GEM THE CENTRE FOR GLOBAL EDUCATION MONITORING

The UNESCO Institute for Statistics Reporting Scales

Concept Note

May 2018

United Nations Educational, Scientific and Cultural Organization

UNESCO INSTITUTE FOR STATISTICS

Australian Government
Department of Foreign Affairs and Trade

ACER

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
LLANS	Longitudinal Literacy and Numeracy Study
LLECE	Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación
LMTF	Learning Metrics Task Force
MTEG	Monitoring Trends in Educational Growth
OECD	Organisation for Economic Co-operation and Development
OLAY	Online Assessment of Year 1
PASEC	Programme d'Analyse des Systèmes Educatifs des Pays de la Confé rence des ministres de l'Éducation des États et gouvernements de la Francophonie (CONFEMEN)
PILNA	Pacific Islands Literacy and Numeracy Assessment
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
SISTA	Solomon Islands Standardized Tests of Achievement
TIMSS	Trends in International Mathematics and Science Study
UIS	UNESCO Institute for Statistics

Why are common reporting scales needed?

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

Measurement of learning achievement is essential to monitor how well education systems are delivering on the promise of universal quality education. This promise is reflected in the United Nations Sustainable Development Goal (SDG) Number 4 (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.

Goal 4 can only be meaningful if there is a shared global understanding of quality education, and relevant and effective learning.

The various indicators associated with SDG Target 4.1 attempt to translate its key constructs into measurable outcomes against which education systems can demonstrate progress. These indicators also require a shared international understanding of their meaning, if they are to inform global efforts to improve the quality of education for all children. The work described in this paper supports monitoring against one such indicator:

> **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.

For this indicator to meaningful across international contexts, a shared understanding must be reached on all of its composite constructs: "reading", "mathematics" and "minimum proficiency", as well as the specific grade levels at which they are to be measured.

Large-scale assessments of student learning are a well-established method of measuring the quality of education systems throughout the developed world. Almost two-thirds of all developing countries also seek to measure their country's education quality: they either implement or participate in regional, national or international learning assessments (Best et al., 2013). However, assessment programs vary in their approach, methodology, reliability, validity and comparability. Despite high levels of participation in learning assessments, clearly defined learning metrics, and comparability within and between and assessments, are currently limited.

Given the diversity of assessments used around the world, this indicator will be most meaningful if it is underpinned by empirically derived common numerical scales that accommodate results from a range of different assessments of learning outcomes.

Scales provide the means to assess the emerging competencies of younger children, and to explore cognitive growth and trends over time. They also allow policymakers, education practitioners and education investors to quantify student proficiency, and describe it meaningfully. At present, *there are no common described scales for reading and mathematics, relevant and applicable to a range of developing country contexts that span learning from basic to more advanced levels.* This is the need that the UNESCO Institute for Statistics reporting scales aim to address.

The UNESCO Institute for Statistics reporting scales

The Global Alliance to Monitor Learning (GAML) is an initiative to support national strategies for measuring learning and enable international reporting. GAML is led by the UNESCO Institute for Statistics (UIS) and 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 outputs are high quality and delivered in a timely way, GAML relies on the technical work from thematic Task Forces. This innovative alliance enables strong links to be forged between all stakeholders and for the creation of 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) have initiated a program to develop and validate common reporting scales in mathematics and reading, and then facilitate and support their use, in partnership with interested countries. The key features of the program are to:

- accommodate results from a range of different assessments of learning outcomes
- yield high-quality data that are nationally relevant and internationally consistent
- emphasise peer-to-peer capacity support and learning opportunities
- have a strong focus on improving data use, and its interface with policy.

The reporting scales do not involve the development of a new test or assessment 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. This document describes the three-phase program to develop the UIS reporting scales (UIS-RS), and to support their use.

