Chapter 3

Component skills and cognitive instruments used in educational assessments

This chapter looks at the frameworks used in PISA and other surveys to assess reading, mathematics and science. In the case of each of the reviewed assessments the chapter outlines the approach used for the following: i) *item development;* ii) *test design;* iii) *psychometric analyses;* iv) *cross-country comparability;* v) *trends;* vi) *proficiency levels;* vii) *translation, adaptation and verification of cognitive instruments; and* viii) *field trials and item selection. Under each of these areas, the implications and lessons for PISA for Development (PISA-D) are discussed.*

Assessment frameworks

An assessment framework is the foundation on which an assessment is based. It provides a clear articulation of what components and subjects make up the assessment, at whom the assessment is targeted, the mode of delivery of the assessment and the length of time the assessment will take to complete.

Different assessments focus on different domains; for example, PIRLS solely focuses on reading, while PISA assesses reading, mathematics, science and problem-solving. This section of the report looks first at reading, mathematics and science domains. For each domain, we look at PISA's assessment frameworks, and then the frameworks of other assessments.

Reading

PISA's reading frameworks

Reading literacy was the major domain tested by PISA in 2000 and 2009. The description of reading literacy was updated for the 2009 test. The PISA definition of reading literacy goes beyond simply understanding text, to include educational and social engagement: "Reading literacy is understanding, using, reflecting on and engaging with written texts, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society" (OECD, 2010).

Reading literacy will be a minor domain in PISA 2015. As such, it will not be assessed as comprehensively as it was in 2009.

For PISA 2015, computer-based assessment will be the primary mode of delivery for all domains, including reading literacy. However, paper-based assessment instruments will continue to be provided for countries choosing not to test their students by computer.

The reading literacy component for both the computer-based and paper-based instruments will comprise the same intact clusters of reading trend items. The number of trend items in both minor domains will be increased, thereby increasing the construct coverage while reducing the number of students responding to each question. This design is intended to both reduce potential bias and stabilise and improve measurement of the trend.

The 2009 report provided separate scales for print reading and digital or electronic reading, although digital reading was not assessed in all participating countries in 2009, and it was not scaled as part of the overall concept of reading literacy. PISA 2015 reporting will not include digital reading scales.

The PISA reading literacy assessment framework is built on three major task characteristics:

- situation – the range of broad contexts or purposes for which reading takes place
- aspect – the cognitive approach that determines how readers engage with a text
- text – the range of material that is read.

In PISA the *situation* in which reading takes place is categorised as personal, public, occupational or educational.

Aspect refers to the mental strategies, approaches or purposes that readers use to negotiate their way into, around and between texts. These aspects are:

- access and retrieve
- integrate and interpret
- reflect and evaluate.

The *text* is the reading material. In an assessment, that material – a text (or a set of texts) related to a particular task – must be coherent within itself. That is, the text must be able to stand alone without requiring additional material to make sense to the proficient reader. There are many different kinds of texts and any assessment should include a broad range. PISA classifies texts by:

- text format – continuous, non-continuous, mixed and multiple
- text type – description, narration, exposition, argumentation, instruction and transaction
- text display space
- environment, such as authored or message-based.

The addition of digital reading in the 2009 framework made text classification more complex. The 2009 reading literacy assessment used a text classification of "medium: print and electronic". With the move to computer-based delivery for 2015, however, this is a potential source of confusion. For 2015 the terminology has been updated to "fixed text" and "dynamic text" to distinguish between delivery mode and the space in which the text is displayed, regardless of whether it is printed or onscreen. Additionally, the "environment" classification was a new variable for the PISA 2009 reading framework, but as it applies only to dynamic texts, it will not be discussed in the 2015 PISA framework. It is important to note that, despite changes to terminology, the constructs of the 2009 framework remain unchanged.

Table 3.1 Target distribution of tasks by situation for PISA 2015

Situation	Percentage of total tasks
Personal	30
Educational	25
Occupational	15
Public	30

Source: OECD, 2013a.

Table 3.2 Target distribution of tasks by text format for PISA 2015

Text format	Percentage of total tasks print
Continuous	60
Non-continuous	30
Mixed	5
Multiple	5

Source: OECD, 2013a.

Table 3.3 Approximate distribution of tasks by aspect for PISA 2015

Aspect	Percentage of total tasks
Access and retrieve	25
Integrate and interpret	50
Reflect and evaluate	25

Source: OECD, 2013a.

One of the greatest challenges for designing assessments is to create a framework and associated items that cover a very wide range of student capacity so that information can be gained about all students participating in the assessment. In PISA 2000, 2003 and 2006 it was noted that, while the level of proficiency of students can be located accurately, there is a shortage of descriptive information about what students at the extremes – particularly at the lower end of the distribution – know and can do as readers. This is because the majority of PISA items tested for skills and knowledge at the proficiency levels relevant to the majority of students. There were still significant numbers of students, however, performing outside these middle proficiency bands: either at a level much lower or much higher than the OECD average. There were few existing PISA tasks at the very easy end and the challenging end of the spectrum of task difficulty. In developing tasks for PISA 2009, therefore, there was an emphasis on including some very easy and some very difficult items. In addition to enhancing the descriptive power of the scale, better matching of the item difficulties to the student achievement distributions in each country improved the reliability of the population parameter estimates. Moreover, the test experience for individual students, particularly those performing at very low levels, has become more tolerable.

Developing items for the lower levels of proficiency was achieved by manipulating elements from PISA's descriptive framework as follows:

- using shorter and simpler texts
- ensuring a closer literal match of terms between the item and the text
- providing more direction to find the relevant information in the text to solve the item
- addressing personal and familiar experiences in reflecting on and evaluating content items, rather than remote, abstract issues
- addressing concrete features in reflecting on and evaluating form items.

Other assessments' reading frameworks

There is a diverse range of approaches used in creating assessment frameworks in reading across the international assessments considered in this report (see Annex C).

Some express a component of future uses of reading. SACMEQ adopted the same definition of reading literacy as PIRLS, which defines reading literacy as: "… the ability to understand and use those written language forms required by society and/or valued by the individual. Readers can construct meaning from texts in a variety of forms. They read to learn, to participate in communities of readers in school and everyday life, and for enjoyment" (Mullis and Martin, 2013: 14).

The STEP reading literacy assessment has been developed specifically for developing country contexts, and it includes sets of questions taken from PIAAC. This overlap allows countries participating in the STEP programme to compare their literacy results with other countries. STEP defines literacy as "understanding, evaluating, using and engaging with written texts to participate in society, to achieve one's goals, and to develop one's knowledge and potential" (Pierre et al., 2014).

