Equating Existing Assessments and Validating the UIS reporting scales

Discussion Document
July 2017

Introduction

This document proposes an approach to the equating of existing regional and international assessments[1], which also serves as the validation phase of the UIS Reporting Scales (UIS RS). The UIS-RS are empirically derived reporting scales in mathematics and reading, which will support international consistency in reporting on Sustainable Development Goal (SDG) Indicator 4.1.1.

The equating and validation process supports the Global Alliance to Monitor Learning (GAML)'s commitment to offering education systems a variety of options for reporting against Indicator 4.1.1, without requiring the adoption of a common assessment program. These options are:

1. *Existing programs for which equating with UIS RS has occurred:* A country may choose to use an existing assessment program to report on Indicator 4.1.1 that has been equated with the UIS RS, through the equating and validation process described in this document.

2. *Programs using items or modules mapped to the UIS RS:* The equating and validation process involves the creation of a pool of items (or modules) that have been mapped to the UIS RS. These items could be incorporated into an assessment program that is not already equated to the UIS RS, and linked items could then be used to locate these assessment on the UIS RS.

3. *Other programs:* The UIS Data Alignment Concept Note sets out a process for countries using other assessment programs to examine their level of alignment with the UIS RS.

The proposed equating and validation process involves item-based, rather than test-based, equating. This allows countries flexibility in incorporating linked items into existing assessment programs, rather than requiring the adoption of a program in its entirety.

The approach involves carrying out multiple linking exercises across, say, 10–15 different countries[2]. In each country, the linking should be of an acceptable technical rigour, and be achieved through methods that are inclusive, efficient and practically feasible.

In each equating exercise, sets of items from the involved assessment program[3] are selected and administered to one or more samples of children. Each sample of children will represent a target population of interest, in terms of their stage of educational development, aligned with the levels in SDG Indicator 4.4.1 (ie grades 2/3, end of primary and end of early secondary).[4]

After all the separate linking exercises are completed, all items that were included will together form a pool of calibrated items – which will become a central tool in the future use of the UIS RS.

The item-based equating of international and regional assessments would also provide data to empirically validate the UIS RS themselves. The UIS RS have been drafted based on the conceptual

[1] Throughout we will use regional and international assessments to cover OECD/PISA, IEA's assessments, LLECE, PASEC, SACMEQ, SEA-PLM, PILNA, EGRA, EGMA and the citizen-led assessments.

[2] This is indicative only and is chosen to both allow broad geographical and cultural representation and to ensure sufficient data for validation purposes while spreading the load across countries.

[3] Potential participants include those assessment programs that contributed with items to drafting of the scales, plus other additional assessment programs that are suitable for the validation work but have not been part of the drafting.

[4] As the operationalisation of schooling varies considerably across countries, out-of-school populations should also be considered. Terms related to schooling are used here as a device to avoid the abstract.

analysis of the relative difficulties of items across assessment programs, and the analysis of already existing datasets, and now require validation at the country level.

The linking approach

Central to the validation is the concept of *linking*. In its most generic term, linking is a procedure through which to the results of two tests can be compared (Linn, 1993; Mislevy, 1992).[5] When the scores of two tests that have been constructed for different purpose and using different content frameworks are put on the same scale, 'the results are said to be comparable, or calibrated' (Feuer et al., 1999, pp. 18-19).

There are two main approaches to calibration that could be considered in this context: *test-based* and *item-based*. In a test-based approach, the goal is to calibrate existing tests so that administration of any of the calibrated tests produces a score that can be mapped to a common metric. In contrast, the goal of an item-based approach is not to calibrate tests but to establish *a pool of calibrated items* from which any country that wishes to could select items and insert them into its own assessment.

Box 1: Test calibration approaches

Three test-linking approaches commonly used are described below (Feuer et al., 1999; Kolen, 1988; Kolen & Brennan, 2014; Muraki, Hombo, & Lee, 2000).

