Assessment Framework and Instrument Development

Ina V.S. Mullis
Kathleen Trong Drucker
Corinna Preuschoff
Alka Arora
Gabrielle M. Stanco

Unique Characteristics of the 2011 TIMSS and PIRLS Assessments

This description of the methods and procedures used to develop the assessment instruments for TIMSS and PIRLS attempts to:

- Describe the overall approach to instrument development for both TIMSS and PIRLS
- Provide specifics about the assessment items and questionnaires developed in 2011 for each program.

Although the general approach to instrument development is similar for TIMSS and PIRLS, and similar from assessment cycle to assessment cycle, there also are differences. Each assessment cycle tends to have some unique characteristics that influence the instrument development approach, and 2011 was no exception. Besides providing additional measures on trend lines monitoring changes in educational achievement, 2011 also was remarkable for two other reasons.

- TIMSS and PIRLS were assessed together. The four-year cycle of the TIMSS assessments begun in 1995 came into alignment with the five-year cycle of the PIRLS assessments begun in 2001. This was the first time the two assessment programs had data collection in the same year, an event that would occur only every twenty years if the programs remain on the same cycles.
- It was the inaugural year of prePIRLS. Countries whose students are not yet prepared to take PIRLS can still participate in this important international project by administering a parallel but less difficult reading assessment.

TIMSS and PIRLS in 2011

The alignment of TIMSS and PIRLS in 2011 provided countries a unique opportunity at the fourth grade to administer one comprehensive assessment of three essential subjects—reading, mathematics, and science. Countries participating in both programs received data about the relative strengths and weaknesses of their fourth grade students in reading, mathematics, and science. However, instrument development—assessment items and an array of background questionnaires—for the two assessments on the same schedule was challenging. Although most of the test item development activities were separate, it was necessary to coordinate background questionnaire development across TIMSS and PIRLS.

The opportunity to compare educational effectiveness across reading, mathematics, and science provided a strong impetus to update the contextual frameworks underlying the background data collection and focus attention on each and every questionnaire item.

As explained later in this publication, 2011 questionnaire development was guided through joint meetings of committees of TIMSS and PIRLS National Research Coordinators (NRCs). Also, a modular approach was adopted for the questionnaires to reduce response burden and facilitate data collection.

The prePIRLS Assessment of Reading Comprehension

For a variety of reasons, there are countries where most children in the fourth grade are still developing fundamental reading skills. So IEA offers several options for matching the PIRLS assessment to the country's educational development. For some countries, participation in PIRLS at the fifth or sixth grade is a better match with students' learning progression. For other countries, prePIRLS offers a stepping stone to participating in PIRLS.

PrePIRLS reflects the same conception of reading as PIRLS except the assessment is less difficult. Designed to test basic reading skills that are prerequisites for success on PIRLS, the prePIRLS assessment is consistent with the PIRLS framework for assessing reading comprehension. However, the passages are shorter, the syntax less complex, and the questions primarily require relatively straightforward retrieval of information. Participation in prePIRLS prepares countries for moving toward participation in PIRLS.

In general, prePIRLS was developed in parallel with PIRLS; It used the same background questionnaires and the expert committees reviewed both the PIRLS and prePIRLS assessment passages, items, and scoring guides.

The challenge was identifying passages with content suitable for fourth grade students, but with relatively straightforward text. Also, the approach was to place questions throughout the passages to enable students to answer questions as they proceeded along through the text. The overall approach was piloted in South Africa, and then development proceeded using the same general overall approach as used for PIRLS 2011.

The TIMSS and PIRLS Approach to Measuring Trends

Because TIMSS and PIRLS are designed to measure trends, the assessments of mathematics, science, and reading cannot change dramatically from cycle to cycle. That is, TIMSS and PIRLS are based on a well-known premise for designing trend assessments (ascribed to John Tukey and Albert Beaton):

> "If you want to measure change, do not change the measure."

However, the achievement tests and questionnaires also need to be updated with each cycle to prevent the assessments from becoming dated and no longer relevant to current learning goals and policy issues. It is important for the content, especially in science, to reflect the most recent discoveries in the field and for the situations presented to be "modern" in terms of students' experiences.

To maintain continuity with past assessments while keeping up with current trends, the TIMSS and PIRLS assessments evolve with each cycle. For the achievement tests, TIMSS and PIRLS each have a specific design for the steady release of items after each cycle and replacing them with newly developed items for the following cycle.

Since it is not necessary to keep the questionnaire items secure in order to measure trends, all the questionnaires are made available to researchers and the public after each assessment cycle. However, as with the achievement tests, there is a tension between maintaining questionnaire trends and addressing new policy-relevant issues. For the questionnaires, each cycle is updated according to the countries' views about the most useful data to collect.

In practice, new questionnaire items and scales are developed for each assessment because countries want particular information about particular issues. However, the questionnaires do represent a considerable response burden for TIMSS and PIRLS participants. To maintain a consistent response burden level, existing questionnaire items must be retired for new items to be included.