Several assessments have a clear list of the different domains of reading that should be included. PASEC lists the domains as comprehension of words, comprehension of sentences, reading/writing, conjugation, grammar and comprehension of text. EGRA, which focuses on the early grades, lists essential components of reading as phonemic awareness, phonics, fluency, vocabulary and comprehension. ASER and Uwezo, both citizen-led assessments, focus on letter recognition, word recognition and passage reading.

Mathematics

PISA's mathematics frameworks

For the purposes of PISA 2015, mathematical literacy is defined as follows:

> *Mathematical literacy is an individual's capacity to formulate, employ, and interpret mathematics in a variety of contexts. It includes reasoning mathematically and using mathematical concepts, procedures, facts and tools to describe, explain and predict phenomena. It assists individuals to recognise the role that mathematics plays in the world and to make the well-founded judgments and decisions needed by constructive, engaged and reflective citizens (OECD, 2013b).*

The PISA mathematical literacy framework is built on three interrelated aspects:

- processes, comprising:
  - formulating situations mathematically
  - employing mathematical concepts, facts, procedures and reasoning
  - interpreting, applying and evaluating mathematical outcomes
- content, comprising:
  - change and relationships
  - space and shape
  - quantity
  - uncertainty and data
- and context, comprising:
  - personal, related to one's self, family or peer group
  - occupational, related to the world of work, such as measuring, costing and ordering materials for building, payroll or accounting, quality control, scheduling or inventory, design or architecture and job-related decision making

- societal, related to one's community (whether local, national or global), such as voting systems, public transport, government, public policies, demographics, advertising, national statistics and economics
- scientific, related to the application of mathematics to the natural world and issues and topics related to science and technology.

Table 3.4 Approximate distribution of score points by process category for PISA 2015

Process category	Percentage of score points
Formulating situations mathematically	Approximately 25
Employing mathematical concepts, facts, procedures	Approximately 50
Interpreting, applying and evaluating mathematical outcomes	Approximately 25

Source: OECD, 2013b.

Table 3.5 Approximate distribution of score points by content category for PISA 2012

Content category	Percentage of score points
Change and relationships	Approximately 25
Space and shape	Approximately 25
Quantity	Approximately 25
Uncertainty and data	Approximately 25

Source: OECD, 2013b.

Table 3.6 Approximate distribution of score points by context category for PISA 2012

Context category	Percentage of score points
Personal	Approximately 25
Occupational	Approximately 25
Societal	Approximately 25
Scientific	Approximately 25

Source: OECD, 2013b.

Other assessments' mathematics frameworks

SACMEQ has a focus on practical application of the knowledge gained through study. It defines mathematics literacy as "the capacity to understand and apply mathematical procedures and make related judgements as an individual and as a member of the wider society" (Ross et al., 2004). LLECE similarly adopts a literacy approach. It defines mathematics literacy as enabling students "to develop their potential, face situations, make decisions using the available information, solve problems, defend and argue their point of view … to integrate into society as full citizens who are critical and responsible" (SERCE, 2009).

The TIMSS definition of mathematics includes a cognitive dimension of "knowing, applying and reasoning". TIMSS 2015 also has a mathematics assessment called TIMSS Numeracy which assesses fundamental mathematical knowledge, procedures and problem-solving strategies by asking students to answer questions and work out problems similar to TIMSS, except with easier numbers and more straightforward procedures. TIMSS Numeracy is designed for countries where most children are still developing fundamental mathematics skills.

The PASEC definition is a list of the domains of interest for both the Grade 2 and Grade 6 populations. EGMA lists the core areas of interest as number identification, number discrimination (which numeral represents a numerical value greater than another), number pattern identification (a precursor to algebra), and addition and subtraction (including word problems).

Most assessments include in their assessment frameworks some notion of numbers, measurement and geometry. Assessments at the higher grades also tend to include algebra and data.

Science

PISA's science framework

Science was the major domain in PISA 2006 and will be the major domain again in PISA 2015.

The PISA definition of scientific literacy outlines what 15-year-old students should know, value and be able to do in order to be "prepared for life in modern society" (OECD, 2013c).

Central to the definition and assessment of scientific literacy are the competencies that characterise science and scientific enquiry. Students' ability to make use of these competencies depends on their scientific knowledge: both their content knowledge of the natural world and their procedural and epistemic knowledge. In addition, it depends on a willingness to engage with science-related topics.

The PISA 2015 framework describes and illustrates the scientific competencies and knowledge that will be assessed, the contexts for test items, and the range of items' "cognitive demand" (their level of difficulty). Test items will be grouped into units, each unit beginning with stimulus material that establishes the context for items. A combination of item types will be used. Computer-based delivery for 2015 offers the opportunity for several novel item formats, including animations and interactive simulations. This will improve the validity of the test and the ease of scoring.

The ratio of items assessing students' content knowledge of science to items assessing procedural *and* epistemic knowledge of science will be about 3:2. Approximately 50% of the items will test students' competency to explain phenomena scientifically, 30% their competency to interpret data and evidence scientifically, and 20% their competency to evaluate and design scientific enquiry. The cognitive demand of items will consist of a range of low, medium and hard. The combination of these weightings and a range of items of varying cognitive demand will enable proficiency levels to be constructed to describe performance in the three competencies that define scientific literacy.

Table 3.7 Major components of the PISA 2015 Framework for Scientific Literacy

Competencies	Knowledge (content)	Attitudes
• Explaining phenomena scientifically • Evaluating and designing scientific enquiry • Interpreting data and evidence scientifically	• Physical systems • Living systems • Earth and space systems • Procedural knowledge • Epistemic knowledge	• Interest in science • Valuing scientific approaches to enquiry • Environmental awareness

Source: OECD, 2013c.

Science in other large-scale assessments

Science is not widely included in large-scale global educational assessments. For TIMSS there is a strong curricular focus and the science frameworks are organised around a content dimension and a cognitive dimension. The content areas for the Grade 4 assessment are life science, earth science and physical science. The content areas for the Grade 8 assessment are biology, earth science, physics and chemistry. The cognitive dimension specifies the domains or thinking processes to be assessed.

The other major international assessment which includes science is LLECE. The science framework for LLECE has a literacy focus. It is based on the notion that the objective of scientific education is to mould students so they know how to fully participate in a world filled with scientific and technological advances. It posits that science education should enable students to adopt responsible attitudes, make fundamental decisions and resolve daily problems with respect for the environment and for future generations that have to live in it.

The LLECE science framework has a content dimension and a process dimension. The content dimension includes living beings and health, earth and environment, and matter and energy. The process dimension includes recognition of concepts, application and interpretation of concepts, and problem solving.

Implications

Assessment frameworks are at the heart of every assessment. Each framework reflects the issues that need to be addressed and formulates a way of going about it.

