Chapter 1

Overview: Lessons from international large-scale assessments in education

The purpose of this chapter is to provide an overview of the main findings of the review of international large-scale learning assessments. In particular, the chapter summarises the practices of these assessments that are recognised as being effective, especially in the context of developing countries and draws lessons from them for the benefit of the PISA for Development (PISA-D) initiative. These findings and lessons are identified and presented in three main areas: i) *component skills and cognitive assessments;* ii) *contextual data collection instruments; and* iii) *implementation procedures, methods and approaches to include out-of-school children, and the use of data.*

This report is the product of a review of a number of large-scale international learning assessments, including school-based surveys and household-based surveys.

The review covered all aspects of the surveys' approaches for assessing and reporting on component skills, from assessment frameworks and item development, through test design and mode of delivery, to analysis and reporting proficiency. Translation, field trialling and final item selection were also covered.

The review also looked at all aspects of the surveys' approaches to collecting and reporting contextual information, including the development of contextual data collection instruments, their translation and adaptation, the main factors and variables used, question formats, scaling, relevant constructs and cross-country comparability.

The review also considered how the surveys were implemented, methods and approaches for including out-of-school children, and the analysis, reporting and use of data.

The review has endeavoured to identify the approaches in these surveys that may be instructive for PISA for Development (PISA-D). The following subsections present the main findings and options for each of the three areas of the review.

Component skills and cognitive assessments

Assessment frameworks

The major international assessments produce clear frameworks to describe the philosophy, content, test design and response styles of their tests. These frameworks not only guide the creation of items (questions or tasks in a test paper) for the test, but also act as a way of communicating information about the assessment to the broader community.

- The majority of the international school-based assessments described in this report have a strong curricular focus, as opposed to the Programme for International Student Assessment (PISA) approach of preparedness for the future. This may also be a reflection of the target group – in PISA it is 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, giving the opportunity to implement remedial interventions where appropriate. It is possible that PISA-D countries might find a curricular approach more suitable to their needs.
- There may be a higher proportion of students not in school at age 15 in the PISA-D countries than in OECD countries. PISA-D could opt to do an assessment at an earlier age, not only to increase the coverage of students, but also to 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.
- A collaborative approach to the development of the assessment frameworks is a characteristic of many of the assessments. If PISA-D were to adopt such an approach, it may lead to a more relevant assessment and encourage better engagement by countries.

Item development

Across the major international assessments there is a well-established procedure for creating new items for a major assessment. This generally follows the steps of item generation, item panelling, cognitive trialling, field trialling and main study selection. Items are reviewed throughout the process by participating countries, but especially before and after the field trial, as preparations are made to choose which items will be included in the main study.

While there will be no new item development in PISA-D, we recommend adopting the process described for any future process to create items. While items could be imported from other assessments, it is important to realise that their characteristics can only by 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.

- The established process in PISA and many assessments involves the steps of item generation, item panelling, cognitive trialling, field trial and main study selection. PISA-D should follow this process when creating new items.
- While items from other assessments were not made available for this review, such as Progress in International Reading Literacy Study (PIRLS), Trends in International Mathematics and Science Study (TIMSS) and the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ), items' characteristics can only by assessed by testing them with the specific target populations for which they are intended. An item that is suitable in one context is not necessarily going to be suitable in another.
- A collaborative approach to item development is a characteristic of many of the assessments. If PISA-D were to adopt such an approach, it may lead to a greater commitment on the part of the countries in the assessments.

Test design

The assessment frameworks developed by the assessments reviewed tend to cover a very wide range of material: more than can be included in one test per student. To cover this range, it has been necessary to incorporate a test design in which each student is assessed on only part of that framework. This has led to a "rotated" booklet design, with common items across the booklets allowing scaling to take place to generate an overall view of student capacity. At this point in time, the assessments are still delivered mostly by paper and pencil, although a move to computer-delivered tests will take place in the next few years in many assessments.

In developing countries assessment frameworks are also expected to cover a wide range of material. This would suggest that PISA-D should also use a rotated booklet design, allowing different students to be assessed on different parts of the framework. While paper-and-pencil tests are more widely accepted and easily administered, the advantages of delivering tests by electronic tablets are also worth considering. Experience has shown that tablets can be used in populations totally unfamiliar with this technology. Delivery via tablet has the advantages of increasing student interest and eliminating expensive data-entry procedures. However, the disadvantages are that there may be extra set-up costs and that strict uniformity across countries is required – which can sometimes be difficult given that countries may be at different stages of technological development.

One of the main attractions of PISA-D is its immediate link to regular PISA. Any difference in the mode of delivery will make this link much more difficult or impossible to establish.