The development of the UIS-RS is one part of GAML's broader work program to improve the monitoring of learning outcomes worldwide. A complementary strand of GAML's work involves the analysis and development of capability in assessment, recognising that education systems are at different stages in their capacity for high-quality educational monitoring. Tools and resources are currently being developed and refined to support this strand of work, including the UIS's Catalogue of Learning Assessments (CLA), and the Principles of Good Practice in Learning Assessment (GP-LA), which will be accompanied by practical guides for improving

assessment processes and programs. The Global Partnership for Education (GPE)'s Assessment for Learning (A4L) initiative, which works to build capacity for national learning assessment systems to measure and improve learning, also complements GAML's work.

Reporting scales: the backbone of effective educational monitoring

Monitoring educational outcomes requires multiple components to be defined. Learning goals and targets need to be set, and indicators need to be defined that enable the evaluation of progress towards these goals. Ideally, the indicators will draw upon accepted reporting scales and benchmarks. As such, reporting scales may be thought of as the backbone of this body of components, which work together to enable learning outcomes to be monitored effectively.

The main components required for monitoring educational outcomes are described below:

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. Targets are typically quantified in relation to benchmarks, and their achievement can be monitored through measurements of progress on indicators within a specified timeframe.
Indicator	An indicator, in this context, is a quantitative expression used to describe the quality, effectiveness, equity or trends of a particular aspect of the education system. Indicators are described through quantitative statements concerning reporting scales, proficiency scores and benchmarks. Scale A scale indicates a dimension, or metric, of educational progression. The scale is depicted as a line with numerical gradations that quantify how much of the measured variable (e.g. reading ability) is present. Where the scale is to be used primarily in reporting, it may be described as a reporting scale.

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 segments of the scale that contain ranges of scores. 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	A benchmark is a point on the scale against which comparisons can be made. This point may be set at a single designated score, or at any point within a designated range of scores (level).

An example of a reporting scale for mathematics is shown in Figure 1. Its central elements are the numerical scale, and the substantive descriptions of the proficiency levels of the scale. The various locations on this scale are proficiency scores. Figure 1 also displays two benchmarks: 'Grade 3 benchmark', and 'Acceptable minimum standard for end of primary school'. These benchmarks have been defined arbitrarily for illustrative purposes, and it is recognised that they may not be appropriate to all education systems.

To illustrate how scales can be used in international reporting, Figure 1 reports the learning outcomes of two countries at Grade 3 and Grade 6. For each grade for each country, a range of results is shown: distribution of performance; mean proficiency scores for all children; and 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.

In the context of SDG 4 monitoring, the target and goal are already described by SDG 4.1, and by Indicator 4.1.1 on learning outcomes in reading and mathematics. The UIS reporting scales are designed to allow for reporting of learning outcomes from different assessment programs against Indicator 4.1.1 – representing a numerical scale and proficiency descriptions. Global consensus will need to be sought to 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 the different stages of developing the UIS-RS, in order to arrive at a fully developed system for monitoring Indicator 4.1.1, and assessing countries' progress towards the associated SDG target and goal.

While the development and use of the UIS-RS are primarily described here in relation to SDG 4.1 reporting, this is not the only purpose for which the UIS-RS may be used. Education systems have their own goals and targets for improving learning, each of

which may be supported by a different set of indicators and benchmarks. The core components of the UIS-RS – the scales themselves, and the described levels of proficiency – can be used as points of reference for education systems to establish their own benchmarks, targets and indicators appropriate to their goals, and can also help to monitor progress towards achieving them. In this way, the UIS-RS provide a backbone for strengthening assessment in a wide variety of contexts, and making the results of assessment more meaningful for informing policy and practice.