PISA focuses on assessing students' preparedness for the future, while the majority of the school-based assessments described here have a strong curricular focus. This may be a reflection of target groups: PISA assesses students at the end of compulsory schooling in most OECD countries, whereas most of the other assessments are given at an earlier time in a student's educational career. Early assessment gives an opportunity to implement remedial interventions based on test results, where appropriate.

Options and opportunities, therefore, exist in a number of different areas. Aligning with PISA's assessment frameworks would allow PISA-D results to be linked to the PISA proficiency levels of OECD and partner countries. While it is possible that PISA-D countries might find a curricular approach more suitable to their needs, this would make the link to existing PISA results more difficult to interpret.

Given that the proportion of students not in school at age 15 is likely to be higher in many PISA-D countries than it is in OECD countries, the PISA-D countries could opt to do an assessment at an earlier age. This would not only increase the coverage of students, but also give the opportunity to implement improvements before the end of students' education.

The inclusion of science as an area of assessment occurs only in a minority of assessments. It may be worth limiting the PISA-D assessment to language and mathematics.

Item development

Organisation and process

The item development process for PISA follows the assessment framework and involves collaboration with participating countries. The development process comprises a number of steps.

The first step in the process is item generation. Many items are created within the categories of the assessment framework. An excess of items is created because many will be dropped during the review and improvement process. To create around 120 items needed for a PISA major domain, the item writers would develop around 480 items to start the process.

Next is the panelling of items. Fellow item writers will review the newly created items and suggest amendments or deletion.

In the cognitive trial, or pilot, the items are tested with a small sample of the target audience. The trial sample will make comments and give feedback on the items' level of language and comprehensibility. Around 240 items will be chosen to go through to the field trial.

In the field trial, all remaining items will be tested in every participating country so that the developers can gain an idea of the items' cultural, geographic or ethnic bias.

Finally, items will be selected for the main study. Following the field trial, the required 120 items will be selected, taking into account coverage of the framework, and ensuring a range of difficulties and item response types are included.

In the creation of test items some of the assessments use a more centrally focused method, while others tend to draw widely. SACMEQ tests, for example, are developed by a panel of subject specialists drawn from all the 15 participating school systems, whereas PASEC tends to create the items centrally and finalise them in association with participating countries. Items are trialled in each of the countries and results analysed.

LLECE uses an approach in which a group of experts creates items, and also calls for submission of items, which are then refined at national co-ordinators' meetings. TERCE is based on a published curriculum analysis, which guided the creation of specification tables, which in turn ensured item development followed the curriculum. Item development was done, in principle, in a participatory fashion, involving specialists from almost all countries.

For EGRA and EGMA, for which results are not internationally comparable, each implementing country develops new versions of the EGRA/EGMA subtasks for its specific implementation. The Research Triangle Institute (RTI) provides guidelines for subtask development, but does not itself supervise or control the quality of the development. In a similar fashion, the ASER reading assessment is developed separately in each of the different assessment languages. The Hindi reading tool is developed at the

ASER Centre in New Delhi, and the reading tools in all other languages are developed by the Pratham and ASER Centre state teams.

The collaborative item development process undertaken by the OECD for PISA, the IEA for PIRLS and TIMSS, and the LLECE for TERCE can lead to a greater commitment on the part of the countries in the assessments.

The degree of centralisation is not related to the quality of the items produced, but more to the purpose of the instrument.

Example items relevant for PISA-D

Secure items from the PIRLS, TIMSS, PASEC and SACMEQ assessments have not been made available for this review. While considering items from other assessments may have been interesting, it is important to realise that items' characteristics can only be assessed by testing them with the specific target populations for which they are intended. An item that is suitable in one context will not necessarily be suitable in another. This is because there will be differences in assessment framework definitions and in the assessment's philosophy. As discussed above, some assessments are curriculum-based, while PISA is future-focused; and even within curriculum-based assessments, items may not be transferable due to differences between the curricula of the countries for whom the item was designed. There will also be differences in the time allowed for the test and differences in response type.

Some assessment programmes, however, publicly release items so that readers can gain an idea of the style of items and their difficulty.

ASER, for example, is designed to give a rapid and global assessment of basic reading skills. Given its orientation to precursor skills for reading literacy that do not include comprehension in any guise, it is judged to not be well aligned for integration with the PISA construct and framework for reading.

Following administration of each PIRLS survey a number of items are released (IEA, 2013a) along with associated item statistics, framework coverage and performance of individual countries (IEA, 2013b).

An example of such an item, 'Fly Eagle Fly', can be seen in Annex B (IEA, 2013a). This item is made up of a stimulus containing text and illustrations, followed by a number of questions. The first question, R21E01M, is a multiple-choice question about what the farmer in the passage was looking for. There were four alternatives. The proportion of students per country who chose the correct answer varied from 58% to 97% (IEA, 2013b). The process of comprehension being assessed in this item was described as "focus on and retrieve explicitly stated information".

The subsequent questions for this PIRLS item vary in their difficulty level. The hardest question was R21E07C, which had percentages for correct responses ranging from 9% in one country to 66% in the highest performing country (IEA, 2013b). The process of comprehension being assessed in this item was "interpret and integrate ideas and information", which is a more demanding task.

Items are also released to illustrate the TIMSS assessment (see Annex B).

In TIMSS 2011 there is an item about fractions (ID: M032166) which was categorised in the content domain of "number" and in the cognitive domain of "knowing". It is a multiple choice item with an average correct rate of 57%. In the highest performing country, Singapore, 92% of the students answered the item correctly, while in the lowest performing country, Ghana, 26% of the students were correct.

One of the most difficult TIMSS items was M032760B, which is in the content domain of "algebra" and the cognitive domain of "reasoning" with the topic area of "patterns". For this item 20% of students across all countries scored correctly. The highest correct rate was 65% in Singapore and the lowest was 3% in Ghana.

The LAMP assessment items shown in Annex B illustrate a prose item – asking students to read a label on a medicine and extract important information; and a numeracy item – asking students to calculate the total number of bottles shown in a diagram. LAMP is aimed at a population of those 15 years and older in developing countries. The items are developed with realistic context.

A sample item from the PIAAC Reading Components assessment is shown in Annex B.

Item developers for PISA-D will be able to use such information to guide the selection of suitable items.

A factor hampering the inclusion of items directly into PISA-D is that the scoring method may not be in line with PISA scoring methods. New response scoring algorithms would need to be developed for the items to be analysed alongside standard PISA items. For example, each element of a multi-part task could be scored in two parts, with some provision for treatment of elements not answered within the time period of the test.

The notion of framework coverage and alignment with existing PISA frameworks is important. When EGMA items, for example, are considered within the PISA framework, all items fit in the "quantity" content category; most fit in the "employ process" category (with only a few in "formulate"); and most of them are presented without any context, whereas context is a key characteristic of PISA items.

