UIS Reporting Scales (UIS-RS)

Concept Note – July 2017

Contents

Abbreviations

ACER-GEM	Australian Council for Educational Research – Centre for Global Education Monitoring
ADEA	Association for the Development of Education in Africa
ASER	Annual Status of Education Report
EGRA	Early Grade Reading Assessment
EGMA	Early Graded Mathematics Assessment
EMIS	Education Management Information System
GAML	Global Alliance to Monitor Learning
GP-LA	Principles of Good Practice in Learning Assessment
GPE	Global Partnership for Education
IRT	Item Response Theory
LEG	Local Education Group
LLECE	Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación
PASEC	Programme d'Analyse des Systèmes Educatifs des Pays de la Confemen
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
SACMEQ	Southern and Eastern Africa Consortium for Monitoring Educational Quality
SEAMEO	South East Asian Ministers of Education Organisation
TIMSS	Trends in International Mathematics and Science Study
UIS	UNESCO Institute for Statistics

> *Poor quality education is jeopardizing the future of millions of children and youth across high-, medium- and low-income countries alike. Yet we do not know the full scale of the crisis because measurement of learning achievement is limited in many countries, and hence difficult to assess at the international level. A global data gap on learning outcomes is holding back progress on education quality.*
>
> (LMTF, 2013, p. 9-10)

Background for the UIS Reporting Scales

Almost two thirds of all developing countries have sought to measure education quality by implementing national or participating in regional or international learning assessment initiatives (Best et al., 2013). However, these assessments vary in approach, methodology, reliability, validity and comparability. Despite the high level of participation in learning assessments, clearly defined learning metrics and intra- as well as inter-assessment comparability remain limited. This presents particular challenges for measuring progress against the global development goals for student learning outcomes

The work described in this paper supports United Nations Sustainable Development Goal Number Four: Quality Education (SDG-4). In particular, it supports monitoring against indicator 4.1.1 of target 4.1:

> *Indicator 4.1.1: Proportion of children and young people: (a) in grades 2/3; (b) at the end of primary; and (c) at the end of lower secondary achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex.*

Target 4.1: By 2030, ensure that all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes.

The learning goal and target will only be meaningful if they are underpinned by empirically derived common numerical scales that accommodate results from a range of different assessments of learning outcomes. The development of common described scales allows policy makers, education practitioners and education investors to not only quantify student proficiency, but also describe it in a meaningful way. A scale provides a means to assess the emerging competencies of younger children, and to explore cognitive growth and trends over time. *A common described scale for reading and mathematics, spanning learning from early primary school to early secondary school, that is relevant and applicable to a range of developing country contexts is currently unavailable.*

The Global Alliance to Monitor Learning (GAML) is an initiative to support national strategies for measuring learning and enable international reporting. Led by the UNESCO Institute for Statistics (UIS), GAML brings together UN Member States, international technical expertise, and a full range of implementation partners—donors, civil society, UN agencies, and the private sector—to improve learning assessment globally. To ensure the quality and timely delivery of GAML expected outputs, GAML relies on the technical work from thematic Task Forces. This innovative alliance enables stronger links to be forged between all stakeholders, to create collaborative solutions to the challenges of monitoring learning worldwide.

As part of GAML the UIS and its technical partner, the ACER Centre for Global Education Monitoring (ACER-GEM) are developing reporting scales in mathematics and reading, and then to facilitate and support their use for monitoring purposes, in partnership with interested countries. This document outlines the three-phase work program for developing the UIS Reporting Scales (UIS-RS), which aims to develop and validate common reporting scales for reading and mathematics, and to support countries to report the results of their assessment activities against these scales. The key features of the program are fourfold:

- It accommodates results from a range of different assessments of learning outcomes.
- It yields high quality data that are nationally relevant and internationally comparable.
- It emphasises peer-to-peer capacity support and learning opportunities.
- It has a strong focus on improving data use and policy interface.

Objectives and Outputs

The objective of the UIS-RS work program is to develop empirically derived reporting scales in mathematics and reading that will support national governments to effectively measure and monitor learning outcomes for policy purposes. The reporting scales do not involve the development of a new test or testing program. Rather, they support the use of existing assessments of various kinds, and a pool of calibrated items that could be used to facilitate measurement and reporting of learning outcomes against common scales.

The key outputs of the work program to develop the UIS-RS will be:

1. reporting scales for reading and mathematics, spanning learning from early primary school to early secondary school;
2. a set of tools and methodologies that permit the broad alignment of existing learning assessments with the reporting scales; and
3. a support (capacity development) framework that enables countries to use the tools and methodologies to report results of national or other assessments against the reporting scales, should they wish to do so.

What is a Reporting Scale?

A reporting scale is one component of an assessment system with multiple integrated parts. Any expression of learning goals should be supported by well-defined indicators, which in turn draw upon accepted reporting scales and benchmarks. The process of setting and monitoring learning goals must have at its core a set of agreed reporting scales so that terms such as *foundation skills* and *acceptable* (in terms of proficiency) can be used with the knowledge that they carry a shared and accepted meaning.