Test-based approach

Single group: In this design, a single group of examinees consecutively completes all instruments to be linked. This design is very demanding in terms of the time needed for test completion, which raises concerns about the effect of examinee fatigue in test results. Long resting periods between administration sessions may alleviate this issue but too long periods may also compromise the results due to changes in examinees' ability. Another concern relates to the order in which instruments are administered. Order effects can be minimised by using a balanced test design.

Random equivalent groups: In this design, equivalent samples are chosen randomly from a larger population of examinees. Each sample completes one of the tests to be linked. Although this design is preferred over the single group design because it is less demanding in terms of administration time (ie fatigue effects are avoided), the level of control needed to randomly assign examinees into assessment instruments is seldom achieved.

Item-based approach

Non-equivalent groups with common items: In this design, modified versions of the tests to be linked are created. These modified versions include a subset of items in common, which represent – as far as possible – the content and properties of the original tests. The modified versions are administered to the non-equivalent groups.

If an item-based approach to calibration is adopted, it will allow for different student assessments to subsequently be linked to the scales in mathematics and reading. If a test-based approach were to be used, then beyond the initial calibration, any additional country that wished to place results

[5] Particular forms of linking include equating, calibration, scores projection, statistical moderation and social moderation. The strictest form of linking is test equating (for details see Feuer, Holland, Green, Bertenthal, & Hemphill, 1999; Kolen & Brennan, 2014; Linn, 1993; Mislevy, 1992).

of its assessment program on this scale would have to undertake a full test-based calibration exercise itself.

Moreover, a common scale developed out of test-based calibration would inevitably bear the imprint of the small number of assessments that were involved in the initial calibration, and may in the end be overly associated with one of those assessments.

In an item-based approach different combinations of items from a range of assessment programs would be administered in different countries, and the aim would be to establish a large bank of calibrated items. This bank of calibrated items would be used to construct the common scale, and any country that wished to do so could insert items from the bank into its own assessment, and therefore have the option of reporting its results against the scale.

There are aspects of the item-based approach that will make it less technically rigorous than the test-based approach. In particular, an item-based approach involves the administration of items removed from their complete and original test forms. The placement of an item in different contexts almost invariably alters the nature of response to the item, for example, making it more easy or difficult. In essence, when placing an item in a novel context it must be considered a somewhat different item to when it was in the original context. However, this difference can be quantified and therefore dealt with analytically.

The item-based approach nevertheless has a number of advantages over the test-based approach in the context of the post-2015 agenda. Some of these advantages are short-term and relate to the development of the common learning metrics. Some of the advantages are longer-term and relate to the ultimate purpose of establishing the metrics – to support countries to conduct assessments that are appropriate for their contexts, while enabling them to report against the metrics should they wish to do so.

The short-term advantages of the item-based approach are:

- fewer items need to be used from each original assessment program
- fewer items need to be administered to children – each child will complete one test that includes items from multiple assessments rather than completing multiple tests
- fewer items need to go through the rigorous translation and adaptation procedures that are necessary to ensure that the quality of calibration is high
- test materials are leaner and less costly
- the test administration burden on countries is minimised
- it is flexible and scalable – more items can always be added to the pool to strengthen the calibration of particular key assessments.

The longer-term advantages of the item-based approach are:

- it will reinforce the fact that the scales developed as part of this work program are new scales, not overly associated with the pre-existing scales of one or another existing assessment program
- it is future proof – countries interested in reporting against the scales in the future can do so by simply inserting items from the calibrated pool into their own assessments.

The rest of this document discusses the equating and validation process. First, the key outputs are listed. Second, an outline of the work program is provided, conceptualised as a series of sometimes overlapping steps, is described. Third, key next steps and an indicative overall timeline is given. Finally, a brief conclusion is provided.

Outputs

The first output will be a pool of calibrated items.

The second output will be an empirically-based update and validation of the draft UIS Reporting Scales that have been developed via conceptual alignment.

The third output will be performance benchmarks set on the scales using an empirical standard-setting exercise.

The fourth output will be mapping of performance on items from those assessments used in this phase onto the common scales.

Work program

Step 1: Identifying and securing involvement of assessment programs and countries

Assessment programs

In this step assessment programs that might be suitable for this work are identified by the GAML Secretariat with input from relevant task forces, and attempts are made to secure their involvement.

