regional	18	0	18	0	0	0
Total	201	0	200	1	0	0

B.8 Canada, Ontario

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<10), special needs schools, native schools, and overseas schools
- Within-school exclusions consisted of children with disabilities (either within regular classrooms or in special education classrooms within regular schools)

Sample Design

- Explicit stratification by language (English, French) for a total of 2 explicit strata
- Implicit stratification by school type (public, Catholic, private) for a total of 6 implicit strata

- Sampled two classrooms per school having at least 100 students (MOS≥100) and one classroom otherwise
- Small schools sampled with equal probabilities (MOS<15)
- Two schools in the French stratum were sampled with certainty

Exhibit B.8 Allocation of School Sample in Canada, Ontario

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
English	120	1	105	2	0	12
French	80	1	68	5	0	6
Total	200	2	173	7	0	18

B.9 Canada, Quebec

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<10), Native schools, non-ministry schools, and special education schools
- Within-school exclusions consisted of children with disabilities or non-native speakers

Sample Design

- Explicit stratification by language (English, French) for a total of 2 explicit strata
- Implicit stratification by school type (public, private) for a total of 4 implicit strata
- Sampled one classroom per school
- Small schools sampled with equal probabilities
- Four schools in the English stratum were sampled with certainty

Exhibit B.9 Allocation of School Sample in Canada, Quebec

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
English	80	2	74	0	0	4
French	120	4	111	0	0	5
Total	200	6	185	0	0	9

B.10 Chinese Taipei

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<15), schools on remote islands, and special needs schools
- Within-school exclusions consisted of disabled students

Sample Design

- Explicit stratification by region (North, Middle, South, East) for a total of 4 explicit strata
- No implicit stratification
- Sampled one classroom per school
- Small schools sampled with equal probabilities

Exhibit B.10 Allocation of School Sample in Chinese Taipei

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
North	68	0	67	1	0	0
Middle	38	0	38	0	0	0
South	40	0	38	2	0	0
East	4	0	4	0	0	0
Total	150	0	147	3	0	0

B.11 Denmark

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<6), and special needs schools
- Within-school exclusions consisted of disabled students

Sample Design

- No explicit stratification
- No implicit stratification

- Sampled two classrooms per school having at least 50 students (MOS≥50) and one classroom otherwise
- Small schools sampled with equal probabilities (MOS<16)

Exhibit B.11 Allocation of School Sample in Denmark

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Denmark	150	150	4	128	16	1
Totals	150	150	4	128	16	1

B.12 England

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<8), and special schools
- Within-school exclusions consisted of pupils with special education needs

Sample Design

- Explicit stratification by school performance for a total of 6 explicit strata
- Implicit stratification by school type (primary, junior, middle, independent) for a total of 23 implicit strata
- Sampled two classrooms per school with at least 100 students (MOS≥100) and one classroom otherwise
- Small schools sampled with equal probabilities

Exhibit B.12 Allocation of School Sample in England

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Low KS2	28	0	20	6	1	1
Low/MID KS2	29	0	27	2	0	0
Mid KS2	29	0	26	3	0	0
Mid/High KS2	29	0	25	3	0	1
High KS2	29	0	25	4	0	0
Unknown KS2	6	0	6	0	0	0
Total	150	0	129	18	1	2

B.13 France

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<4), schools found in La Réunion and in Guyana, private schools without contracts, French schools in foreign countries, and specialized schools (note that schools found in Overseas Territories (TOM) were considered out of scope and therefore were not considered)

Sample Design

- Explicit stratification by school size (large schools, small schools (MOS<15)) for a total of 2 explicit strata
- Implicit stratification by school type (public not Priority Education Zone (ZEP), private, public ZEP) for a total of 6 implicit strata
- Sampled two classrooms per school whenever possible
- Schools within the 'small schools' stratum sampled with equal probabilities

Exhibit B.13 Allocation of School Sample in France

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Large	115	0	108	4	0	3
Small	60	0	56	0	1	3
Total	175	0	164	4	1	6

B.14 Georgia

Coverage and Exclusions

- Coverage was restricted to students whose language of instruction was Georgian
- School-level exclusions consisted of very small schools (MOS<4), and special education schools

Sample Design

- Explicit stratification by region for a total of 12 explicit strata
- Implicit stratification by school type (urban, rural) for a total of 23 implicit strata

- Sampled two classrooms per school with at least 65 students (MOS≥65), one classroom otherwise
- Small schools sampled with equal probabilities (small school definition vary by region)

Exhibit B.14 Allocation of School Sample in Georgia

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Kvemo Kartli	10	1	8	1	0	0
Adjara	14	0	11	2	1	0
Apxazeti	2	0	1	1	0	0
Guria	6	1	5	0	0	0
Imereti	27	0	27	0	0	0
Kaxeti	14	1	13	0	0	0
Mckheta-Tianeti	5	0	5	0	0	0
Racha-Lechkhumi	2	0	1	0	1	0
Samckhe-Javakheti	5	0	4	0	1	0
Shida Kartli	12	0	11	1	0	0
Tbilisi	39	0	38	1	0	0
Samegrelo	16	0	15	0	1	0
Total	152	3	139	6	4	0

B.15 Germany

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<6), and schools for students with disabilities
- Within-school exclusions consisted of special needs students and non-native language speakers

Sample Design

- Explicit stratification by state for a total of 16 explicit strata
- Implicit stratification by school type (primary, special education) and by region (North, South, West, East, Northwest, etc.) within 'primary schools' strata for a total of 45 implicit strata
- Sampled one classroom per school

- Small schools sampled with equal probabilities (small school definition vary by state)

Exhibit B.15 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	25	0	22	0	1	2
Bayern	25	0	24	1	0	0
Berlin	25	0	25	0	0	0
Brandenburg	25	0	25	0	0	0
Bremen	25	0	25	0	0	0
Hamburg	25	0	23	2	0	0
Hessen	25	0	24	1	0	0
Mecklenburg-Vorpommern	25	0	25	0	0	0
Niedersachsen	25	0	24	0	1	0
Nordrhein-Westfalen	35	0	35	0	0	0
Rheinland-Pfalz	25	1	23	1	0	0
Saarland	25	0	25	0	0	0
Sachsen	25	0	25	0	0	0
Sachsen-Anhalt	25	0	25	0	0	0
Schleswig-Holstein	25	0	25	0	0	0
Thüringen	25	2	22	1	0	0
Total	410	3	397	6	2	2

B.16 Hong Kong SAR

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<10), and students from international schools
- Within-school exclusions consisted of students in special education classes from regular schools

Sample Design

- Explicit stratification by financial sources (aided, private, government, direct subsidies) and session within the 'aided schools' stratum (AM, PM, whole day) for a total of 6 explicit strata

- Implicit stratification by region groups (high / medium / low performing regions) for a total of 18 implicit strata
- Sampled one classroom per school
- Small schools sampled with equal probabilities (small school definition vary by explicit stratum)

Exhibit B.16 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	
Aided - AM	19	0	16	3	0	0
Aided - PM	17	0	15	1	1	0
Aided - Whole Day	91	6	77	8	0	0
Private	10	0	9	1	0	0
Government	9	0	9	0	0	0
Direct Subsidies	4	0	4	0	0	0
Total	150	6	130	13	1	0

B.17 Hungary

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<6), and special education schools (SEN schools)
- Within-school exclusions consisted of SEN students

Sample Design

- Explicit stratification by type of community (capital, county town, town, rural area) for a total of 4 explicit strata
- Implicit stratification by performance level (high, medium, low, unknown) and by region for a total of 75 implicit strata
- Sampled two classrooms per school having at least 75 students (MOS≥75), and one classroom otherwise
- Small schools sampled with equal probabilities

Exhibit B.17 Allocation of School Sample in Hungary

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Capital	22	0	21	1	0	0
Country Town	28	1	26	1	0	0
Town	50	0	50	0	0	0
Rural Area	50	0	50	0	0	0
Total	150	1	147	2	0	0

B.18 Iceland

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<5)
- Within-school exclusions consisted of disabled students and non-native language speakers

Sample Design

- No explicit stratification
- Implicit stratification by region for a total of 5 implicit strata
- Sampled all schools and all classrooms

Exhibit B.18 Allocation of School Sample in Iceland

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Island (Grade 4)	136	5	128	0	0	3
Total	136	5	128	0	0	3

B.19 Indonesia

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<6), schools from Nanggroe Aceh Darussalam (political reasons), schools from Papua (geographical reasons), and special education schools

Sample Design

- Explicit stratification by school type (general primary school & Islamic primary school) and school status (public, private) for a total of 4 explicit strata
- Implicit stratification by group of province (Western Indonesia, Central Java, Eastern Java & Banten, Central Indonesia, Eastern Indonesia) and urbanization (village, town) for a total of 40 implicit strata
- Sampled one classroom per school
- Small schools sampled with equal probabilities

Exhibit B.19 Allocation of School Sample in Indonesia

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
General Elementary Public	140	2	136	2	0	0
General Elementary Private	12	0	12	0	0	0
Islamic Elementary Public	2	0	2	0	0	0
Islamic Elementary Private	16	0	16	0	0	0
Total	170	2	166	2	0	0

B.20 Iran, Islamic Rep. of

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<5), and schools from Bam

Sample Design

- Explicit stratification by school type (public, private) and school gender (boys, girls, mixed) for a total of 5 explicit strata
- Implicit stratification by province for a total of 145 implicit strata
- Sampled one classroom per school
- Small schools sampled with equal probabilities

Exhibit B.20 Allocation of School Sample in Iran, Islamic Rep. of

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Public - Girls	74	1	73	0	0	0
Public - Boys	68	0	68	0	0	0
Public - Mixed	48	2	46	0	0	0
Private - Girls	30	1	29	0	0	0
Private - Boys	20	0	19	1	0	0
Total	240	4	235	1	0	0

B.21 Israel

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<13), ultra-orthodox schools, schools with unknown SES, and special education schools
- Within-school exclusions consisted of students found in special classes within regular schools and special needs students within regular classes

Sample Design

- Explicit stratification by school type (Hebrew religious, Hebrew secular, Arab secular) for a total of 3 explicit strata
- Implicit stratification by SES indicator (low, medium, high) for a total of 9 implicit strata
- Sampled one classroom per school
- Small schools sampled with equal probabilities

Exhibit B.21 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 schools	40	0	39	0	1	0
Hebrew secular schools	70	1	67	1	1	0
Arab secular schools	40	0	40	0	0	0
Total	150	1	146	1	2	0

B.22 Italy

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<8)
- Within-school exclusions consisted of disabled students and non-native language speakers

Sample Design

- No explicit stratification
- Implicit stratification by region and urbanization (province capital towns, small towns) for a total of 40 implicit strata
- Sampled two classrooms in most larger school, and one classroom otherwise
- All schools sampled with probability proportional to the size of the school

Exhibit B.22 Allocation of School Sample in Italy

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Italy	150	0	136	11	3	0
Total	150	0	136	11	3	0

B.23 Kuwait

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of international schools and special education schools
- Within-school exclusions consisted of disabled students

Sample Design

- No explicit stratification
- Implicit stratification by region (Asema, Hawalli, Farwaniya, Ahmadi, Jahra, Mubarak) and gender (boys, girls) for a total of 12 implicit strata
- Sampled two classrooms per school having at least 175 students (MOS≥175), and one classroom otherwise
- The largest 25 schools were sampled with certainty

Exhibit B.23 Allocation of School Sample in Kuwait

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Kuwait	150	0	149	0	0	1
Total	150	0	149	0	0	1

B.24 Latvia

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<4), language schools (other than Latvian or Russian), and special education schools
- Within-school exclusions consisted of disabled students