Developing reporting scales therefore requires definition of each of the following components:

- Scale: This term is used to indicate a dimension, or metric, of educational progression. For example, a developmental scale of reading or mathematics would be considered a

reporting scale. The scale is depicted as a line with numerical gradations that quantify how much of the measured variable (e.g. reading ability) is present.

- Proficiency: Student proficiency on a reporting scale may be described numerically (*proficiency* scores) or substantively (*proficiency descriptions*). It is not practical to develop a proficiency description for each proficiency score on the numerical scale, so proficiency descriptions are usually developed to cover particular segments of the scale. These segments are called *levels*. The *proficiency description* for a particular level can then be understood as describing the skills and proficiencies of students who attained proficiency scores that are within that particular segment of the scale. Those students would also have the skills described for lower levels.
- Benchmark: When a location is set on a scale this is referred to as a *benchmark*, which is a point on the scale against which comparisons can be made. The point may be set at a single designated score, or at any point within a designated range of scores (*level*).
- Indicator: An *indicator*, in this context, is a quantitative expression that is used to describe the quality, the effectiveness, the equity or the trends of a particular aspect of the education system. It does so through mathematical statements concerning reporting scales, proficiency scores and benchmarks.
- Goal and target: A *goal* is often a broad aspirational statement of desired outcomes. A *target* is a specific statement of intended improvement in some particular outcome for a particular population or sub-population of interest, quantified in relation to the benchmarks, and the achievement of which can be monitored through measurements of progress on the indicators within a specified timeframe.

An example of a reporting scale for mathematics is shown in Figure 1. Its central elements are the numerical scale, and the descriptions of the levels of the scale in meaningful substantive terms. The various locations on this scale are proficiency scores. Given agreement on the scale, assessment tools can be developed and locations on the scale can be chosen as benchmarks, of which two have been displayed: *Grade 3 benchmark* (which may be an appropriate yardstick for some countries), and *Acceptable minimum standard for end of primary school*.

Against the reporting scale in Figure 1, the learning outcomes of two countries at Grade 3 and Grade 6 are reported. For each grade for each country, a range of indicators is shown: the distribution of performance; the mean proficiency scores for all children; and the mean proficiency scores for girls, boys, urban children and rural children. A range of other indicators could also be highlighted – growth over years, differences between subgroups and so on.

An example of a reporting scale for mathematics is shown in Figure 1. Its central elements are the numerical scale, and the descriptions of the levels of the scale in meaningful substantive terms. The various locations on this scale are proficiency scores. Given agreement on the scale, assessment tools can be developed and locations on the scale can be chosen as benchmarks, of which two have been displayed: *Grade 3 benchmark* (which may be an appropriate yardstick for some countries), and *Acceptable minimum standard for end of primary school*.

In the case of the UIS-RS, the indicator, target and goal are already described by SDG-4.1. It therefore remains to define the reporting scales themselves; allocate proficiency scores and descriptions; and identify points or levels among the range of possible proficiency scores that

constitute meaningful benchmarks of student ability. The work program outlined in this paper addresses each of these stages, to arrive at a fully-developed system for monitoring the 4.1.1 indicator, and assessing countries' progress towards the associated SDG target and goal.