For assessment programs, the benefit of participating in the validation is twofold. First, contributing items for the linking studies will add value to the original programs because, as soon as the linking is completed, their participants will be able to report against the common learning metrics should they wish to do so. Second, the validation activities will support capacity building within the involved programs because experts will be able to share and compare item development processes, and be exposed to a wider pool of items. In addition, access to new assessment design and to research in advanced technology and methodology will be available to the assessment community involved in this project.

In identifying potential assessment programs for the validation work, the GAML Secretariat needs to ensure that they together cover the range of learning from foundation/reception to early secondary schooling, if not higher, because this is the range that the metrics intend to span. In order to examine whether or not a set of assessment programs spans an adequate range of learning, it is convenient to map target populations side-by-side and then to consider whether, in the context of the equating and validation work, the programs might be able to provide information on learning levels of populations other than their target populations.

Table 1 and Table 2 below show illustrative mappings of sets of assessment programs that could feasibly be involved in the validation. With the support of experts from the assessment programs

that have been approached to be involved in the work, the GAML Secretariat, with input from the Strategic Planning Committee, could establish the most accurate mapping.

The final choice of the assessments contributing items for the validation is made by the GAML Secretariat with input from the relevant task forces, and depends on:

- which assessment programs are willing to contribute items and the technical merit of those items
- which set of assessment programs covers in the most balanced and effective way the full range of learning that the metrics intend to span
- particular gaps identified during the Phase I development work.

Table 1: An illustrative set of literacy assessments for the validation, their target populations and expected other populations that may be covered

	Assessments that could contribute items							
Grade	Community-based assessments*	EGRA	LLECE	PASEC	PILNA	PIRLS & PIRLS Literacy	PISA^	SACMEQ
Grade 11								
Grade 10							Gr 10	
Grade 9				Gr 9			Gr 9	
Grade 8								
Grade 7								
Grade 6			Gr 6		Gr 6	Gr 6		Gr 6
Grade 5				Gr 5		Gr 5		
Grade 4		Gr 4			Gr 4	Gr 4		
Grade 3		Gr 3	Gr 3					
Grade 2	Gr 2	Gr 2		Gr 2				
Grade 1	Gr 1	Gr 1						
Reception								
Foundat'n								

Table 2: An illustrative set of numeracy assessments, their target populations and expected other populations that may be covered

	Assessments that could contribute items							
Grade	Community-based assessments*	EGMA	LLECE	PASEC	PILNA	TIMSS & TIMSS Numeracy	PISA^	SACMEQ
Grade 11								
Grade 10							Gr 10	
Grade 9				Gr 9			Gr 9	
Grade 8						Gr 8		
Grade 7								
Grade 6			Gr 6		Gr 6	Gr 6		Gr 6
Grade 5				Gr 5		Gr 5		
Grade 4		Gr 4			Gr 4	Gr 4		
Grade 3		Gr 3	Gr 3					
Grade 2	Gr 2	Gr 2		Gr 2				
Grade 1	Gr 1	Gr 1						
Reception								
Foundat'n								

Target population by grade, as given in assessment program documentation

Additional grades for which the assessments may provide information

* Community-based assessments could include ASER in India and Uwezo in East Africa

^ PISA is an age based survey of 15 year olds. In most countries 15 year olds attend Grade 9 and/or Grade 10

Countries

In this step countries that are suitable sites for the validation linking studies are identified by the GAML Secretariat with input from relevant task forces and attempts are made to secure the participation of an appropriate number of them.

A list of regions considered as important for this work is given below.

- Africa – Northern
- Africa – Sub-Saharan
- Asia – Eastern
- Asia – South-Eastern
- Asia – Southern
- Asia – Western
- Oceania
- Latin America and the Caribbean
- Caucasus and Central Asia

To ensure that data are obtained from an adequate representation of countries in terms of geography, culture and language, it is suggested that at least two countries per region participate in the validation activities. The following points should guide the process of identifying and approaching potential countries:

- Countries that might make use of the metrics or that are at least representative of the types of countries that might make use of the metrics should be approached. Factors such as membership of the Global Partnership for Education (GPE) should therefore inform the decision about which countries to approach.
- Countries that have expressed interest in developing their capacity in learning assessment should be approached. Where possible, the validation activities should aim to make use of in-country expertise.