- 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 that includes some common items, to allow linking between the booklets.
- Regard should be given to the mode of delivery of the test. Many of the tests examined here are paper-and-pencil tests. However, the Australian Council for Educational Research (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.

Psychometric analyses, scaling, calibration and equating methods

Major international assessments have adopted "item response theory" scaling as the means of analysing student responses to an assessment. This theory, built on the Rasch model,[1] allows a clear picture of student capacity to be drawn, see the details provided in section 3.4. In developing countries, item response theory will deliver an accurate picture of student capacity across a wide range of item difficulties.

It is recommended that the parameters used in scaling standard PISA should be adopted for PISA-D. This will allow countries to compare their own results with PISA more easily.

- Item response theory scaling is the preferred method of analysing student data. This type of scaling is based on continuous interaction between the student's capacity and an item's difficulty. This gives a clear picture of the students' capacity.
- Item response theory scaling allows one test to be linked to another 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 uses a one-parameter model based on the item difficulty. The International Association for the Evaluation of Educational Achievement (IEA) in PIRLS and TIMSS employs a three-parameter model. Use of a one-parameter model in PISA-D would facilitate comparisons to PISA.

Cross-country comparability

To be able to establish the student capacity of one country – and then for that country to be able to compare results with other countries – is a central aim of the large-scale assessments. This allows countries to share information and techniques to improve learning for their students. A "differential item functioning process" is usually undertaken at the field trial stage to identify any item-by-country interactions. This will identify any items that work to a particular country's advantage or disadvantage. How confident a country is to get involved in the process may depend on how fairly they feel they are being treated. When developing countries get involved in internationally comparable assessments they must be confident that their students are being compared in an unbiased manner to all the other countries in the assessment.

- We recommend undertaking a differential item functioning process in PISA-D to identify any item-by-country interactions, in a similar way to the process used in PISA. This will identify any items that work to a particular country's advantage or disadvantage. How confident a country is to become involved in the process depends on the perception that they are being treated fairly.

Trends

The different assessments use a variety of approaches to measure change over time. In PIRLS, a number of blocks of items are used from one assessment to another. PISA keeps most items secure from one survey to the next so that they can be re-used.

The PISA-D countries will be able to access the normal PISA measurement of trends if the surveys are administered regularly.

- One of the biggest attractions to countries wanting to participate is being able to monitor changes over time. PISA-D will need to include a selection of the same items from one survey administration to the next. This has implications for maintaining security for those items, which if they enter the public domain cannot be used confidently for this purpose.

Proficiency levels

Student results reported as a single number or grade do little to describe the capacity of the student population. Closely examining the items that a student can do will provide a much more accurate and useful measure of the individual's capacity. Nearly all the global and regional assessments undertake the process of dividing the students into a number of different levels of proficiency so that participating countries will obtain a better picture of their own students' strengths and weaknesses. The profile of percentages of students at the different levels gives valuable direction to the countries in deciding between possible intervention strategies. Arriving at described proficiency levels involves examining the items grouped according to their difficulty and then describing the tasks that are needed to complete these items.

For developing countries, an appropriately targeted test will give them much more information than a test that is poorly targeted and contains too many difficult items for their students. This can lead to a situation where a substantial percentage of their students are below the lowest described proficiency level. If the test is appropriately targeted then the countries will receive valuable information about their students' capabilities and where they need to focus resources to bring about improvements.

- It is highly desirable to define students' proficiency levels as well as assigning them a numerical value for their results. Described proficiency levels are based on the items' level of difficulty and the tasks associated with the items. Proficiency levels highlight students' strengths and weaknesses.

Translation, adaptation and verification of cognitive instruments

There are a variety of approaches to translating test material across the different assessments. Approaches include single translation, back translation and double translation. In back translation the material is translated from one language to another, then translated back to the original language, and the two versions compared and validated. The double translation method means that two source versions of the test in one (or, preferably, two) languages will first be translated within the country separately, then those versions reconciled, and the resulting version verified by an independent international expert language organisation.

For all countries, including developing countries, the biggest challenge is often to find people with sufficiently high skills in both the language of the source version of the test and the language the test is administered in.

- To maintain the highest standards for translation it is recommended that the PISA-D project adopt a two-source-version approach. This involves independent translations of each source version and verification of that process by an expert language organisation. This process will also give better comparability with results from existing PISA surveys.

Field trial and item selection

Most of the international assessments reviewed in this report employ a field trial, which is done after item development has taken place but before the main study. The field trial item analysis data gives valuable information about the quality of the translations used.

For developing countries without previous experience in international assessments, the field trial provides essential practice, not only for assessing the logistical needs of the assessment, but also in how to manage the review and translation of the cognitive and contextual instruments.