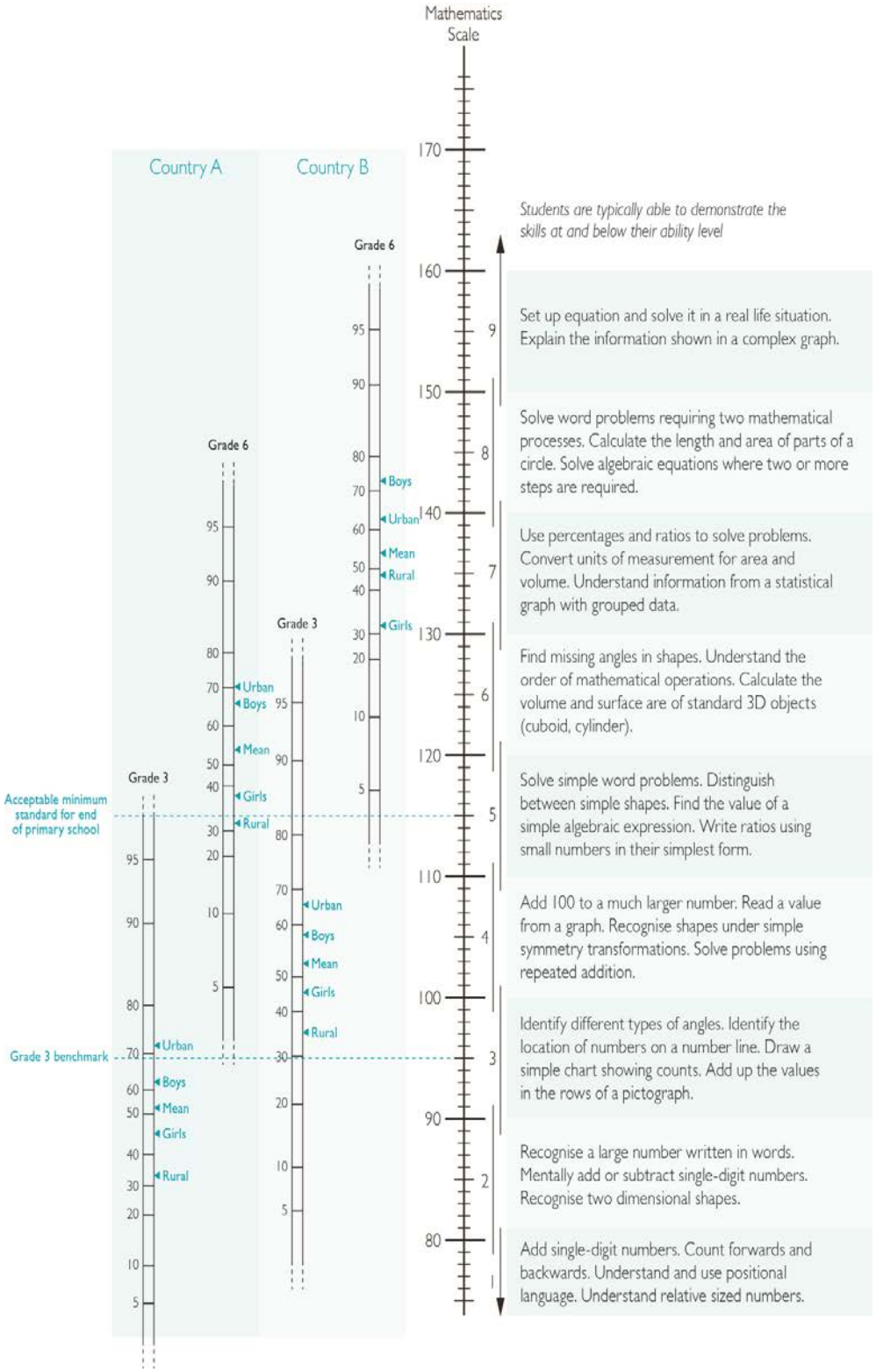

Figure 1: Example reporting scale for mathematics

Description of the UIS-RS

The UIS-RS aims to balance two seemingly competing necessities: the necessity for common learning metrics to underpin meaningful learning goals; and the necessity for a global framework for monitoring learning outcomes that recognises and can accommodate country-specific contexts and activities. While reconciling these necessities presents complex challenges, the work is driven by a shared belief that a workable, useful set of scales can be built, suitable for providing a global perspective on growth in reading and mathematics. Although the assumptions of equivalence underlying the reporting scales may never be perfectly realised across diverse international contexts, the process outlined in this paper is fit-for-purpose to achieve the best-possible approximation of international comparability.

A key element of the UIS-RS is that it draws from existing student assessments and country-level experiences, ensuring that any reporting scales developed will be relevant for different countries' educational needs. Developing the UIS-RS does not involve the development of a new assessment. Rather through a technically rigorous process of linking existing learning metrics, empirically validating, trialling, testing and reviewing in a set of pilot countries it is planned that the set of reporting scales developed will be locally relevant and reflective of varying country contexts, whilst offering comparability between contexts and over time.

Three specific deliverables will be developed through this work.

1. Common Reporting Scales

The UIS with ACER-GEM will develop, through its partnership approach, common scales in two key domains – reading and mathematics. This will take a two pronged approach: the first will be a conceptual and the second an empirical linking exercise. The metrics will cover the range of skills and abilities tested by large-scale international and regional assessments such as PISA, PIRLS, TIMSS, SACMEQ, LLECE and PASEC, but also extend down to more foundational levels of competence that are tested by ASER, Uwezo, EGRA, EGMA. National and sub-national assessments can also be used to broaden the scope of items for comparison.

2. Alignment of Existing Learning Assessments with the Common Metric

Technical work with other regional or national assessment programs will support their alignment to the common metrics. The preferred mechanism to achieve this is to form a pool of items from which a selection could be made for incorporation into existing assessments, and using those items as the basis for linking the other assessments with the common metric.

3. Country Level Implementation and Support Framework

This deliverable is focussed on the application of the metrics in conjunction with in-country system strengthening in learning assessment. UIS and ACER-GEM will focus on in-country and inter-country capacity support and development with a view to sharing technical assistance, experiences and perspectives and developing a set of tools and methodologies to systematically report results against the common metrics as part of the ongoing implementation of existing national, regional, or international assessments.

Implementation Phases and Duration

The development and implementation of the UIS-RS comprises three key phases. The first phase has been completed (see forthcoming Technical Report and Learning Progression Explorer prototype). Phase II is due to commence in October 2017, pending Task Force 4.1's endorsement of the process outlined in this paper. The UIS-RS Secretariat will work closely with relevant GAML Task Forces and in-country Task Teams (see Figure 1 below), to develop a more detailed plan of Phase II activities.

Phase I: Drafting the Reporting Scales

The purpose of this phase is to develop a set of draft reading and mathematics reporting scales from the earliest available developmental levels to the end of lower secondary school. Each comprises a graduated scale and a set of descriptions of what individuals at various locations on the scale are typically able to do, illustrated by a selection of items spread along the scale. In the interest of timeliness this first phase has been undertaken without the collection of new data from students – that is it draws upon pre-existing performance data.