Implications

Key points of relevance to PISA-D are that a collaborative item development process can lead to a greater commitment among countries to the assessments, and that a centralised approach (with country input) allows for items to be more efficiently developed to reflect a given assessment framework.

In terms of using existing items from other assessments in PISA-D, the approach used in TIMSS Numeracy and prePIRLS Reading is relevant. TIMSS Numeracy asks students to answer questions and work out problems similar to TIMSS, except with easier numbers and more straightforward procedures.

Where a clear link to existing PISA assessments is required, items will need to fit into the framework structure of PISA, and be implemented in a similar way to PISA. In addition, care needs to be taken that the scoring procedures adopted in the items match that of PISA. PISA uses multiple choice and open or constructed and short response items. Response items are scored according to a two-point (0, 1), three-point (0, 1, 2) or, rarely, four-point (0, 1, 2, 3,) categorisation scheme.

Test design

Organisation, item difficulty, test targeting and mode of delivery

When designing a test to adequately ensure coverage over a range of difficulty levels, the item pool must contain many more items than could fit into a single test. As a result, not every student will sit the same test. Each student is exposed to a part of the total item pool. For this reason it is necessary to construct different booklets. PISA tests have traditionally been constructed in booklets composed of four clusters of items, which could include reading, mathematics or science items. For scaling to take place, some items must be common across the booklets. PISA focuses on one major domain and two minor domains each iteration, cycling through the domains from one survey implementation to the next. Every booklet will include at least one cluster of the major domain.

The basic PISA design had 13 booklets, enhanced by including items at the lower and upper extremes to a further 13 booklets. Countries opt to do just one of the sets of booklets.

Across the clusters there is a variety of items with different difficulty levels and modes of response.

In PISA the main contractor has been responsible for proposing the design to the PGB for approval.

PIRLS

In both TIMSS and PIRLS, approximately half the items are constructed response and half are multiple-choice (Mullis et al., 2012: 10).

Each multiple choice question is worth one point. Constructed response questions are worth one, two or three points, depending on the depth of understanding required. In the development of comprehension questions, the decision to use either a multiple choice or a constructed response format is based on the process being assessed, and on which format best enables test takers to demonstrate their reading comprehension.

Multiple choice questions provide students with four response options, of which only one is correct. For students who may be unfamiliar with this test question format, the instructions given at the beginning of the test include a sample multiple choice item that illustrates how to select and mark an answer (Mullis and Martin, 2013: 62).

Each constructed response question has an accompanying scoring guide that describes the essential features of appropriate and complete responses. Scoring guides focus on evidence of the type of comprehension the questions assess. The guides describe evidence of partial understanding and evidence of complete or extensive understanding. In addition, sample student responses at each level of understanding provide important guidance to scoring staff. In scoring students' responses to constructed response questions, the focus is solely on students' understanding of the text, not on their ability to write well. Also, scoring takes into account the possibility of various interpretations that may be acceptable, given appropriate textual support. Consequently, a wide range of answers and writing ability may appear in the responses that receive full credit for any one question (Martin, Mullis and Foy, 2013a: 63).

Significantly for the PISA-D initiative, the prePIRLS items use multiple choice and constructed response formats, as in PIRLS, but with several differences to accommodate the lower proficiency levels of the test takers. Constructed response items usually are worth only one or two points. However, there is a slightly higher percentage of constructed response items in the prePIRLS assessment, comprising up to 60% of the total score points. This decision was made because constructed response items that require a very short response often are easier for early readers due to the lighter reading demand, as compared with multiple choice items that require students to read and evaluate four response options. In addition, multiple choice items may lose some of their effectiveness in passages as short as those used in prePIRLS, because there are fewer plausible distracters that can be drawn from the text (Martin, Mullis and Foy, 2013a: 66).

Each domain contains items representing a full range of difficulty (Jones, Wheeler and Centurino, 2013: 55)

In educational measurement, analysis is most informative when the difficulty of the items used to assess student achievement matches the ability of the students taking the assessment. In the context of assessing mathematics/science achievement, measurement is most efficient when there is a reasonable match between the mathematics/science ability level of the student population being assessed and the difficulty of the assessment items. The greater the mismatch, the more difficult it becomes to achieve reliable measurement. In particular, when the assessment tasks are much too challenging for most students, to the extent that many students are responding at chance level, it is extremely difficult to achieve acceptable measurement quality.

PIRLS and prePIRLS are currently paper-and-pencil assessments.

TIMSS

The TIMSS assessments primarily use multiple choice and constructed response items. At least half of the total number of points represented by all the items will come from multiple choice items. Each multiple choice item is worth one score point. Constructed response items generally are worth one or two score points, depending on the nature of the task and the skills required to complete it. In developing assessment items, the choice of item format depends on the mathematics or science being assessed, and the format that best enables students to demonstrate their proficiency (Martin, Mullis and Foy, 2013b: 92).

In the context of assessing mathematics/science achievement, measurement is most efficient when there is a reasonable match between the mathematics/science ability level of the student population being assessed and the difficulty of the assessment items.

The mode of test delivery for TIMSS has been paper-and-pencil.

SACMEQ

SACMEQ tests were developed by a panel of subject specialists drawn from all the 15 SACMEQ school systems to identify those elements of curriculum outcomes that were considered important and which were to be assessed in the tests. The subject specialists also reviewed the test items to ensure that they conformed to the national syllabuses of SACMEQ countries (Hungi, 2011: 3).

SACMEQ is a paper-and-pencil assessment.

PASEC

The PASEC Grade 2 assessment is administered at the beginning of the school year. There are no rotated booklets but only one booklet: all the students have the same items.

Tests are taken individually with the assistance of a test administrator in charge of oral instructions and coding. The administrator is given a notebook for each student. Each notebook contains instructions and correction tables. The administrator directly corrects each student's answers in that student's notebook after the administration of the exercise. For most exercises, the administrator provides the students with a student support resource, containing images, letters and words grids and texts that students must browse and read in order to answer the various exercises. In mathematics, students are also given a slate and chalk to help them solve operations and problems.

Students can answer questions with very brief answers, by pointing to an image or an item with their finger on the student support, by reading letters, numbers, words or sentences aloud or by showing their written answer on their slate. Some examples are given at the beginning of each exercise to ensure that all students understand the meaning of the question.

The full PASEC test is administered to Grade 6 students. The 2014 PASEC test used an item pool of 92 reading items, divided into four blocks; and 81 mathematics items, also in four blocks. These items were then arranged into four test booklets (booklet A/B/C/D). Each booklet contained two blocks of reading items and two blocks of maths items. Each student only answers 46 reading items and 40 mathematics items.

