Annex C

Overview tables related to cognitive data collection instruments

Table C.1 Reading frameworks for other assessments

	Assessment	Reading definition
Large-scale international surveys	**PIRLS**	Currently, the PIRLS definition of reading literacy is (Mullis, Martin, and Sainsbury, 2013: 13): “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 et al., 2013: 14). PIRLS assesses students’ reading achievement within the two overarching purposes for reading that account for most of the reading done by young students both in and out of school: • reading for literary experience • reading to acquire and use information. The prePIRLS 2016 assessment reflects the same conception of reading as PIRLS 2016, except it is less difficult and is designed to test basic reading skills that are a prerequisite for PIRLS. The reading passages are shorter, with easier vocabulary and syntax. Students’ ability to read and answer questions about these passages can provide valuable information about their strengths and weaknesses in reading comprehension (Mullis and Martin, 2013: 8).
	TIMSS	Not relevant – mathematics and science only.
	SACMEQ	Reading literacy is defined as: “the ability to understand and use those written language forms required by society and/or valued by the individual.” Narrative prose: Continuous texts in which the writer aims to tell a story – whether this be fact or fiction. Expository prose: Continuous text in which the writer aims to describe, explain, or otherwise convey factual information or opinion to the reader. Documents: Structured information organised by the writer in a manner that requires the reader to search, locate, and process selected facts, rather than to read every word of a continuous text (Ross et al., 2004: 46).

<table>
<tr><th></th><th>Assessment</th><th>Reading definition</th></tr>
<tr><td><b>Large-scale international surveys</b> (cont.)</td><td><b>PASEC</b></td><td>The 2014 PASEC international assessment puts forward a new methodological framework, taking into consideration:
• scientific research in reading, reading comprehension and mathematics
• internationally shared common skills standards in reading and mathematics
• students' skills level in reading mathematics, but also the environmental context of the countries assessed and the effective curricula of these countries
• international standards for measuring reading comprehension and mathematics.

Grade 2: skills assessed in the language of instruction are used to measure students' abilities at an early stage of learning to read.
Oral comprehension: understand vocabulary, recognise vocabulary and word families, understand a text.
Familiarisation with writing, phonological awareness and reading decoding: read invented words, read letters, recognise syllables, read words, recognise invented words.
Reading comprehension: decode the meaning of words, read and understand sentences, understand a text.
Grade 6: levels of decoding and reading comprehension should be assessed, for the diagnostic of students' daily reading abilities (in and out of school), in order for them to learn, understand and entertain themselves. The tests focus on two major domains of reading competencies: decoding words and isolated sentences and texts comprehension (extract explicit information, make simple inferences, interpret and combine information).</td></tr>
<tr><td></td><td><b>LLECE</b></td><td>Reading: Correctly interpret and resolve communicative problems based on written information contained in various authentic texts. Authentic texts could be news articles, encyclopaedia articles, fiction, entertainment, educational, functional and others.

<table><tr><td></td><td>Domain content</td><td>Process</td></tr><tr><td>Reading</td><td>Reading of paragraphs and texts
Reading of statements and words</td><td>Literal
Simple inference
Complex inference</td></tr></table></td></tr>
<tr><td></td><td><b>WEI</b></td><td>There are no cognitive items in this assessment</td></tr>
<tr><td><b>School-based surveys</b></td><td><b>EGRA</b></td><td>In the EGRA toolkit, five essential components of effective reading are listed:
• Phonemic awareness
• Phonics
• Fluency
• Vocabulary
• Comprehension</td></tr>
</table>

Household-based surveys	**PIAAC**	Task characteristics (categorising texts) The following variables have been used to categorise texts for the purposes of the PIAAC assessment: • Medium (print and digital) • Format (continuous and non-continuous) • Type (rhetorical stance) • Physical layout (type of matrix organisation) • Features unique to digital texts • Social context (OECD, 2013: 20). Reading components PIAAC includes a component test intended to provide more information on the abilities of those with low levels of literacy (OECD, 2013: 27). The PIAAC components assessment includes test of print vocabulary, sentence processing and basic passage comprehension. In skilled reading, these components are integrated to support literacy performance. During acquisition, even by adults, they may be measured separately, with different profiles having implications for learning, instruction and policy (OECD, 2013: 28).
	STEP	Reading literacy assessment The STEP reading literacy assessment has been developed specifically for use in the context of developing countries, and it includes sets of questions taken from PIAAC, the International Adult Literacy Survey, and the Adult Literacy and Life Skills Survey. This overlap allows countries participating in the STEP programme to compare their literacy results with those of over 30 other countries (Pierre et al., 2014: 35). Definition of literacy: “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” (PIAAC Literacy Framework, Pierre et al., 2014: 36).
	LAMP	• Alphanumeric perceptual knowledge and familiarity • Word recognition • Decoding and sight recognition • Sentence processing • Passage reading
	ASER	• Letter recognition • Word recognition • Passage reading (4 sentences, approx. 19 words) • Passage reading (7-10 sentences, approx. 60 words)
	Uwezo	The framework retained the levels used by ASER in literacy and numeracy and was informed by the EGRA design. This framework documents the various competencies to be tested, levels of competencies and steps to be used in developing the tests. It also lays out the rules governing each test. The framework was first developed in 2008/9 and critiqued and improved in 2010” (Uwezo, 2011: 17).