Each of the countries participating in PISA-D have had international experience in either the Conference of the Ministers of Education of French speaking countries (CONFEMEN) Programme for the Analysis of Education Systems (PASEC), Latin American Laboratory for Assessment of the Quality of Education (LLECE) or SACMEQ. This is excellent experience for those countries, provided that the personnel involved are still available.

- A field trial should take place to test the suitability of the items for the target sample and to see if the participating country has the capacity to implement the

assessment. A large number of items are usually discarded following the field trial.

- It is vital that the countries participating in PISA-D gain as much experience as possible in the procedures associated with international testing, and this is best done with a field trial.

Contextual data collection instruments

Types of contextual data collection instruments and mode of delivery

With regard to the questionnaire type, Willms and Tramonte (2014: 20) underline the importance of discerning the best informant or respondent for measuring the relevant constructs (the conceptual element that is being measured). All surveys reviewed collect contextual data. International large-scale surveys use questionnaires for students, teachers and principals. In addition, some surveys collect data from parents.

Most of the questionnaires and interviews used for contextual data collection in the surveys reviewed are administered in paper-and-pencil mode. Electronic means could be considered, as discussed in the section above. Such an option would allow "spoken" and "visual" language components to be incorporated for struggling readers.

Regarding a teacher questionnaire, it is not clear how the information collected at the classroom level will relate to student achievement in PISA-D. It is worth noting that performance in PISA is seen as an accumulation of the student's educational experience and that PISA does not sample from intact or whole student classes. For a parent questionnaire, an interview approach could be considered in PISA-D.

- PISA-D should give careful consideration to the types of questionnaires implemented, in order to collect the most essential contextual information in the most efficient way. It will be important to calculate a cost/value ratio for various contextual data collection instruments.

- PISA-D should consider implementing a parent questionnaire as a core instrument in its assessment. Implementing a parent questionnaire will require significant effort, for example, through an interview approach or other methods to secure response rates. Student contextual questionnaires may be able to collect some of the desired data. Comparisons between student and parent questionnaire responses in PISA have shown that students are a reliable source of data about family-related topics such as language use, parental occupation and education.

- Similarly, we recommend considering the benefits of a teacher questionnaire, compared to collecting the aggregated school-level data through the principal questionnaire. At present, it is not clear how factors captured in a teacher questionnaire will be analysed. It may not be appropriate to relate information collected at the classroom level to student achievement, especially because performance in PISA is seen as an accumulation of the student's educational experience, and the sample does not use intact classes.

- The benefits of principal and teacher contextual questionnaires should also be weighed against the possibility of using system-level, administrative or agency-collected data. If some contextual data can be garnered at the system level, it will reduce contextual data collection through teachers and principals

(and most likely through students). For example, questions about instruction time could be administered at system level.

Development of contextual data collection instruments

Most large-scale international surveys follow a very similar questionnaire development process as PISA. The process defines policy priorities and/or research questions, and constructs a context framework. The context framework provides the theoretical underpinning of the context variables and factors implemented in the survey, as well as how they relate to achievement. This process is used in PISA, PIRLS and TIMSS, World Education Indicators' Survey of Primary Schools (WEI-SPS) and the Programme for the International Assessment of Adult Competencies (PIAAC). Alternatively, some surveys (such as SAQMEC and LLECE) construct analytical models to describe the relationship between the surveyed contextual factors and achievement.

In constructing context indices, items should be in a format that allows self-reported measures to be adjusted, to further explore and potentially increase cross-country comparability. Also, PISA-D should analyse the extent of different patterns of response styles in developing countries.

It is of utmost importance for PISA-D to field trial contextual questionnaires in all participating countries, in order to gain data for item statistics, validate new questionnaire items and constructs and test contextual data collection procedures.

Data analyses after field trial and the main study need to capture the validity of questionnaire items across countries and ensure that items work in the same way in all countries. This is relevant for cognitive as well as contextual items.

- It is crucial that PISA-D participating countries be involved in all phases of the contextual questionnaire development process, including framework development. It is also crucial that the countries be involved on different levels, including school, teacher and operational levels. Countries should also be involved in education policy, such as participation on the PISA Governing Board (PGB), and especially with respect to identifying and addressing the main education policy. Country involvement in education research could include: participation in the Questionnaire Expert Group; identifying and addressing questions for developing country contexts as part of the framework development; and development and review of specific questionnaire items.
- In regards to capacity building, PISA-D participating countries should be actively involved in item development activities to enable them to create and implement items of specific national interest.
- It is of utmost importance for all PISA-D countries to participate in: field trialling of contextual questionnaires in order to gain data for item statistics; validation of new questionnaire items and constructs; and testing contextual data collection procedures.