Each block is found twice in the four booklets (A/B/C/D). A total of eight blocks (four in reading: "L" and four in mathematics: "M") are located in the four booklets so that each block appears once at the beginning and once at the end. The eight blocks are located in the four booklets as follows:

Booklet A	Block 1 L	Block 2 L	Block 1 M	Block 2 M
Booklet B	Block 2 L	Block 3 L	Block 2 M	Block 3 M
Booklet C	Block 3 L	Block 4 L	Block 3 M	Block 4 M
Booklet D	Block 4 L	Block 1 L	Block 4 M	Block 1 M

PASEC is a paper-and-pencil delivered assessment.

LLECE

A UNESCO panel of experts developed the test and booklet design. There are six clusters of items per domain and two clusters are used per booklet.

The Second Regional Comparative and Explanatory Study (SERCE) and TERCE were paper-and-pencil tests.

Table 3.8 SERCE test and booklet design

Grade	Domain	Multiple choice items	Open items	Total items
3	Reading	11, in each cluster	0	11
	Maths	10, in clusters 2, 4 and 6 12, in clusters 1, 3 and 5	2, in clusters 2, 4 and 6	12
6	Reading	16, in each clusters	0	16
	Maths	13, in clusters 2, 4 and 6 16, in clusters 1, 3 and 5	3 in clusters 1, 3 and 5	16
	Science	14, in each clusters	1, in all clusters	15

Source: SERCE, 2010.

PIAAC

The set of items for the PIAAC main study was balanced in terms of construct representation, based on the overall distributions recommendations in the framework. A total of 58 items was selected for literacy and numeracy, with the distribution across linking and new paper and computer versions shown in Table 3.9 (Louise and Tamassia, 2013: 24). The test design for PIAAC was based on a variant of matrix sampling (using different sets of items, multi-stage adaptive testing and different assessment modes) where each respondent was administered a subset of items from the total item pool. Different groups of respondents therefore answered different sets of items.

PIAAC can be taken as a paper-based survey or as a computer-based survey.

Table 3.9 Literacy and numeracy items in the PIAAC main study

	Literacy		Numeracy	
	Linking	New	Linking	New
Paper-based	18	6	19	6
Computer-based	30 (including computer versions of the 18 above linking items)	22	28 (including computer versions of 14 of the above linking items)	22 (including computer versions of 3 of the above linking items)

Source: Louise and Tamassia, 2013: 25.

LAMP

LAMP is one of the few assessments that includes an adaptive process – one that filters or directs students on the basis of previous responses.

Students first sit a filter test. This is a brief booklet intended to establish if the respondent would most likely possess lower or higher levels of literacy skills. It therefore helps in deciding what sort of instruments should be used to gain a more in-depth picture of the respondent’s skills.

Students with relatively low filter test results are given the module for those with lower performance. This module is composed of two instruments. One instrument supplements the information produced by the filter test with more detail and establishes more precisely where the respondent stands in relation to the lower skill levels. The other enables an in-depth exploration of the operations (reading components) that might be preventing the respondent from achieving a better performance.

Students with relatively high filter test results are given a module for those with higher performance. This module comprises one booklet (in two versions) that supplements the information produced by the filter test with more detail and establishes more precisely where the respondent stands in relation to the higher skill levels.

LAMP is a paper-and-pencil delivered assessment.

ASER

Each year there are four test forms for each domain. There are no common items across any two forms and no systematic method for rotating forms during test administration.

Regarding form rotation during test administration, ASER Centre says:

> The test administers are instructed not to use the sample form for the children from the same household; especially if they are around each other as the test is being administered. In rural household settings, it is often the case that the siblings of the child hang around out of curiosity, while [the child] is being tested. To avoid imitation of responses from one child to another, this instruction was incorporated. (Banerji, R., personal communication, 27 April 2014)

In ASER the test is designed so that there are a given number of items per task in each domain. For example, in the reading domain, for the task for 'letters', any five letters from a set of ten letters are selected and read aloud. Items are selected to cover the other tasks of "words" (five items), "paragraph" (reading text of four sentences) and "story" (reading aloud seven to ten sentences).

Similarly in the mathematics domain, five items are included for each task of "number recognition (1 to 9)" and "number recognition (10 to 99)", two items for "subtraction" and one item for "division".

There are varying degrees of difficulty.

ASER is a paper-and-pencil delivered assessment.

Uwezo

Each year there are four test forms for each domain. There are no common items across any two forms.

In literacy, the tests include items about letter/syllable recognition, reading aloud and comprehension. In numeracy the tests include items about counting, number recognition and understanding of terms such as "greater than" as well as knowledge of the operations, addition, subtraction, multiplication and division.

There is no systematic method for rotating forms at the time of test administration, but the Uwezo standards state that to avoid familiarity a different set of tests should be administered to each child in a household (Uwezo, 2012: 8).

Implications

To cater for the expected wide range of student capacity, a test design must include sufficient items across all levels of difficulty. Experience with PISA has shown that, in the context of developing countries, the tests can be too hard for the majority of the students. In some countries over 50% of the students score below Proficiency Level 1, meaning that there is no description of the capacity of these students. This is despite the fact that from 2009 onwards PISA tests have been extended to include a greater number of easy items.

A large range of item types and difficulties needs to be included in the test. This will be best done with a multi booklet approach. The booklets should include some common items to allow linking between them.

Regard should be given to the mode of delivery of the test. Many of the tests examined here are paper-and-pencil tests. However, ACER has recently successfully implemented tests using tablet computers, in Lesotho, Afghanistan and remote Indigenous communities in Australia. This form of test delivery is worth considering. There are advantages to this approach:

- Students are more stimulated by the test experience.
- Students easily master the equipment, even when they have never seen a tablet before.
- Innovations such as sound can be easily introduced, thereby accommodating students with sight difficulty.
- Student responses are captured instantly, alleviating the need for an expensive data-entry process.
- Data-entry errors are eliminated.
- Data management is much easier and more secure; data loss is reduced; and data can be uploaded whenever administrators have a reliable Internet connection.
- Tablets can be re-used many times.

At the same time it should be acknowledged there are some potential obstacles to the introduction of tablets. ACER's experience suggests a number of challenges:

- The design processes for the platform and the app could be costly and time-consuming, especially if starting from scratch.
- Translation processes could be difficult. (Translation is built into ACER's existing translation management system.)
- It may be difficult to ensure all countries use the same model tablet. If they do not, the app and directions will generally have to change for each model. (This would not be a significant issue if only simple multiple choice items were used.)
- Some countries may face problems with theft; a tablet is a more attractive item than a standard test booklet.

Using tablets would also be in line with PISA's latest implementation, a mostly computer-based test. However, this technology is not currently widely used by the assessments included in this report.

Psychometric analyses, scaling, calibration and equating methods

For scaling, PISA employs "item response theory" in the form of a one-parameter Rasch model. Open student responses (as opposed to multiple choice) are coded by trained coders to ensure consistency across countries. The codes for the responses to both open and multiple choice items are entered into a custom designed software package by trained data entry personnel.