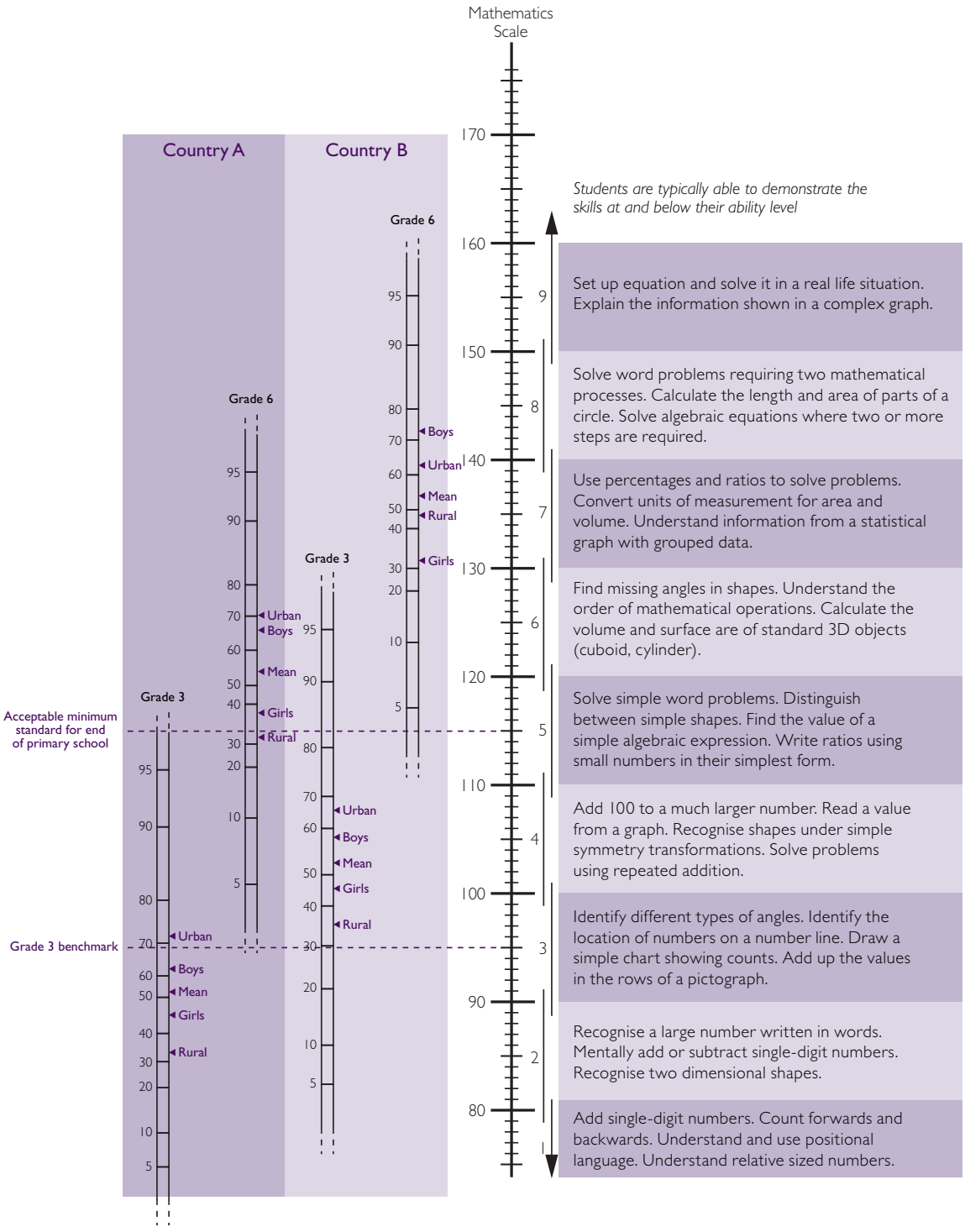

Figure 1: Example reporting scale for mathematics

The UIS-RS work program

The UIS-RS aim to balance two seemingly competing necessities: the need for common learning metrics to underpin meaningful learning goals, and the need 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 UIS-RS work program is driven by a shared belief that a workable, useful set of scales can be built and will be suitable for providing a global perspective on growth in reading and mathematics. Although the assumptions of equivalence that underlie the reporting scales may never be perfectly realised across diverse international contexts, the work program outlined in this paper is fit-for-purpose to achieve the best-possible approximation of international comparability.