Sample Design

- Explicit stratification by urbanization (Riga, other cities, rural) for a total of 3 explicit strata
- Implicit stratification by language (Latvian, Mixed, Russian) for a total of 9 implicit strata
- Sampled two classrooms per school having at least 50 students (MOS≥50), and one classroom otherwise
- The largest 9 schools were sampled with certainty
- Small schools sampled with equal probabilities

Exhibit B.24 Allocation of School Sample in Latvia

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Riga	42	0	41	0	0	1
Other Cities	64	0	62	0	0	2
Rural	44	0	42	2	0	0
Total	150	0	145	2	0	3

B.25 Lithuania

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<4), and special education schools
- Within-school exclusions consisted of disabled students

Sample Design

- Explicit stratification by county for a total of 10 explicit strata
- Implicit stratification by urbanization (Vilnius, other major cities, regional centers, towns and villages) for a total of 26 implicit strata
- Sampled two classrooms per school whenever possible
- Small schools sampled with equal probabilities

Exhibit B.25 Allocation of School Sample in Lithuania

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Vilnius county	26	0	25	1	0	0
Kauno county	33	1	32	0	0	0
Klaipedos county	16	0	15	1	0	0
Liauliu county	17	0	17	0	0	0
Panevelio county	14	0	14	0	0	0
Alytaus county	9	2	7	0	0	0
Marijampoles county	10	0	10	0	0	0
Taurages county	8	1	7	0	0	0
Telliu county	10	0	10	0	0	0
Utenos county	7	0	7	0	0	0
Total	150	4	144	2	0	0

B.26 Luxembourg

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<3), and special education schools
- Within-school exclusions consisted of students in special education classes within regular schools, non-native language speakers, and disabled students within regular classes

Sample Design

- Explicit stratification by urbanization (urban, rural) for a total of 2 explicit strata
- No implicit stratification
- Sampled all schools and all classrooms

Exhibit B.26 Allocation of School Sample in Luxembourg

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Rural	125	4	121	0	0	0
Urbain	46	1	45	0	0	0
New schools	12	0	12	0	0	0
Total	183	5	178	0	0	0

B.27 Macedonia, Rep. of

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<15), Turkish and Serbian schools, and special education schools
- Within-school exclusions consisted of disabled students

Sample Design

- Explicit stratification by language (Macedonian, Albanian) and region (Skopje, outside Skopje) for a total of 4 explicit strata
- Implicit stratification by urbanization (urban, rural) for a total of 8 implicit strata
- Parts of school (Macedonian, Albanian) were sampled rather than schools

- Sampled two classrooms per part of school having at least 152 students (MOS≥152), and one classroom otherwise
- The largest 39 parts of school were sampled with certainty
- Small parts of school sampled with equal probabilities

Exhibit B.27 Allocation of School Sample in Macedonia, Rep. of

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Macedonian - Skopje	30	0	30	0	0	0
Macedonian - Outside Skopje	67	0	67	0	0	0
Albanian - Skopje	14	0	13	0	1	0
Albanian - Outside Skopje	39	0	39	0	0	0
Total	150	0	149	0	1	0

B.28 Moldova

Coverage and Exclusions

- Coverage in Moldova is restricted to students living outside the Transnistria region
- School-level exclusions consisted of very small schools (MOS<6), Ukrainian schools, and special education schools

Sample Design

- Explicit stratification by urbanization (urban, rural) and language (national, mixed, Russian) for a total of 6 explicit strata
- Implicit stratification by school type (lyceum, gymnasium, primary, general) for a total of 23 implicit strata
- Sampled two classrooms per school having at least 75 students (MOS≥75), and one classroom otherwise
- Small schools sampled with equal probabilities

Exhibit B.28 Allocation of School Sample in Moldova, Rep of

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Urban - National	34	0	32	1	1	0
Urban - Mixed	4	0	4	0	0	0
Urban - Russian	14	0	14	0	0	0
Rural - National	84	0	84	0	0	0
Rural - Mixed	4	0	4	0	0	0
Rural - Russian	10	0	10	0	0	0
Total	150	0	148	1	1	0

B.29 Morocco

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<7)

Sample Design

- Explicit stratification by school type (autonomous, centre, satellite, private) for a total of 4 explicit strata
- Implicit stratification by urbanization (urban, rural) for a total of 7 implicit strata
- Sampled one classroom per school
- Sampled 25 students within sampled classrooms
- Small schools sampled with equal probabilities

Exhibit B.29 Allocation of School Sample in Morocco

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
École Autonome	80	0	79	0	0	1
Secteur Scolaire Centre	33	0	33	0	0	0
École satellite	37	0	37	0	0	0
École Privée	10	0	7	3	0	0
Total	160	0	156	3	0	1

B.30 Netherlands

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<6), and special education schools
- Within-school exclusions consisted of non-native language speakers and children with disabilities

Sample Design

- Explicit stratification by mean student weight indicator (low, medium, high) for a total of 3 explicit strata
- Implicit stratification by urbanization (very high, high, moderate, low, very low) for a total of 15 implicit strata
- Sampled all classrooms within sampled schools
- Small schools sampled with equal probabilities (MOS<22)

Exhibit B.30 Allocation of School Sample in the Netherlands

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Low mean student weights	44	0	31	6	4	3
Medium mean student weights	62	0	49	10	2	1
High mean student weights	44	0	24	9	4	7
Total	150	0	104	25	10	11

B.31 New Zealand

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<4), Rudolf Steiner schools, correspondence schools, Māoris in bilingual schools with less than 4 Māoris, and special education schools

- Within-school exclusions consisted of foreign fee paying students, special needs students, and students with insufficient instruction in test language

Sample Design

- Explicit stratification by language (Māori, Māori & English, English) for a total of 3 explicit strata
- Implicit stratification by Targeted Funding for Educational Achievement (TFEA) in English stratum (high, medium, low, unknown) and urbanization in English stratum (urban, rural) for a total of 9 implicit strata
- Sampled one classroom in the Māori stratum and one Māori classroom in the Māori & English stratum
- Sampled two English classrooms in the Māori & English stratum and the English stratum in schools having at least 60 students (MOS≥60), and one classroom otherwise
- School sampled with equal probabilities in the Māori and the Māori & English strata
- Small schools sampled with equal probabilities in the English stratum (MOS<16)

Exhibit B.31 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 Immersion	25	0	10	5	4	6
Maori & English	25	0	24	1	0	0
English Only	200	0	186	12	1	1
Total	250	0	220	18	5	7

B.32 Norway

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<3), Sami schools, and schools with missing stratification data
- Within-school exclusions consisted of foreign language speakers

Sample Design

- Explicit stratification by immigrant status, language (Bokmål, Nynorsk), and municipal expenditures (low, medium, high, large cities) for a total of 9 explicit strata
- Implicit stratification by municipal expenditures for immigrant school stratum (low, large cities) and immigrant status in all other explicit strata for a total of 18 implicit strata
- Sampled two classrooms per school
- Sampled all schools in the immigrant stratum
- Small schools sampled with equal probabilities in other strata

Exhibit B.32 Allocation of School Sample in Norway

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Immigrant schools	30	0	14	0	0	16
Bokmål - Low expenditures	74	1	55	8	0	10
Bokmål - Medium expenditures	20	0	12	3	0	5
Bokmål - High expenditures	4	0	4	0	0	0
Bokmål - Large cities	26	0	16	2	0	8
Nynorsk - Low expenditures	6	0	4	1	0	1
Nynorsk - Medium expenditures	12	0	9	2	0	1
Nynorsk - High expenditures	4	0	2	1	0	1
Nynorsk - Large cities	2	0	2	0	0	0
Total	178	1	118	17	0	42

B.33 Poland

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<6), and small schools in distant villages
- Within-school exclusions consisted of disabled students in mainstream school

Sample Design

- Explicit stratification by urbanization (villages, towns, cities) for a total of 3 explicit strata
- Implicit stratification by region (16 regions: Dolnoslaskie, Kujawsko-Pomorskie, Lubelskie, Lubuskie, Lodzkie, Malopolskie, Mazowieckie, Opolskie, Podkarpackie, Podlaskie, Pomorskie, Slaskie, Swietokrzyskie, Warminsko-Mazurskie, Wielkopolskie, Zachodniopomorskie) for a total of 48 implicit strata
- Sampled two classrooms per school whenever possible
- Small schools sampled with equal probabilities

Exhibit B.33 Allocation of School Sample in Poland

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Villages	62	2	59	1	0	0
Towns	22	0	22	0	0	0
Cities	66	0	66	0	0	0
Total	150	2	147	1	0	0

B.34 Qatar

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<10)

Sample Design

- Explicit stratification by school type (government, private Arabic, independent) and gender (girls, boys) for a total of 6 explicit strata
- No implicit stratification
- Sampled all schools
- Sampled all classrooms

Exhibit B.34 Allocation of School Sample in Qatar

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Ministry of Education - Girls	42	1	41	0	0	0
Ministry of Education - Boys	35	1	34	0	0	0
Private Arabic - Girls	16	1	15	0	0	0
Private Arabic - Boys	11	1	10	0	0	0
Independent - Girls	8	0	8	0	0	0
Independent - Boys	11	0	11	0	0	0
Total	123	4	119	0	0	0

B.35 Romania

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<5), unidentified schools, mobile Gypsy schools, and special education schools
- Within-school exclusions consisted of disabled students

Sample Design

- Explicit stratification by region for a total of 7 explicit strata
- Implicit stratification by urbanization (urban, rural) for a total of 14 implicit strata
- Sampled two classrooms per school having at least 60 students (MOS≥60), and one classroom otherwise
- Small schools sampled with equal probabilities

Exhibit B.35 Allocation of School Sample in Romania

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Ardeal	44	1	43	0	0	0
Banat	9	0	9	0	0	0
Bucuresti	11	0	11	0	0	0
Dobrogea	7	1	6	0	0	0
Moldova	30	0	30	0	0	0
Muntenia	34	1	32	0	0	1
Oltenia	15	0	15	0	0	0
Total	150	3	146	0	0	1

B.36 Russian Federation

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<6), evening schools, and special education schools
- Within-school exclusions consisted of disabled students

Sample Design

- A sample of 45 regions out of 89 is first drawn with PPS. The largest 17 regions were sampled with certainty (identified by a * in the next table). A sample of schools was then drawn within each region
- Implicit stratification by school location (rural settlement, cities with less than 50,000 people, cities between 50,000 and 100,000 people, cities between 100,000 and 450,000 people, cities between 450,000 and 680,000 people, cities with more than 680,000 people, St. Petersburg, Moscow) for a total of 233 implicit strata
- Sampled one classroom per school
- Small schools sampled with equal probabilities

Exhibit B.36 Allocation of School Sample in the Russian Federation

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Belgorod oblast	4	0	4	0	0	0
Rostov oblast	8	0	8	0	0	0
Adygea	4	0	4	0	0	0
Komi	4	0	4	0	0	0
Hakasia	4	0	4	0	0	0
Razan oblast	4	0	4	0	0	0
Marii Al	4	0	4	0	0	0
Tula oblast	4	0	4	0	0	0
Kaliningrad oblast	4	0	4	0	0	0
Altai kr	6	0	6	0	0	0
Kabardino oblast	4	0	4	0	0	0
Kurst oblast	4	0	4	0	0	0
Dagestan	8	0	8	0	0	0
Kirov oblast	4	0	4	0	0	0
Lipstek oblast	4	0	4	0	0	0
N Novgorod oblast	6	0	6	0	0	0
Orenburg oblast	6	0	6	0	0	0
Amur oblast	4	0	4	0	0	0
Pskov oblast	4	0	4	0	0	0
Irkutsk oblast	6	0	6	0	0	0
Saratov oblast	4	0	4	0	0	0
Tatarstan	10	0	10	0	0	0
Volvograd oblast	4	0	4	0	0	0
Bashkortostan	12	0	12	0	0	0
Kurgan oblast	4	0	4	0	0	0
Krasnodar kr	8	0	8	0	0	0
Novosibirsk oblast	4	0	4	0	0	0
St. Petersburg	4	0	4	0	0	0
Sverdlovsk oblast	8	0	8	0	0	0
Alania	4	0	4	0	0	0
Tambov oblast	4	0	4	0	0	0
Udmurtia	4	0	4	0	0	0
Perm oblast	6	0	6	0	0	0
Stavropol kr	4	0	4	0	0	0