Overview of the Achievement Tests and Questionnaires

Although the majority of the assessment items, passages, and questionnaires are carried forward from the previous assessment cycle to measure trends, the task of updating the instruments for each new cycle—every four years for TIMSS since 1995 and every five years for PIRLS since 2001—is a substantial undertaking.

Considering that TIMSS assesses two subjects at two grades, it actually encompasses four achievement tests: mathematics at the fourth and eighth grades and science at the fourth and eighth grades. Each of the two fourth grade tests includes approximately 170 items, and each of the eighth grade tests includes approximately 200 items.

PIRLS, which assesses reading comprehension at the fourth grade, has two achievement tests: PIRLS, for students who have made the transition from learning to read and are now reading to learn, and prePIRLS, a less difficult version of PIRLS for countries where many fourth grade students are still learning to read. Development for both PIRLS and prePIRLS involves the extremely challenging aspect of selecting reading passages that are agreed upon by the diverse participating countries.

In addition, because the central goal of TIMSS and PIRLS is to provide information for improving educational outcomes, the assessments collect considerable amounts of descriptive data about the contexts for teaching and learning in participating countries. (See the box on Page 5.)

Thus, TIMSS includes a set of questionnaires for the fourth grade and a corresponding set for the eighth grade that are administered to the students, their mathematics and science teachers, and their school principals. Besides questionnaires for students, teachers, and principals that parallel those in TIMSS, PIRLS includes an additional questionnaire for students' parents because of the home environment's strong impact on early literacy skills.

Finally, countries provide information about their educational systems and curricula in mathematics, science, and reading by completing a country-level curriculum questionnaire and writing a descriptive chapter based on an outline created as part of the development process. This information is compiled in the ***TIMSS 2007 Encyclopedia*** and ***PIRLS 2006 Encyclopedia***, which, respectively, describe the major aspects of teaching and learning in mathematics and science and in reading in each participant country.

The TIMSS & PIRLS International Study Center prepares an international version of all the assessment instruments for TIMSS and PIRLS in English.

Questionnaires Explore Contexts for Teaching and Learning

Typically, each assessment cycle includes the following questionnaires:

The *Student Questionnaire* asks students about their home and school experiences in learning mathematics and science (TIMSS) and how to read (PIRLS).

The *PIRLS Learning to Read Survey (Home Questionnaire)* asks students' parents or guardians about home resources for fostering literacy and about their reading habits, highest level of education, and employment situations. It also asks about their child's attendance in preschool and literacy-centered activities in the home before the child attended school, such as reading books, singing songs, and writing alphabet letters or words.

The *Teacher Questionnaire* asks students' teachers about their education, professional development, and experience in teaching. It also asks about the instructional activities and materials used in the class of students selected for the TIMSS or PIRLS assessment.

The *School Questionnaire* asks the principals (or their designees) of the students' schools about the availability of resources, types of programs, and environments for learning in their schools.

The *Curriculum Questionnaire* collects information from participating countries about the organization and content of the curriculum in mathematics and science (TIMSS) and reading (PIRLS).

Subsequently the test and questionnaire instruments are translated by participating countries into their languages of instruction with the goal of creating high quality translations that are appropriately adapted for the national context and at the same time are internationally comparable. Therefore, a significant portion of the development and review effort by National Research Coordinators is dedicated to ensuring that the passages, test items, and questionnaires can be translated accurately.

The Item Development Process

The TIMSS & PIRLS International Study Center at Boston College uses a collaborative process to develop the new items needed for the mathematics,

science, and reading achievement tests and questionnaires for each cycle. To provide a broad overview, the process includes the following:

- updating the frameworks for the upcoming assessment;
- for PIRLS, identifying and selecting appropriate reading passages;
- developing items and their scoring guides in accordance with the frameworks;
- conducting a full-scale field test;
- selecting the assessment items based on the frameworks, field test results, and existing items from previous cycles; and
- conducting training in how to reliably score responses to constructed-response items (i.e., questions to which students provide a written response rather than choosing from a set of options).

The development process is directed and managed by the staff of the TIMSS & PIRLS International Study Center at Boston College, who collectively have considerable experience in the measurement and assessment of mathematics, science, and reading achievement and in developing questionnaires.

Also playing a key role in test and questionnaire development are the National Research Coordinators, who are designated by the participating countries to be responsible for the complex tasks involved in implementing the studies in their countries. For PIRLS, the NRCs submit, review, and approve reading passages. For both TIMSS and PIRLS, the NRCs and experts from the countries develop the test items together with the scoring guides for constructed-response items. They also review the items prior to the field test and, after the field test, select the items for the assessment.

To provide additional subject-matter expertise and support, an external consultant in each subject area—a TIMSS Mathematics Coordinator, TIMSS Science Coordinator, and PIRLS Reading Coordinator—works with staff on development activities. Additional advice and guidance are provided through periodic reviews by TIMSS and PIRLS committees of international experts. (See the box on Page 7.) During busy periods, the external Coordinator and several committee members serve as a task force to assist in completing a specific task, such as updating scoring guides after the field test.

