2

Developing the PIRLS Reading Assessment

Marian Sainsbury

Jay Campbell

2.1 Overview

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

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

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

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

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

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

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

Exhibit 2.1: Overview of the Test Development Process

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

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

2.2 The PIRLS Assessment Framework

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

> *The abi i der a d a d e h e ri e a g age f r re ired b cie a d/ r a ed b he i di id a . Y g reader ca c r c ea i g fr a arie f e . The read ear , ar ici a e i c i ie f reader , a d f r e j e .*

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

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

The following sections provide a description of the text types (purposes for reading) and the item types (processes and strate-

gies) that were included in the framework – and that guided the instrument development process.

2.2.1 Text Types

Readers interact with text in different ways to construct meaning. Their approach to constructing meaning varies by the purpose for reading and the type of text being read. Certain purposes for reading are associated with certain types of text. For nine- and ten-year-old students, the two most common purposes for reading are reading for enjoyment and reading to learn. As such, the PIRLS framework specifies the inclusion of two broad types of text in the assessment: literary texts read for *i erar e eri- e ce r e j e* , and informative texts read to *ac ire a d e i f r a i* .

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

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

2.2.2 Processes and Strategies

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

ten to engage students in one of four processes: 1) focus on and retrieve explicitly stated information, 2) make straightforward inferences, 3) interpret and integrate ideas and information, and 4) examine and evaluate content, language, and textual elements. A brief description of each process is provided below.

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

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

Exhibit 2.2: Distribution of Literacy and Informational Blocks Across Booklets

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

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

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

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

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

2.3 Finding and Selecting the Passages

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

2.3.1 The Initial Search for Passages

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

All passages:

- Must be suited in their content and reading level to 9- and 10-year-olds
- Should be well written in order to foster authentic engagement in the reader and to facilitate questioning across the PIRLS processes and strategies
- Could be either literary or informational, and should include as wide a range as possible within these two broad categories
- Should not exceed 1200 words in length
- Should avoid specific cultural references and material offensive to particular cultural or religious groups.

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

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

At the first NRC meeting in July 1999, participants considered 68 passages that had been contributed by 11 different countries: Albania, Australia, Austria, Cyprus, France, Italy, Hungary, New Zealand, Russia, Singapore, and the United Kingdom. These comprised passages sent in advance or brought to the meeting by the NRCs themselves; texts suggested by members of the RDG; and passages found by the NFER
r

already represented a wide range of material, it was agreed, at that meeting, that further texts should be sought and reviewed by the NRCs following the meeting.

2.3.2 Reviewing the Passages

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

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

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

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

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

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

2.4.2 Item Review and Revision

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

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

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

2.4.3 Finalizing the Items

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

2.5 Field Test

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

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

2.6 Selection of Blocks for Main Survey

The results of the field test were reviewed at a meeting of the RDG in December 2000, and the assessment blocks for the main survey were selected. These were reviewed and approved (with minor modifications) at a m

and reviewers would view the scoring criteria as an essential component of developing a reliable and valid constructed-response question. Drafting of scoring criteria must be part of constructed-response item development and review processes, so that thoughtful and ongoing considerations of how student responses will be scored can sharpen the focus and increase the measurement value of these open-ended item types.

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

2.7.2 Student Responses

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

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

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

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

2.7.4 Training Scorers

National Research Coordinators were responsible for training scoring staff and for conducting scoring in their countrit mafft60.36 0 TDrioL,o