Step 1: Developing a conceptual growth framework

The UIS-RS will be informed by ongoing development and refinement of conceptual growth frameworks, based on well-established educational learning theory and informed by curriculum scope and sequence documents. Work on the draft scale development commenced with establishing a broad conceptual understanding of reading and mathematics progressions, based on a synthesis of the literature, and how these domains are typically organised in curriculum and assessment. This conceptual framework will continue to be refined throughout subsequent phases of the work program, drawing on the content reference list concurrently under development by UIS.

Step 2: Identifying suitable existing assessment programs

The UIS-RS initiative does not aim to develop new test items but rather conduct a comprehensive analysis of existing items from a suitable range of assessment programs, mapping these items against the draft mathematics and reading scales and then calibrate these items across assessments. In order to do this, development of the draft scales involved working closely with a range of assessment programs in order to jointly review these instruments. Assessment programs were selected to cover learning from foundation/reception to early secondary and represent a range of item difficulties and the knowledge, skills, contexts and abilities each program attempts to measure.

Items from some potentially suitable candidates were already on hand or in the public domain (for example, ASER, Uwezo, and the EG*A instruments). Where permission could be gained and timelines permitted, instruments from programs including PASEC, SACMEQ, LLECE, PILNA, TIMSS Numeracy, PIRLS Literacy, and any others deemed relevant were also included. In addition, some national and sub-national assessments were available (LLANS, MTEG, SISTA and OLAY Northern Territory) which provided useful information.

Step 3: Conceptual and empirical analysis of assessment items

The first part of the analysis involved conceptual mapping of the cognitive demand of an agreed set of items used in a variety of existing assessments. A pairwise comparison of items was then conducted to enable the different assessments to be approximately aligned. Pairwise comparison in this context refers to a process where item development specialists compare pairs of test items and judge their relative difficulties. Well-established procedures (Bradley and Terry 1952; Luce 1959) were applied to develop an estimated alignment of all available items along a single scale. By using many comparisons and many raters, a numeric scale is yielded that estimates item difficulties with properties similar to those from other IRT models (e.g. scalar).

To support the drafting of the reporting scales, existing data from assessments was also used where available to align items from each source assessment program. Some assessments using different methods of administration, such as one-on-one oral administration, or paper-based group administration, provided comparative analysis that can be mapped against a scale using Item Response Theory (IRT) techniques.

Step 4: Formulating draft proficiency descriptions

In this step information from the previous steps informed the formulation of descriptions of growth according to the empirical difficulty of tasks used to assess elements of the conceptual framework. This step therefore constructed *separate draft reporting scales for reading and mathematics*. They were connected to some or all of PISA, PIRLS, TIMSS, SACMEQ, LLECE and PASEC scales, but extended down to more foundational levels of competence. Existing within-test calibrations were used to order items sourced from the same tests, with the outcomes of the pairwise comparison used to determine between-test item difficulty.

Phase II: Validating the Scales

The draft scales developed during Phase I are based on the conceptual analysis of the relative difficulties of items across assessment programs, and the analysis of already existing datasets. In Phase II, the draft scales will be validated at the country level. Data will be collected by administering combinations of items to children, which will enable the empirical determination of the relative difficulties of items across assessment programs. An item-based approach to linking the student data is preferred to a test-based approach[1], as it will result in a pool of calibrated test items from which any country that wished to could select items and insert them into its own assessment. This means that participating countries will have the option of reporting their results against the common scales.

This phase of activities will therefore involve multiple linking exercises of items from existing assessments against the draft scales across different countries, including assessments used in

[1] There are two main approaches to equating student data: *test based* and *item-based*. The *test-based approach* is considered the most technically rigorous as assessments are administered in their complete and original test form. However, any additional country that wished to place results of its assessment program against a metric that has been validated in this manner will need to undertake a full test-based equating exercise. An alternative is an *item-based approach* where different combinations of items from a range of assessment programs are administered in different countries with the aim to establish a large bank of equated items. It is the item-based approach that is being advocated here.

Phase I and other assessments not yet used. The start-up of activities in this phase will see extensive consultation with the view to working with at least 15 countries across different continents. A clearly defined coordination mechanism will be established to facilitate strong cross-country peer support. In-country technical teams will be identified and through a process of cross-country consultation and collaboration, country-specific plans for test administration will be developed.

Phase II will have five outputs. The first will be a pool of calibrated items. The second will be an empirically-based update and validation of the draft reporting scales that were developed via conceptual alignment in Phase I. The third will be performance benchmarks set on the scales using an empirical standard-setting exercise. The fourth will be a mapping of performance on items from the assessments used in this phase onto the reporting scales. The fifth will be the establishment of a peer-to-peer capacity support coordination mechanism across multiple country locations.

The validation phase is expected to take approximately 30 months, commencing once the draft scales have been developed in Phase I. A series of steps to implement Phase II are proposed as follows:

Step 1: Participation and coordination

UIS and their technical partner ACER-GEM will identify assessment programs suitable to participate in Phase II work and attempts will be made to secure their involvement. UIS and ACER-GEM will work with existing international coordination bodies involved in the development of learning metrics including the LMTF, SACMEQ, LLECE, PASEC, SEAMEO, ADEA and others to seek country-level interest in participating in Phase II. Analysis of current assessment systems in potential participating countries, alongside targeted consultation with Ministries of Education, will help to identify opportunities to align Phase II with local policy goals and capacity development needs.