Advisory Committees Deliver Additional Expertise

Both TIMSS and PIRLS have two expert advisory committees to help guide and review the instrument development process: one for test items and another for the questionnaires. These are described below:

TIMSS COMMITTEES

SCIENCE AND MATHEMATICS ITEM REVIEW COMMITTEE (SMIRC)

The SMIRC typically has 20 members: 10 experts in mathematics and mathematics education and 10 experts in science and science education. Members are nominated by countries participating in TIMSS. It sometimes is necessary to have more science members to ensure expertise across the fields of biology, chemistry, and physics.

QUESTIONNAIRE ITEM REVIEW COMMITTEE (QIRC)

The QIRC comprises 10–12 experienced NRCs from different participating countries who have analyzed TIMSS data and use it in their countries. For example, members include NRCs who research factors related to mathematics achievement, science achievement, school effectiveness in general, issues of equity, and other relevant policy issues. (Please click to view the full TIMSS 2011 Development Team and Committees.)

PIRLS COMMITTEES

READING DEVELOPMENT GROUP (RDG)

The RDG typically comprises 8–10 experts in reading education, research, and assessment. The experts are nominated by the countries participating in PIRLS.

QUESTIONNAIRE DEVELOPMENT GROUP (QDG)

The QDG typically consists of 8–10 experienced PIRLS NRCs from different participating countries. These NRCs have analyzed PIRLS data from the perspective of improving teaching and learning in reading in their own countries and internationally. (Please click to view the full PIRLS 2011 Development Team and Committees.)

TIMSS and PIRLS Assessment Frameworks

Developing new items for each assessment cycle begins with updating the assessment framework that specifies the content to be assessed, the contexts for teaching and learning to be covered in the background questionnaires, and the design of the blocks and booklets for the assessment. The TIMSS and PIRLS frameworks for each assessment cycle are presented in full on the TIMSS and PIRLS website, together with the information for each cycle.

The Mathematics and Science Content Frameworks

The current version of the TIMSS mathematics and science frameworks can be found in the ***TIMSS 2011 Assessment Frameworks***. The first two chapters of the *TIMSS 2011 Assessment Frameworks,* respectively, describe the mathematics and science frameworks in detail.

The basic structure of the TIMSS mathematics and TIMSS science assessment frameworks consists of two dimensions: content and cognitive.

The content domains for mathematics at the fourth grade are number, geometric shapes and measures, and data display; at the eighth grade, the content areas are number, algebra, geometry, and data and chance. For science, content domains at the fourth grade are life science, physical science, and earth science; at the eighth grade, they are biology, chemistry, physics, and earth science.

Separately for the fourth and eighth grades, the TIMSS mathematics and science frameworks specify several topic areas within each content domain. For example, patterns, algebraic expressions, and equations/formulas and functions are topic areas within the algebra content domain.

The cognitive domains are the same for mathematics and science: knowing, applying, and reasoning. However, the descriptions of the cognitive skills to be assessed differ somewhat between mathematics and science.

The Reading Assessment Frameworks

The current version of the framework for assessing reading comprehension, which serves as the foundation for both the PIRLS and prePIRLS assessments, can be found in the ***PIRLS 2011 Assessment Framework***. Although the PIRLS framework for assessing reading comprehension has been refined for each assessment cycle since 2001, its major components have remained the same. It consists of a definition of reading literacy together with a detailed description of how PIRLS should assess reading comprehension.

Updating the TIMSS and PIRLS Frameworks: A Multiphase Process

The process of updating the TIMSS and PIRLS frameworks for each assessment cycle includes the following stages:

Several months before the first NRC meeting for the upcoming TIMSS or PIRLS cycle, staff at the TIMSS & PIRLS International Study Center with direct responsibility for the assessment conduct an extensive search of research articles, reports, and papers relevant to updating the descriptions of the content topics and educational learning contexts covered in the frameworks.

TIMSS & PIRLS International Study Center staff, together with the subject coordinator (reading, mathematics, or science), review any clarifications or updates that should be made to the framework based on the results of the prior assessment experience.

At the first NRC meeting of the assessment cycle, the TIMSS & PIRLS International Study Center presents its recommendations for updating the content/cognitive and contextual frameworks and solicits recommendations from NRCs.

After the meeting, update ideas are distributed to NRCs in the form of an online survey on the relative importance of making each update. This enables NRCs to garner opinions from various constituencies in their countries and take these into consideration in completing the survey.

Based on the results of the survey, TIMSS & PIRLS International Study Center staff work with the subject coordinator to create an updated draft of the framework.

The expert content area committee—the SMIRC or RDG—reviews the updated draft frameworks for the content/cognitive areas and makes suggestions for any further modifications.

The NRC questionnaire committee—the QIRC or QDG—reviews the updated draft contextual frameworks and previous questionnaires, making recommendations for further revisions.