Translating, adapting and verifying contextual data collection instruments

In relation to translation, adaptation and verification, country involvement in all stages of reviewing the context framework and questionnaires is essential for checking the "face-validity" (or face value) and cultural appropriateness of the content, as well as for identifying possible issues with translation.

Standardised procedures are provided in most of the international large-scale surveys, as well as the household-based surveys that aim for international comparison. Most surveys acknowledge the importance of adapting questionnaires to match national contexts, to provide key elements for analysis and, therefore, to accomplish the goals set at the national level.

- In regard to languages, it is important to consider which languages are the most appropriate ones for the different groups of respondents. Questionnaires are preferably translated into the languages in which students, teachers, principals and parents are expected to be proficient. This may not always match with the defined "language of assessment" (for example, the languages most often spoken at home for the parent questionnaire).
- The issues around language of instruction are very well documented for prePIRLS in South Africa. Results show that in most languages used in prePIRLS, achievement was significantly higher when children wrote in their home language as opposed to the language of instruction (Howie et al., 2012: 31). We suggest considering language issues during field trial analyses, to rule out discrimination based on the language of assessment.
- Translation, adaptation and verification procedures are already highly elaborate for PISA and comply with very high standards. PISA-D needs to ensure that PISA-D countries can satisfy these standards. A capacity needs analysis might reveal what is necessary in this regard. It is also necessary to enable national centres to: perform adequate adaptations and to document accurately; to understand and interpret field trial analyses; and to create national options. PISA-D needs to build capacity around methodology of contextual data collection instruments. This will enable participating countries to create national questionnaire options.

Main factors and variables

Most of the international surveys articulated a theoretical underpinning of the context factors collected and understood the relationship between these factors and achievement. This combines educational research questions based on a model of learning and policy questions. The surveys offer a wide range of factors and variables that are relevant, including early learning opportunities, language at home and at school, socio-economic measures, quality of instruction, learning time, school resources, family and community support, and health and wellbeing.

In developing countries this range of variables would provide valuable information for policymakers and practitioners.

The PISA-D questionnaires should contain similar content to the standard PISA questionnaires to allow a genuine comparison. However, some modifications will be needed according to the prevailing conditions in each of the participating countries.

- Regarding *early learning opportunities,* the PIRLS and TIMSS Learning to Read Survey (for parents), the LLECE questions about early reading and how often someone at home reads aloud to the child, and the questions about out-of-school status from the Annual Status of Education Report (ASER) and Uwezo may all be of interest to PISA-D.

- Regarding *language at home and school*, a number of assessments contain items that may be relevant for PISA-D. The PISA 2012 educational career questionnaire contains language-related questions. PIRLS and TIMSS contain questions about the frequency of speaking the language of the test at home and the language spoken by the student before school enrolment. PIRLS and TIMSS also ask if the books at home ("books at home" as used as an indicator for socio-economic status) are mainly in the test language. Questions from Skills Toward Employment and Productivity (STEP) and the Literacy Assessment and Monitoring Programme (LAMP) may also be useful to help PISA-D gain a full picture of language use, because these questions differentiate between language that is spoken and language that is read and written. For a teacher questionnaire, PISA-D should consider questions about the language spoken by the teacher, from PASEC. Additionally, teachers could be asked to estimate how many students have difficulties understanding the spoken language of the test, as done in PIRLS and TIMSS. Questions to address language of instruction should also be included at the school level, such as those from PISA 2009, PIRLS and TIMSS. Questions from LLECE about language of instruction (for partial or all instruction) and indigenous language services and resources may also be of interest. It may be useful to ask about the official time used for teaching the language of instruction, as for example in WEI-SPS, as well as about the languages in which textbooks are provided, as in Uwezo.
- *Socio-economic status* indicators relevant to children living in poverty cover areas of parental education, home facilities and possessions, educational materials and resources, and main source of income. Relevant questions are included in SACMEQ, PASEC, LLECE, Early Grade Reading Assessment (EGRA) and Early Grade Mathematics Assessment (EGMA), ASER and Uwezo. STEP employs a particular asset index (Pierre et al., 2014: 15) that may be useful for PISA-D. Together STEP, LAMP, ASER and Uwezo provide a pool of items about household characteristics that PISA-D can draw from. This will allow PISA-D to identify and use relevant variables for extending the PISA index of economic, social and cultural status and for developing poverty-related measures. Employment information as captured in LAMP, STEP or PIAAC may be of interest for the PISA parent questionnaire, in regard to extending existing measures of the parents' employment status. For example, STEP module 4 obtains basic employment information, such as the labour force status (employed, unemployed or inactive; including self-employed – with and without pay; underemployed or holding low-productivity jobs).
- Regarding *quality of instruction*, the reviewed surveys cover a range of topics of particular interest to PISA-D. These concern general aspects of quality of instruction, including pedagogical practices, teacher limitations, assessing and monitoring academic progress, classroom organisation and management, homework, evaluation and professional development of teachers. The surveys also cover domain-related aspects of quality of instruction, including strategies for reading instruction, and training for specific subject teaching.
- Regarding *learning time*, PASEC and LLECE student questionnaires ask about working outside of school. Topics include the type of work, such as whether work is in the household, in agriculture or in retail; whether work occurs in or outside the home; and if the students are paid for working. Topics also cover the amount