TIMSS, PIRLS, SACMEQ, LLECE, PIAAC and STEP all use item response theory.

PASEC provides an interesting example of scaling methodology evolution. They had employed classic test theory, but used an item response theory analysis (Rasch measurement) from 2012. This item response theory analysis was initially for cognitive tests only in Mali, Vietnam, Cambodia and PDR Lao, but is now being extended to both tests and contextual data.

For EGRA, the situation is different in each country, so there is no overall international scaling. However, item response theory is sometimes used to analyse field trial data (to ensure, for example, that the reading passages cover the whole ability range and discriminate well between different ability levels).

While item response theory is not the only method employed, it is widely used in scaling and is the preferred method of all the assessments under review, as discussed below.

Implications

Item response theory scaling is the preferred method for analysing student results of all the assessments under review. This type of scaling is based on a continuous interaction between the student's capacity and an item's difficulty, and gives a clear picture of students' capacity.

It also allows a particular test to be linked to any other test by including common items in both. This can be done over successive years to gain an accurate picture of a student's educational growth.

PISA has used a one-parameter model based on item difficulty, and will be modifying the approach slightly for PISA 2015. PIRLS and TIMSS each employ a three-parameter model.

Given the wide range of student capacity across the countries, PISA-D might incorporate a process to determine what type of test would be most appropriate for particular students to do. Some form of adaptive testing may be considered. This would be done with the aim of targeting tests at students' level of skill. The approach used in PIAAC is relevant for PISA-D. PIAAC's test design is based on a variant of matrix sampling (using different sets of items, multi-stage adaptive testing, and different assessment modes) with each respondent administered a subset of items from the total item pool. Different groups of respondents therefore answer different sets of items, making it inappropriate to use any scaling system based on the number of correct responses.

Cross-country comparability

The first step in establishing cross-country comparability in PISA takes place during the item development process, when each participating country is given the opportunity to review the items to ensure that they are relevant to their student body.

A further step involves an analysis of countries' results on all items from the field trial and comparing performance on each item with the expected performance. Any deviation – that is, where the item appears to be too easy or too hard – is investigated, to see if the cause is the translation, the presentation of the item or some cultural or geographic factor that changes the expected difficulty. This is known as differential item functioning. In PISA this is done both for the field trial and the main study.

In assessments such as PIRLS, TIMSS, PIAAC and LAMP, where cross-country comparisons are routinely made, item-by-country interactions are analysed. It is essential that the items do not behave in an incongruous manner from one country to another.

Until 2014, PASEC did not have a focus on international comparisons, but from 2014 it will undertake a study of item-by-country interaction.

No such measures are needed in assessments such as EGRA or EGMA, where comparisons are not made, nor in ASER which is focused on one country only.

Uwezo, which is based on ASER, occurs across three countries. The tests are not identical because they are based on curriculum expectations of the respective countries. Including results from questions seen as equivalents across the countries achieves a measure of comparability.

Implications

PISA-D should undertake a differential item functioning process to identify any item-by-country interactions. This will identify any items that may advantage or disadvantage a particular country. How confident a country is to become involved in the process depends on their perception of being treated fairly.

This is best done in a two-step approach using the results from a field trial to identify any problems in cross-country comparability, and then acting to remove these problems for the main study. Following the main study, analysis should take place again to verify the cross-country comparability of the assessment.

Trends

One of the most important uses of assessment data for countries is to observe any changes occurring over time. Changes over time, or trends, give a measure of growth and improvement in student capacity. To facilitate this process it is necessary to include a proportion of the same items from one survey administration to the next; to measure change in results all other variables must remain fixed, such as the method of measurement.

Growth is measured when the same cohort is measured at different stages of their educational career. For most assessments it is not feasible to test the same students, but a measure of growth can be obtained if a representative sample is taken of the same cohort in a country as that cohort moves through the education system. TIMSS achieves this to a degree, by assessing students at both Grade 4 and at Grade 8.

Improvement (or declining performance) is indicated when there is an increase (or decrease) in student capacity at the same level in successive administrations of the survey – so change can be indicated for a country when comparing student performance in PISA 2003 and PISA 2012, for example.

In each PISA cycle, items are kept secure for future use and are deployed as link items appearing in a number of different survey cycles. This allows a measure of change to be calculated.

Change calculations include an estimate of error in linking from one survey cycle to the next. This estimate, known as the linking error, is built into the calculation of standard errors associated with the difference between the results of two PISA surveys.

The different assessments use a variety of approaches to measure change over time. In PIRLS, six of the ten 40-minute blocks of items were included in previous PIRLS assessments: two in all three assessments (2001, 2006 and 2011), two in both PIRLS 2006 and PIRLS 2011, and two in PIRLS 2011 only. Four new blocks will be developed for use for the first time in the 2016 assessment.

SACMEQ includes not only items from past implementations of its test, but also includes some items from other assessments such as TIMSS and PIRLS. The data from the combination of these items can be used to analyse change over time. In the first phase of PASEC, each country had been tested with the same booklets. Results from PASEC 2014, and from the next cycle in 2018, will be directly comparable.

The first two administrations of the LLECE assessments, namely the First Regional Comparative and Explanatory Study (PERCE) and SERCE, are not comparable because SERCE introduced a series of modifications resulting from the experience and knowledge gained from the implementation of PERCE. Some of the changes are related to sampling, test design, target population and knowledge domains covered by the assessment. However, by aligning the methodology, SERCE and the third implementation (TERCE) are considered comparable studies. There will be two scales: a comparable scale (already published), and a TERCE scale, to be used as the baseline from now on.

In ASER, care is taken to ensure that one year's reading tool is comparable with previous years' tools in terms of word count, sentence count, types of words and conjoint letters in words. They don't necessarily use the same items from one year to the next.

In Uwezo, an attempt is made to ensure that the level of difficulty and comparability across the years is retained. In each year one new aspect will be added, while keeping the core the same to enable comparability across years.

Implications

One key to attracting participants for a large-scale educational assessment such as PISA-D will be to allow countries to monitor changes to educational standards over time, such as what proportion of students are achieving at a given level from one administration to the next. The assessments will need to include a selection of the same items from one survey administration to the next. This will allow the scaling process to link the two test cycles. This has implications for maintaining security of those items; if they enter the public domain they cannot be used confidently for this purpose. Items' security is paramount if a reliable measure of trends is to be achieved.

In addition, it is also likely that countries will want some measure of the growth of students' skills and knowledge as each cohort progresses in their education. This is done by administering the test to different year levels, but using some common items, so that the results can be mapped on the same scale.