TIMSS & PIRLS International Study Center staff prepare the penultimate drafts of the updated frameworks and disseminate them to NRCs for review prior to the second NRC meeting.

At the second NRC meeting, there is a final review, and the frameworks are adopted. Subsequently, the frameworks are published and distributed to all interested parties.

In brief, the PIRLS framework defines the two major aspects of students' reading: 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 reading experiences of young children. Thus, half the PIRLS and prePIRLS assessments are based on literary passages, and half are based on informational passages.

PIRLS assesses the following four processes of comprehension:

- focusing on and retrieving explicitly stated information;
- making straightforward inferences;
- interpreting and integrating ideas and information; and
- examining and evaluating content, language, and textual elements.

These processes are the basis for developing the reading comprehension questions in PIRLS.

Contextual Frameworks

To improve student achievement, it is important to understand the home, school, and social contexts in which students learn and how these relate to their achievement. Therefore, TIMSS and PIRLS collect a range of contextual information about teaching and learning in mathematics, science, and reading. The publications containing the TIMSS and PIRLS assessment frameworks also include a chapter on each study's contextual framework, describing the information that should be collected in the background questionnaires.

Just as the mathematics, science, and reading frameworks describe the content and cognitive processes to be assessed, the contextual frameworks describe the types of student educational experiences and the major characteristics of schools and classrooms to cover in the background questionnaires.

Assessment Designs

The final chapter of each assessment framework publication describes the assessment design. The TIMSS and PIRLS assessment designs are developed by the TIMSS & PIRLS International Study Center and are included in the frameworks publications because the designs determine the available assessment time and types of items to be developed. Both TIMSS and PIRLS are composed of approximately half constructed-response and half multiple-choice items.

At both the fourth and eighth grades, TIMSS is divided into 28 blocks of items: 14 for mathematics and 14 for science. The blocks are assigned to 14 booklets according to a specific plan, such that each booklet has four blocks of items—two in mathematics and two in science. Each block is in two booklets to enable linking among booklets, and two mathematics blocks or two science blocks alternately begin every other booklet to balance position effects.

The assessment time for each student is 72 minutes (18 minutes per block) at the fourth grade and 90 minutes (22.5 minutes per block) at the eighth grade. During administration of the assessment, a break is given between the first two and second two blocks.

PIRLS is divided into 10 blocks of items, while prePIRLS is divided into 8 blocks. PIRLS and prePIRLS booklets each have two blocks, with each block containing a reading passage and 12–17 associated items. Half the blocks are devoted to measuring the literary purpose, and half to the informational purpose. In both cases, eight of the blocks are assigned to 12 booklets according to a specific plan that enables linking among booklets and balances position

effects. The remaining two PIRLS blocks (one literary and one informational) are presented in a magazine format in the *PIRLS Reader.* The *PIRLS Reader* has been in color since 2001, and, beginning in 2011, the entire assessment is moving to color to better mirror students' reading experiences.

Updating the Frameworks

Updating the frameworks is central to planning for each assessment cycle. For example, recommendations for updating content and cognitive domains can involve modifying content areas and their weightings (but no more than 5 percent); adding, deleting, or modifying topics within content areas to keep current with research findings and ensure that the number of topics reflects the content area weighting; rewriting to improve clarity for item writers; and perhaps combining some topic areas to reduce redundancy. (Please click to view the **TIMSS 2011** or **PIRLS/prePIRLS 2011 Schedule of Development Activities**.)

Recommendations about updating the contextual areas for teaching and learning that should be emphasized in collecting background data typically revolve around 1) addressing areas of emerging interest in educational improvement and 2) reducing efforts related to information that has not proved particularly useful.

Because TIMSS and PIRLS were both assessed in 2011 and many countries took the opportunity to participate in both studies, ensuring the relevance and quality of background data was a special priority for 2011. The complexity of this undertaking for each participating country provided significant impetus for a careful and rigorous updating of the TIMSS and PIRLS frameworks for collecting background information. The goal was to collect data about teaching and learning in mathematics, science, and reading that would help participating countries improve the effectiveness of their educational systems. In particular, the process of updating the frameworks involved asking National Research Coordinators to contribute publications and articles describing current research and policy issues in their countries, comparing the strengths of the existing TIMSS and PIRLS frameworks to emphasize the best of "two worlds," and a series of iterative reviews by the NRCs.

During the initial review of the PIRLS Assessment Framework for

2011, some PIRLS countries expressed a desire to expand PIRLS to include a component assessing web-based reading. Considerable development work was accomplished on the web-based reading initiative, with two web-based texts and questions undergoing development and review. However, not enough countries were able to secure funding for the project to continue. (Please click to view a description of the **PIRLS 2011 Web-based Reading Initiative**.)