Table C.2 Mathematics frameworks from other assessments

	Assessment	Mathematics definition
Large-scale international surveys	**TIMSS and TIMSS Numeracy**	There is no specific definition for mathematics, because it is somewhat dependent on the curricula of the participating countries. At each grade the mathematics framework is organised around two dimensions: a *content dimension* specifying the subject matter to be assessed and a *cognitive dimension* specifying the thinking processes to be assessed. The Grade 4 content dimension includes number, geometry and data display, while the Grade 8 content dimension includes number, algebra, geometry and data and chance. The cognitive dimension includes knowing, applying and reasoning. TIMSS 2015 also has a new, less difficult mathematics assessment called TIMSS Numeracy. TIMSS Numeracy assesses fundamental mathematical knowledge, procedures, and problem-solving strategies that are prerequisites for success on TIMSS. TIMSS Numeracy asks students to answer questions and work problems similar to TIMSS, except with easier numbers and more straightforward procedures. TIMSS Numeracy is designed to assess mathematics at the end of the primary school cycle (Grades 4, 5 or 6) for countries where most children are still developing fundamental mathematics skills (Mullis and Martin, 2013: 7, 8).
	SACMEQ	Mathematics literacy is defined as “the capacity to understand and apply mathematical procedures and make related judgements as an individual and as a member of the wider society”. Mathematics subdomains: Number: operations and number line, square roots, rounding and place value, significant figures, fractions, percentages, and ratios. Measurement: measurements related to distance, length, area, capacity, money, and time. Space-data: geometric shapes, charts (bar, pie, and line), and tables of data (Ross et al., 2004: 49).
	PASEC	The 2014 PASEC international assessment puts forward a new methodological framework, taking into consideration: • scientific research in reading, reading comprehension and mathematics • internationally shared common skills standards in reading and mathematics • students’ skills level in reading mathematics, but also the environmental context of the countries assessed and the effective curricula of these countries • international standards for measuring reading comprehension and mathematics. *Grade 2:* PASEC tests are used to measure students’ knowledge and competencies in their early stages of learning mathematics. Arithmetic: Count to 100, recognise numbers, count objects, determine quantities, sort numbers, continue sequences of numbers 1, continue sequences of numbers 2, add and subtract, solve problems. Geometry: space and measures: recognise geometric shapes, situate oneself in space, evaluate sizes. *Grade 6:* • Arithmetic (numbers and operations): • whole numbers, factions and decimals • the 4 operations • sentences and numerical models (numerical sentences, operation signs, sequences of operations). Measurement: Measurement units and properties learnt in primary school (perimeter, calculations of surfaces, etc.). Spatial geometry: 2 or 3-dimensional geometric figures learnt in primary school.

<table>
<tr><th></th><th>Assessment</th><th>Mathematics definition</th></tr>
<tr><td><b>Large-scale international surveys</b> (cont.)</td><td><b>LLECE</b></td><td>Definition of mathematical literacy: “a permanent process throughout existence that includes such knowledge, technical skills, abilities, principles, values and attitudes necessary to include in the mathematics school curriculum so that Latin American students learn to develop their potential, face situations, make decisions using the available information, solve problems, defend and argue their point of view amongst many other key aspects that enable them to integrate into society as full citizens who are critical and responsible” Page 14 (12) of maths results report (SERCE, 2009).
<table>
<tr><td>Domain content</td><td>Process</td></tr>
<tr><td>Numbers, geometry, measurement, statistics, change.</td><td>Recognition of objects and elements
Solving simple problems
Solving complex problems</td></tr>
</table></td></tr>
<tr><td><b>School-based surveys</b></td><td><b>EGRA/ EGMA</b></td><td>The Core EGMA measures foundational mathematical skills.
Mathematical subdomains in the core EGMA:
• number identification
• number discrimination (which numeral represents a numerical value greater than another)
• number pattern identification (a precursor to algebra)
• addition and subtraction (including word problems).</td></tr>
<tr><td><b>Household-based surveys</b></td><td><b>PIAAC</b></td><td>The ability to access, use, interpret and communicate mathematical information and ideas, in order to engage in and manage the mathematical demands of a range of situations in adult life.
Definition of numerate behaviour: “Numerate behaviour involves managing a situation or solving a problem in a real context, by responding to mathematical content/information/ideas represented in multiple ways” (OECD, 2013: 34).</td></tr>
<tr><td></td><td><b>LAMP</b></td><td>The LAMP defines numeracy skills as skills that enable individuals to perform short mathematical tasks that required computing; estimating; and understanding notions of shape, length, volume, currency and other measures.</td></tr>
<tr><td></td><td><b>ASER</b></td><td>ASER defines numeracy skills in the following ways:
Number recognition (1-digit numbers)
Number recognition (2-digit numbers)
Subtraction (2-digit by 2-digit with borrowing)
Division (3-digit by 1-digit with carry over)</td></tr>
<tr><td></td><td><b>Uwezo</b></td><td>Numeracy: number recognition, place value and operations shall be tested in the numeracy tests. The highest level of number operations in Kenya and Uganda shall be divisions, while in Tanzania multiplication.</td></tr>
</table>