Efficiencies would be gained where countries participating in the linking exercises have already participated in the assessments from which the linking items are drawn, because these countries have existing materials on hand. However, these efficiencies must not dictate the range of countries that participate and a balance may need to be struck between efficiency and representation.

With the support of development partners, the GAML Secretariat with input from the relevant task forces – will canvass interest in potential countries. It is hoped that the governance structures from the involved assessment programs will support the efforts to secure country participation.

Once country participation has been secured, in-country Task Teams should be established. The work of these teams, which is central to the activities in step 5 (see page 9 below), is coordinated by the GAML Secretariat with input from relevant task forces. Where possible, the in-country Task Teams should be housed in existing educational research/learning assessment units within government, research institutions or other suitable bodies.

Step 2: Selecting the items

Once the involvement of assessment programs has been secured, a subset of items from each program needs to be selected. This step is overseen by the GAML Secretariat and guided by experts from the involved assessment programs, and should proceed via negotiation and collaboration. It is envisaged that this stage will provide opportunities for fruitful exchange and peer-to-peer learning. A possible process for selecting the items might be as follows:

- Following the guiding principles provided by the GAML Secretariat, experts from each involved assessment program select a set of items that represents the range of item difficulties and the knowledge, skills, contexts and abilities the program attempts to measure.
- Experts from each involved assessment program submit their selected set of items to the GAML Secretariat.
- Experts from each involved assessment program and the GAML Secretariat review the items together with the aims of:
    - ensuring the GAML Secretariat is adequately familiar with the items
    - ensuring there is minimal duplication in the items, in terms of the skills, knowledge and abilities they assess.
- Experts from each involved assessment program and the GAML Secretariat together establish a final item set for the program.
- Experts from all involved assessment programs and the GAML Secretariat come together to review the final items sets for all programs.

One key consideration in the selection of items for calibration is whether or not open-ended items are included.[6] Scoring is much easier, quicker and less costly if the number of open-ended items is minimised. However, this may not be a valid approach, because for many assessment programs, the range of skills that the assessments aim to measure will not be adequately covered without test-takers having to construct responses to some questions.

Step 3: Designing the tests

The previous steps will yield sets of items from a range of assessment programs as well as a group of participating countries. The first activity in this step will be to determine which sets of items from which assessment programs will be administered in each participating country. For example, in which participating country or countries might the set of items from SACMEQ be administered, and in which participating country or countries might the set of items from PIRLS be administered?

It will also be necessary to determine how many populations will be assessed in each participating country.

[6] An open-ended item is one where the test-taker is required to construct a response (as distinct from items requiring selection from given response options). For example, a test-taker may be required to give an opinion, summarise an argument, or show mathematical working.

These decisions will be based on the following questions:

- What are the existing patterns of assessment participation within the participating countries and can these patterns be used to enhance efficiency?
- What are the age or ability ranges that are of particular interest to each of the participating countries?
- How many responses to each item will be required to ensure robust linking?

After determining the mix of items to be calibrated in each participating country and the populations to be assessed, an appropriate technical test design should be developed by the GAML technical partner with input from the relevant task forces. For each population–country combination, the test design gives the testing time per child and the sequence of items in different test forms. It also shows how items will appear in multiple test forms, both within one population–country combination and across different population–country combinations, to facilitate linking.

Sample sizes for each population–country combination will not be known until after the technical test design has been completed. At this stage it is expected that sample sizes will be in the range of 500–1000 per population–country combination.

Step 4: Preparing test materials

After test designs have been developed for all population–country combinations, the next step involves test preparation activities.