of work, measured in days per week and hours per day, and whether working hinders learning or school attendance, or causes fatigue during instruction.

- Regarding *school resources*, relevant factors relate to basic services, didactic facilities and didactic resources. Basic services include the conditions of the school building and school infrastructure; the availability of electricity, toilets and water sources; and the provision of school meals, transportation and medical and clothing programmes. Didactic facilities include the teachers' workspace, classroom resources and infrastructure, such as tables and chairs, blackboard, chalk, pen, notebook and adequate lighting in classroom. Didactic resources include teaching resources such as: television, photocopier or computer; availability and quality of educational material; availability of a library; and student learning materials such as textbooks, pencils and other writing materials. Relevant questions were found in SACMEQ, PASEC, EGRA and EGMA, ASER and Uwezo. Other relevant topics are school safety, teacher satisfaction (including factors such as travel distance, if teacher housing is provided and level of salary), staff stability, and issues regarding funding and grants.
- Regarding *family and community support*, information about parental involvement is captured on all levels: student, parent, teacher and school level. Factors about parents' involvement that may be relevant for PISA-D are found in PIRLS and TIMSS, SACMEQ, LLECE, WEI-SPS, PASEC and EGRA and EGMA. Information about community support is mainly captured through the principal. Useful factors and variables can be found in SACMEQ, WEI-SPS, PIRLS and TIMSS and PASEC. Specific measures of cultural and social capital, which are of relevance for PISA-D, are included in PIAAC and LAMP.
- Factors measuring *health and wellbeing* that may be of particular interest for PISA-D are included in several surveys. Uwezo asks about health and other services, such as the presence of a nurse, the main health issue keeping children out of school, provision of sanitary items for girls, availability of drinking water and the presence of feeding services. PASEC asks about wellbeing at school. LAMP asks about personal wellbeing and health-related literacy.

Technical aspects of contextual data collection instruments

A number of different question formats were used across all contextual data collection instruments in the surveys reviewed. These included:

- dichotomous questions: mostly yes/no; particularly in ASER, Uwezo
- nominal variables
- Likert scales: three, four, five and ten-point scales
- open-ended questions: also largely used in ASER and Uwezo, but not very cost or time-effective for data capture, analyses and aggregation, and information grouping
- rankings: for example, the Uwezo household survey sheet includes a ranking item about major issues facing the community; the respondent is asked to choose three of nine options and rank the three chosen ones in order of importance.

Including a wide range of appropriate formats will enhance the quality of information derived from the questionnaires.

In regard to scaling and computing relevant context constructs, there are generally two kinds of indices created from context questionnaires. These are simple indices, created through transforming and/or recoding, and scale indices, which are constructed by scaling multiple items.

Developing countries could pursue the same combination of methods, and for PISA-D it would be logical to use the same scaling technology – item response theory scaling – as standard PISA.

PISA contains context constructs relevant for PISA-D. Other relevant context constructs can be found in PIRLS and TIMSS for early learning opportunities, quality of instruction and school resources. LLECE includes indices for educational opportunity, accessibility of basic school services and school infrastructure. The SACMEQ school community contribution factor is also considered valuable for developing countries.

- Regarding question formats, PISA-D should include item formats that allow for an adjustment of self-reported measures. This will allow analyses to further explore and potentially increase cross-country comparability. PISA-D could undertake, for example, correlation analyses at the between-country level between adjusted measures and scales or indices other than performance, in order to examine the impact of such adjustments in terms of construct validity. PISA-D should also undertake analyses to examine the extent of different patterns of response styles in participating countries.
- Regarding scaling and computing of relevant contextual constructs, including socio-economic measures, PISA-D should follow the procedures used for the scaling of context questionnaires in PISA. These procedures employ item response theory scaling methodology (for example, see OECD, 2009). PIRLS and TIMSS context questionnaire scaling could be of particular interest for PISA-D. Given that PIRLS and TIMSS have used Conquest, the algorithm underlying this particular scaling would probably be similar to what's been done in PISA.
- Relevant context constructs from international surveys of interest for PISA-D can be found in PIRLS and TIMSS in regards to early learning opportunities, quality of instruction and school resources. LLECE uses indices of educational opportunity, accessibility of basic school services and school infrastructure that may be of interest to PISA-D. The SACMEQ school community contribution factor may also be valuable for PISA-D.