Table C.3 Science frameworks from other assessments

<table>
<tr><th></th><th>Assessment</th><th>Science definition</th></tr>
<tr><td><b>Large-scale international surveys</b></td><td><b>TIMSS</b></td><td>The TIMSS science assessment is based on a comprehensive framework for each domain, developed collaboratively with the participating countries. At each grade the science frameworks are organised around two dimensions: a <i>content dimension</i> specifying the domains or subject matter to be assessed within science, and a <i>cognitive dimension</i> specifying the domains or thinking processes to be assessed. The content domains and the topic areas within the domains are described separately for the fourth and eighth grades, with each topic area elaborated with specific objectives (Martin et al., 2012: 11). TIMSS 2015-Science also assesses science practices.
There is a strong curricular focus. At Grade 4 the content areas are Life Science, Earth Science and Physical Science. At Grade 8 the content areas are Biology. Earth Science, Physics and Chemistry.</td></tr>
<tr><td></td><td><b>LLECE</b></td><td>Science: the basic objective of science education is to mould students – future citizens – to know how to fully participate in a world filled with scientific and technological advances, and so they can adopt responsible attitudes, make fundamental decisions and resolve daily problems with a view to respecting others, the environment and future generations that have to live in the environment. For this, questions need to be asked that orient the student towards science for life and for the citizen.
<table>
<tr><th>Domain content</th><th>Process</th></tr>
<tr><td>Living beings and health
Earth and environment
Matter and energy</td><td>Recognition of concepts
Application and interpretation of concepts
Problem-solving</td></tr>
</table></td></tr>
</table>

Table C.4 Item development in other assessments

	Assessment	Item development
Large-scale international surveys	**PIRLS** **TIMSS**	The TIMSS and 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 • 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).
	SACMEQ	Main responsibility and involvement of participating countries: the 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).
	PASEC	For PASEC 2014, items were generated at the PASEC centre with the support of specialists, and finalised in close association with the countries. Items receive final approval from the Scientific Committee. Items are calibrated thanks to a trial test in each country.
	LLECE	UNESCO formed expert groups for each domain, consisting of UNESCO specialists and consultants as well as some members invited from the country technical teams. Using the submitted items as a base, each expert group selected some of these items to include in the test as well as developing their own new items to ensure that the framework specifications were met. The first set of items was presented to the national co-ordinators at a meeting in Havana. At this meeting, it was decided that the "language" domain should be split in two – reading and writing. A second set of items was presented to national co-ordinators at a meeting in Managua 6 months later. Countries had 3 weeks to review the items and provide comment. For an item to be removed from the pool, at least 70% of countries needed to reject it. TERCE is based on a published curriculum analysis. Using this basis, specification tables were developed to form a blueprint for the item development phase. Item development was done in a participatory fashion, in principle, involving specialists from almost all countries.
	WEI	OECD led the questionnaire development, with support from UIS and international experts, and OECD incorporated the experience from other large-scale surveys/questionnaires. Numerous consultations took place with OECD, UIS, international experts and countries. Once a draft list of indicators was created, it was sent to national project managers, who rated indicators by priority and relevance to their national contexts. This was done in 2003 and taken into account over the course of several more meetings with stakeholders, and with the project steering committee until the questionnaire frameworks, with draft questionnaires finalised by November 2003. OECD decided to try to use items from other surveys where possible, in order to source items that had already been tested and validated in international contexts. Specifically, items were drawn from the IEA.