The UIS-RS work program aims to achieve the following outputs:

1. **A set of reporting scales for each of the key domains of reading and mathematics, which span from early learning to more advanced levels.** Development of the scales will occur through a two-pronged approach: a conceptual exercise, and an empirical linking and validation exercise in a set of pilot countries.
2. **A set of tools and methods to broadly align existing learning assessments with the reporting scales.** The preferred mechanism is to form a calibrated pool of items, from which a selection could be made for incorporation into existing assessments. These items could then be used as the basis for linking the existing assessment with the common scales.
3. **A support (capacity development) framework to support the application of the reporting scales, in conjunction with in-country system strengthening in learning assessment.** This involves in-country and inter-country capacity support and development, with a view to sharing technical assistance, experiences and perspectives with any countries that have an interest in using the scales. It will include developing a set of tools and methods to systematically report results against the scales as part of the ongoing implementation of existing national, regional, or international assessments.

The development and implementation of these outputs comprises three key phases, which are detailed below.

Phase I: Draft the reporting scales

Phase I of the development of UIS-RS leverages work commenced under the former Learning Metrics Partnership, which informs the work of UIS. A full description of the Phase I process and outcomes can be found in the forthcoming Technical Report.

The purpose of Phase I was 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 set 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. For timeliness, Phase 1 was undertaken without collecting new data from students, that is, it drew upon pre-existing performance data. The four steps in this phase are described below.

Step 1: Develop a conceptual growth framework

The development of the UIS-RS is informed by well-established educational learning theories, curriculum scope and sequence documents, and ongoing development and refinement of conceptual growth frameworks. Development work on the draft scale started 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 curricula 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: Identify suitable existing assessment programs

The next step involved conducting a comprehensive analysis of existing items from a suitable range of assessment programs, mapping them against the conceptually-developed mathematics and reading progression from step 1, and then calibrating them across assessments. A range of assessment programs was jointly reviewed to identify suitable programs for analysis. The ones that were selected covered learning from foundation/reception to early secondary, and represented a range of the item difficulties and knowledge, skills, contexts and abilities that each program attempts to measure.

Items from some programs were already on hand or in the public domain (e.g. 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) and provided useful information. Some of the assessments selected use different methods of administration, such as one-on-one oral administration, or paper-based group administration. Please refer to the List of Abbreviations at the start of this paper.

Step 3: Analyse assessment items conceptually and empirically

The first part of the analysis involved conceptual mapping of the cognitive demand of an agreed set of items that had been used in a variety of existing assessments. Next, a pairwise comparison of items was conducted, to enable the different assessments to be approximately aligned. Pairwise comparison in this context refers to a process where item development specialists ("raters") compare pairs of test items and judge their relative difficulties. Well-established procedures (Bradley & Terry, 1952; Luce, 1959) were applied to develop an estimated alignment of all available items along a single scale. Using many comparisons and many raters yields a numeric scale, which

estimates item difficulties with properties similar to those from other item response theory (IRT) models (e.g. scalar). To support the drafting of the reporting scales, existing data from assessments were also used, where available, to align items from each source assessment program.

Step 4: Formulate draft proficiency descriptions

In this step, information from steps 1–3 helped to formulate descriptions of growth according to the empirical difficulty of tasks used to assess elements of the conceptual framework. Step 4 constructed *separate draft reporting scales for reading and mathematics.* They were connected to some or all of the existing scales for PISA, PIRLS, TIMSS, SACMEQ, LLECE and PASEC, but reached 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: Empirical linking and validation

The draft scales developed in 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 empirically linked to existing assessments, and validated at the country level. Data will be collected by administering combinations of items to children, which will enable the relative difficulties of items across assessment programs to be empirically determined. 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 have the option of reporting their results against the common scales.

Phase II involves multiple linking of items from existing assessments against the draft scales across different countries, including assessments used in Phase I and other assessments not yet used (such as national assessments). The commencement of work in this phase will involve extensive consultation, with the intention of identifying at least 15 countries across different continents to be involved in the linking and validation exercise. A clearly defined coordination mechanism will be established to facilitate strong cross-country peer support. In-country Task Teams will be identified and through a process of cross-country consultation and collaboration, countryspecific plans for test administration will be developed. The in-country Task Teams will work with relevant Task Forces and the UIS-RS Secretariat in a Reference Group (see section on 'Governance and coordination of the UIS-RS work program').

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 wishes 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 of establishing a large bank of equated items. The item-based approach is advocated in this paper.