Exhibit B.36 Allocation of School Sample in the Russian Federation *(continued)*

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Hanty-Mansii ok	4	0	4	0	0	0
Krasnoyarsk kr	6	0	6	0	0	0
Chuvashia	4	0	4	0	0	0
Sakha	4	0	4	0	0	0
Kemerovo oblast	4	0	4	0	0	0
Moscow	8	0	8	0	0	0
Moskva oblast	8	0	8	0	0	0
Orel oblast	4	0	4	0	0	0
Chelyabinsk oblast	6	0	6	0	0	0
Chita oblast	4	0	4	0	0	0
Omsk oblast	4	0	4	0	0	0
Total	232	0	232	0	0	0

B.37 Scotland

Coverage and Exclusions

Coverage is 100%

- School-level exclusions consisted of very small schools (MOS<5), Gaelic schools, and special education schools
- Within-school exclusions consisted of pupils with special education needs

Sample Design

- Explicit stratification by school location for a total of 6 explicit strata
- Implicit stratification by school deprivation index (low FSM, medium FSM, high FSM, Unknown FSM, independent) for a total of 29 implicit strata
- Sampled two classrooms per school having at least 50 students (MOS≥50), one classroom otherwise
- Small schools sampled with equal probabilities

Exhibit B.37 Allocation of School Sample in Scotland

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Large urban area	55	0	34	9	1	11
Other urban area	46	0	33	5	2	6
Accessible small town	17	0	14	1	1	1
Remore small town	5	0	3	0	0	2
Accessible rural area	19	0	13	3	3	0
Remore rural area	8	0	4	4	0	0
Total	150	0	101	22	7	20

B.38 Singapore

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of religious schools, private schools, and special education schools

Sample Design

- No explicit stratification
- No implicit stratification
- Sampled two classrooms per school. Classrooms were sampled with PPS. A sample of 19 students was drawn in each class
- All schools were sampled

Exhibit B.38 Allocation of School Sample in Singapore

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Singapore	178	0	178	0	0	0
Total	178	0	178	0	0	0

B.39 Slovak Republic

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<5), private schools, district board schools, civil association schools, and foreign language schools
- Within-school exclusions consisted of disabled students

Sample Design

- Explicit stratification by language (Slovak, Hungarian) and region within the 'Slovak schools' stratum for a total of 9 explicit strata
- Implicit stratification by region for Hungarian schools (Bratislavsky, Trnavsky, Nitriansky, Banskobystricky, Kosicky), school type for Slovak schools (public, church) and by school size for all strata (small, medium, large) for a total of 54 implicit strata
- Sampled two classrooms per school whenever possible
- Small schools sampled with equal probabilities

Exhibit B.39 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 - Bratislavsky	16	1	14	0	0	1
Slovak - Trnavsky	16	0	14	1	0	1
Slovak - Trenciansky	16	0	14	0	1	1
Slovak - Nitriansky	16	0	15	1	0	0
Slovak - Zilinsky	22	0	22	0	0	0
Slovak - Banskobystricky	16	0	16	0	0	0
Slovak - Presovsky	28	1	25	2	0	0
Slovak - Kosicky	22	1	17	3	0	1
Hungarian	22	0	18	3	1	0
Total	174	3	155	10	2	4

B.40 Slovenia

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<8), and Italian schools

- Within-school exclusions consisted of disabled students

Sample Design

- Explicit stratification by school system (old, new, old & new) for a total of 3 explicit strata
- No implicit stratification
- Sampled two classrooms per school whenever possible
- Small schools sampled with equal probabilities
- Twelve schools were sampled with certainty due to their (large) size

Exhibit B.40 Allocation of School Sample in Slovenia

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Old only	66	0	64	0	0	2
New only	80	0	72	4	1	3
Old and New	4	0	4	0	0	0
Total	150	0	140	4	1	5

B.41 South Africa

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<14), other language schools, schools in very small strata, and special education schools

Sample Design

- Explicit stratification by province and language for a total of 62 explicit strata
- Implicit stratification by region (32 regions) for a total of 250 implicit strata
- Sampled one classroom per school
- Small schools sampled with equal probabilities
- Seven schools were sampled with certainty due to their (large) size in Mpumalanga-Isindebele

Exhibit B.41 Allocation of School Sample in South Africa

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Eastern Cape - Afrikaans	3	0	3	0	0	0
Eastern Cape - English	17	1	14	1	0	1
Eastern Cape - Isixhosa	26	0	25	0	0	1
Eastern Cape - Sesotho	2	0	1	0	0	1
Eastern Cape - Eng. & Afr.	2	0	2	0	0	0
Eastern Cape - Bilingual	2	0	2	0	0	0
Eastern Cape - Missing	2	0	2	0	0	0
Free State - Afrikaans	2	0	2	0	0	0
Free State - English	3	0	3	0	0	0
Free State - Sesotho	21	1	20	0	0	0
Free State - Setswana	2	0	2	0	0	0
Free State - Eng. & Afr.	2	0	2	0	0	0
Free State - Bilingual	2	1	1	0	0	0
Free State - Missing	2	0	2	0	0	0
Gauteng - Afrikaans	5	0	3	1	0	1
Gauteng - English	18	1	17	0	0	0
Gauteng - Isixhosa	2	0	2	0	0	0
Gauteng - Isizulu	2	0	2	0	0	0
Gauteng - Sepedi	2	0	2	0	0	0
Gauteng - Sesotho	3	0	3	0	0	0
Gauteng - Setswana	2	0	2	0	0	0
Gauteng - Eng. & Afr.	3	0	2	1	0	0
Gauteng - Bilingual	3	0	3	0	0	0
Gauteng - Missing	2	0	1	0	1	0
Kwazulu Natal - Afrikaans	2	0	2	0	0	0
Kwazulu Natal - English	19	2	16	1	0	0
Kwazulu Natal - Isizulu	32	6	24	0	0	2
Kwazulu Natal - Eng. & Afr.	2	0	2	0	0	0
Kwazulu Natal - Bilingual	2	0	2	0	0	0
Kwazulu Natal - Missing	2	1	1	0	0	0
Limpopo - Afrikaans	2	0	2	0	0	0
Limpopo - English	11	0	11	0	0	0
Limpopo - Sepedi	22	2	19	0	0	1
Limpopo - Tshivenda	25	5	18	0	0	2

Exhibit B.41 Allocation of School Sample in South Africa *(continued)*

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Limpopo - Xitsonga	25	6	18	0	0	1
Limpopo - Eng. & Afr.	2	0	2	0	0	0
Limpopo - Bilingual	2	0	2	0	0	0
Limpopo - Missing	2	0	2	0	0	0
Mpumalanga - Afrikaans	2	0	2	0	0	0
Mpumalanga - English	11	0	9	0	0	2
Mpumalanga - Isindebele	25	0	25	0	0	0
Mpumalanga - Isizulu	4	0	3	0	0	1
Mpumalanga - Sepedi	2	0	2	0	0	0
Mpumalanga - Setswana	2	0	2	0	0	0
Mpumalanga - Siswati	25	1	24	0	0	0
Mpumalanga - Eng. & Afr.	2	0	2	0	0	0
Mpumalanga - Bilingual	3	0	3	0	0	0
Mpumalanga - Missing	2	1	1	0	0	0
Northern Cape - Afrikaans	6	0	6	0	0	0
Northern Cape - English	2	0	2	0	0	0
Northern Cape - Setswana	2	1	1	0	0	0
Northern Cape - Eng. & Afr.	3	0	3	0	0	0
Northern Cape - Missing	12	1	11	0	0	0
North West - Afrikaans	2	0	2	0	0	0
North West - English	2	0	2	0	0	0
North West - Setswana	22	1	21	0	0	0
North West - Eng. & Afr.	2	0	2	0	0	0
North West - Missing	2	0	2	0	0	0
Western Cape - Afrikaans	12	0	12	0	0	0
Western Cape - English	4	0	3	1	0	0
Western Cape - Isixhosa	5	0	5	0	0	0
Western Cape - Eng. & Afr.	7	0	7	0	0	0
Total	441	31	391	5	1	13

B.42 Spain

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<7), and special education schools
- Within-school exclusions consisted of disabled students and non-native language speakers (less than a year of instruction in the language of test)

Sample Design

- Explicit stratification by autonomous communities for a total of 18 explicit strata
- Implicit stratification by school type (public, private) for a total of 36 implicit strata
- Sampled two classrooms per school with at least 55 students (MOS≥55) and one classroom otherwise
- Small schools sampled with equal probabilities

Exhibit B.42 Allocation of School Sample in Spain

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Andalucia	32	0	32	0	0	0
Aragon	4	0	4	0	0	0
Asturias	2	0	2	0	0	0
Baleares (Islas)	3	0	3	0	0	0
Canarias	7	0	6	1	0	0
Cantabria	2	0	1	1	0	0
Castilla-La Mancha	7	0	7	0	0	0
Castilla y Leon	7	0	7	0	0	0
Cataluna	22	0	22	0	0	0
Comunidad Valenciana	16	0	16	0	0	0
Extremadura	4	0	4	0	0	0
Galicia	8	0	8	0	0	0
Madrid	20	0	20	0	0	0
Murcia (Región de)	6	0	6	0	0	0
Navarra	2	0	2	0	0	0
Pais Vasco	6	0	6	0	0	0
La Rioja	2	0	1	0	1	0
Ceuta y Melilla	2	0	2	0	0	0
Total	152	0	149	2	1	0

B.43 Sweden

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<5), non-Swedish speaking schools, hospital and refugee schools, Sami schools, and special education schools
- Within-school exclusions consisted of disabled students and non native language speakers (one year or less of Swedish instruction)

Sample Design

- Explicit stratification by school type (private, public) for a total of 2 explicit strata
- No implicit stratification
- Sampled two classrooms per school whenever possible
- Small schools sampled with equal probabilities

Exhibit B.43 Allocation of School Sample in Sweden

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Public	120	2	118	0	0	0
Private	30	1	29	0	0	0
Total	150	3	147	0	0	0

B.44 Trinidad and Tobago

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<6)

Sample Design

- Explicit stratification by region for a total of 8 explicit strata
- Implicit stratification by school type (private, government, denominational) and gender (mixed, girls, boys) for a total of 38 implicit strata
- Sampled two classrooms per school having at least 75 students (MOS≥75), and one classroom otherwise

- Small schools sampled with equal probabilities
- 17 schools were sampled with certainty due to their (large) size

Exhibit B.44 Allocation of School Sample in Trinidad & Tobago

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
Caroni	22	0	22	0	0	0
North Eastern	9	0	9	0	0	0
Port of Spain & Environs	29	0	27	0	0	2
St George East	34	0	34	0	0	0
St Patrick	15	0	15	0	0	0
South Eastern	12	0	12	0	0	0
Victoria	21	1	20	0	0	0
Tobago	8	0	8	0	0	0
Total	150	1	147	0	0	2

B.45 United States

Coverage and Exclusions

- Coverage is 100%
- School-level exclusions consisted of very small schools (MOS<11), special education, vocational and alternative public schools, and special education, vocational and alternative private schools. Note that students in the five U.S. Territories of American Samoa, Guam, Northern Marianas, Puerto Rico, and the Virgin Islands were considered out of scope. Students enrolled in foreign Department of Defense schools were also considered out of scope.
- Within-school exclusions consisted of special education students, and English language learners (students with < 1 year of English instruction)

Sample Design

- Explicit stratification by Metropolitan Statistical Area (MSA) status (the 10 largest MSA versus all other MSAs) for a total of two explicit strata
- Within the 10 largest MSA, implicit stratification by MSA, Common Core of Data/ Private School Survey (CCDPSS) (1 or 2), poverty indicator (high, low), and school size. A sample of 70 schools was drawn with PPS where schools with CCDPSS=2 and schools with a high poverty status were given more chances to be drawn in the sample.