	Assessment	Item development
School-based surveys	**EGRA** **EGMA**	There is no one item development process – each implementing country develops new versions of the EGRA/EGMA subtasks for its specific implementation. RTI provides guidelines for subtask development in various documentation, but does not itself supervise or control the quality of the development.
Household-based surveys	**PIAAC**	The selection of items from IALS and ALL to serve as linking items in literacy and numeracy and the development of new items took place in parallel with the development of the frameworks. Final selection of items for the Field Test took place in March 2009 (Kirsch and Thorn, 2013: 11).
	LAMP	• Each item in the test poses a task for the individual to perform. • These tasks are developed taking into account the following criteria, which also happen to translate into the expected difficulty level of each item. Task classification for prose, document and numeracy: • Tasks are developed in relation to a specific context and content that is relevant to a particular situation. These include home and family issues; health and safety issues; community and citizenship; consumer economic situations; work-related situations; and leisure and recreation.
	ASER	The 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 (ASER Centre, 2013; R. Banerji, personal communication, 27 April 2014). In all cases the reading tools are developed by people who have spent considerable time in teaching and learning activities in reading with children (R. Banerji, personal communication, 27 April 2014).
	Uwezo	According to the Uwezo standards, “Test item development will be guided by the following principles” (Uwezo Uganda, 2010): • The Uwezo approach is to develop and use assessment tools that are effective, low cost, simple and easy to use to ensure that the sampled households are at ease with the assessment. • Three different panels will be constituted to develop the three areas of assessment, namely: English, local language and numeracy tests. • All items should be constructed to the level found in the Primary 2 curriculum and recommended text books for Primary 2 in Ugandan schools. • Every test item should address a specific competence. • Every item shall stand alone and not be dependent on an understanding of a previous item. • Items should not be lifted directly from textbooks, especially for literacy, in order to make the assessment fair. • Items should take into consideration concerns with environment, gender, culture and religious biases.

Table C.5 Scaling methodology in other assessments

	Assessment	Scaling methodology
Large-scale international surveys	**PIRLS/TIMSS**	TIMSS and PIRLS rely on item response theory (IRT) scaling to describe student achievement on the assessments and to provide accurate measures of trends. As each student responds to only a part of the assessment item pool, the TIMSS and PIRLS scaling approach uses multiple imputation – or "plausible values" – methodology to obtain proficiency scores in reading (for PIRLS) and in mathematics and science (for TIMSS) for all students.
	SACMEQ	During the SACMEQ II study, the Rasch scores on the final pupil reading and mathematics tests were transformed to have a mean of 500 and a standard deviation of 100 (for the pooled data with equal weight given to each country). During the SACMEQ III study, Rasch measurement procedures were employed to equate the SACMEQ II and SACMEQ III scores (Hungi, 2011: 3).
	PASEC	PASEC used classic test theory until 2012; since then it uses IRT analysis (Rasch measurement). This IRT analysis has been used for the Mali, Vietnam, Cambodia and PDR Lao studies (tests only) and is being used for the first international assessment (tests and contextual data). PASEC's scaling approach will use plausible values and multidimensional methodology to obtain proficiency scores in reading and in mathematics (main domain and subdomains) for all students.
	LLECE	LLECE reports assessment results using a single continuous scale obtained from applying the Rasch model or IRT approach for each subject. The Rasch model is used with complementary IRT models of 2 and 3 parameters. Subscales – reported as percentage correct.
Household-based surveys	**PIAAC**	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. That is, different groups of respondents answered different sets of items, making it inappropriate to use any scaling system based on the number of correct responses. Differences in total scores (or statistics based on them) among respondents who took different sets of items may be due to variations in difficulty in the adaptively administered test forms. Unless one makes very strong assumptions – for example, that the different test forms are perfectly parallel – the performance of the two groups assessed in a matrix sampling arrangement cannot be directly compared using total score statistics. To overcome the limitations of conventional scoring methods, and to increase the accuracy of the cognitive measurement, PIAAC used plausible values (which are multiple imputations) drawn from a posterior distribution, by combining the IRT scaling of the cognitive items with a latent regression model using information from the background questionnaire in a population model (Yamamoto, Khorramdel and Davier, 2013a: 1).
	STEP	Once the data had been cleaned and weighted, ETS undertook the IRT scaling of the reading literacy data to provide the estimation of item parameters and the proficiency distribution of the population. The latter was then used to calculate a posterior distribution together with the household questionnaire variables, using latent regression. From this distribution, plausible values (multiple imputations) were obtained to provide a more accurate and reliable proficiency estimation than the proficiency estimation of the IRT scaling alone. Similar to the approach used by PIAAC, STEP used the two-parameter logistic model for dichotomously scored responses (Pierre et al., 2014: 60).
	LAMP	• Traditional item statistics to identify initial problems in comparability, particularly printing and translation errors. • Factor analysis to check unidimensionality of scales • Within each country, IRT scaling is compared with across-country scaling • Differential item functioning used for questions across participating countries. Tasks not operating in the same way in all participating countries are excluded in summary statistics.