Phase II will have five outputs, which will be:

1. empirically-based validation of, and refinement of, the draft reporting scales
2. a pool of calibrated items
3. identification of performance benchmarks on the scales for use in reporting against Indicator 4.1.1, based on an empirical standard-setting exercise
4. a mapping of performance on items from the assessments used in Phase II onto the reporting scales
5. a tool for education systems to use to align their assessment programs with the UIS-RS, in cases where empirical equating has not occurred (see Data Alignment Concept Note).

The proposed steps for implementing Phase II are listed below.

Step 1: Identify assessment programs and secure participation

Step 1 will be to identify suitable assessment programs, seek country-level interest and secure participation in the empirical linking and validation. Collaboration with international, regional and national assessment bodies will be essential for this undertaking. In addition, analyses of current assessment programs 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 from each of the 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.

An additional group of existing assessments will be used to pilot the process for conceptually aligning assessments with the UIS-RS, in cases where empirical equating has not occurred. This process can be done using existing test data, along with assessment frameworks and other reference documents that can assist in conceptual alignment between existing assessments and the UIS-RS. The pilot process will inform the development of a tool for education systems to use to align an assessment program with the UIS-RS.

Step 2: Select items

Once assessment programs have been identified for empirical linking and validation, items will be selected for use in the equating process. The item selection will be based on specific criteria to ensure adequate coverage of the skills, knowledge and abilities.

The pool of items compiled at this step will constitute the first iteration of the UIS-RS item pool – a resource that will be maintained and refined throughout the subsequent stages of the work program. As more assessment programs are equated with the UIS-RS in Phase II, the pool of items that can be used to assist in equating will be

expanded. This expansion can continue in Phase III, as education systems begin to use the UIS-RS in their reporting. By contributing items from their assessment programs to the UIS-RS pool, education systems have an opportunity to strengthen the alignment between their program and the UIS-RS, and to contribute to a global shared resource to improve consistency of educational monitoring.

Step 3: Determine item sets, assessment populations and develop test design

Step 3 in empirical linking and validation will determine which combinations of item sets from different assessment programs will be administered in each participating country, and how many sub-populations will be assessed. For example, which grade levels, or whether regional populations should be considered (such as when different regions use different languages).

After this has been confirmed, an appropriate technical test design for each country will be developed. The test design will indicate the time required for testing each child and the item sequence in different test forms. It will also show how items will appear in multiple test forms to facilitate linking.

Sample sizes are expected to 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 and accommodate language coverage. The sample size therefore will not be as large as for a national student assessment initiative. Assistance will be provided through the Reference Group to support decision-making about the sample size in each country as required.

Step 4: Prepare and administer test materials

The test materials are likely to be different for each country and will depend on the items that are being administered, and the method of administration. If a population–country combination uses 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 uses items that children must answer independently, then the test materials might comprise a test booklet on which the child writes their 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 Reference Group members providing guidance and support in relation to the incorporation of items, as required.

Step 5: Collect data

Step 5 comprises the in-country data collection activities. These 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 complete different test forms, the training for test administration and the administration itself will vary. It will nevertheless be important to ensure that the preparations made for test administration 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, the methods for data capture could vary across the population–country combinations.

The in-country Task Teams will lead the activities in this step and be supported by the Reference Group. In-country training programs will be agreed between the Task Teams and the Reference Group members before data collection commences.

Step 6: Analyse data and set 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 the Reference Group. Modern psychometric techniques, such as item response modelling, will be used.

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, De Jong, Gebhardt, Berezner, Halleux-Monseur, 2007). Engagement through GAML and relevant Task Forces 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 and validation exercise, 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: Map and disseminate 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. The tool for alignment of non-equated assessment programs will also be disseminated.

Phase III: Country-level implementation

In Phase III, the focus of the work program shifts from developing methods and tools, to putting them into practice. In this phase, education systems will begin to use the UIS-RS to report against Indicator 4.1.1. Education systems will be able to report learning outcomes from an assessment program against Indicator 4.1.1, if the assessment program has been equated with the UIS-RS (for example in Phase II or through incorporating items from the calibrated item pool). In cases where empirical equating has not occurred, countries can use the conceptual alignment tool to analyse the alignment between an assessment program and the UIS-RS, to report against Indicator 4.1.1.