Socio-economic status and poverty-related measures

The review of international surveys shows that measures of socio-economic status applied in international surveys conducted in developing country contexts commonly include indicators relevant to children living in poverty, but do not measure them distinctly from socio-economic status (SES). Such indicators are mainly based on home resources, household characteristics, and possessions and assets. Developing countries will need to draw on other countries' experiences for their own variables to measure socio-economic status. Already, some countries participating in PISA-D have a history of effective data collection on socio-economic status in their cultural and geographical contexts.

- The surveys reviewed contain several good examples for SES and poverty-related measures relevant to PISA-D. SACMEQ, PASEC and LLECE include

SES-related indices. SACMEQ and WEI-SPS include school and classroom measures that are related to SES.

- PISA-D should consider constructing an asset index, such as that created for STEP (Pierre et al., 2014: 15). Given the breadth of the countries participating in PISA-D, the challenge would be to find assets that differentiate levels of possessions equally well across these countries.
- PISA-D should also consider options for a finer differentiation of socio-economic status. One approach would be to ask, not just whether or not respondents have an item, but also whether the respondents would actually like to have an item they do not own.
- PISA-D also can draw on experiences of different countries regarding their own variables for measuring socio-economic status. Countries participating in PISA-D have a history of data collection and valuable experience on how to effectively assess socio-economic status in their cultural and geographical contexts.
- In regards to cross-cultural comparability, three aspects have been identified as crucial:
  - In relation to translation, adaptation and verification, country involvement in review of context framework and questionnaires is essential to check the face-validity of face value and cultural appropriateness of the content and identify possible issues with translation.
  - With respect to constructing context indices, it will be useful for PISA-D to include item formats that allow for an adjustment of self-reported measures to further explore and potentially increase cross-country comparability. PISA-D should undertake analyses to examine the extent of different patterns of response styles in participating countries.
  - Data analyses after the field trial and main study needs to capture the validity of questionnaire items across countries and ensure that items work in the same way in all countries. This applies to cognitive as well as contextual items. Country involvement is crucial in this regard.

Implementation procedures, methods and approaches to include out-of-school children, and use of data

Implementation procedures

Generally the international institutional arrangements for the reviewed large-scale international assessments involve a governing group, or steering committee, to set overarching policies and priorities, and one or more groups to provide technical guidance.

- The role and mandate of the PISA-D International Advisory Group is broad and varied. PISA-D needs to consider how it can accommodate the interests of the different stakeholder groups represented on it.
- The capacity-building and peer-to-peer learning emphases of PISA-D should be formalised in the institutional arrangements at the international and national levels. For example, partnerships could be established between PISA-D countries and PISA countries that have similar capacity needs. PISA-D countries should be

encouraged to establish their national centres to maximise capacity-building support.

- The OECD should clearly describe the roles and responsibilities of the national committee and guide each participating country to ensure that a productive relationship is established between the national committee and the national centre.
- The OECD should be prepared to encounter a variety of national level arrangements, from full responsibility concentrated on one group to a range of activities being outsourced. National centres should be supported to manage in-country relationships. Quality assurance requirements should be effectively communicated so that, in the case of some in-country outsourcing, all involved parties understand their responsibilities.

Survey implementation

Sampling

All the reviewed surveys employ a multi-stage sampling methodology. This involves choosing a school sample first, and then selecting students from the school. This happens in a variety of ways across the assessments, including sampling subsets of children across all classes in the target grades of sample schools (SACMEQ); sampling one classroom for each target grade (PASEC, Third Regional Comparative and Explanatory Study [TERCE]); and sampling intact classes from the target grades in sample schools (PIRLS, TIMSS). PISA, on the other hand, samples 15-year-old students in Grade 7 and above. The household-based surveys sample households and then select a sample from individuals within the target population in the sampled households.

In PISA-D, it is worth considering how to construct a school sampling frame that satisfies PISA's technical standards in countries that do not maintain a complete list of schools. Additionally, if up-to-date and complete lists of students are difficult to obtain from schools in advance, alternative methods for sampling students should be considered (for example, the SACMEQ and PASEC approach of sampling children on the day of testing).