Table C.6 Cross-country comparability measures in other assessments

	Assessment	Cross-country comparability
Large-scale international surveys	**PIRLS**	PIRLS undertakes a study of item-by-country interaction.
	TIMSS	TIMSS undertakes a study of item-by-country interaction.
	PASEC	Up to 2014, international comparisons were not emphasised. A tentative comparison is provided in Chapter 6 of the national reports (before 2012) and in the synthesis of PASEC reports, using classical test theory. From 2014, PASEC will undertake a study of item-by-country interaction.
	WEI	The UIS produced tables containing univariate statistics for each variable, broken down by country. These tables were verified at the UIS for any exceptional results, which were addressed in close co-operation with the national programme manager. In a few cases, the exceptional results were due to ambiguous translations or concepts that were not well understood by teachers or school heads. In such cases, the variable was recoded as "not applicable". Any such occurrences were documented in the national deviations database.
School-based Surveys	**EGRA** **EGMA**	Assessments are not designed for cross-country comparability.
Household-based surveys	**PIAAC**	PIAAC undertakes a study of item-by-country interaction.
	Uwezo	With respect to test comparability across the three countries: The 2013 regional report states that tests are not identical because they are based on curriculum expectations of the respective countries, but that to aid comparability across countries, results only include those questions that are "equivalent" across the countries (Uwezo, 2014).
	LAMP	• First steps in LAMP implementation involves a discussion on how the operational definition of literacy relates to the conceptions in a particular country given its language(s) and cultural characteristics. • Traditional item statistics are used to identify initial problems in comparability, especially those that might occur due to printing or translation errors. • Factor analysis is used to check for the unidimensionality of proposed scales. • Within each country, IRT scaling is compared with across-country scaling in order to check that measures are comparable among countries or language groups. Differential item functioning techniques are used to see if any questions are operating differently (e.g. they are unusually hard or easy, compared to other items) across participating countries. Tasks that are not operating in the same way in all participating countries are not included in summary statistics. This is item-by-country interaction.

Table C.7 Measuring trends in other assessments

	Assessment	Measuring trends
Large-scale international surveys	**PIRLS**	Test design: six of the ten 40-minute blocks 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. These "trend" blocks provide a foundation for measuring trends in reading achievement. Four new blocks will be developed for use for the first time in the 2016 assessment (Martin, Mullis, and Foy, 2013a: 60).
	TIMSS	In most booklets two of the blocks contain trend items from 2011 and two contain items newly developed for TIMSS 2015 (Martin, Mullis, and Foy, 2013b: 89, 90).
	SACMEQ	The SACMEQ II tests for pupils and teachers included linked items selected from five earlier studies: the Zimbabwe Indicators of the Quality of Education Study (Ross, 1995), the SACMEQ I and SACMEQ II projects, the IEA's Third International Mathematics and Science Study (TIMSS) (Mullis et al., 2001), and the IEA's International Study of Reading Literacy (PIRLS) (Elley, 1992). These "overlaps", when combined with Rasch item analysis and test scoring techniques, made it possible to make valid comparisons among the following groups of respondents: pupils with teachers in the SACMEQ II project, pupils in the SACMEQ I project with pupils in the SACMEQ II project, and pupils in both SACMEQ projects with pupils in the IEA's TIMSS and IRL studies. See tables below for the test items that were used for calibrating test items.
	PASEC	In PASEC's previous national assessment, each country was tested with the same booklets. PASEC 2014 has only had one implementation; it will link with the next cycle scheduled for 2018.
	LLECE	PERCE and SERCE are not comparable, because SERCE introduced a series of modifications resulting from the experience and knowledge gained from implementing PERCE. Some of the changes are related to sampling, test design, target population and knowledge domains covered by the assessment. However, by aligning methodology, SERCE and TERCE are comparable studies. In fact, there will be two scales: a comparable scale (already published) and a TERCE scale, to be used as the baseline from now on.
Household-based surveys	**PIAAC**	To date PIAAC has only had one implementation.
	ASER	The ASER Centre takes care to ensure that assessment tools are comparable in the same language across different years; one year's reading tool is comparable with previous years' tools in terms of "word count, sentence count, type of word and conjoint letters in words" (ASER Centre, 2014: 22).
	Uwezo	Uwezo tries to ensure that the level of difficulty and comparability across the years is retained. In each year one new aspect will be considered, while keeping the core the same, to enable cross-country comparability across years" (Uwezo, 2011: 17).