Alongside reporting against SDG 4.1.1, Phase III of the UIS-RS work program will also have a focus on developing capacity support plans to strengthen assessment and reporting at the country level. Capacity support will be based on the key quality concepts for learning assessment and the 14 key areas of a robust assessment program, as described in the Principles of Good Practice in Learning Assessment (Figure 2).

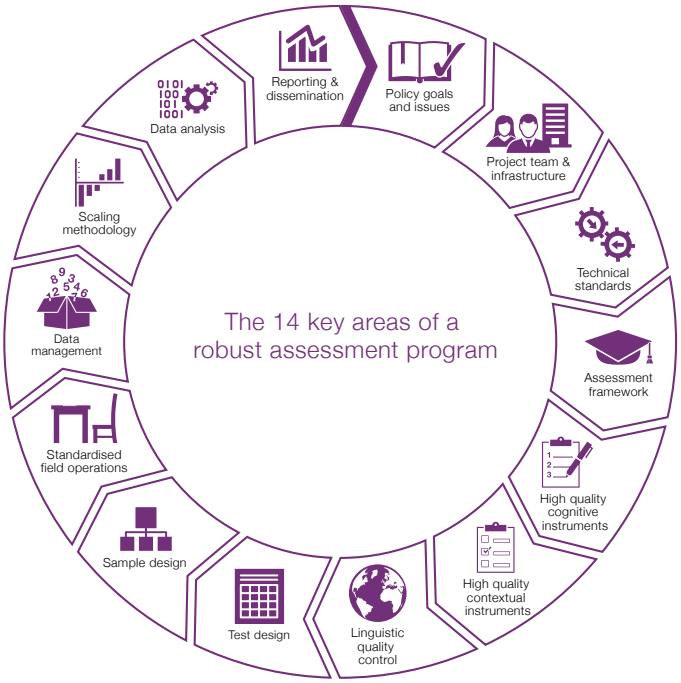

Figure 2: Key areas of a robust assessment program

While reporting against Indicator 4.1.1 is a key focus of the UIS-RS work program, it is intended as a global program to strengthen the use of learning assessment to influence education policy in numerous ways. The use of the UIS-RS and related tools and methods will allow governments to compare data across contexts and against benchmarks, and monitor educational growth and trends over time. GAML's capacity-development strand will support education systems to analyse how well assessment programs are being used to inform policy in their context, and how well assessment is integrated throughout the education system as a key driver of improvement. This reflects GAML's understanding that robust educational monitoring is not an end in itself, but a tool for driving and informing system-wide efforts to improve educational quality and outcomes.

The timing and nature of Phase III implementation will vary across education systems. As most assessment programs typically require one to two years to prepare, it is anticipated that education systems will adopt the UIS-RS-related tools and methods gradually and iteratively, over multiple 4.1.1 reporting cycles. Each iteration will provide an opportunity to strengthen the UIS-RS tools and methods, and improve their relevance and usefulness for education systems in diverse international contexts. Phase III should therefore be seen as a period of continuous improvement, with the UIS-RS providing a common reference point for ongoing dialogue and collaborative support to improve educational monitoring worldwide. The strength of GAML's collaborative approach is that it can provide tailored country-level technical support to build on and strengthen existing student assessment programs, while 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.

Governance and coordination of the UIS-RS work program

The first phase of UIS-RS development has been coordinated through a partnership between UIS and ACER-GEM, drawing on expert input from the GAML Task Forces and network and the wider education community. The next phases of the UIS-RS work program will involve the establishment of a new governance and coordination structure, to leverage country-level expertise to ensure that the work program responds to diverse education systems and contexts. This governance structure will oversee the equating and validation exercise in Phase II, and will also provide a framework for ongoing oversight of the UIS-RS's application in practice in Phase III. The governance and coordination framework will involve (see also Figure 3):