- PISA-D should consider subnational arrangements to enable participation by countries with stable and unstable areas.
- Some countries do not maintain complete and up-to-date lists of schools. PISA-D will need to construct a school sampling frame that satisfies PISA's technical standards in these countries.
- We recommend considering whether PISA's approach to student sampling is appropriate in contexts where schools do not maintain complete and up-to-date lists of students. SACMEQ's approach – where children are sampled on the day of testing – may be worth exploring.

Data collection

In terms of cognitive data collection, the reviewed surveys can be broadly categorised. In several surveys, the cognitive assessment is a paper-based instrument that is administered in schools to groups of children, and each respondent completes the assessment independently by reading questions and recording responses on paper. Surveys that fit this category are PIRLS and prePIRLS, TIMSS, LLECE, SACMEQ and

PASEC Grade 6. In other surveys, the cognitive assessment is a paper-based or computer-based instrument that is administered one-on-one, either in households or in schools. Surveys that fit this category are ASER, EGRA, EGMA, STEP, LAMP, Uwezo and PASEC Grade 2. Similarly, for collecting contextual data, some ask respondents to complete questionnaires – LLECE, PASEC Grade 6, PIAAC, prePIRLS, PIRLS, SACMEQ, TIMSS, WEI-SPS; and for others, data collectors interview the respondents – ASER, EGRA, EGMA, STEP, LAMP, Uwezo, PASEC Grade 2.

- PISA-D should consider interview sessions to collect contextual data from respondents other than students. These respondents might include principals and teachers. It may be useful to implement:
  - A tablet-based data collection tool to eliminate recording errors.
  - Cognitive test administration over multiple days.
  - Permitting extra time to complete cognitive assessments.
  - Establishing on-site test administrator checks of student booklets to reduce the incidence of missing/discrepant data.
  - Sourcing test administrators who are local to the sites of test administration as a means of securing community engagement and buy-in.

Data processing

In regard to coding, or marking students' responses with codes once tests are complete, the reviewed surveys devote considerable time and resources to coder training and coding itself – including the steps taken to confirm that coding is being undertaken with acceptable reliability. In PIRLS, prePIRLS and TIMSS, comprehensive coder training is provided including actual responses from children. In LLECE, coder training is provided centrally to national representatives, who then return to their countries and replicate the training with their national coding teams. Responses to constructed response items (items requiring a written response rather than choosing from a set of options) are sometimes coded twice.

In EGRA and EGMA, coding is undertaken at the time of test administration, and coding training forms part of test administrator training. In PIAAC, participating countries that used a paper-based assessment were required to undertake in-country reliability studies in both the field trial and the main survey. In these studies, a second coder coded a predefined number of responses, and the level of agreement had to be at least 95%. Cross-country reliability studies were also conducted to identify any systematic coding bias across countries.

Services such as the PISA Coder Enquiry Service would be very useful for developing countries. This service entitles a country that has expended all efforts to arrive at an agreed code, but has failed to agree one, to write to the contractor, whose advice will then be recorded for all countries to see.

In PISA-D, constructed response items will be coded within the participating countries. Coding quality will be ensured by different procedures, including coding verification by expert coders and a coder reliability study across all participating countries (OECD, 2014: 43).

- For coding, PISA-D should consider services such as the PISA Coder Enquiry Service.
- For data capture, PISA-D should consider data entry application. It will need to be of adequate rigour, but it should not be so complex or unusual that it does not really serve the project's articulated aims about sustainable capacity development. PISA-D should consider more stringent requirements for double data entry than are currently implemented in PISA.
- For data cleaning, PISA-D should consider undertaking validation steps *before the test administrators leave the schools* (as is done in SACMEQ). Including these steps may simplify processes and reduce subsequent data cleaning activities.

Standardising implementation

In most of the reviewed surveys, standards are typically articulated through specific standards documentation, or through the instructional materials that are prepared to guide implementation. Standards should be included in a project implementation plan as well as in a dedicated standards document.

Some of the reviewed assessments have highlighted the difficulty of establishing standardised procedures when the participating countries are geographically, culturally and economically diverse. They also refine the standardised processes after a field trial.

PISA has a range of technical and operational standards that are articulated in a specific standards document. These standards cover aspects of implementation that have a direct impact on data quality, management standards that address operational objectives, and national involvement standards. To ensure comparability with standard PISA, it will be necessary for PISA-D countries to adhere to the accepted PISA standards.

All the large-scale international assessments produce manuals and use training meetings to familiarise the participating countries with the standard processes, and with the international and national quality monitors for monitoring assessment implementation.