Table C.8 Use of proficiency levels in other assessments

	Assessment	Proficiency levels
Large-scale international surveys	PIRLS TIMSS	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 expert international committees (the Science and Mathematics Item Review Committee for TIMSS and the Reading Development Group for PIRLS) to conduct a scale anchoring analysis in order 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. Thus, the scale anchoring procedure yields a content-referenced interpretation of the achievement results that can be considered in light of the TIMSS and PIRLS frameworks for assessing mathematics, science, and reading (Mullis, 2012).
	SACMEQ	SACMEQ has created proficiency levels for reading and mathematics. Four main steps were used in the SACMEQ II project to define levels of competence: 1) Rasch item response theory was used to establish the difficulty value for each test item; 2) national research co-ordinators (NRCs) subjected each test item to an intensive "skills audit" in order to identify the required problem-solving mechanisms for each item "through a Grade 6 pupil's eyes"; 3) the items were clustered into eight groups or "levels" that had similar difficulties and that required similar skills; 4) the NRCs wrote descriptive accounts of the competencies associated with each cluster of test items, using terminology that was familiar to ordinary classroom teachers (Ross et al., 2004: 18).
	PASEC	Results were not reported according to proficiency levels until 2012; proficiency levels were constructed for the last four national assessments. PASEC 2014 will define proficiency levels for each grade and domain. PASEC will also define levels, after examining the items and carrying out statistical processes.
	LLECE	A group of items was selected with the lowest anchor point and the lowest skill level, as described in the framework. All other items were given an anchor point that corresponded to the point on the scale where students had a probability equal to or greater than 0.6 of responding correctly. To establish the second anchor point, items were selected for which the probability of responding correctly was equal to or greater than 0.6, and in the previous anchor point, the probability of responding correctly was less than 0.5 and the difference of the probabilities was more than 0.2. For a panel of experts to establish items for each anchor point and to describe these points within the framework, the skill level of each student was obtained. A student was assigned a high skill level when the probability of correctly responding to items in this point was equal to or greater than 0.6. For TERCE, these levels were redefined using the bookmark methodology.
School-based surveys	EGRA EGMA	No proficiency levels. Results are generally reported separately for each task.
Household-based surveys	PIAAC	The proficiency scale in each of the domains assessed can be described in relation to the items located at different points on a scale according to difficulty. To help interpret the results, the reporting scales have been divided into "proficiency levels" defined by particular score-point ranges. Six proficiency levels are defined for literacy and numeracy (Levels 1-5, plus below Level 1) and four for problem solving in technology-rich environments (Levels 1-3, plus below Level 1).
	STEP	Six literacy scale proficiency levels are defined, provided on the same five-level scale as PIAAC (Pierre et al., 2014: 44, 83).
	LAMP	Levels are created by ETS; LAMP item difficulty levels are defined by countries developing the items, and later on verified and/or adjusted by ETS.
	ASER	No proficiency levels. Children's proficiency is understood in terms of the highest level task they completed successfully.
	Uwezo	No proficiency levels.

Table C.9 Translation procedures in other assessments

	Assessment	Translation procedure
Large-scale international surveys	**PIRLS** **TIMSS**	The TIMSS and PIRLS International Study Center prepares an international version of all the assessment instruments for TIMSS and PIRLS in English. The test and questionnaire instruments are then translated by participating countries into their languages of instruction; the goal is to create high quality translations that are appropriately adapted for the national context, and at the same time are internationally comparable.
	SACMEQ	SACMEQ suggests that independent translations should be made by at least two different expert translators familiar with age-appropriate linguistic demands. In cases of disagreement, consensus should be achieved either by direct negotiation between the two translators or by a third expert making the final choice (SACMEQ, 2007: 29).
	PASEC	Tests are developed in French. Procedures follow double independent group translation plus external reconciliation. The translation process is outsourced to specialist consultants and overseen by the PASEC technical team with control of the Scientific Committee.
	LLECE	Spanish is the language most commonly used for LLECE materials. A back-translation process was used to create the Portuguese version: the Spanish source was translated into Portuguese, which was then translated back into Spanish. The source Spanish version and back-translated version were compared and validated before the test.
	WEI	The WEI-SPS translation processes were based on materials and procedures used in OECD's PISA assessment. WEI-SPS outlines procedures for translating into the language of administration for the national context, and for the process of making adaptations. The UIS instructed countries to submit translations and adaptations to them for approval before the survey was administered (i.e. before the questionnaires were printed for administration). As Spanish is a language commonly used for assessments, a "standard" version was produced by UNESCO Santiago, then adapted for individual countries. Assessments were administered in Tamil in two countries, but no standard version developed.
School-based surveys	**EGRA** **EGMA**	Translation procedures depend to a degree on the way that the assessment is implemented. Since EGRA is in the public domain it can be borrowed and adapted at will. When RTI implements it, however, it applies a great deal of quality control during a one-week adaptation workshop.
Household-based surveys	**PIAAC**	Participating countries were responsible for translating the assessment instruments and the background questionnaire. Any national adaptations of either the instruments or the questionnaire was subject to strict guidelines, and to review and approval by the international consortium. The recommended translation procedure was for a double translation from the English source version by two independent translators, followed by reconciliation by a third translator. All national versions of the instruments were subject to a full verification before the field test.
	STEP	Two independent translators separately translated the household questionnaire and reading literacy assessment, before a third translator reconciled, and documented, any discrepancies. The STEP team and ETS checked the translations and worked closely with the survey firms to finalise the instruments. In English-speaking countries, the instruments were adapted to reflect local idioms (Pierre et al., 2014: 58).