Identifying Reading Passages for PIRLS

Readers make meaning from a text in a variety of ways, depending not only on the purpose for reading but also on the difficulty of the text and the readers' prior knowledge. Thus, identifying appropriate passages for the PIRLS and prePIRLS assessments is critical. Examples of literary texts include contemporary short stories as well as traditional tales and fables. Informational texts can be from a variety of sources, such as textbooks, "how to" activities, biographies, brochures, or advertisements, and may include graphic support in the form of charts, tables, or diagrams.

At the beginning of the assessment cycle, a call for passages is sent to all National Research Coordinators. At the same time, TIMSS & PIRLS International Study Center staff and the Reading Coordinator also begin the search for suitable materials. In general, passages should:

- Be suitable for fourth grade students in content, interest, and reading ability.
- Be well written in terms of depth and complexity to allow for a sufficient number of questions.
- Avoid bias in that they are sensitive to cultural differences and are likely to be equally familiar or unfamiliar to all students.

Text length and readability formulas are also used to guide passage selection, in conjunction with a qualitative evaluation of each text's characteristics and appropriateness in different languages and cultures. Reviewing the passages is an iterative activity involving the NRCs and the Reading Development Group and is conducted at meetings and online. During the year or so allocated to find texts, hundreds of passages are often reviewed in order to identify about eight that are needed to develop items for the field test. (Please click to view descriptions of the **PIRLS/prePIRLS 2011 Passages**, including word counts and readability.)

Writing, Reviewing Field Test Items and Scoring Guides

The TIMSS & PIRLS International Study Center uses a collaborative process involving the participating countries to develop test items and scoring guides for the TIMSS and PIRLS assessments. Most of the second NRC meeting is devoted to a workshop for developing the field test items. The NRCs, together with experienced item writers from participating countries, create the newly developed items for each assessment.

Prior to the workshop, TIMSS & PIRLS International Study Center staff members identify the scope of the item writing task for the field test, examining the weight given to each topic in each of the updated frameworks. Considerations include the total items needed, given the percentage of weight assigned to a particular area (for example, geometry), and the number of topics in that area (four, for example), as well as how many items exist from previous assessments. Subtracting the number of existing items from the total needed gives the number of new items required for the upcoming assessment. Because generally twice the number of TIMSS and PIRLS items are field-tested than are actually required, multiplying the target number by two indicates how many field test items need to be developed.

Also, in preparation for the item writing workshop, the TIMSS & PIRLS International Study Center updates the TIMSS and PIRLS item writing manuals that have been specifically developed for each assessment. Most recently, the ***TIMSS 2011 Item Writing Guidelines*** and the ***PIRLS 2011 Item Writing Guidelines*** were updated and used for 2011.

The item writing guidelines contain general information about procedures for obtaining good measurement (for instance, items should be independent and not provide clues to the correct responses of other items) as well as specific information on how to deal with translation and comparability issues (for example, using TIMSS' fictitious unit of currency, the "zed," for any money items). The item writing guidelines also describe the necessary steps for developing scoring guides, and they have checklists for reviewing items.

At the item writing workshop, country representatives are divided into teams and given specific item writing assignments to ensure that enough field test items are developed in each of the content areas and processes specified in the frameworks. (Please click **TIMSS 2011** or **PIRLS/prePIRLS 2011 Field Test Item Development** to view the number of participants in the item writing workshops as well as the number of items written and passages considered.)

The Coordinators and TIMSS & PIRLS International Study Center staff use the item writing guidelines in providing training to the teams on item writing procedures for the TIMSS and PIRLS assessments. Also, the staff stresses that although the international versions of items are written in English, the items selected for the field test will be translated from English into the languages of instruction of participating countries. Therefore, item writers must be sensitive to issues that might affect how well items can be translated to produce internationally comparable items. For example, idioms and expressions that are difficult to translate must be avoided.

Once teams have completed their item writing assignments, they begin reviewing the items drafted by other teams. In addition, some teams continue to send items to the TIMSS & PIRLS International Study Center for several weeks after the item writing workshop.

Following the item writing workshop, the respective internal and external TIMSS or PIRLS Coordinators are responsible for a thorough review of the draft field test items. Depending on the extent of the required revisions, the Coordinators arrange for task force meetings or to obtain additional input from experienced item writers to implement the necessary changes. The goal is a set of draft field test items that adequately covers all areas of the framework.

This draft set of field test items first receives a thorough review by the TIMSS & PIRLS International Study Center. Reviewers include staff; the external Coordinator; and consultants experienced in developing assessment items, such as those from Educational Testing Service, the National Foundation for Educational Research in England, and the Australian Council for Educational Research. Reviewers may also include SMIRC members with particular item writing skills.

Sometimes TIMSS and PIRLS incorporate pilot tests into the item development process to explore the possibilities of new areas of assessment, innovative item types, or items measuring higher order reasoning skills. These exploratory efforts, often involving only one or two countries, are invaluable for informing the item development process. The development process for 2011 involved several such efforts.