- GAML has a collaborative structure that will enable the flow of information to all relevant stakeholders, including donors, assessment organisations and education systems. The collaborative structure will also allow GAML to draw on international expertise, and to initiate and sustain cross-country peer support and capacity exchange opportunities. GAML will continue to oversee strategic alignment of the UIS-RS work program with the broader SDG 4 monitoring program.
- Relevant GAML Task Forces will be the key steering groups for technical and implementation decisions at relevant points throughout the UIS-RS work program.
- The UIS-RS program activity will be managed through a UIS-RS Secretariat, comprising membership from UIS and ACER.
- The UIS-RS work program will require an in-country Task Team to be assembled in each participating country.[2] Task Teams comprise 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 3.[3]

2 This refers to countries participating in Phase II, and countries using the UIS-RS for reporting in Phase III.

3 For simplicity, only three in-country Task Teams are shown in the diagram, although there may be many more.

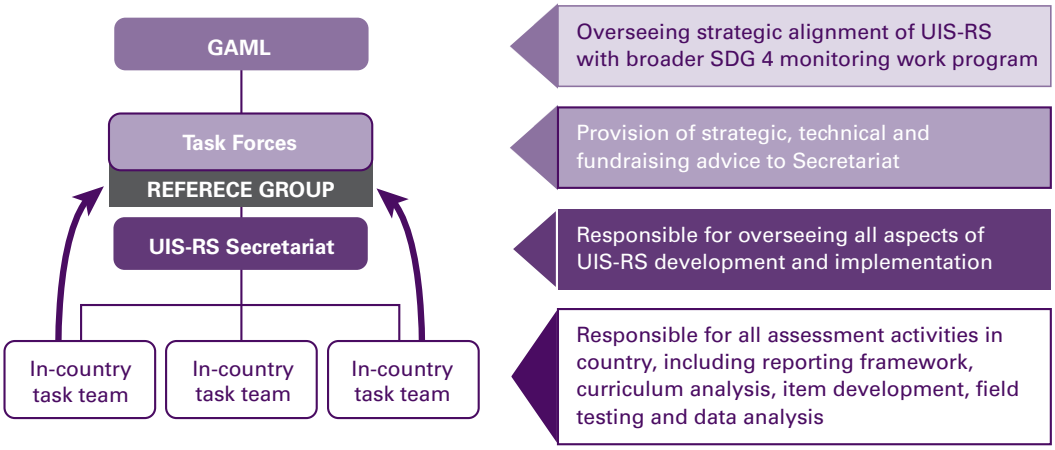

Figure 3: Proposed coordination framework for the UIS-RS work program

Risk management

As an innovative, substantial international collaboration, the UIS-RS work program inevitably carries some level of risk. Strategies for managing the 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 just one of many possible approaches. All methods 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 to follow agreed processes with timeliness and fidelity in their implementation. 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.

Conclusion

This paper has described the rationale and process for developing common reporting scales in reading and mathematics, to support consistency in reporting against SDG 4 Indicator 4.1.1. The UIS-RS will not only enable Indicator 4.1.1 reporting to better accommodate a range of assessment programs, but will also provide a valuable reference point for international dialogue about learning in these domains. The shared understanding of learning supported by the scales will provide a strong foundation for collaborative efforts to improve education quality around the world.

A key element of the UIS-RS work program is that it draws on existing assessments and country-level experiences: through the empirical linking and validation exercise, the collaborative approach to designing and trialling methods and tools, and the focus on capacity development. The draft scales created in Phase I cover the range of skills and abilities tested by existing 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 assessments such as ASER, Uwezo, EGRA and EGMA. The empirical linking and validation exercise will strengthen the alignment of the scales to different assessment programs – potentially including national and sub-national assessments – to further ensure that the scales will be relevant for different countries' specific educational needs.

The work program for developing the UIS-RS will continue to require extensive collaboration between GAML and its Task Forces, UIS and ACER-GEM, participating education systems, assessment organisations and many other education stakeholders. This collaboration will include ongoing discussion about the conceptual and theoretical foundations of the scales, as well as shared engagement in empirical work to refine and validate the scales, and develop tools and methods for their application. Collaborative engagement throughout its development will enhance the UIS-RS's value across diverse international contexts, as a shared global good to enhance assessment for learning.

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 University, 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.