To ensure geographical, cultural and language representation, UIS hopes to work with one to two countries each from the following nine regions:

Africa (Northern); Africa (Sub-Saharan); Africa (Eastern); Asia (Eastern); Asia (South-Eastern); Asia (Western); Oceania; Latin America and the Caribbean; Caucasus and Central Asia. GAML provides a collaborative structure to improve the flow of information, draw on international expertise, and to take advantage of cross-country peer support and capacity exchange opportunities. GAML 4.1, Assessment Implementation, and Capability Development Task Forces, will be the key steering groups for technical and implementation decisions at relevant points throughout the work program. Coordination of program activity will be managed through a UIS-RS Secretariat, comprising membership from UIS and ACER-GEM.

The work program also requires an in-country Task Team to be assembled in each participating country, comprising technical and grade-level specialists, as well as Ministry of Education representatives and specialists from other institutions as required. Once country-level participation has commenced, representatives of in-country Task Teams will work with relevant Task Forces and the UIS-RS Secretariat in a Reference Group capacity. This will ensure that the international collaborative expertise of the Task Forces is complemented by the detailed

understanding of each national context provided by key Task Team members. An outline of the coordination framework is presented in Figure 2.

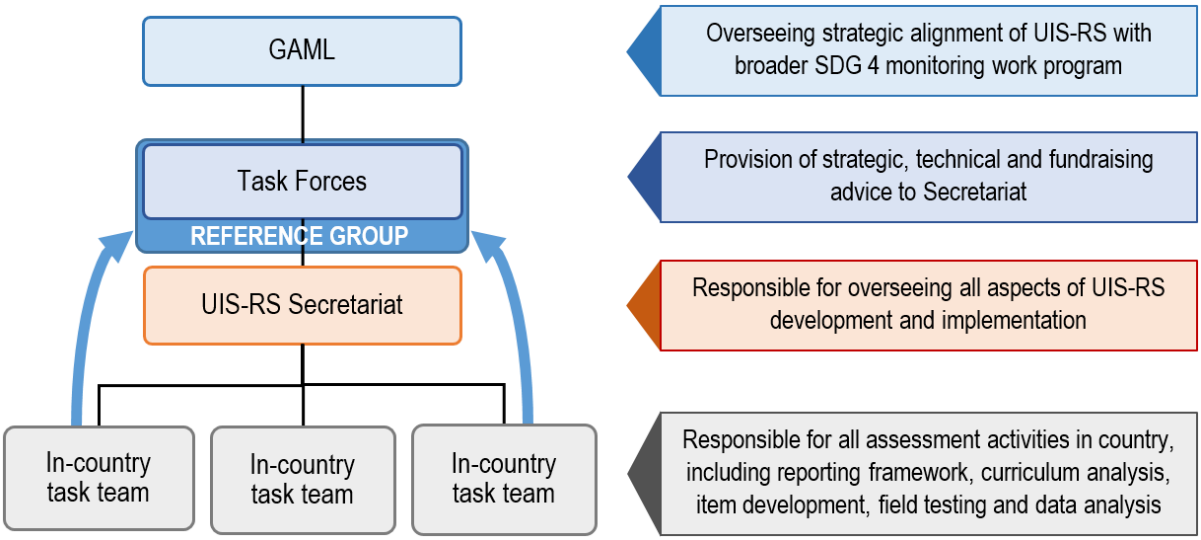

Figure 2: Proposed coordination framework for the UIS-RS work program

Step 2: Selecting the items

Items selected by the experts from the assessment programs in Phase I will be again reviewed by the relevant Reference Group members to ensure there is adequate coverage of the skills, knowledge and abilities. The Reference Group will also assess what additional items should be considered for inclusion. Preliminary discussions with regional assessment agencies indicated broad support for items from tests used in Phase I to be made available for use in Phase II.

Step 3: Designing the tests

In this step it will be necessary to determine which combinations of item sets from different assessment programs will be administered in each participating country.

It will also be necessary to determine how many sub-populations will be assessed in each participating country. For example, which grade levels, or whether regional populations should be considered (such as when different regions use different languages).

After the mix of items to be calibrated in any one country and the sub-populations to be assessed have been determined, an appropriate technical test design for each country will be developed. The test design will give the testing time per child and the sequence of items in different test forms. It will also show how items will appear in multiple test forms to facilitate linking.

At this stage it is expected that sample sizes will be in the range of 500–1000 per population–country combination. The student sample size is not intended to be representative, but rather provide the means to empirically calibrate the relevant test items, including accommodating language coverage. The population size for the sample therefore will not be as large as for a national student assessment initiative. The expertise and knowledge residing in the Reference Group will be tapped to make the decisions needed.