Firstly, where a participating country does not have access to an appropriately adapted or translated version of an item within their country-specific linking pool, that item needs to undergo translation/adaptation. These processes should be subject to rigorous quality assurance to ensure linguistic and psychometric equivalence. A translation/adaptation partner selected by the GAML Secretariat with input from the relevant task forces, will undertake necessary translation and adaptation in consultation with experts from the involved assessment programs and personnel from the in-country teams in participating countries.

Once all items are available, test materials are prepared for each participating country according to the test design developed in the previous step. Note that what constitutes the test materials for each test implementation will be dependent 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. In some cases, materials may be computer-based and therefore appropriate rendering will be required.

Test materials are then printed according to predetermined quality standards. Printing could be undertaken separately in each participating country or centrally for all participating countries.

Step 5: Preparing for and undertaking data collection

The in-country Task Teams established in step 2 will be heavily involved in data collection. To ensure appropriate data collection procedures, they are supported and monitored by the GAML Secretariat with input from the relevant task forces.

In this step, preparations are made for the in-country activities. These preparations include:

- sourcing and training test administrators
- sampling children to sit the assessment
- taking steps to identify and secure appropriate sites for test administration
- sourcing and training data entry personnel (if applicable)
- sourcing and training coders (if applicable).

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

Sampled children will sit 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. If applicable, open-ended responses will be scored according to a suitable method.

Once all data have been captured and scored, they will be returned for analysis.

Step 6: Analysing data and setting benchmarks

In this step, the GAML technical partner with input from the relevant task forces analyses the performance data. It may be necessary for the GAML technical partners to undertake some preliminary data cleaning in collaboration with the in-country project teams before data analysis can begin.

Various modern psychometric techniques such as item response modelling will need to be employed.

Box 2: Method options for Test Calibration

The two most common methods that can be used to calibrate different assessments in non-equivalent groups with common items designs are equipercentile calibration (Kolen & Brennan, 2014) and item response theory (IRT) (Kolen & Brennan, 2014; Muraki et al., 2000; van der Linden & Hamleton, 1997). In non-equivalent groups with common items designs pairs of tests share common items but they also contain unique items. Further, they are administered to differing groups of students that may not have been sampled from a common population.

Equipercentile calibration

An equipercentile linking function is estimated and used to map the raw test scores of tests that contain subsets of common items to a shared scale. The linking function is estimated by ensuring

the distribution of mapped scores is equivalent for groups of students that are matched according to their scores on the common items.

This methodology is based upon raw test scores and allows the results from a network of tests sharing subsets of common items to be mapped to a shared scale.

Item Response Theory (IRT) calibration (Kolen & Brennan, 2014; Muraki et al., 2000; van der Linden & Hamleton, 1997)

IRT focuses in analysing data at the item level, rather than at the test level. IRT models assume that the ability of an examinee (the theta value, θ) is an unobserved variable that can be estimated from the responses to each item. IRT models relate the probability of correctly answering an item to examinee ability (θ) by estimating an item response function (IRF). Typically the IRF is characterised by one, two or three parameters. In the more restricted IRT model – known as the Rasch model or one-parameter model (1PL) – the probability of correctly answering an item depends only on examinee ability and item difficulty. Therefore, examinees with the same estimated ability should have the same probability of getting a given item correct.

If tests share items then IRT approaches can be used to calibrate the items so they are located on a common scale If item parameters are estimated by separate calibration, the common items are used to estimate the necessary scale transformation. If a concurrent calibration is performed, no transformations are needed and parameters are on the same scale in a single run. Once all item parameters have been placed in the same scale, they are referred as being calibrated. Items. Any test instrument that contains items in the pool of calibrated items can then be reported on same metric (scale).

An analysis plan that describes models to be used for scaling and sets standards for acceptable results will have been previously prepared by the GAML technical partner. This plan will explain how items whose traditional scoring is not readily amenable to item response modelling (eg the items from ASER) will be treated during the analysis. When this step begins it may be necessary to adjust the plan and experiment with alternative methodologies to achieve the desired outcomes. This step may provide a good opportunity for the GAML Secretariat with input from the relevant task forces – to offer workshops for capacity development to participating assessment programs/countries.