Proficiency levels

A numerical student score, while providing a basis for comparison to other students, does not provide a guide to the student's strengths and weaknesses. This can be done by examining the test items that the student is capable of and those that the student finds too difficult. Item response theory gives a means of doing this by dividing students into like groups and describing the characteristics of those groups. These characteristics are known as described proficiency levels.

In PISA, creating proficiency levels involves statistical processes and examination of item content. Following the main study, student results are scaled and items are divided into proficiency levels according to how many students answered each item correctly. The items are then examined for content and descriptions are created for each of the proficiency levels based on the tasks that are included in the items.

In PISA there are typically six levels, from Level One (the most basic), to Level Six (the most advanced). The proportion of students in each level provides valuable information to the participating countries about their students. If a country has a large proportion of students in the lower levels, for example, this might inspire the country to implement policy interventions aimed at remediation in an attempt to help the students catch up.

On the other hand, a country may have the vast bulk of students in the middle levels with few at the extremes. This might suggest that the students at the top end of the scale need strategies to extend their skills.

TIMSS and PIRLS have identified four points along the achievement scales to use as international benchmarks of achievement – Advanced International Benchmark (625), High International Benchmark (550), Intermediate International Benchmark (475), and Low International Benchmark (400). With each successive assessment, TIMSS and PIRLS work with the expert international committees to describe student competencies at the benchmarks. Experts then summarise the detailed list of item competencies in a brief description of achievement at each international benchmark.

For SACMEQ proficiency levels are created for reading and mathematics. Rasch item response theory was used to establish the difficulty value for each test item, national research co-ordinators subjected each test item to an intensive "skills audit", and then wrote descriptive accounts of the competencies associated with each cluster of test items by using terminology that was familiar to ordinary classroom teachers.

Similar processes have been in used in LLECE assessments and in PIAAC, STEP and LAMP.

Defining proficiency levels has not been a feature of the ASER, EGRA/EGMA and Uwezo surveys.

The different assessments use a range of different strategies when describing the proficiencies of the students – some assessments use a process not dissimilar to the one described for PISA above, while others do not report proficiency levels at all.

Implications

It is highly desirable to define the proficiency levels of the students in addition to assigning them a numerical value for their results. Described proficiency levels based on the level of difficulty of the items and the tasks associated with the items give a better idea of students' strengths and weaknesses.

Translating, adapting and verifying cognitive instruments

For a test to be taken across different countries in different languages, a systematic method of translation needs to be established to ensure that the test is a true reflection of student capacity and not a reflection of the language in which the test is administered.

The standard PISA survey is administered in more than 65 countries in approximately 44 languages.

Table 3.10 Translated languages in other assessments

	Assessment	Number of translated languages
Large-scale international surveys	PIRLS	48
	TIMSS	58
	SACMEQ	3
	PASEC	15
	LLECE	2
School-based surveys	EGRA	70
Household-based surveys	PIAAC	20
	STEP	8
	LAMP	15
	ASER	18
	Uwezo	6

Source: Author's analysis of the technical manuals of each assessment.

In PISA, test items are created originally in English and then a second parallel version in French is prepared. Countries are supplied with these two source versions of every item, which they then organise for translation into their own language. In some countries this can be more than one language – for example in Switzerland, the assessment is administered in French, German and Italian. The test is administered in the language of instruction. In some countries this can be more than one language – for example, in Luxembourg there are different languages for different subjects. The same applies in Qatar where an extra complication arises because the languages (English and Arabic) are read in different directions – English is left to right and Arabic is right to left.

The two versions of the test emanating from the two source languages are then reconciled to ensure that they are the same. Following this a linguistic quality company will verify that the translated test is indeed the same as the one intended, so that the students receive no advantage nor disadvantage by undergoing the test in their own language.

In SACMEQ independent translations are made by at least two different expert translators familiar with age-appropriate linguistic demands. In cases of disagreement, consensus is achieved either by direct negotiation between the two translators or by a third expert making the final choice.

In PIRLS, TIMSS and PASEC, procedures also follow double independent translation plus external reconciliation. In PASEC the translation process is outsourced to specialised consultants and overseen by the PASEC technical team.

Translation processes for WEI-SPS and PIAAC were based on the materials and procedures used in PISA; that is, two independent translations from source versions followed by reconciliation and verification.

For the LLECE assessments, Spanish is the common language for all participating countries except for Brazil, where the test is in Portuguese. LLECE uses the back translation process: the Spanish source version is translated into Portuguese and then translated back into Spanish. The source Spanish version and back-translated version are compared and validated before the test.

For EGRA and EGMA there is no specified translation process.

For ASER and Uwezo the tools are developed independently in the separate languages.

Assessments where international comparisons are of prime importance usually undertake a double translation process with an independent verifier.

Implications

To maintain the highest standards for translation, an assessment should adopt a two-source version approach with independent translations of each source version, which are then reconciled and verified by an expert language organisation.

Field trial and item selection

In PISA, a field trial is administered in all participating countries in the year before the main study takes place. A country cannot participate in the main study if it has not done the field trial. Generally about 1 000 students of the target population undertake the field trial.

There are two main purposes of a field trial.

The first is to test the suitability of the item pool. Generally speaking, each newly created item needs to be administered to a minimum of 200 students per country so that the characteristics of the item can be fully described. This includes the item's difficulty level, discrimination, point biserial[1] (a figure which indicates if the better students are getting the more difficult questions right) and an indicator of how closely the item fits a model proposed by the Rasch analysis. This also establishes whether the item performs differently for boys and girls.

The second purpose of a field trial is to test the country's logistical capacity to carry out the assessment. The administration of a large-scale international assessment is a complex task. The field trial allows the country to test the various procedures that are necessary: for example, sampling the students, contacting schools, and training test administrators, coders and data-entry personnel.

There are other benefits of a field trial: for example, the coding guide for some items may need to be modified or extended slightly based on the field trial results. In PISA generally, twice as many items are field trialled than are needed for the main study. Final item selection is based on ensuring framework coverage, a diverse range of difficulty levels appropriately targeted at the student sample, different item response formats, minimal cultural bias, including items which take the appropriate amount of time allocated and ensuring that there is a balance of gender effects. This doesn't mean that all items need to be gender-neutral, but that there should be a balance of items that tend to favour both boys and girls. In this way the different response patterns can be explored.

All global assessments reviewed undertook some form of field trial or pilot study before the main study. Most assessments use the field trial to select the most appropriate items to go forward to the main study, keeping in mind the need to ensure that a wide range of difficulty and response types are included.

In PIRLS and TIMSS the field trial sample size is approximately 30 schools in each country, yielding at least 200 student responses to each item. To lessen the load on schools, the samples for the field trial and the main study are drawn simultaneously, using the same random sampling procedures. This ensures that field trial sample closely approximates the main study samples, and that a school is selected for either the field trial or the main study, but not both.