- In the other explicit stratum (all other MSAs), implicit stratification by PSU code (counties or contiguous counties). A sample of 38 PSUs was drawn with PPS. Within each sampled PSU, another implicit stratification was done by CCDPSS, poverty indicator, and school size. Schools with CCDPSS=2 and schools with a high poverty status also were given a greater chance of being sampled. Finally, a PPS sample of 4 schools was drawn within each selected PSU.
- Sampled one or two classrooms per school

Exhibit B.45 Allocation of School Sample in the United States

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
USA - Self-Representative Units	70	1	32	14	9	14
USA - PSU 23	4	1	1	1	1	0
USA - PSU 42	4	0	3	1	0	0
USA - PSU 53	4	0	4	0	0	0
USA - PSU 68	4	0	2	1	0	1
USA - PSU 76	4	1	2	1	0	0
USA - PSU 92	4	0	4	0	0	0
USA - PSU 115	4	0	2	1	1	0
USA - PSU 127	4	1	1	0	2	0
USA - PSU 134	4	0	1	0	2	1
USA - PSU 159	4	0	2	1	0	1
USA - PSU 172	4	0	1	0	0	3
USA - PSU 200	4	0	3	0	1	0
USA - PSU 210	4	0	1	0	1	2
USA - PSU 215	4	0	0	1	2	1
USA - PSU 224	4	0	4	0	0	0
USA - PSU 244	4	0	0	0	3	1
USA - PSU 251	4	0	0	1	1	2
USA - PSU 264	4	0	3	0	1	0
USA - PSU 288	4	1	2	1	0	0
USA - PSU 294	4	0	3	0	1	0
USA - PSU 300	4	1	2	1	0	0
USA - PSU 314	4	0	4	0	0	0
USA - PSU 322	4	0	3	0	0	1

Exhibit B.45 Allocation of School Sample in the United States *(continued)*

Explicit Stratum	Total Sampled Schools	Ineligible Schools	Participating Schools			Non-Participating Schools
			Sampled	1st Replacement	2nd Replacement	
USA - PSU 336	4	0	4	0	0	0
USA - PSU 343	4	0	3	0	0	1
USA - PSU 366	4	0	2	2	0	0
USA - PSU 374	4	0	3	0	1	0
USA - PSU 381	4	0	3	0	0	1
USA - PSU 386	4	0	3	0	1	0
USA - PSU 397	4	0	1	2	1	0
USA - PSU 404	4	0	2	1	1	0
USA - PSU 410	4	0	3	1	0	0
USA - PSU 417	4	1	3	0	0	0
USA - PSU 422	4	0	2	0	0	2
USA - PSU 428	4	1	2	1	0	0
USA - PSU 434	4	0	3	0	1	0
USA - PSU 439	4	0	4	0	0	0
USA - PSU 446	4	0	2	2	0	0
Total	222	8	120	33	30	31

Appendix C

Country Adaptations to Items and Item Scoring

Items to Be Deleted

All Countries

R021S08C (scaling did not converge/ item discrimination too low for many countries)

Indonesia

R021E08M (item mistranslated)

Kuwait

R021K10C (item mistranslated)

Qatar

R021U06C (item mistranslated)

R021S05C (item mistranslated)

Items Needing Options Changed

R021U04M (recoded A to D and D to A)

Constructed-response Items Needing Category Recoding

All Countries

R011F12C (Recoded 3 into 2)

Kuwait

R021K10C (Recoded 3 into 2)

R021S13C (Recoded 2 into 1)

Appendix D

Parameters for IRT Analysis of PIRLS Achievement Data

Exhibit D.1 IRT Parameters for PIRLS Joint 2001-2006 Scaling of Overall Reading

Item	Slope (a_j)	Location (b_j)	Guessing (c_j)	Step 1 (d_{j1})	Step 2 (d_{j2})	Step 3 (d_{j3})
R011H01M	0.719 (0.063)	-1.516 (0.203)	0.315 (0.068)			
R011H02M	0.992 (0.067)	-1.410 (0.108)	0.207 (0.048)			
R011H03C	0.369 (0.020)	0.676 (0.047)		0.644 (0.065)	-0.644 (0.079)	
R011H04C	0.934 (0.041)	-1.117 (0.047)				
R011H05M	1.255 (0.078)	-1.002 (0.072)	0.208 (0.037)			
R011H06M	0.879 (0.060)	-0.503 (0.081)	0.159 (0.035)			
R011H07C	0.636 (0.025)	-0.613 (0.031)		0.276 (0.053)	-0.276 (0.040)	
R011H08C	0.821 (0.038)	-0.069 (0.030)				
R011H09C	0.774 (0.028)	-0.703 (0.029)		0.064 (0.049)	-0.064 (0.036)	
R011H10C	0.675 (0.019)	0.300 (0.018)		-0.164 (0.055)	1.105 (0.052)	-0.941 (0.047)
R011H11M	1.405 (0.086)	-0.483 (0.053)	0.195 (0.031)			
R011M01M	1.336 (0.093)	-0.658 (0.071)	0.334 (0.035)			
R011M02M	1.224 (0.085)	-1.234 (0.094)	0.311 (0.046)			
R011M03M	1.407 (0.093)	0.161 (0.038)	0.201 (0.020)			
R011M04C	0.807 (0.040)	0.591 (0.034)				
R011M05M	1.272 (0.084)	-0.523 (0.064)	0.267 (0.032)			
R011M06C	1.067 (0.036)	-0.555 (0.020)		0.281 (0.035)	-0.281 (0.025)	
R011M07C	1.128 (0.045)	-0.705 (0.031)				
R011M08M	1.198 (0.119)	0.657 (0.047)	0.270 (0.021)			
R011M09M	1.205 (0.073)	-0.680 (0.063)	0.189 (0.033)			
R011M10C	1.183 (0.056)	-1.566 (0.052)				
R011M11C	0.832 (0.040)	0.393 (0.030)				
R011M12C	0.621 (0.023)	0.541 (0.022)		0.768 (0.044)	-0.119 (0.048)	-0.649 (0.063)
R011M13M	0.958 (0.086)	-0.227 (0.098)	0.340 (0.039)			

( ) Standard errors appear in parentheses

Exhibit D.1 IRT Parameters for PIRLS Joint 2001-2006 Scaling of Overall Reading *(continued)*

Item	Slope (a_j)	Location (b_j)	Guessing (c_j)	Step 1 (d_{j1})	Step 2 (d_{j2})	Step 3 (d_{j3})
R011M14C	0.936 (0.040)	-0.402 (0.031)				
R021E01M	1.386 (0.114)	-1.132 (0.098)	0.381 (0.046)			
R021E02M	1.177 (0.092)	-0.422 (0.078)	0.298 (0.036)			
R021E03M	0.545 (0.059)	-0.174 (0.152)	0.163 (0.047)			
R021E04M	1.345 (0.103)	-1.204 (0.093)	0.303 (0.046)			
R021E05C	0.691 (0.024)	-0.446 (0.028)		-0.355 (0.056)	0.355 (0.047)	
R021E06M	1.355 (0.096)	-0.249 (0.056)	0.236 (0.029)			
R021E07C	0.677 (0.028)	-0.202 (0.027)		0.225 (0.049)	-0.225 (0.042)	
R021E08M	1.437 (0.110)	0.432 (0.035)	0.162 (0.019)			
R021E09C	0.607 (0.029)	0.604 (0.032)		0.531 (0.045)	-0.531 (0.056)	
R021E10C	1.072 (0.049)	-0.220 (0.029)				
R021E11M	0.874 (0.082)	0.126 (0.081)	0.203 (0.034)			
R021E12C	0.844 (0.035)	0.147 (0.021)		0.335 (0.037)	-0.335 (0.036)	
R021U01M	0.667 (0.076)	-0.248 (0.156)	0.278 (0.050)			
R021U02M	1.125 (0.079)	-0.920 (0.086)	0.213 (0.040)			
R021U03M	0.677 (0.072)	-0.099 (0.127)	0.217 (0.044)			
R021U04M	0.755 (0.077)	0.110 (0.097)	0.202 (0.037)			
R021U05C	1.008 (0.047)	-0.704 (0.039)				
R021U06C	0.922 (0.044)	-0.648 (0.040)				
R021U07M	0.740 (0.072)	-0.862 (0.170)	0.301 (0.058)			
R021U08C	0.985 (0.038)	-0.276 (0.021)		0.333 (0.038)	-0.333 (0.030)	
R021U09M	0.994 (0.090)	-0.264 (0.095)	0.304 (0.039)			
R021U10C	0.803 (0.042)	-0.966 (0.054)				
R021U11C	0.616 (0.026)	0.395 (0.022)		0.519 (0.052)	-0.345 (0.058)	-0.174 (0.065)
R021U12C	0.811 (0.036)	-0.156 (0.025)		0.455 (0.044)	-0.455 (0.036)	
R021Y01M	1.223 (0.105)	0.197 (0.055)	0.258 (0.027)			
R021Y02M	1.696 (0.121)	-0.191 (0.046)	0.281 (0.027)			
R021Y03C	0.897 (0.047)	0.490 (0.032)				
R021Y04M	1.282 (0.096)	0.098 (0.048)	0.203 (0.025)			
R021Y05M	1.806 (0.123)	0.098 (0.034)	0.214 (0.021)			
R021Y06M	1.719 (0.120)	0.121 (0.036)	0.218 (0.022)			
R021Y07M	0.889 (0.065)	-0.946 (0.106)	0.181 (0.044)			
R021Y08M	1.526 (0.109)	-0.272 (0.052)	0.269 (0.029)			
R021Y09C	1.005 (0.036)	-0.557 (0.024)		0.061 (0.043)	-0.061 (0.032)	
R021Y10C	0.821 (0.045)	0.465 (0.034)				
R021Y11M	1.553 (0.120)	0.035 (0.048)	0.288 (0.026)			
R021Y12C	0.752 (0.022)	0.003 (0.021)		-1.014 (0.059)	1.014 (0.057)	
R021Y13C	0.778 (0.030)	0.380 (0.018)		0.553 (0.042)	-0.237 (0.045)	-0.316 (0.052)
R021Y14C	0.625 (0.023)	0.206 (0.025)		-0.477 (0.055)	0.477 (0.055)	
R011C01C	1.366 (0.038)	-0.328 (0.017)				
R011C02C	0.857 (0.029)	0.239 (0.020)				

( ) Standard errors appear in parentheses

Exhibit D.1 IRT Parameters for PIRLS Joint 2001-2006 Scaling of Overall Reading *(continued)*