This stage will also involve the setting of the benchmarks. This is an activity that requires collaboration between the GAML Secretariat, the organisations responsible for the assessments and curriculum experts 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 at this stage to include representatives from other countries that intend to make use of the metrics. A discussion paper concerning benchmarking has been prepared as: *Setting benchmarks on the UIS reporting scales.*

Step 7: Mapping assessment results onto the metrics and dissemination of results

After the data have been analysed, the metrics validated and the benchmarks set, the final step in the validation is the preparation of materials enabling countries to report against the metrics. These materials describe how both countries that are involved in the original assessment programs and countries that contributed items to the linking studies can map their results onto the metrics should they wish to do so.

The GAML Secretariat with input from relevant task forces, will prepare this material in collaboration with the involved assessment programs.

Key next steps and indicative overall timeline

Before the validation can be planned or costed in more detail, there are two decisions that must be made:

- Which assessment programs will be involved? (Step 1)
- Which countries will participate? (Step 1)

Activities directed towards making these two decisions therefore feature at the start of the indicative timeline given in Table 3.

Table 3: Indicative overall timeline for

Step	Activity within step	Target date
Step 1	Extend requests to the governance structures of assessment programs that could contribute items to the linking item pool and negotiate their involvement	March Y1
	Establish final list of involved assessment programs	31 March Y1
	Identify possible participating countries and canvass interest	March – April Y1
	Establish final list of participating countries	30 April Y1
Step 2	Consult with experts from assessment programs to determine item pool	April – July Y1
	Establish final item pool	31 July Y1
Step 3	Design tests for each participating country ie determine items from which assessments will be administered in each participating country	August – September Y1
Step 4	Translate and adapt items[7]	September Y1 – February 2018
	Prepare and print test 'booklets' and associated assessment materials (eg training and administration manuals)	November Y1 – April Y2
	Develop processes for data capture (eg create data entry application or plan for scanning)	November – December Y1
Step 5	Source and train in-country test administrators	January – August Y2
	Select and secure suitable locations for test administration	January– August Y2
	Sample children	January– August Y2
	Administer tests	February – September Y2
	Code responses	March – October Y2
	Capture data (ie data entry or scanning)	April – November Y2
Step 6	Clean data	JuneY2 – January Y3
	Analyse data	August Y2 – March Y3
	Consult on setting of benchmarks	March– June Y3
	Set final benchmarks	30 June Y3
Step 7	Prepare document describing how results of involved assessment programs can map onto the metrics	June – July Y3
Completion		**Target date**
Complete validated learning metrics for mathematics and reading, including benchmarks and descriptions of how assessment results can map onto the metrics		31 July Y3

[7] It is expected that from this point onwards activities will be staggered to reflect the fact that different participating countries are likely to have different translation and adaptation loads.

References

Cizek, G., & Bunch, M. (2007). *Standard Setting: A guide to establishing and evaluating performance standards on tests*. California, London, India: SAGE.

Cizek, G., Bunch, M., & Koons, H. (2004). Setting Performance Standards: contemporary Methods. *Educational Measurement: Issues and Practice, 23*(4).

Feuer, M., Holland, P., Green, B., Bertenthal, M., & Hemphill, F. (1999). Uncommon measures: Equivalence and linkage among educational tests. Washington, DC: National Academy of Sciences - National Research Council, Washington, DC. Board on Testing and Assessment.

Kolen, M. (1988). Traditional Equating Methodology. *Educational Measurement: Issues and Practice, 7*(4), 29-37.

Kolen, M., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. New York: Springer.

Linn, R. L. (1993). Linking Results of Distinct Assessments. *Applied Measurement in Educational Measurement, 6*(1), 83-102.

Mislevy, R. (1992). Linking educational assessments: Concepts, issues, methods, and prospects. Princeton, NJ: ETS Policy Information Center.

Muraki, E., Hombo, C., & Lee, Y.-V. (2000). Equating and Linking of Performance Assessments. *Applied Psychological Measurement, 24*(4), 325–337.

van der Linden, W., & Hamleton, R. (Eds.). (1997). *Handbook of Modern Item Response Theory*. New York: Springer.