The PASEC field trial sample is around 20 schools per country.

In the WEI-SPS, field trial analysis looked at the feasibility and cross-cultural validity of questions across the countries.

In TERCE, item behaviour in the pilot study was analysed based on an analytical plan by the implementation partner.

In EGRA, countries are encouraged to do a field trial to ensure the tool is accurately measuring what children know in the specific context and language(s) of assessment. It also allows verification of the validity and reliability of the instruments and gives the EGRA team an opportunity to address technical issues before the main study.

In LAMP, the field trial involves administering the entire battery of survey instruments to a carefully selected sample (not random) of roughly 500 adults in each test language.

In Uwezo, pre-tests involving six sample forms for each domain are conducted in several districts with different geographical characteristics. During pre-tests the test administrators note the tasks that are difficult for the children. After each pre-test there is a revision meeting in which feedback from test administration is shared. Revisions are made based on this feedback and recorded in the test tracking tool. The forms are then sent into the next pre-test. At the pre-testing stage, the data collected to inform test development are anecdotal data from the test administrators, whereas at the district-wide pilot stage assessment data are collected and analysed as they are in the main administration.

Implications

In the vast majority of assessments, some form of field trial takes place to ensure that the instrument is appropriately targeted. Most of the assessments also use the field test to examine the procedures needed to carry out the assessment.

For PISA-D, a field trial should take place to test the items' suitability for the target sample and to see if each participating country has the capacity to implement the assessment. It is normal for a large number of items to be discarded following the field trial.

None of the countries to be undertaking PISA-D has participated previously in PISA. It is also possible that future participants may not have taken part in any international assessments and may not have well-developed national assessments. It is vital, therefore, that they gain as much experience as possible in the procedures associated with international testing, and this is best done with a field trial.

A field trial is also needed to ensure that the items used in the assessment are effectively targeted at the participating countries.

Notes

1. Biserial correlation is a measure of association between a continuous variable and a binary variable. It is constrained to be between -1 and +1. The point biserial correlation is positive when large values of X are associated with Y=1 and small values of X are associated with Y=0.

References

Hungi, N. (2011), *Accounting for Variations in the Quality of Primary School Education*, SACMEQ, Paris, www.sacmeq.org/?q=publications.

IEA (2013a), *PIRLS 2011 User Guide for the International Database: PIRLS Released Passages and Items*, TIMSS and PIRLS International Study Center, Boston College, Chestnut Hill, MA, and International Association for the Evaluation of Educational Achievement (IEA), Amsterdam.

IEA (2013b), *PIRLS 2011 User Guide for the International Database: PIRLS Percent Correct Statistics for the Released Items*, TIMSS and PIRLS International Study Center and IEA, Chestnut Hill, MA and Amsterdam.

Jones, L.R., G. Wheeler, and V.A.S. Centurino (2013), "TIMSS 2015 science framework", in I.V.S. Mullis and M.O. Martin (eds.), *TIMSS 2015 Assessment Frameworks,* TIMSS and PIRLS International Study Center and IEA, Chestnut Hill, MA and Amsterdam, pp. 29-58.

Louise, M. and C. Tamassia (2013), "Chapter 2: The development of the PIAAC cognitive instruments", *Technical report of the Survey of Adult Skills (PIAAC),* pre-publication copy, OECD, Paris.

Martin, M.O., I.V.S. Mullis and P. Foy (2013a), "PIRLS 2016 assessment design and specifications", in I. V. S. Mullis and M. O. Martin (eds.) *PIRLS 2016 Assessment Frameworks,* TIMSS and PIRLS International Study Center and IEA, Chestnut Hill, MA and Amsterdam, pp. 57-69.

Martin, M.O., I.V.S. Mullis and P. Foy (2013b), "TIMSS 2015 assessment design", in I.V.S. Mullis and M.O. Martin (eds.), *TIMSS 2015 Assessment Frameworks*, TIMSS and PIRLS International Study Center and IEA, Chestnut Hill, MA and Amsterdam.

Mullis, I.V.S. et al. (2012), "Assessment framework and instrument development", in M.O. Martin and I.V.S. Mullis (eds.), *Methods and Procedures in TIMSS and PIRLS 2011*, TIMSS and PIRLS International Study Center, Chestnut Hill, MA.

Mullis, I.V.S. and M.O. Martin (eds.) (2013), *PIRLS 2016 Assessment Framework*, TIMSS and PIRLS International Study Center and IEA, Chestnut Hill, MA and Amsterdam.

OECD (2013a), *PISA 2015 Draft Reading Literacy Framework*, www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Reading%20Framework%20.pdf.

OECD (2013b), *PISA 2015 Draft Mathematics Framework*, www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Mathematics%20Framework%20.pdf.

OECD (2013c), *PISA 2015 Draft Science Framework*, www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Science%20Framework%20.pdf.

OECD (2010), *PISA 2009 Assessment Framework: Key Competencies in Reading, Mathematics and Science*, PISA, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264062658-en

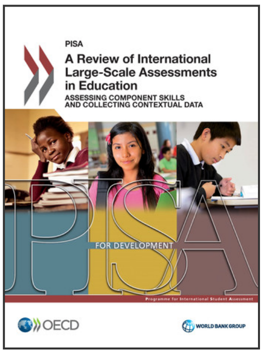
Pierre, G. et al. (2014), *STEP Skills Measurement Surveys: Innovative Tools for Assessing Skills*, working paper, World Bank Human Development Network, Washington DC.

Ross, K. et al. (2004), "Chapter 2: Methodology for SACMEQ II Study", IIEP, UNESCO, Paris.

SERCE (2010), *Segundo Estudio Regional Comparativo y Explicativo: Compendio de los manuales*, S. Block (ed.), A. Atorresi, C. Pardo, D. Glejberman, G. Espinosa, L. Toranzos, M. Rocha, M. Castro, Santiago: UNESCO/OREALC.

SERCE (2009), *Segundo Estudio Regional Comparativo y Explicativo: Aportes para la enseñanza de la matemática*, L. Bronzina, G. Chemello, M. Agrasar, Santiago: UNESCO/OREALC.

Uwezo (2012), *Standards Manual,* Uwezo, Nairobi.



From:

A Review of International Large-Scale Assessments in Education

Assessing Component Skills and Collecting Contextual Data

Access the complete publication at:

https://doi.org/10.1787/9789264248373-en

Please cite this chapter as:

Cresswell, John, Ursula Schwantner and Charlotte Waters (2015), “Component skills and cognitive instruments used in educational assessments”, in *A Review of International Large-Scale Assessments in Education: Assessing Component Skills and Collecting Contextual Data*, The World Bank, Washington, D.C./ OECD Publishing, Paris.

DOI: https://doi.org/10.1787/9789264248373-6-en

This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of OECD member countries.

This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.