Item	Slope (a_j)	Location (b_j)	Guessing (c_j)	Step 1 (d_{j1})	Step 2 (d_{j2})	Step 3 (d_{j3})
R011C03C	1.367 (0.039)	-0.608 (0.020)				
R011C04M	1.373 (0.068)	0.197 (0.028)	0.182 (0.015)			
R011C05M	0.892 (0.064)	-0.186 (0.081)	0.359 (0.029)			
R011C06C	1.134 (0.033)	-0.253 (0.018)				
R011C07M	1.184 (0.061)	-0.464 (0.052)	0.268 (0.026)			
R011C08C	0.613 (0.016)	0.143 (0.017)		-0.310 (0.035)	0.310 (0.035)	
R011C09M	1.289 (0.073)	0.528 (0.027)	0.159 (0.014)			
R011C10C	0.657 (0.016)	0.124 (0.013)		0.192 (0.035)	-0.266 (0.038)	0.073 (0.037)
R011C11C	0.812 (0.021)	0.027 (0.016)		0.737 (0.027)	-0.737 (0.025)	
R011C12M	0.847 (0.058)	0.097 (0.064)	0.225 (0.027)			
R011C13M	0.954 (0.065)	0.261 (0.051)	0.227 (0.023)			
R011F01M	1.357 (0.059)	-0.601 (0.040)	0.192 (0.022)			
R011F02M	0.666 (0.041)	-0.757 (0.109)	0.188 (0.039)			
R011F03M	0.931 (0.043)	-0.778 (0.061)	0.143 (0.027)			
R011F04M	1.285 (0.060)	-0.921 (0.054)	0.240 (0.028)			
R011F05M	0.971 (0.053)	-0.377 (0.061)	0.236 (0.027)			
R011F06C	0.841 (0.027)	-0.409 (0.025)				
R011F07C	0.538 (0.013)	0.355 (0.019)		-0.799 (0.042)	0.799 (0.044)	
R011F08C	1.149 (0.034)	-0.280 (0.018)				
R011F09C	1.055 (0.026)	-0.697 (0.017)		0.071 (0.030)	-0.071 (0.021)	
R011F10C	0.913 (0.031)	-1.246 (0.039)				
R011F11M	0.735 (0.052)	0.199 (0.068)	0.175 (0.027)			
R011F12C	0.679 (0.018)	0.571 (0.017)		-0.303 (0.031)	0.303 (0.035)	
R011F13M	1.102 (0.063)	-0.035 (0.048)	0.240 (0.023)			
R011N01M	0.937 (0.074)	-0.342 (0.090)	0.269 (0.037)			
R011N02M	0.846 (0.079)	0.115 (0.085)	0.251 (0.034)			
R011N03M	1.082 (0.075)	-0.780 (0.086)	0.266 (0.040)			
R011N04M	1.234 (0.089)	0.324 (0.040)	0.170 (0.021)			
R011N05M	1.432 (0.096)	0.110 (0.040)	0.220 (0.022)			
R011N06M	1.872 (0.144)	0.708 (0.026)	0.175 (0.013)			
R011N07C	0.643 (0.027)	0.492 (0.026)		0.277 (0.040)	-0.277 (0.046)	
R011N08C	0.640 (0.025)	0.086 (0.026)		0.743 (0.043)	-0.743 (0.042)	
R011N09M	1.283 (0.080)	-0.485 (0.057)	0.208 (0.031)			
R011N10C	0.962 (0.049)	0.887 (0.037)				
R011N11M	1.185 (0.089)	0.260 (0.047)	0.198 (0.024)			
R011N12C	0.842 (0.029)	0.370 (0.019)		-0.035 (0.034)	0.035 (0.036)	
R011N13C	0.651 (0.037)	0.038 (0.037)				
R011R01M	0.828 (0.064)	-0.115 (0.079)	0.178 (0.033)			
R011R02M	1.262 (0.097)	0.400 (0.041)	0.213 (0.020)			
R011R03M	0.817 (0.062)	-1.216 (0.142)	0.268 (0.055)			
R011R04C	0.922 (0.040)	-1.175 (0.047)				

( ) Standard errors appear in parentheses

Exhibit D.1 IRT Parameters for PIRLS Joint 2001-2006 Scaling of Overall Reading *(continued)*

Item	Slope (a_j)	Location (b_j)	Guessing (c_j)	Step 1 (d_{j1})	Step 2 (d_{j2})	Step 3 (d_{j3})
R011R05C	1.130 (0.045)	-0.798 (0.032)				
R011R06C	0.651 (0.020)	-0.356 (0.024)		-0.492 (0.051)	0.492 (0.046)	
R011R07C	1.042 (0.043)	0.010 (0.024)				
R011R08C	0.861 (0.031)	0.130 (0.019)		0.313 (0.032)	-0.313 (0.032)	
R011R09C	0.671 (0.025)	-0.174 (0.024)		0.101 (0.044)	-0.101 (0.039)	
R011R10C	0.391 (0.015)	0.189 (0.031)		0.993 (0.082)	0.544 (0.069)	-1.536 (0.083)
R011R11C	0.635 (0.024)	-0.223 (0.021)		0.374 (0.060)	0.185 (0.051)	-0.559 (0.044)
R021K01C	0.432 (0.022)	-1.007 (0.056)		0.171 (0.087)	-0.171 (0.064)	
R021K02C	0.794 (0.039)	-0.748 (0.046)				
R021K03M	1.063 (0.086)	0.074 (0.061)	0.201 (0.029)			
R021K04M	0.848 (0.177)	1.192 (0.109)	0.377 (0.029)			
R021K05C	1.072 (0.048)	-0.002 (0.026)				
R021K06M	1.409 (0.100)	-0.127 (0.050)	0.230 (0.027)			
R021K07C	0.745 (0.030)	-0.034 (0.023)		0.113 (0.044)	-0.113 (0.040)	
R021K08M	1.098 (0.095)	0.304 (0.055)	0.195 (0.026)			
R021K09M	1.149 (0.093)	-0.016 (0.063)	0.235 (0.031)			
R021K10C	0.803 (0.030)	0.696 (0.025)		-0.305 (0.042)	0.305 (0.049)	
R021K11M	1.119 (0.101)	0.250 (0.061)	0.235 (0.029)			
R021K12C	0.621 (0.024)	-0.123 (0.022)		0.292 (0.063)	-0.053 (0.059)	-0.239 (0.053)
R021N01M	0.918 (0.083)	-0.575 (0.119)	0.320 (0.046)			
R021N02C	0.868 (0.042)	-0.466 (0.037)				
R021N03C	0.755 (0.036)	0.857 (0.032)		0.275 (0.038)	-0.275 (0.052)	
R021N04M	1.436 (0.107)	0.226 (0.041)	0.197 (0.022)			
R021N05M	1.630 (0.127)	-0.760 (0.070)	0.351 (0.038)			
R021N06M	1.633 (0.104)	-0.457 (0.046)	0.186 (0.028)			
R021N07M	1.205 (0.089)	-0.086 (0.057)	0.202 (0.029)			
R021N08C	0.988 (0.046)	-0.291 (0.031)				
R021N09M	1.260 (0.099)	-0.207 (0.066)	0.275 (0.033)			
R021N10M	0.947 (0.107)	0.288 (0.088)	0.320 (0.035)			
R021N11C	0.633 (0.023)	0.036 (0.025)		-0.415 (0.054)	0.415 (0.052)	
R021N12C	0.663 (0.030)	0.057 (0.026)		0.165 (0.048)	-0.165 (0.046)	
R021S01M	0.846 (0.071)	-0.014 (0.076)	0.152 (0.032)			
R021S02M	0.510 (0.054)	-0.605 (0.197)	0.164 (0.057)			
R021S03M	0.974 (0.079)	-0.559 (0.097)	0.255 (0.041)			
R021S04M	1.424 (0.112)	0.070 (0.049)	0.268 (0.026)			
R021S05C	0.803 (0.041)	-0.142 (0.034)				
R021S06M	1.279 (0.101)	-0.657 (0.083)	0.328 (0.039)			
R021S07C	0.484 (0.018)	0.735 (0.036)		-1.148 (0.075)	1.148 (0.083)	
R021S09C	0.886 (0.043)	-0.302 (0.034)				
R021S10C	0.845 (0.043)	-0.162 (0.034)				
R021S11C	0.727 (0.045)	0.757 (0.047)				

( ) Standard errors appear in parentheses

Exhibit D.1 IRT Parameters for PIRLS Joint 2001-2006 Scaling of Overall Reading *(continued)*

Item	Slope (a_j)	Location (b_j)	Guessing (c_j)	Step 1 (d_{j1})	Step 2 (d_{j2})	Step 3 (d_{j3})
R021S12C	0.631 (0.045)	1.074 (0.070)				
R021S13C	0.687 (0.047)	1.063 (0.065)				
R021S14M	1.617 (0.135)	0.489 (0.035)	0.207 (0.020)			
R021S15C	0.574 (0.021)	0.327 (0.027)		-0.782 (0.063)	0.782 (0.065)	
R011A01C	0.905 (0.031)	-1.367 (0.041)				
R011A02M	1.174 (0.066)	0.066 (0.041)	0.250 (0.020)			
R011A03C	0.798 (0.027)	-0.987 (0.036)				
R011A04C	0.787 (0.018)	-0.161 (0.018)		1.035 (0.030)	-1.035 (0.025)	
R011A05M	1.053 (0.054)	-1.215 (0.078)	0.250 (0.036)			
R011A06M	1.060 (0.056)	-1.254 (0.083)	0.279 (0.038)			
R011A07C	0.719 (0.017)	-0.555 (0.016)		0.167 (0.044)	-0.008 (0.038)	-0.159 (0.029)
R011A08C	0.606 (0.019)	-0.888 (0.028)		0.480 (0.046)	-0.480 (0.030)	
R011A09C	0.724 (0.021)	-0.129 (0.017)		0.532 (0.030)	-0.532 (0.026)	
R011A10M	1.426 (0.060)	-0.104 (0.028)	0.127 (0.016)			
R011A11C	0.844 (0.029)	-0.026 (0.021)				
R011L01M	0.539 (0.038)	-2.304 (0.260)	0.256 (0.080)			
R011L02M	0.836 (0.066)	0.554 (0.052)	0.221 (0.021)			
R011L03C	0.661 (0.024)	-0.494 (0.031)				
R011L04C	0.625 (0.013)	0.329 (0.017)		1.455 (0.035)	-0.984 (0.036)	-0.471 (0.048)
R011L05M	1.283 (0.078)	0.541 (0.029)	0.208 (0.014)			
R011L06C	0.765 (0.027)	0.055 (0.023)				
R011L07M	0.841 (0.057)	0.426 (0.048)	0.167 (0.021)			
R011L08C	0.822 (0.022)	0.563 (0.017)		0.652 (0.023)	-0.652 (0.029)	
R011L09M	0.994 (0.051)	-0.888 (0.072)	0.229 (0.033)			
R011L10C	0.798 (0.023)	0.581 (0.016)		0.092 (0.025)	-0.092 (0.030)	
R011L11M	0.909 (0.051)	-0.333 (0.064)	0.201 (0.028)			
R011L12C	0.827 (0.022)	0.510 (0.017)		0.702 (0.023)	-0.702 (0.029)	

( ) Standard errors appear in parentheses

Exhibit D.2 IRT Parameters for PIRLS Joint 2001-2006 Scaling of Literary Purposes