	Assessment	Translation procedure
Household-based surveys (cont.)	**LAMP**	The UIS provides a set of instruments in English, French or Spanish (eventually this will be available in other languages, if the countries that originally produced those versions kindly authorise their use), which might need to be translated and adapted to the particular characteristics of each country and its language usage. After the cognitive instruments are adapted, a verification process is required to ensure that what is being measured reflects the original design. Typically, verification is a two-stage process that allows for a detailed discussion of changes and ends with an agreement on the final version of every instrument. The adaptation of the background questionnaire is of utmost importance as it will provide key elements for analysis and, therefore, for accomplishing the goals set at the national level (UIS, 2009: 37).
	ASER	Translation procedures are not really applicable – reading tools are developed separately in each language and the maths tool is so simple that is does not really require translation as such.
	Uwezo	No “translation” as such – tests are developed separately in each language.

Table C.10 Field trial processes in other assessments

	Assessment	Field trial process
Large-scale international surveys	**PIRLS** **TIMSS**	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. 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 a larger sample of 180 schools is selected and a systematic sample of 30 schools is selected from the 180 schools (Mullis et al., 2012: 18).
	SACMEQ	SACMEQ II: the data from the trial-testing phase were subjected to both Rasch and classical item analyses in order to detect items that did not “fit” the relevant scales, or that were “behaving differently” across subgroups of respondents defined by gender and country. The poor quality test items were rejected – keeping in mind the need to prepare a “balanced” test across skill levels and domains (Ross et al., 2004: 52). There is a similar description for SACMEQ III in Hungi (2011: 3).
	PASEC	Test instruments are systematically trialled on a sample of 20 schools in each country. Trial testing confirms the quality of the tests and questionnaires and the relevance of the procedures.
	LLECE	In SERCE, expert groups met to discuss the findings, to select the best performing items that match the framework, and to design the clusters and booklets for the main survey. Expert groups also adjusted coding guides for open-ended maths and science items, according to the field trial results and the reliability between coders and supervisors. The language expert group also had to assess the level of consistency in the coding for writing, considering the difficulties in the range of rating criteria and the limited experience in the region of such large-scale writing evaluations. There was at least one domain expert, one assessment expert and one psychometrics expert at each meeting. In TERCE, item behaviour in the pilot study was analysed based on an analytical plan by the implementation partner.
	WEI	Field trial analysis looked at the feasibility and cross-cultural validity of questions across the countries. The type of questions that raised cross-cultural validity concerns were mostly in the Opportunity to Learn questionnaire.

	Assessment	Field trial process
School-based surveys	**EGRA** **EGMA**	Implementing countries are encouraged to field trial the cognitive instruments they develop for their specific implementations, but there is not much information about this in the specific implementation reports. From the guidelines for planning and implementing EGRA (see RTI International and International Rescue Committee, 2011), the pilot test will help to ensure the tool is accurately measuring what children know in the specific context and language(s) of assessment. It will also allow verification of the validity and reliability of the instrument(s) and give the EGRA team an opportunity to address technical issues before the cost-intensive data collection phase. The main issue is that in the lower grades, the orthographic transparency of the language matters a great deal in how quickly children develop skills. Thus, the assessments are not translated but adapted to reflect orthography. In later grades, quality of instruction trumps (to a very large degree) orthographic transparency. This is also a reason why technicians working on EGRA tend to discourage inter-language comparisons of summary measures such as oral reading fluency.
Household-based surveys	**PIAAC**	The field test addressed three main areas: 1) operational (in terms of feasibility); 2) instrumentation; and 3) scaling and psychometric characteristics. It was important for the successful implementation of the main study, especially given that the PIAAC results had to be linked to previous assessments, while also being implemented in both PBA and CBA modes (including an adaptive aspect) (Yamamoto, Khorramdel and Davier, 2013b: 1).
	STEP	Once a final proposal was complete, pilots were conducted in several countries by the World Bank to identify administration problems and suggest item wording calibration. Feedback from these pilots led to several important adjustments, particularly in the rewording of items that had proven to be difficult for participants to understand, and the general reframing of all items as questions instead of statements.
	LAMP	The field test involves administering the entire battery of survey instruments to a carefully selected sample (not probability/random) of roughly 500 adults in each test language.
	ASER	Qualitative and quantitative data are collected during the pilot, and refinements may be made to the instructions for administering the tools. Quantitative data are presented to the district in which a pilot is conducted as a “block report card” (R. Banerji, personal communication, 27 April 2014).
	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.