Most significantly, 2011 was the inaugural year for prePIRLS. This version of the PIRLS reading comprehension assessment is based on the PIRLS framework, but has shorter, less difficult passages and a greater focus on the processes of retrieval and straightforward inferencing. (Please click for more on the **prePIRLS Pilot Test** including the number of items tested and student

responses collected, as well as how the results were used.)

TIMSS item development for 2011 included a small pilot test of the most complex reasoning items (Please click for more on the **TIMSS Pilot of Reasoning Items**, including the number of items tested and student responses collected, as well as how the results were used).

Once the issues raised by the reviews and pilot tests have been addressed, the next step is to identify the most promising PIRLS passage and item sets to propose as field test blocks and to assemble mathematics and science blocks that reflect the desired assessment as much as possible. These draft item blocks then undergo a thorough review by the appropriate expert committee—either the RDG or SMIRC, depending on the assessment.

Finally, the proposed field test blocks are reviewed by the National Research Coordinators prior to instrument production. Typically, there is general agreement with the recommendations made by the expert committees. However, the NRC review of the field test blocks is very important in minimizing any problems with translation and comparability. In particular, NRCs identify any potential issues with words or phrases that would change in meaning or difficulty in their languages of instruction. In addition, the NRCs provide suggestions for improving the clarity of the questions, options (for multiple-choice items), and scoring guides (for constructed-response items). Once the items have been finalized and approved by the NRCs, the TIMSS & PIRLS International Study Center implements the changes and provides the final international version of the field test booklets to the NRCs so that they can begin translating the field test materials into their languages of instruction.

Updating and Reviewing the Questionnaire Items

With each new assessment cycle, part of one of the first National Research Coordinators meeting is devoted to thoroughly reviewing each questionnaire's contents. In conjunction with the goals of the updated contextual frameworks for the background data, and in light of reporting trends in the upcoming assessment, the NRCs review each questionnaire item by item, sharing their views about the usefulness of items and response categories.

Afterward, the TIMSS & PIRLS International Study Center updates the questionnaires in accordance with the NRCs' recommendations, keeping in mind that countries can elect to complete the teacher, school, and country questionnaires either online or via paper and pencil.

The questionnaire committee—the QIRC for TIMSS or the QDG for

PIRLS—then reviews the updated drafts of the field test questionnaires for appropriate coverage of the topics specified in the contextual frameworks, the analytic potential of the items and reporting scales, and the clarity of the specific questions. The questionnaire committees make recommendations based on knowledge of ongoing educational research in a number of areas.

Committee members also make recommendations about the existing items and scales found most useful for analytic purposes and, on the other hand, those found to be less useful and that can be eliminated. The members suggest improvements and provide additional new items as necessary to meet the goals of the upcoming assessment. Representatives from the IEA Data Processing and Research Center (DPC) also attend this meeting to ensure that the questionnaire committee's recommendations are amenable to efficient and error-free data collection, whether online, by key entry, or by optical scanning.

The TIMSS & PIRLS International Study Center implements the committee's recommendations, and the draft field test questionnaires are reviewed again by the NRCs. The NRCs make suggestions for final revisions, which are then implemented by the TIMSS & PIRLS International Study Center. The field test questionnaires are then provided to the NRCs for translation, production, and data collection (Please click to view the **TIMSS and PIRLS 2011 Schedule of Development Activities for the Background Questionnaires**.)

Questionnaire Development Approach

Because TIMSS and PIRLS both were assessed in 2011, much of the questionnaire development work was accomplished during joint meetings of the TIMSS 2011 Questionnaire Item Review Committee (QIRC) and the PIRLS 2011 Questionnaire Development Group (QDG). The National Research Coordinators for TIMSS and those for PIRLS had multiple opportunities for reviewing the full array of questionnaires and making modifications. The goals for updating the TIMSS and PIRLS questionnaires for 2011 are included:

- Updating frameworks and questionnaires to reflect current policies and practices for teaching and learning
- Enhancing the validity of TIMSS and PIRLS 2011 background data by better aligning contextual frameworks and background questionnaires
- Maximizing the opportunity for comparisons across the three subject areas of reading, mathematics, and science

- Including scales that measure effective contexts for learning and meet the criteria for rigorous measurement
- Retaining questions for reporting trends on background contexts for learning
- Minimizing the response burden

The goals for updating background questionnaires in TIMSS and PIRLS 2011 reflect the desire to advance the questionnaires conceptually and empirically with each assessment cycle. Background questionnaire development begins with updating the contextual frameworks to reflect recently published research literature about effective educational policies and practices. The questionnaires are then updated in alignment with the frameworks, so that they measure important aspects of effective learning environments.

As a 2011 initiative, renewed emphasis was placed on developing reliable scales that would provide valid measurement of effective home, school, and classroom environments for learning. Because of the need to keep response burden to a minimum, attention was restricted to constructs expected to have a positive relationship with student achievement in reading, mathematics, or science in the TIMSS or PIRLS data. The questionnaire development effort included adding questionnaire items to strengthen existing measures, such as the student self-confidence in learning scales for reading, mathematics, and science, and the index of school discipline and attendance problems. However, considerable effort also was invested in developing new scales to measure important aspects of effective classrooms, such as student engagement, teachers' efforts to stimulate student engagement, and teachers' confidence in teaching mathematics and science to their class.