Item	Slope (a_j)	Location (b_j)	Guessing (c_j)	Step 1 (d_{j1})	Step 2 (d_{j2})	Step 3 (d_{j3})
R011H01M	0.665 (0.059)	-1.438 (0.226)	0.313 (0.071)			
R011H02M	0.953 (0.065)	-1.277 (0.115)	0.214 (0.048)			
R011H03C	0.339 (0.018)	0.942 (0.051)		0.700 (0.071)	-0.700 (0.086)	
R011H04C	0.895 (0.039)	-0.977 (0.048)				
R011H05M	1.246 (0.077)	-0.829 (0.071)	0.218 (0.036)			
R011H06M	0.881 (0.059)	-0.275 (0.081)	0.176 (0.034)			
R011H07C	0.621 (0.024)	-0.436 (0.032)		0.312 (0.055)	-0.312 (0.041)	
R011H08C	0.837 (0.037)	0.145 (0.030)				
R011H09C	0.789 (0.028)	-0.520 (0.028)		0.114 (0.049)	-0.114 (0.036)	
R011H10C	0.642 (0.018)	0.526 (0.019)		-0.148 (0.058)	1.159 (0.055)	-1.012 (0.050)
R011H11M	1.253 (0.077)	-0.330 (0.060)	0.198 (0.031)			
R011M01M	1.171 (0.080)	-0.564 (0.079)	0.319 (0.035)			
R011M02M	1.112 (0.078)	-1.174 (0.104)	0.305 (0.044)			
R011M03M	1.317 (0.086)	0.377 (0.040)	0.198 (0.020)			
R011M04C	0.770 (0.037)	0.835 (0.035)				
R011M05M	1.197 (0.077)	-0.357 (0.066)	0.269 (0.031)			
R011M06C	1.049 (0.035)	-0.392 (0.021)		0.332 (0.037)	-0.332 (0.026)	
R011M07C	1.000 (0.040)	-0.589 (0.035)				
R011M08M	1.167 (0.115)	0.932 (0.048)	0.278 (0.020)			
R011M09M	1.153 (0.068)	-0.522 (0.064)	0.192 (0.031)			
R011M10C	1.074 (0.050)	-1.546 (0.057)				
R011M11C	0.786 (0.037)	0.627 (0.031)				
R011M12C	0.575 (0.021)	0.792 (0.024)		0.846 (0.048)	-0.135 (0.051)	-0.710 (0.068)
R011M13M	0.898 (0.080)	-0.019 (0.102)	0.349 (0.037)			
R011M14C	0.853 (0.037)	-0.243 (0.034)				
R021E01M	1.279 (0.105)	-1.040 (0.105)	0.365 (0.046)			
R021E02M	1.057 (0.080)	-0.309 (0.084)	0.267 (0.036)			
R021E03M	0.531 (0.056)	0.025 (0.153)	0.160 (0.045)			
R021E04M	1.332 (0.101)	-1.072 (0.091)	0.286 (0.043)			
R021E05C	0.644 (0.022)	-0.268 (0.030)		-0.379 (0.060)	0.379 (0.051)	
R021E06M	1.313 (0.090)	-0.066 (0.056)	0.223 (0.028)			
R021E07C	0.637 (0.027)	-0.003 (0.029)		0.246 (0.052)	-0.246 (0.045)	
R021E08M	1.417 (0.107)	0.681 (0.036)	0.167 (0.018)			
R021E09C	0.562 (0.027)	0.864 (0.035)		0.573 (0.049)	-0.573 (0.060)	
R021E10C	0.977 (0.045)	-0.031 (0.031)				
R021E11M	0.857 (0.079)	0.361 (0.081)	0.208 (0.033)			
R021E12C	0.788 (0.033)	0.369 (0.023)		0.363 (0.040)	-0.363 (0.039)	
R021U01M	0.617 (0.070)	-0.067 (0.171)	0.276 (0.051)			
R021U02M	1.125 (0.077)	-0.746 (0.083)	0.214 (0.038)			
R021U03M	0.623 (0.065)	0.082 (0.139)	0.211 (0.045)			
R021U04M	0.701 (0.070)	0.312 (0.104)	0.197 (0.037)			

( ) Standard errors appear in parentheses

Exhibit D.2 IRT Parameters for PIRLS Joint 2001-2006 Scaling of Literary Purposes *(continued)*

Item	Slope (a_j)	Location (b_j)	Guessing (c_j)	Step 1 (d_{j1})	Step 2 (d_{j2})	Step 3 (d_{j3})
R021U05C	0.982 (0.045)	-0.532 (0.040)				
R021U06C	0.907 (0.042)	-0.466 (0.040)				
R021U07M	0.682 (0.067)	-0.724 (0.188)	0.304 (0.059)			
R021U08C	1.018 (0.039)	-0.076 (0.021)		0.377 (0.037)	-0.377 (0.029)	
R021U09M	0.865 (0.078)	-0.147 (0.109)	0.283 (0.041)			
R021U10C	0.723 (0.038)	-0.866 (0.060)				
R021U11C	0.523 (0.022)	0.643 (0.026)		0.570 (0.060)	-0.402 (0.069)	-0.168 (0.077)
R021U12C	0.729 (0.032)	0.021 (0.028)		0.504 (0.049)	-0.504 (0.040)	
R021Y01M	1.158 (0.099)	0.430 (0.058)	0.260 (0.026)			
R021Y02M	1.624 (0.116)	0.027 (0.048)	0.289 (0.026)			
R021Y03C	0.827 (0.043)	0.743 (0.035)				
R021Y04M	1.193 (0.089)	0.319 (0.052)	0.203 (0.025)			
R021Y05M	1.712 (0.117)	0.325 (0.036)	0.215 (0.021)			
R021Y06M	1.629 (0.114)	0.353 (0.038)	0.221 (0.021)			
R021Y07M	0.828 (0.061)	-0.809 (0.114)	0.182 (0.043)			
R021Y08M	1.424 (0.100)	-0.090 (0.056)	0.262 (0.029)			
R021Y09C	0.938 (0.034)	-0.390 (0.026)		0.076 (0.046)	-0.076 (0.034)	
R021Y10C	0.770 (0.042)	0.712 (0.037)				
R021Y11M	1.458 (0.114)	0.262 (0.051)	0.292 (0.026)			
R021Y12C	0.695 (0.020)	0.216 (0.023)		-1.095 (0.064)	1.095 (0.062)	
R021Y13C	0.721 (0.028)	0.623 (0.020)		0.601 (0.045)	-0.260 (0.049)	-0.342 (0.056)
R021Y14C	0.575 (0.022)	0.435 (0.027)		-0.518 (0.059)	0.518 (0.060)	
R011C01C	1.296 (0.036)	-0.148 (0.018)				
R011C02C	0.790 (0.027)	0.463 (0.022)				
R011C03C	1.262 (0.036)	-0.464 (0.021)				
R011C04M	1.259 (0.062)	0.418 (0.030)	0.181 (0.015)			
R011C05M	0.866 (0.061)	0.054 (0.080)	0.375 (0.027)			
R011C06C	1.069 (0.031)	-0.068 (0.020)				
R011C07M	1.103 (0.056)	-0.286 (0.056)	0.277 (0.024)			
R011C08C	0.567 (0.014)	0.358 (0.018)		-0.328 (0.038)	0.328 (0.038)	
R011C09M	1.166 (0.067)	0.790 (0.029)	0.161 (0.014)			
R011C10C	0.591 (0.014)	0.336 (0.014)		0.214 (0.039)	-0.299 (0.042)	0.086 (0.041)
R011C11C	0.739 (0.019)	0.228 (0.018)		0.815 (0.030)	-0.815 (0.027)	
R011C12M	0.781 (0.054)	0.321 (0.069)	0.231 (0.026)			
R011C13M	0.863 (0.060)	0.503 (0.056)	0.233 (0.023)			
R011F01M	1.283 (0.055)	-0.442 (0.042)	0.193 (0.022)			
R011F02M	0.638 (0.040)	-0.557 (0.116)	0.207 (0.038)			
R011F03M	0.884 (0.041)	-0.615 (0.064)	0.152 (0.027)			
R011F04M	1.205 (0.057)	-0.785 (0.057)	0.246 (0.027)			
R011F05M	0.907 (0.050)	-0.185 (0.065)	0.245 (0.026)			
R011F06C	0.794 (0.025)	-0.234 (0.026)				

( ) Standard errors appear in parentheses

Exhibit D.2 IRT Parameters for PIRLS Joint 2001-2006 Scaling of Literary Purposes *(continued)*

Item	Slope (a_j)	Location (b_j)	Guessing (c_j)	Step 1 (d_{j1})	Step 2 (d_{j2})	Step 3 (d_{j3})
R011F07C	0.515 (0.012)	0.582 (0.019)		-0.819 (0.044)	0.819 (0.046)	
R011F08C	1.095 (0.032)	-0.095 (0.019)				
R011F09C	1.030 (0.025)	-0.546 (0.018)		0.110 (0.031)	-0.110 (0.022)	
R011F10C	0.858 (0.029)	-1.144 (0.041)				
R011F11M	0.662 (0.049)	0.432 (0.076)	0.180 (0.027)			
R011F12C	0.607 (0.016)	0.830 (0.019)		-0.347 (0.035)	0.347 (0.039)	
R011F13M	1.017 (0.058)	0.167 (0.052)	0.242 (0.023)			

( ) Standard errors appear in parentheses

Exhibit D.3 IRT Parameters for PIRLS Joint 2001-2006 Scaling of Informational Purposes

Item	Slope (a_j)	Location (b_j)	Guessing (c_j)	Step 1 (d_{j1})	Step 2 (d_{j2})	Step 3 (d_{j3})
R011N01M	0.887 (0.067)	-0.293 (0.089)	0.248 (0.035)			
R011N02M	0.864 (0.073)	0.193 (0.076)	0.233 (0.030)			
R011N03M	1.138 (0.081)	-0.579 (0.082)	0.318 (0.035)			
R011N04M	1.293 (0.089)	0.455 (0.037)	0.173 (0.019)			
R011N05M	1.566 (0.101)	0.240 (0.036)	0.220 (0.020)			
R011N06M	1.926 (0.141)	0.825 (0.025)	0.168 (0.013)			
R011N07C	0.654 (0.027)	0.614 (0.025)		0.299 (0.040)	-0.299 (0.046)	
R011N08C	0.624 (0.024)	0.196 (0.027)		0.783 (0.044)	-0.783 (0.043)	
R011N09M	1.267 (0.077)	-0.399 (0.056)	0.201 (0.028)			
R011N10C	0.947 (0.048)	1.027 (0.038)				
R011N11M	1.152 (0.079)	0.319 (0.044)	0.166 (0.022)			
R011N12C	0.694 (0.025)	0.514 (0.022)		-0.090 (0.041)	0.090 (0.044)	
R011N13C	0.551 (0.033)	0.119 (0.043)				
R011R01M	0.818 (0.062)	0.041 (0.078)	0.173 (0.032)			
R011R02M	1.317 (0.096)	0.535 (0.038)	0.201 (0.020)			
R011R03M	0.884 (0.067)	-0.930 (0.128)	0.303 (0.050)			
R011R04C	0.942 (0.041)	-0.998 (0.046)				
R011R05C	1.220 (0.049)	-0.600 (0.030)				
R011R06C	0.718 (0.021)	-0.169 (0.022)		-0.406 (0.047)	0.406 (0.042)	
R011R07C	1.108 (0.044)	0.178 (0.022)				
R011R08C	0.880 (0.032)	0.293 (0.018)		0.320 (0.032)	-0.320 (0.031)	
R011R09C	0.641 (0.024)	-0.024 (0.025)		0.096 (0.046)	-0.096 (0.041)	
R011R10C	0.388 (0.015)	0.348 (0.031)		1.011 (0.083)	0.545 (0.070)	-1.556 (0.084)
R011R11C	0.661 (0.024)	-0.062 (0.021)		0.400 (0.059)	0.171 (0.049)	-0.571 (0.042)
R021K01C	0.448 (0.022)	-0.824 (0.054)		0.184 (0.084)	-0.184 (0.061)	
R021K02C	0.836 (0.041)	-0.561 (0.044)				
R021K03M	1.051 (0.084)	0.222 (0.061)	0.197 (0.029)			
R021K04M	0.878 (0.172)	1.333 (0.103)	0.378 (0.028)			
R021K05C	1.097 (0.049)	0.155 (0.025)				
R021K06M	1.476 (0.104)	0.044 (0.048)	0.237 (0.027)			
R021K07C	0.772 (0.030)	0.123 (0.023)		0.125 (0.042)	-0.125 (0.039)	
R021K08M	1.110 (0.093)	0.450 (0.054)	0.192 (0.026)			
R021K09M	1.172 (0.092)	0.126 (0.061)	0.227 (0.030)			
R021K10C	0.797 (0.030)	0.855 (0.025)		-0.303 (0.042)	0.303 (0.050)	
R021K11M	1.123 (0.099)	0.394 (0.060)	0.231 (0.029)			
R021K12C	0.650 (0.025)	0.036 (0.022)		0.312 (0.061)	-0.053 (0.057)	-0.259 (0.051)
R021N01M	0.848 (0.075)	-0.537 (0.126)	0.295 (0.047)			
R021N02C	0.848 (0.041)	-0.348 (0.038)				
R021N03C	0.743 (0.035)	1.008 (0.032)		0.286 (0.038)	-0.286 (0.053)	
R021N04M	1.444 (0.106)	0.361 (0.040)	0.197 (0.022)			
R021N05M	1.793 (0.141)	-0.590 (0.063)	0.365 (0.035)			
R021N06M	1.676 (0.106)	-0.321 (0.045)	0.191 (0.027)			
R021N07M	1.157 (0.085)	0.035 (0.059)	0.200 (0.029)			