References

ASER Centre (2014), *Annual Status of Education Report (Rural) 2013*, ASER Centre, New Delhi.

ASER Centre (2013), *Guidelines for Development of ASER Tools*, ASER Centre, New Delhi.

Elley, W. (1992), *How in the World Do Students Read? IEA Study of Reading Literacy*, IEA, Amsterdam.

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

Kirsch, I. and W. Thorn (2013), "Foreword: The Programme for International Assessment of Adult Competencies - an overview", in Technical report of the Survey of Adult Skills (PIAAC), OECD, Paris.

Martin, M. O. et al. (2012), *TIMSS 2011 International Results in Science*, TIMSS and PIRLS International Study Center, Chestnut Hill, MA.

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. (2012), "Using scale anchoring to interpret the TIMSS and PIRLS 2011 achievement scales", 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. et al. (eds.) (2013), *TIMSS 2011 Encyclopedia: Education Policy and Curriculum in Mathematics and Science*, Volume 1: A–K, TIMSS and PIRLS International Study Center, Chestnut Hill, MA.

Mullis, I.V.S. et al. (2001), *The Mathematics Benchmarking Report*, TIMSS and PIRLS International Study Center and IEA, Chestnut Hill, MA, and Amsterdam.

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.

Mullis, I.V.S., Martin, M.O. and Sainsbury, M. (2013), "PIRLS 2016 reading framework", in I.V.S. Mullis and M. O. Martin (eds.), *PIRLS 2016 Assessment Framework*, TIMSS and PIRLS International Study Center and IEA, Chestnut Hill, MA and Amsterdam, pp. 13-31.

OECD (2013), *Literacy, Numeracy and Problem Solving in Technology-Rich Environments: Framework for the OECD Survey of Adult Skills*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264128859-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.

Ross, K.N. (ed.) (1995), "From educational research to educational policy: An example from Zimbabwe", *International Journal of Educational Research*, 23(4), Sage Publications, Thousand Oaks, CA, pp. 301-401.

RTI International and International Rescue Committee (2011), *Guidance Notes for Planning and Implementing EGRA*, RTI International, North Carolina.

SACMEQ (2007), *SACMEQ III: Manual for National Research Co-ordinators: Main Study*, SACMEQ, Paris.

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

UIS (2009), The Next Generation of Literacy Statistics: Implementing the Literacy Assessment and Monitoring Programme (LAMP), UNESCO Institute for Statistics, Montreal.

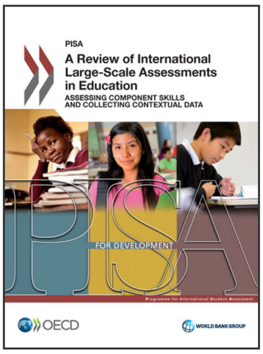
Uwezo (2014), *Are Our Children Learning? Literacy and Numeracy across East Africa 2013*, Uwezo and Hivos/Twaweza, Nairobi.

Uwezo (2011), "Improving learning outcomes in East Africa 2009-2013: Strategy update", www.uwezo.net/strategies.

Uwezo Uganda (2010), *Test Development Framework 2010-2014*, Uwezo Uganda, Kampala.

Yamamoto, K., L. Khorramdel and M.v. Davier (2013a), "Chapter 17: Scaling PIAAC cognitive data" in Technical report of the Survey of Adult Skills (PIAAC), pre-publication copy, OECD, Paris.

Yamamoto, K., L. Khorramdel and M.v. Davier (2013b), "Chapter 19: Proficiency scale construction" in *Technical report of the Survey of Adult Skills (PIAAC)*, OECD, Paris.



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), “Overview tables related to cognitive data collection instruments”, 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/10.1787/9789264248373-12-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.