As a further 2011 initiative, questionnaire items were chosen so that the response data from students, teachers, principals, and parents could be scaled using the 1-parameter IRT (Rasch) partial credit measurement model. Sufficient items were drafted for each scale so that the Rasch model could be reliably applied. Many questionnaire items had four response categories (e.g., "agree a lot," "agree a little," "disagree a little," and "disagree a lot"); a scale comprised of such items required 6-7 items to meet the minimum requirements of Rasch scaling. Scales comprised of items with three response categories required more

items.[1] Generally, one or two extra items were developed for each scale to allow for attrition after field testing.

Almost 40 background scales were developed for the TIMSS and PIRLS 2011 field tests. At the same time, however, many questions from the 2006 and 2007 cycles were retained, to provide countries with trend measures on background contexts for student learning. To prevent an increase in response burden, the TIMSS & PIRLS International Study Center together with QIRC, QDG, and NRCs had the task of updating or replacing questionnaire items rather than adding more and more items.

To reduce the questionnaire response burden for countries assessing both TIMSS and PIRLS in 2011, the TIMSS & PIRLS International Study Center adopted a modular approach for the background questionnaires at fourth and eighth grade (Please click for a description of the **TIMSS and PIRLS 2011 Modular Design for Background Questionnaires**.)

The Field Test

Development of the TIMSS and PIRLS assessments includes conducting a full-scale field test in each participating country that is designed to be a "dress rehearsal" operationally for the assessment. That is, the data collection and scoring procedures to be employed in the assessment are practiced in the field test. The field test also provides important information about how well the test items function and whether the questionnaires are appropriate for the measurement purposes for which they were designed.

The field test is designed to yield at least 200 student responses to each reading, mathematics, and science item, as well as sufficient data to evaluate the validity and reliability of the various questionnaire scales. The field test sample size is approximately 30 schools in each country. (Please click for more on **TIMSS 2011 Field Test of Mathematics and Science Items** and **PIRLS/prePIRLS 2011 Field Tests of Reading Comprehension Items**.)

Generally, the samples for the field test and the assessment are drawn simultaneously, using the same random sampling procedures. This ensures that field test samples closely approximate assessment samples, and that a school is selected for either the field test or the assessment, but not both. For example, if 150 schools are needed for the assessment and another 30 for the field test, then

1 As a guideline, Suen's (1990) formula was used to determine the minimum requirements for Rasch scaling: (number of items)*(number of response categories – 1) ≥ 20.

a larger sample of 180 schools is selected and a systematic sample of 30 schools is selected from the 180 schools.

While countries are translating and finalizing their field test instruments for data collection, the TIMSS and PIRLS International Study Center prepares materials for training countries in how to evaluate students' answers to constructed-response items in the field test.

It is important to prepare scoring training materials for the newly developed constructed-response field test items in advance of the field test so field test scoring can occur immediately upon completion of data collection. To provide these materials, approximately six English-speaking countries administer the newly developed constructed-response field test items in a small selection of classrooms. The goal is to collect a total of at least 200 responses to each newly developed constructed-response field test item. The TIMSS and PIRLS task forces use these responses to provide examples of student responses in the field test scoring guides and to develop sets of training materials. (Please click **TIMSS 2011** or **PIRLS/prePIRLS 2011 Obtaining Student Responses for Developing Field-test Scoring Training Materials**.)

The scoring training materials for each constructed-response item include a scoring guide, approximately 8–10 example papers illustrating the categories in the scoring guide, and approximately 8–10 practice papers so that country representatives can practice making the distinctions among categories and reach agreement about how to make consistent decisions across countries.

A large portion of the fourth NRC meeting is devoted to scoring training for the field test. Led by the Reading, Mathematics, or Science Coordinator, respectively, staff and consultants train representatives from each country in how to reliably apply the scoring guide for each PIRLS or TIMSS field test item.

The trainer explains the purpose of the item and reads it aloud. The trainer then describes the scoring guide, explaining each category and the example papers. Rationales for the score assigned to each example paper are included in the scoring materials and discussed during each paper's presentation. After review of the scoring guide and example papers, the country representatives score the practice papers. Any inconsistencies in scoring are discussed, and, as necessary, the field test guides are clarified and sometimes categories are revised.

Corresponding to the procedures for the assessment, countries collect their field test data, score students' responses, and prepare their data files. The field test data files are sent to the IEA DPC, where they are checked for accuracy as well as for within-country and international consistency.

Finalizing the Assessment Instruments

The TIMSS & PIRLS International Study Center reviews and analyzes the field test data. Data almanacs containing summary item statistics are prepared for each test item and questionnaire item. The data almanac for a test question contains, row by row for each country: the sample size, the item difficulty and discrimination, the percentage of students answering each option (multiple-choice) or in each score category (constructed-response), and the point-biserial correlation for each multiple-choice option or constructed-response category. For constructed-response items, the degree of scoring agreement also is provided.