( ) Standard errors appear in parentheses

Exhibit D.3 IRT Parameters for PIRLS Joint 2001-2006 Scaling of Informational Purposes *(continued)*

Item	Slope (a_j)	Location (b_j)	Guessing (c_j)	Step 1 (d_{j1})	Step 2 (d_{j2})	Step 3 (d_{j3})
R021N08C	0.983 (0.046)	-0.164 (0.031)				
R021N09M	1.268 (0.097)	-0.083 (0.064)	0.271 (0.032)			
R021N10M	0.945 (0.104)	0.424 (0.087)	0.320 (0.034)			
R021N11C	0.621 (0.023)	0.167 (0.025)		-0.417 (0.055)	0.417 (0.053)	
R021N12C	0.656 (0.029)	0.188 (0.026)		0.176 (0.049)	-0.176 (0.046)	
R021S01M	0.823 (0.071)	0.138 (0.079)	0.168 (0.032)			
R021S02M	0.510 (0.056)	-0.438 (0.203)	0.185 (0.057)			
R021S03M	0.963 (0.078)	-0.438 (0.096)	0.265 (0.039)			
R021S04M	1.324 (0.100)	0.137 (0.052)	0.241 (0.026)			
R021S05C	0.778 (0.039)	-0.030 (0.036)				
R021S06M	1.365 (0.104)	-0.525 (0.075)	0.331 (0.036)			
R021S07C	0.466 (0.017)	0.883 (0.038)		-1.188 (0.078)	1.188 (0.087)	
R021S09C	0.873 (0.042)	-0.192 (0.035)				
R021S10C	0.851 (0.042)	-0.044 (0.034)				
R021S11C	0.707 (0.043)	0.899 (0.048)				
R021S12C	0.632 (0.043)	1.202 (0.068)				
R021S13C	0.660 (0.045)	1.223 (0.067)				
R021S14M	1.583 (0.130)	0.618 (0.036)	0.205 (0.019)			
R021S15C	0.535 (0.019)	0.457 (0.029)		-0.841 (0.067)	0.841 (0.069)	
R011A01C	0.812 (0.027)	-1.444 (0.046)				
R011A02M	1.206 (0.065)	0.179 (0.039)	0.260 (0.018)			
R011A03C	0.734 (0.025)	-1.001 (0.039)				
R011A04C	0.741 (0.017)	-0.098 (0.019)		1.115 (0.032)	-1.115 (0.026)	
R011A05M	0.916 (0.049)	-1.293 (0.095)	0.254 (0.038)			
R011A06M	0.994 (0.054)	-1.259 (0.090)	0.292 (0.037)			
R011A07C	0.728 (0.017)	-0.502 (0.016)		0.247 (0.045)	-0.025 (0.038)	-0.222 (0.028)
R011A08C	0.594 (0.018)	-0.856 (0.028)		0.532 (0.047)	-0.532 (0.031)	
R011A09C	0.690 (0.019)	-0.060 (0.018)		0.576 (0.032)	-0.576 (0.027)	
R011A10M	1.365 (0.057)	-0.022 (0.028)	0.130 (0.015)			
R011A11C	0.793 (0.027)	0.050 (0.023)				
R011L01M	0.525 (0.037)	-2.251 (0.267)	0.253 (0.079)			
R011L02M	0.803 (0.062)	0.680 (0.054)	0.214 (0.021)			
R011L03C	0.643 (0.023)	-0.382 (0.032)				
R011L04C	0.627 (0.013)	0.463 (0.017)		1.479 (0.035)	-0.985 (0.036)	-0.494 (0.048)
R011L05M	1.203 (0.073)	0.678 (0.031)	0.203 (0.015)			
R011L06C	0.758 (0.027)	0.184 (0.023)				
R011L07M	0.812 (0.055)	0.570 (0.049)	0.168 (0.021)			
R011L08C	0.812 (0.022)	0.703 (0.017)		0.669 (0.023)	-0.669 (0.030)	
R011L09M	0.986 (0.051)	-0.762 (0.072)	0.240 (0.031)			
R011L10C	0.801 (0.022)	0.718 (0.016)		0.105 (0.025)	-0.105 (0.030)	
R011L11M	0.901 (0.051)	-0.179 (0.064)	0.219 (0.027)			
R011L12C	0.829 (0.022)	0.645 (0.017)		0.716 (0.023)	-0.716 (0.029)	

( ) Standard errors appear in parentheses

Exhibit D.4 IRT Parameters for PIRLS Joint 2001-2006 Scaling of Retrieving and Straightforward Inferencing Processes

Item	Slope (a_j)	Location (b_j)	Guessing (c_j)	Step 1 (d_{j1})	Step 2 (d_{j2})	Step 3 (d_{j3})
R011H01M	0.667 (0.055)	-1.686 (0.213)	0.267 (0.071)			
R011H02M	0.962 (0.063)	-1.476 (0.109)	0.181 (0.046)			
R011H04C	0.876 (0.038)	-1.157 (0.049)				
R011H05M	1.097 (0.065)	-1.152 (0.077)	0.153 (0.036)			
R011H07C	0.508 (0.021)	-0.694 (0.039)		0.272 (0.065)	-0.272 (0.049)	
R011M01M	1.292 (0.087)	-0.659 (0.071)	0.325 (0.034)			
R011M02M	1.179 (0.082)	-1.267 (0.097)	0.304 (0.045)			
R011M03M	1.318 (0.085)	0.213 (0.040)	0.192 (0.020)			
R011M05M	1.204 (0.078)	-0.505 (0.067)	0.268 (0.032)			
R011M07C	0.989 (0.040)	-0.738 (0.035)				
R011M09M	1.043 (0.062)	-0.746 (0.072)	0.169 (0.033)			
R011M10C	1.088 (0.051)	-1.649 (0.056)				
R011N01M	0.985 (0.075)	-0.255 (0.083)	0.287 (0.034)			
R011N02M	0.950 (0.081)	0.212 (0.072)	0.267 (0.029)			
R011N03M	1.201 (0.081)	-0.672 (0.076)	0.293 (0.036)			
R011N05M	1.535 (0.102)	0.198 (0.038)	0.237 (0.020)			
R011N09M	1.177 (0.072)	-0.509 (0.061)	0.195 (0.031)			
R011R03M	0.728 (0.054)	-1.363 (0.159)	0.233 (0.058)			
R011R04C	0.840 (0.037)	-1.236 (0.051)				
R011R05C	1.033 (0.042)	-0.824 (0.035)				
R011R06C	0.575 (0.017)	-0.350 (0.027)		-0.570 (0.058)	0.570 (0.051)	
R011R07C	0.934 (0.038)	0.058 (0.026)				
R021E01M	1.434 (0.116)	-1.163 (0.092)	0.355 (0.045)			
R021E02M	1.112 (0.082)	-0.476 (0.080)	0.261 (0.036)			
R021E03M	0.521 (0.056)	-0.135 (0.163)	0.164 (0.048)			
R021E04M	1.294 (0.096)	-1.304 (0.093)	0.253 (0.045)			
R021E05C	0.607 (0.021)	-0.462 (0.032)		-0.428 (0.063)	0.428 (0.054)	
R021E06M	1.153 (0.078)	-0.326 (0.063)	0.192 (0.030)			
R021K01C	0.381 (0.019)	-1.085 (0.064)		0.167 (0.098)	-0.167 (0.072)	
R021K02C	0.703 (0.035)	-0.798 (0.052)				
R021K03M	1.068 (0.084)	0.134 (0.060)	0.206 (0.027)			
R021K04M	0.600 (0.135)	1.368 (0.150)	0.348 (0.038)			
R021K05C	0.954 (0.043)	0.031 (0.029)				
R021K06M	1.338 (0.095)	-0.091 (0.053)	0.234 (0.027)			
R021K08M	0.989 (0.085)	0.354 (0.061)	0.189 (0.027)			
R021K11M	0.990 (0.091)	0.297 (0.070)	0.231 (0.030)			
R021N01M	0.856 (0.075)	-0.624 (0.126)	0.301 (0.047)			
R021N02C	0.769 (0.038)	-0.488 (0.042)				
R021N04M	1.371 (0.101)	0.278 (0.043)	0.196 (0.022)			
R021N05M	1.773 (0.137)	-0.748 (0.063)	0.346 (0.035)			
R021N06M	1.655 (0.102)	-0.452 (0.045)	0.179 (0.026)			

( ) Standard errors appear in parentheses

Exhibit D.4 IRT Parameters for PIRLS Joint 2001-2006 Scaling of Retrieving and Straightforward Inferencing Processes *(continued)*

Item	Slope (a_j)	Location (b_j)	Guessing (c_j)	Step 1 (d_{j1})	Step 2 (d_{j2})	Step 3 (d_{j3})
R021N07M	1.068 (0.077)	-0.101 (0.063)	0.183 (0.029)			
R021N08C	0.892 (0.042)	-0.289 (0.034)				
R021N09M	1.136 (0.089)	-0.197 (0.074)	0.275 (0.034)			
R021N10M	0.789 (0.092)	0.309 (0.109)	0.304 (0.038)			
R021S02M	0.511 (0.052)	-0.558 (0.194)	0.164 (0.057)			
R021S03M	0.941 (0.075)	-0.533 (0.101)	0.261 (0.042)			
R021S04M	1.246 (0.098)	0.087 (0.058)	0.257 (0.028)			
R021S06M	1.246 (0.095)	-0.666 (0.084)	0.318 (0.039)			
R021S11C	0.667 (0.041)	0.866 (0.051)				
R021U01M	0.640 (0.069)	-0.252 (0.162)	0.263 (0.051)			
R021U02M	1.071 (0.073)	-0.964 (0.087)	0.190 (0.040)			
R021U03M	0.607 (0.063)	-0.128 (0.144)	0.194 (0.047)			
R021U04M	0.717 (0.069)	0.121 (0.101)	0.186 (0.037)			
R021U05C	1.034 (0.047)	-0.667 (0.037)				
R021U06C	0.868 (0.041)	-0.649 (0.042)				
R021U07M	0.735 (0.069)	-0.834 (0.169)	0.302 (0.057)			
R021U09M	0.838 (0.073)	-0.389 (0.112)	0.250 (0.043)			
R021U10C	0.720 (0.038)	-1.027 (0.060)				
R021Y01M	1.126 (0.097)	0.249 (0.061)	0.259 (0.027)			
R021Y04M	1.301 (0.096)	0.162 (0.048)	0.213 (0.024)			
R021Y05M	1.668 (0.113)	0.127 (0.037)	0.207 (0.021)			
R021Y06M	1.655 (0.114)	0.159 (0.037)	0.215 (0.021)			
R021Y07M	0.830 (0.061)	-0.988 (0.115)	0.176 (0.045)			
R021Y08M	1.357 (0.094)	-0.313 (0.058)	0.244 (0.030)			
R021Y09C	0.872 (0.031)	-0.591 (0.027)		0.038 (0.049)	-0.038 (0.037)	
R011A01C	0.860 (0.029)	-1.404 (0.042)				
R011A02M	1.119 (0.061)	0.104 (0.043)	0.245 (0.020)			
R011A03C	0.751 (0.025)	-1.009 (0.038)				
R011A05M	0.964 (0.050)	-1.302 (0.087)	0.231 (0.038)			
R011A06M	1.042 (0.054)	-1.293 (0.083)	0.257 (0.038)			
R011A07C	0.673 (0.016)	-0.556 (0.017)		0.177 (0.047)	-0.000 (0.041)	-0.177 (0.031)
R011A08C	0.552 (0.017)	-0.924 (0.030)		0.512 (0.050)	-0.512 (0.033)	
R011C01C	1.327 (0.037)	-0.304 (0.017)				
R011C02C	0.819 (0.027)	0.299 (0.022)				
R011C03C	1.269 (0.036)	-0.616 (0.021)				
R011C04M	1.284 (0.062)	0.241 (0.030)	0.175 (0.015)			
R011C05M	0.810 (0.057)	-0.200 (0.091)	0.344 (0.031)			
R011C07M	1.148 (0.058)	-0.450 (0.053)	0.264 (0.025)			
R011C08C	0.554 (0.014)	0.197 (0.019)		-0.350 (0.039)	0.350 (0.039)	
R011C09M	1.155 (0.066)	0.612 (0.030)	0.155 (0.014)			
R011F02M	0.627 (0.039)	-0.780 (0.118)	0.183 (0.040)			