Step 4: Preparing test materials

The test materials are likely to be different for each country and will depend on the items that are being administered. If a population–country combination is using items that are delivered one-on-one and orally, the test materials might comprise a test administrator's stimulus booklet, a data collection sheet on which the test administrator can record the children's answers, and an associated manual to support test administration. If a population–country combination is using items that children must answer independently, then the test materials might comprise a test booklet on which a child writes his or her answers directly, and an associated manual to support test administration.

The development of test materials for each country will depend upon the extent to which items from the UIS-RS item bank can be incorporated into existing materials. Development of any new materials will be managed by in-country Task Teams, with other Reference Group members providing guidance and support in relation to the incorporation of items, as required.

Step 5: Preparing for and undertaking data collection

In this step preparations will be made for the in-country activities. These preparations will include:

- sourcing and training test administrators
- obtaining a sampling frame and sampling children to undertake the assessment
- taking steps to identify and secure appropriate sites for test administration
- sourcing and training data entry personnel (if applicable)
- sourcing and training personnel to code student responses (if applicable).

Since each population–country combination will be completing different test forms, training for test administration and the administration itself will vary from one population–country combination to the next. It will nevertheless be important to ensure in this step that preparations are made for test administration methods that are of an agreed level of standardisation where appropriate.

Sampled children will undertake the assessments and the resulting data will be captured. Methods for data capture could include data entry into a tailored software application or scanning. Again, it may be that the methods for data capture vary across the population–country combinations.

The in-country Task Teams will lead the activities in this step and be supported by relevant test development and administration experts from the Reference Group (such as regional assessment agencies or Task Force members), and other agencies where appropriate. In-country training programs will be agreed between the Task Teams and the Reference Group members prior to the start of this step.

Step 6: Analysing data and setting benchmarks

Once all data have been captured and scored, analysis will be undertaken in partnership with the in-country Task Teams with the support of relevant members of the Reference Group. Various modern psychometric techniques such as item response modelling will be employed.

The analytic process will be iterative over time, and will depend on the number of countries participating, the scale of the process within each country, and the spread of countries across economic, geographic and language groups. For example, if a number of similar countries participate early in the validation process, sufficient data may be obtained to confirm the validity of the scales in other countries with comparable profiles, but further validation may be required to confirm their fitness-for-purpose in more dissimilar contexts. Prior analysis of international assessments suggests that the reporting scales are likely to retain some variation across geographic and linguistic contexts, even after validation (Grisay et al 2007). Engagement through Task Force 4.1 and GAML will enable decisions about acceptable levels of consistency to be made, using both empirical methods and expert professional judgement.

This stage will also involve setting international benchmarks to enable reporting against the SDG 4.1.1 indicator. This will require establishing clear definitions of grade levels and minimum proficiency, and agreeing a method for benchmark calculation, using a combination of content referencing and normative data where available. A panel of experts will be convened from within the Task Force to develop advice on a preferred approach.

Trust and goodwill in international benchmarking depends upon shared understanding around what is valued in a monitoring program (for example, a focus on improvements within countries over time rather than necessarily on cross-country comparisons). Finalisation of the benchmarks will therefore require collaboration between the Reference Group and the in-country Task Teams with relevant curriculum experts and Ministry of Education representatives from the participating countries. In order to ensure that the benchmarks are valid for countries beyond those that participated in the linking exercises, the consultation process could be widened to include representatives from other countries that intend to make use of the scales. Individual countries may request additional training programs by the Reference Group to support data analysis work.

Step 7: Mapping and dissemination of results

Analysis will provide evidence of the coverage of the individual assessment programs against the UIS-RS. UIS and their technical partners will prepare this material in collaboration with the involved assessment programs. It will be the beginning of the suite of tools and methodologies that will be further developed in Phase III.

It is intended that the results relating to the development of UIS-RS will be disseminated as widely as possible to best inform the start-up of activities related to Phase III of the program.

Phase III: Country Level Implementation

Phase III is the development of a set of tools and methodologies that permit the broad alignment of existing learning assessments and country-developed assessments with the UIS-RS. In addition to enabling reporting against SDG 4.1, this phase involves capacity development in participating countries to improve assessment systems and classroom practice.

This phase of activities will have as its major objective the development of a strategy to support country-level activities through a longer-term capacity-building partnership. Every country context will have different needs from a student assessment monitoring program. For many countries, test materials and methodologies will already be well established at the country level

and only slight adjustments may be needed so that reporting can be made against the common scales. In other cases, a range of testing materials and methodologies can be available to countries who may wish to review and extend their own programs.

Establishment or strengthening an educational monitoring program, which is a central and ongoing focus of Phase III, will recognise that the most important element of any assessment program is that it is designed so that it can inform key policy issues. The use of the UIS-RS and related tools and materials will allow governments to make comparisons of data across contexts, against benchmarks, monitoring trends over time and monitoring educational growth. Opportunities for system strengthening through the UIS-RS will be supported by complementary initiatives, including the GPE's international Assessment for Learning (A4L) platform, and the UIS's Catalogue of Learning Assessments (CLA). An additional component will employ a framework based on the 14 key assessment areas indicated in the Principles of Good Practice in Learning Assessment (GP-LA), to identify areas of strength and improvement in a country's assessment and examination systems.