- Articulation of standards could be included in memoranda of understanding or project implementation plans, as well as in a dedicated standards document. Including the standards in documents that are specific to each participating country, rather than general documents, may assist each country to be fully aware of its responsibilities with respect to the standards. A description of standards could be used as an opportunity to reflect the project's underlying values and ideology in a way that will help to secure local commitment to the project and acceptance of its results.
- With respect to training and quality assurance, the methods and processes of the reviewed large-scale international assessments should be explored in more detail. In particular, information should be sought about measures taken to ensure the quality of test administration.

Methods and approaches to include out-of-school children

Of the reviewed surveys, only PIAAC, STEP, LAMP, ASER and Uwezo include out-of-school children. They achieve this by having target population definitions that are age-based and make no reference to the enrolment or schooling status of individuals.

All five of these assessments sample households. STEP samples households in urban areas only; LAMP, PIAAC and Uwezo sample households across the participating countries (both urban and rural); and ASER samples households in rural districts only. In ASER households are sampled on the same day that tests are administered. In urban slum areas it may be difficult to establish a household list because it may be not be easy to distinguish between households. Uwezo tests children in urban areas, some of which would qualify as informal settlements or slums.

The language of the assessment is of critical importance when testing out-of school children. In ASER, for example, out-of-school children are allowed to choose which language they complete the reading assessment in, and in Uwezo, all children are allowed to receive the instructions for the mathematics test in whichever language they are most comfortable using.

- Input should be sought from ASER and Uwezo, and perhaps the other household-based assessments about how often they encounter problems with outdated sampling frames and how these are dealt with.
- Input should be sought from ASER (and perhaps Uwezo) about how to deal with multiple-occupancy households, as well as how to approach children who might be shy because they cannot read and children who are perhaps considered adults in their households.
- PISA-D should review the ways ASER and Uwezo obtain local buy-in to the survey. Some of these approaches may be applicable for the PISA-D out-of-school children strand.
- PISA-D should pursue an adaptive design for testing out-of-school children. Training and quality assurance measures will need to account for the additional burden adaptive design places on test administrators.

Analysis, reporting and use of data

It may be worth incorporating benchmarks in PISA-D analysis and reporting. Benchmarks that define minimum expected levels of performance may become increasingly relevant in the context of the post-2015 development goals and targets for education quality.

Countries may need considerable support in preparing national results reports. PISA-D should consider supporting participating countries to develop dissemination plans. Without the preparation and dissemination of national-level material that decision makers judge to be useful and relevant, a survey can only ever have a limited impact.

Data should be freely available to allow secondary analysis to take place. Ministry staff's active involvement in implementing research can be the key to linking results and actions.

- The use of benchmarks in the reviewed surveys should be examined. PISA-D should consider whether benchmarks might be incorporated into PISA-D analysis and reporting. Benchmarks that define minimum expected levels of performance may become increasingly relevant in the context of the post-2015 development goals and targets for education quality.
- Steps should be taken to ensure that questionnaire scales developed and used in reporting are considered relevant to policy in the participating countries.

- In regard to analytical approaches used for reporting, national-level reports from relevant countries may be useful. PISA-D should examine national level reports from countries that have participated in the reviewed large-scale assessments (such as South Africa in prePIRLS and PIRLS 2011, the SACMEQ countries) to get a sense of the kinds of analysis and reporting options that these countries have deemed relevant for their contexts.
- Regarding reports and communicating results, it may be valuable for PISA-D to present information on participating country contexts. The TIMSS and PIRLS encyclopaedias provide an example.
- The OECD and the international contractors for Strand A and Strand B of PISA-D should be prepared to offer considerable support to countries for the important work of preparing national results reports.
- PISA-D should consider supporting participating countries to develop and implement dissemination plans. National level material must be useful and relevant for decision makers if the survey is to have a significant impact.
- Regarding use of data and results, observations from SACMEQ highlight that active involvement of ministry staff in the research implementation is key to linking results and actions. We recommend considering how to ensure that government buy-in leads to similar success with PISA-D.

Notes

1. See www.oecd.org/pisa/pisafaq/.

References

Howie, S. et al. (2012), *PIRLS 2011: South African Children's Reading Literacy Achievement, Summary Report*, Centre for Evaluation and Assessment, University of Pretoria, Pretoria, www.up.ac.za/media/shared/Legacy/sitefiles/file/publications/2013/pirls_2011_report_12_dec.pdf.

OECD (2014), "OECD list of ODA recipients", www.oecd.org/dac/stats/daclistofodarecipients.htm (accessed 4 August 2014).

OECD (2009), *PISA 2006 Technical Report*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264048096-en.

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

Willms, J.D. and L. Tramonte (2014), "Towards the development of contextual questionnaires for the PISA for development study", *OECD Education Working Papers*, No. 118, OECD Publishing, Paris, http://dx.doi.org/10.1787/5js1kv8crsjf-en.