The data almanacs for questionnaire items contain the percentage of students responding to each question option, together with the corresponding average student achievement in mathematics, science, or reading, respectively. In addition, the background questionnaire item sets (scales) are evaluated for unidimensionality, internal consistency, and relationship with achievement. (Please click for a description of the **TIMSS and PIRLS 2011 Field Test Analysis of Background Questionnaire Scale Items**.)

The field test data are used by the TIMSS & PIRLS International Study Center, expert committees, and National Research Coordinators to assess the quality of the field test instruments.

The TIMSS & PIRLS International Study Center staff members, together with the external Coordinators, first review the field test data to make an initial judgment about the quality of each item's measurement properties. Items are eliminated from further consideration if they have poor measurement properties, such as being too difficult or easy or having low discrimination. Particular attention is paid to unusual item statistics in individual countries since these could give an indication of errors in translation.

Afterward, TIMSS & PIRLS International Study Center staff collaborate with the Coordinator and the task force to assemble a set of recommended assessment blocks for review by the expert committee (the SMIRC or RDG, as appropriate). The expert committees scrutinize the recommendations for the newly developed assessment blocks item by item. The items and scoring guides are reviewed for content accuracy, clarity, and adherence to the frameworks. In addition, the newly developed items are considered in relation to the trend item blocks for overall coherence as a complete assessment.

Similarly, the TIMSS & PIRLS International Study Center prepares a set of questionnaires along with the field test data for review by the appropriate

questionnaire committee (the QIRC or QDG). These expert committees review each questionnaire item for clarity, examine the data to make sure the options are providing useful information, and make suggestions for refinements in preparation for data collection.

The expert committees' recommendations are implemented by staff, and the penultimate assessment blocks and questionnaires are sent to the NRCs for review. NRCs have the opportunity to review the recommended materials in light of the field test results and within the security of their own countries. Each country also checks any unusual national results that might be an indication of translation errors and corrects the translation as necessary or recommends revisions to accommodate translation. Finally, there is an NRC meeting to review and approve all the assessment instruments. (Please click **TIMSS 2011** or **PIRLS/prePIRLS 2011 Number of Items in the Assessments** for more information about the final assessment specifications.)

The TIMSS & PIRLS International Study Center makes the final revisions and sends the newly developed assessment blocks and updated questionnaires to the countries for translation and adaptation. Countries that have participated in previous assessments have their trend assessment materials from previous cycles, but countries joining TIMSS or PIRLS for the first time also need to translate the trend assessment blocks. The TIMSS & PIRLS International Study Center also implements the suggested revisions to the scoring guides for the selected items and, in some cases, updates the scoring training materials.

Prior to the main data collection, two intensive scoring training sessions are held; the first is for countries conducting the assessment on the Southern Hemisphere schedule, and the second for countries assessing on the Northern Hemisphere schedule. The sessions include scoring training on the final guides for the newly developed constructed-response items and also on the scoring guides for the constructed-response items included in the trend blocks. For the trend scoring training, the guides and training papers are exactly the same as those used in prior assessments. (Please click **TIMSS 2011** or **PIRLS/prePIRLS 2011 Scoring Training Sessions** for more information, including the number of training items and participants in scoring training sessions.)

The Process Following Instrument Development

In general, after the participating countries receive the international version of the assessment instruments, they begin the process of translation and cultural adaptation (some adaptation to local usage typically is necessary

even in English-speaking countries) and production of the materials for printing. At the same time, countries should be making final arrangements for data collection, including the host of activities necessary to obtain school participation, implement test administration, and score the responses to the tests and questionnaires.

After each cycle approximately 40 percent of the assessment items are released to the public to illustrate the content of the assessments and for educational purposes, such as further research into students' strengths and weaknesses in reading or as part of instructional materials about topics in mathematics or science. The remaining assessment items are kept secure to be readministered in subsequent cycles as the basis for measuring trends.

The release plans provide for each assessment to incude items from several cycles—three for TIMSS and four for PIRLS. In TIMSS, approximately 40 percent of the items are newly developed for each cycle, 40 percent from the previous cycle, and 20 percent from two cycles before. For PIRLS also, approximately 40 percent of the items are newly developed for each cycle. However, beginning in 2016, PIRLS will have 20 percent of its items from the previous cycle, 20 percent from two cycles before, and 20 percent from three cycles before. Thus, for each upcoming cycle, it is necessary to replace a specific portion of the achievement items. (Please click to learn more about the **TIMSS 2011** or **PIRLS/prePIRLS 2011 Item Release Plan**, including when each assessment block was first used and when it will be released to the public.) All of the TIMSS and PIRLS procedures following instrument development, overall and for 2011 in particular, are described in subsequent sections of this publication.