A policy-focused approach allows users to attach real meaning to assessment outcomes, informing the next steps needed to drive improvement. For example, the inclusion of multiple grade levels in an assessment allows for information about cohort **growth between grade levels**. Information about cohort growth sheds lights on how much value is being added to students' education at different stages of their schooling, and can help education practitioners and policy makers identify the stages at which policy interventions may be required.

Additionally, an ongoing assessment program yields information about **trends over time**. This information can come in a variety of forms, including information about changes in achievement outcomes at specific grade levels or within particular sub-populations. If the program assesses multiple grade levels, then the trend information can also include details of changes in growth between grade levels over time. Trend information such as this can assist in tracking the impact of educational reforms, and guide the development of new policy.

A student assessment program must be designed to meaningfully inform policy and sector reform initiatives. In order for this to occur, it is recommended that countries undertake a policy mapping exercise prior to commencing any work on an assessment program. The aim of a policy mapping exercise is to undertake a stock take of current education policies and levels of education provision at the national, sub-national and school level. Policies related to teacher support and professional training, curriculum, school financing and school fees, provision of learning materials, hours of learning, examination systems, school quality assurance, school feeding, and school management councils, are all areas that can have an impact on learning outcomes. Whilst a comprehensive policy mapping exercise may be difficult to implement, a broad understanding of the education policy context and educational statistics is critical. Once the needs of a student assessment program become clearer so will the capacity requirements to undertake the task at hand.

Step 1: Capacity analysis

A key step in understanding different countries' strengths and program priorities as well as opportunities for peer-to-peer capacity support is an in-depth, country-specific, capacity analysis. A capacity analysis could consider areas such as:

- leadership and vision
- institutional roles and responsibilities
- staff capacity (match between staff assigned and tasks required)
- work environment (physical capacity of the workplace to service program needs)
- technical capacity and needs analysis (including more detailed capacity support plan)
- sustainability (institutional, technical, financial).

A capacity analysis can form the basis of an assessment plan that outlines all aspects of program design (technical and financial requirements), plan of activities, timelines, roles and responsibilities, and expected outcomes. Assessment plans will build on, support and strengthen existing activities in each country, drawing on insights from the CLA.

Step 2: Capacity support for assessment programs

Whilst every program will be different, to ensure that assessment results are able to meaningfully inform policy there are a number of technical elements of the program that need to be considered. These may be summarised in Figure 3.

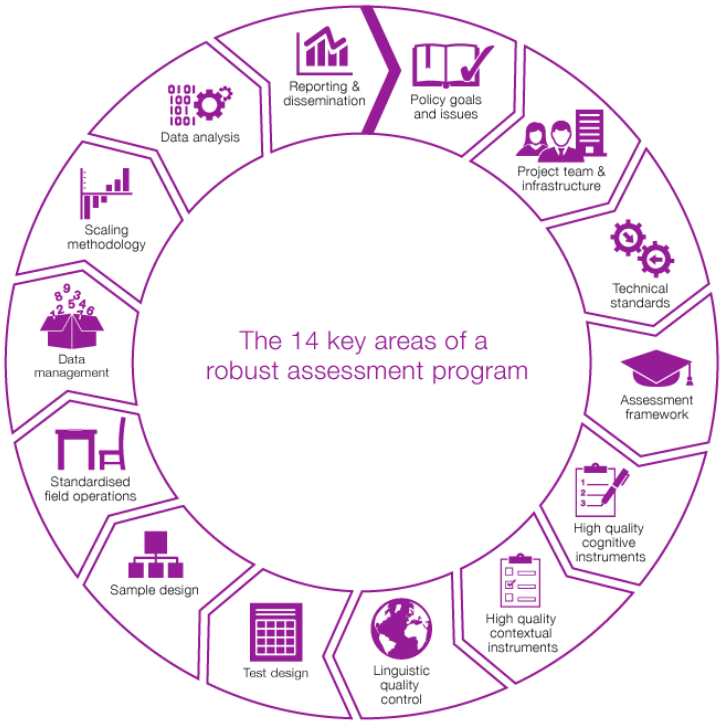

Figure 3: Key areas of a robust assessment program

The UIS-RS work program aims to provide the mechanisms to support individual countries in any one of, or all elements of the above mentioned areas. This request for support can be made through in-country education coordination bodies, such as the Local Education Group (LEG) in the case of GPE members. The Task Forces and Reference Group members can provide support to country level education coordination bodies to help define the type and scope of support required. Direct support may take the form of tendering for large-scale programs, specific short-term technical assistance or longer-term tailor-made training programs. The advantage of the UIS-RS governance structure is that relevant technical expertise through the

Task Forces or the in-country task teams will have been mapped, which will provide opportunities for country twinning and/or peer-support initiatives.

Most assessment programs typically require one to two years to prepare. Assessment frameworks, capacity analysis, policy mapping, technical teams, test design, item development, field operations manuals, piloting, data collection and analysis, all need to be developed and completed prior to the roll-out of a full-scale assessment. Data for the first year of the assessment form the baseline for ongoing assessment. Assessment programs should aim to be integrated into national planning and monitoring frameworks of education sector plans in the same manner as (and if possible linked to) education management information systems (EMIS).