( ) Standard errors appear in parentheses

Exhibit D.4 IRT Parameters for PIRLS Joint 2001-2006 Scaling of Retrieving and Straightforward Inferencing Processes *(continued)*

Item	Slope (a_j)	Location (b_j)	Guessing (c_j)	Step 1 (d_{j1})	Step 2 (d_{j2})	Step 3 (d_{j3})
R011F03M	0.885 (0.041)	-0.794 (0.064)	0.139 (0.028)			
R011F04M	1.230 (0.057)	-0.952 (0.056)	0.232 (0.028)			
R011F05M	0.889 (0.048)	-0.396 (0.067)	0.225 (0.028)			
R011F06C	0.779 (0.025)	-0.406 (0.026)				
R011F08C	1.071 (0.031)	-0.268 (0.020)				
R011F09C	0.993 (0.024)	-0.715 (0.018)		0.081 (0.031)	-0.081 (0.022)	
R011F10C	0.850 (0.029)	-1.305 (0.041)				
R011L01M	0.525 (0.036)	-2.355 (0.260)	0.244 (0.080)			
R011L02M	0.802 (0.061)	0.634 (0.054)	0.220 (0.021)			
R011L03C	0.604 (0.022)	-0.493 (0.034)				
R011L05M	1.215 (0.075)	0.656 (0.031)	0.219 (0.014)			
R011L06C	0.662 (0.024)	0.102 (0.026)				
R011L08C	0.688 (0.019)	0.686 (0.020)		0.744 (0.027)	-0.744 (0.035)	
R011L09M	0.892 (0.045)	-0.989 (0.080)	0.196 (0.034)			

( ) Standard errors appear in parentheses

Exhibit D.5 IRT Parameters for PIRLS Joint 2001-2006 Scaling of Interpreting, Integrating, and Evaluating Processes

Item	Slope (a_j)	Location (b_j)	Guessing (c_j)	Step 1 (d_{j1})	Step 2 (d_{j2})	Step 3 (d_{j3})
R011H03C	0.354 (0.019)	0.658 (0.049)		0.674 (0.068)	-0.674 (0.083)	
R011H06M	0.837 (0.064)	-0.485 (0.093)	0.210 (0.036)			
R011H08C	0.798 (0.037)	-0.119 (0.031)				
R011H09C	0.722 (0.026)	-0.805 (0.031)		0.075 (0.053)	-0.075 (0.038)	
R011H10C	0.687 (0.020)	0.257 (0.018)		-0.121 (0.054)	1.076 (0.051)	-0.956 (0.047)
R011H11M	1.483 (0.096)	-0.430 (0.052)	0.262 (0.029)			
R011M04C	0.823 (0.040)	0.545 (0.033)				
R011M06C	0.948 (0.032)	-0.647 (0.023)		0.280 (0.040)	-0.280 (0.028)	
R011M08M	1.188 (0.117)	0.643 (0.048)	0.276 (0.020)			
R011M11C	0.866 (0.040)	0.341 (0.029)				
R011M12C	0.663 (0.024)	0.485 (0.021)		0.782 (0.042)	-0.122 (0.045)	-0.660 (0.059)
R011M13M	0.938 (0.083)	-0.275 (0.096)	0.343 (0.036)			
R011M14C	0.951 (0.041)	-0.453 (0.031)				
R011N04M	1.184 (0.089)	0.341 (0.043)	0.192 (0.021)			
R011N06M	1.704 (0.132)	0.696 (0.028)	0.172 (0.014)			
R011N07C	0.630 (0.026)	0.461 (0.026)		0.292 (0.041)	-0.292 (0.047)	
R011N08C	0.633 (0.024)	0.040 (0.026)		0.770 (0.044)	-0.770 (0.043)	
R011N10C	0.971 (0.049)	0.852 (0.037)				
R011N11M	1.089 (0.087)	0.271 (0.052)	0.218 (0.024)			
R011N12C	0.847 (0.029)	0.335 (0.019)		-0.019 (0.034)	0.019 (0.036)	
R011N13C	0.668 (0.036)	0.001 (0.036)				
R011R01M	0.882 (0.070)	-0.056 (0.073)	0.218 (0.030)			
R011R02M	1.221 (0.093)	0.375 (0.042)	0.213 (0.020)			
R011R08C	0.805 (0.030)	0.089 (0.020)		0.327 (0.035)	-0.327 (0.034)	
R011R09C	0.648 (0.024)	-0.230 (0.025)		0.114 (0.046)	-0.114 (0.041)	
R011R10C	0.362 (0.014)	0.145 (0.033)		1.069 (0.089)	0.582 (0.075)	-1.651 (0.090)
R011R11C	0.575 (0.022)	-0.301 (0.024)		0.401 (0.067)	0.192 (0.056)	-0.594 (0.048)
R021E07C	0.678 (0.028)	-0.252 (0.027)		0.245 (0.049)	-0.245 (0.042)	
R021E08M	1.232 (0.099)	0.421 (0.042)	0.166 (0.020)			
R021E09C	0.580 (0.028)	0.582 (0.033)		0.557 (0.047)	-0.557 (0.058)	
R021E10C	1.031 (0.047)	-0.279 (0.030)				
R021E11M	0.915 (0.085)	0.139 (0.075)	0.227 (0.032)			
R021E12C	0.898 (0.036)	0.104 (0.021)		0.353 (0.036)	-0.353 (0.034)	
R021K07C	0.759 (0.030)	-0.075 (0.023)		0.133 (0.043)	-0.133 (0.040)	
R021K09M	1.073 (0.088)	-0.061 (0.067)	0.238 (0.031)			
R021K10C	0.820 (0.031)	0.659 (0.024)		-0.281 (0.041)	0.281 (0.048)	
R021K12C	0.612 (0.024)	-0.174 (0.023)		0.318 (0.065)	-0.062 (0.060)	-0.256 (0.054)
R021N03C	0.690 (0.033)	0.869 (0.035)		0.281 (0.041)	-0.281 (0.057)	
R021N11C	0.625 (0.023)	-0.010 (0.025)		-0.407 (0.055)	0.407 (0.053)	
R021N12C	0.640 (0.029)	0.005 (0.027)		0.179 (0.050)	-0.179 (0.047)	
R021S01M	0.769 (0.069)	-0.038 (0.086)	0.165 (0.033)			
R021S05C	0.771 (0.039)	-0.195 (0.036)				

( ) Standard errors appear in parentheses

Exhibit D.5: IRT Parameters for PIRLS Joint 2001-2006 Scaling of Interpreting, Integrating, and Evaluating Processes *(continued)*

Item	Slope (a_j)	Location (b_j)	Guessing (c_j)	Step 1 (d_{j1})	Step 2 (d_{j2})	Step 3 (d_{j3})
R021S07C	0.509 (0.019)	0.682 (0.034)		-1.065 (0.072)	1.065 (0.079)	
R021S09C	0.851 (0.042)	-0.363 (0.036)				
R021S10C	0.796 (0.040)	-0.221 (0.036)				
R021S12C	0.640 (0.044)	1.028 (0.067)				
R021S13C	0.719 (0.047)	0.998 (0.060)				
R021S14M	1.410 (0.122)	0.476 (0.041)	0.207 (0.021)			
R021S15C	0.611 (0.022)	0.285 (0.026)		-0.703 (0.059)	0.703 (0.061)	
R021U08C	0.877 (0.035)	-0.352 (0.024)		0.347 (0.042)	-0.347 (0.033)	
R021U11C	0.656 (0.027)	0.350 (0.021)		0.552 (0.049)	-0.338 (0.055)	-0.214 (0.062)
R021U12C	0.748 (0.033)	-0.224 (0.027)		0.485 (0.048)	-0.485 (0.039)	
R021Y02M	1.499 (0.110)	-0.228 (0.052)	0.296 (0.027)			
R021Y03C	0.894 (0.046)	0.460 (0.032)				
R021Y10C	0.825 (0.044)	0.429 (0.034)				
R021Y11M	1.315 (0.106)	-0.018 (0.056)	0.289 (0.028)			
R021Y12C	0.782 (0.023)	-0.033 (0.021)		-0.948 (0.057)	0.948 (0.055)	
R021Y13C	0.822 (0.031)	0.342 (0.018)		0.584 (0.040)	-0.236 (0.043)	-0.348 (0.049)
R021Y14C	0.663 (0.024)	0.169 (0.024)		-0.415 (0.052)	0.415 (0.052)	
R011A04C	0.710 (0.016)	-0.230 (0.019)		1.126 (0.033)	-1.126 (0.027)	
R011A09C	0.722 (0.020)	-0.185 (0.017)		0.561 (0.031)	-0.561 (0.026)	
R011A10M	1.259 (0.055)	-0.144 (0.032)	0.146 (0.017)			
R011A11C	0.832 (0.028)	-0.076 (0.022)				
R011C06C	0.929 (0.029)	-0.356 (0.022)				
R011C10C	0.704 (0.016)	0.081 (0.012)		0.248 (0.033)	-0.262 (0.036)	0.014 (0.034)
R011C11C	0.862 (0.022)	-0.021 (0.016)		0.745 (0.026)	-0.745 (0.023)	
R011C12M	0.739 (0.052)	-0.003 (0.073)	0.207 (0.027)			
R011C13M	0.892 (0.061)	0.243 (0.054)	0.235 (0.022)			
R011F01M	1.303 (0.061)	-0.580 (0.044)	0.249 (0.022)			
R011F07C	0.486 (0.012)	0.340 (0.020)		-0.902 (0.047)	0.902 (0.049)	
R011F11M	0.740 (0.055)	0.250 (0.066)	0.208 (0.025)			
R011F12C	0.672 (0.018)	0.546 (0.018)		-0.295 (0.032)	0.295 (0.036)	
R011F13M	1.076 (0.062)	-0.032 (0.048)	0.263 (0.022)			
R011L04C	0.596 (0.013)	0.291 (0.017)		1.529 (0.036)	-1.038 (0.038)	-0.491 (0.050)
R011L07M	0.822 (0.057)	0.428 (0.048)	0.180 (0.020)			
R011L10C	0.803 (0.023)	0.545 (0.016)		0.106 (0.025)	-0.106 (0.030)	
R011L11M	0.887 (0.051)	-0.347 (0.065)	0.227 (0.026)			
R011L12C	0.865 (0.023)	0.461 (0.016)		0.705 (0.023)	-0.705 (0.028)	

( ) Standard errors appear in parentheses

TIMSS & PIRLS
International Study Center
Lynch School of Education, Boston College

BOSTON
COLLEGE

ISBN 1-889938-46-7

pirls.bc.edu