Implementing a national student assessment program therefore will take on different forms in different countries, depending on the policy requirements, the available capacity in-country, the level of financial resources and the student population size. It is recognised that every country's requirements will be different. The strength of UIS-RS partnership approach is that it can provide tailored country level technical support to build on and strengthen existing student assessment programs, whilst allowing each country to use the products of Phases I and II to report learning assessment results against an internationally recognised set of metrics for mathematics and reading.

Risk Management

As an innovative, substantial international collaboration, the UIS-RS work program inevitably carries some level of risk. Strategies for managing major identified risks are outlined below:

Conceptual risks: Development of international learning metrics has been critiqued based on whether they are realistic representations of actual students' learning growth, and whether such representations are applicable across diverse education systems. The UIS-RS addresses these concerns by responding directly to a real need for international assessment tools, driven by a shared commitment to the SDG-4 learning goals and targets. This commitment necessitates a joint effort to confront the conceptual limits of assessment in rigorous, innovative ways.

Methodological risks: The proposed method for developing the reporting scales is one among many possible approaches, all of which have strengths as well as limitations that may place the validity of the scales at risk. The suitability of the proposed approach is supported by its origins in a well-established body of assessment theory and practice, which has been applied internationally by OECD (PISA) and IEA (PIRLS, TIMSS, ICCS), and in many large-scale national assessments. These methods have proven to be effective in enabling the development of comparable international tests, and are also fit-for-purpose for empirically deriving common numerical scales that accommodate results from a range of different assessments.

Implementation risks: All phases of the UIS-RS work program depend on a high level of international cooperation in their implementation, to follow agreed processes with timeliness and fidelity. The successful completion of Phase I in the context of changing international governance arrangements demonstrates the durable commitment of all partners to the work program, and their ability to collaborate to deliver quality results. The establishment of GAML will strengthen the basis for international collaboration to sustain the UIS-RS work program through the next phases of its implementation.

Political risks: International assessments carry a level of political risk, as some countries will inevitably score more highly than others. This risk will be mitigated in the UIS-RS work program by close engagement with Ministries of Education and assessment experts in participating countries, and clear agreement on the purpose of assessment to guide system improvement.

Proposed Budget

The budget required to implement the UIS-RS work program will be dependent upon the number of countries involved and the level of existing funding in each country project.

Phase I was funded by the Australian Department of Foreign Affairs and Trade's Australian Aid Program and ACER through ACER-GEM. Additional funding is sought to implement Phase II. Funding for Phase III is expected to be sourced from in-country government and donor-supported funding allocations. Phase III activities will vary widely by country depending on the extent of assistance required and the scope of the assessment activities. [2]

The following financial requests therefore relate to the second phase of the work program. Approximately USD500,000 will be required for technical assistance costs related to the validation of the draft reporting scales. The linking and comparative analysis work is planned to commence from the beginning of 2015, and will entail an intensive level of time-on-task.

The proposed budget for Phase II in-country work will be dependent upon which countries wish to be involved and to what extent they request an expansion of their existing in-country activities. On average however, it is anticipated that approximately USD150,000 per country per year will be required for technical assistance, with in-country costs calculated additionally (noting that in-country costs again will vary depending on logistical requirements and existing infrastructure arrangements). Robust validation will require participation of countries from each of the nine regions listed above.

Phase III costs are more difficult to estimate and will depend upon the specifics of each country's approach to implementation. It is expected, however, that an amount of approximately USD1,000,000 over a period of three years would be required for in-country capacity development and training initiatives.

[2] Costs of technical support projects for student assessment programs can range from between USD 200,000-1,000,000 per year of support, with the majority of projects between USD 400,000–700,000 per year per grade level. Other short-term training initiatives can normally be budgeted for USD 40,000 and above.

References

Best M, Knight P, Lietz P, Lockwood C, Nugroho D, Tobin M (2013). *The impact of national and international assessment programmes on education policy, particularly policies regarding resource allocation and teaching and learning practices in developing countries*. Final report, London: EPPI-Centre, Social Science Research Uni, Institute of Education, University of London.

Bradley, R.A. and Terry, M.E. (1952). Rank analysis of incomplete block designs, I. the method of paired comparisons. *Biometrika*, 39, 324–345.

Grisay, A, De Jong, J.L., Gebhardt, E, Berezner, A, Halleux-Monseur, B (2007). Translation Equivalence across PISA Countries. *Journal of Applied Measurement*, 8(3), 249–266.

LMTF (Learning Metrics Task Force) (2013). *Toward Universal Learning: Recommendations from the Learning Metrics Task Force*. Montreal and Washington, D. C.: UNESCO Institute for Statistics and Center for Universal Education at the Brookings Institution.

Luce, R.D. (1959). *Individual Choice Behaviours: A Theoretical Analysis*. New York: J. Wiley.

Wagner D (2011). *Smaller, Quicker, Cheaper: Improving Learning Assessment for Developing Countries*, UNESCO/IIEP.

World Bank (2006). *World Development Report 2007: Development and the Next Generation,* Washington, DC: World Bank.