Mary Lou Fulton Teachers College
Arizona State University

Center for Advanced Studies in Global Education

Options in achieving global comparability for reporting on SDG 4

Silvia Montoya and Brenda Tay-Lim

CASGE WORKING PAPER #3 | 2018

The views expressed herein are those of the authors and do not necessarily reflect the views of CASGE.

Options in achieving global comparability for reporting on SDG 4

http://dx.doi.org/10.14507/casge3.2018

CASGE working papers are circulated for discussion and comment purposes. They have not been peer-reviewed.

Table of Contents

Introduction

The Sustainable Development Goals for Education (SDG 4) indicator 4.1.1 calls on member states to report on the "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."

There is ample research about optimal approaches for achieving global comparability for reporting on SDG 4, and the factors that could influence these choices. It is now abundantly clear that selecting a global data collection strategy is a technically complex matter, with serious financial and behavioural implications at various levels, and could easily be contested. Previous approaches put forward differ primarily in terms of their technical complexity, financial cost, and implied comparability of national statistics. Less obvious differences relate to their sustainability over time, their impact on the politics, planning, and operations of national education authorities, their ability to contribute to capacity building within these authorities, and their persuasive power in the media and policy debates. Existing proposals could be taken forward in several ways. Hybrid approaches are feasible and may be the most practical.

Many more countries have participated in cross-national assessments and conducted national learning assessments in recent years. Among countries that conduct learning assessments, variations in content coverage and quality of reporting exist. How do countries report on SDG 4 indicator 4.1.1 when there are diverse content, quality and reporting metrics? What are the standard features of assessment systems that could help to produce comparable results? What are the implementing issues that even well-designed assessment programs could not overcome? Finally, how do we build a pragmatic system that could produce comparable results that allow for trend reporting? Some key features guide the reporting framework. This reporting framework should accommodate results from different methods and metrics while at the same time respect country ownership, meet national needs. and be sensitive to country and cultural contexts.

Current Landscapes

The UNESCO Institute for Statistics (UIS) has the mandate to "work with partners to develop new indicators, statistical approaches and monitoring tools to better assess progress across the targets related to UNESCO's mandate, working in coordination with the Education 2030 Steering Committee." In particular, as the custodian agency for SDG indicator 4.1.1, the UIS is coordinating the development of methodologies and data reporting tools to ensure the capturing provision of "equitable and quality primary and secondary education leading to relevant and effective learning outcomes."

According to UIS 2018 estimates, 81% of countries have participated in a cross-national initiative in the last five years. Given the amount of information available for indicator 4.1.1, the international community has discussed strategy to use existing data to report on SDG 4. Even though many countries have conducted national assessments, the quality and scope of national assessment data varies considerably. A potential solution to the challenge of learning outcomes data from different sources is to link the different assessments in some manner, rather than insisting on a single universal assessment. Linking is understood as the statistical machinery that enables countries and the larger education community to interpret evidence about student achievement in a coherent manner. UIS's objective for indicator 4.1.1 implies, among others, finding ways to link different assessment results and to report them in a globally comparable way.

Currently, the UIS has an approach to report on indicator 4.1.1 in the short term (2018, and likely 2019) based on the following principles, recognizing that linked assessments is the goal for the longer term:

- Be as pragmatic as possible while being as rigorous as possible; and
- Build on existing work and what work is available.

In practice, UIS will accept data in the short term that is not perfectly aligned with 4.1.1 or comparable with other countries, but that broadly meets the needs for reporting against 4.1.1. UIS starts with data from international and regional assessments, but allows countries to submit national assessment data if they choose. If a country does not respond to UIS's request for data then UIS will decide which data source to use. It also accepts other learning assessment data or even unofficial data, like Citizen-Lead Assessment, in the short term in order to cover data gaps.

Currently, UIS reporting approach includes the following assumptions[1]:

- report the definition of reading and mathematics as proposed by each assessment;

[1] Reference document in the Fifth Global Alliance to Monitoring Learning meeting, 17-18 October Hamburg, Germany – SDG Data Reporting: Proposal of a protocol for reporting indicator 4.1.1 - http://gaml.uis.unesco.org/wp-content/uploads/sites/2/2018/10/4.1.1_27_Interim-reporting-strategy-protocol.pdf, dated 22 October 2018.

- report on in-school students, with the exclusions taken by each assessment, as well as the target grade (with -1/+1 grade at end of primary and -2/+2 grade at end of lower secondary, if the target grade is not cleanly defined);
- identify any assessment used in 4.1.1 reporting that includes children or young people outside of school;
- preface Indicator 4.1.1 reporting with a clear explanation that assessment programs may measure varying levels of learning progress;
- report any major operational issues, in consultation with education systems;
- report the results of this analysis, which will be collected through Catalogue of Learning Assessments Module 2;
- work with education systems to ensure that technical documentation about scaling is available in the public domain, for any assessment programs used in 4.1.1 reporting (e.g., via the Catalogue of Learning Assessments);
- preface Indicator 4.1.1 reporting with a clear explanation that assessment programs may define minimum standards of proficiency in different ways; and
- report on the periodicity of each assessment, and if it is longitudinally equated.

The proposed interim approach to reporting includes recommendations for how the data on the percentage of students meeting minimum proficiency standards (for the relevant domain and measuring point) will be footnoted. Footnotes will denote the data source and how it was selected, population covered, whether data is based on an assessment that is longitudinally equated, and whether out-of-school youth are included in the estimate.

Four suggested options in linking

In order to improve coverage and comparability for reporting indicator 4.1.1, statistically linking international and regional assessments was identified as a viable first step. The first baseline for this approach will be in 2019 when TIMSS and many regional assessments will be implemented.

Four suggested options in linking learning assessments

In the context of finding ways to link different assessment results and to report them in a globally comparable way, the UIS convened a meeting with the technical experts of the regional and international assessment agencies. Three methodologies were proposed and discussed with the last strategy (2c) suggested by the UIS as a form of validation of other approaches.

Strategy 1 Non-Statistical approach

1. **Policy linking**: pedagogically informed recalibration of existing data

Strategy 2 Statistical approach

2a. **Item-based linking**: Psychometrically informed recalibration based on common items

2b. **Test-based linking**: Recalibration by running a parallel test on a representative sample of students

2c. **Statistical alignment**: Recalibration of existing data base on countries who have participated in more than one cross-national assessment

Strategy 1: Non-Statistical approach - Policy linking[2]

A non-statistical approach called social moderation (SM) or policy linking was proposed as an option by MSI. Although a non-statistical approach, it is not a qualitative method and does provide statistics to allow judgement on the quality of alignment. This procedure uses definitions of proficiency levels for reading and mathematics to produce a reporting scale, called a proficiency scale (for mathematics and reading respectively), and a mechanism for linking existing assessments and their performance levels to this scale. Several steps are involved in constructing proficiency scales, called the UIS Proficiency Scales, or UIS-PSs. A toolkit and mechanism is under-development to facilitate the linking of national assessments (NAs) and cross-national assessments (CNAs) to that scale.
In brief, the six steps involved:

---

[2] Reference document in the Fifth Global Alliance to Monitoring Learning meeting, 17-18 October Hamburg, Germany - SDG 4 Reporting: Linking to the UIS Reporting Scale through Social Moderation - http://gaml.uis.unesco.org/wp-content/uploads/sites/2/2018/10/4.1.1_22_Linking-to-the-UIS-Reporting-Scale-through-Social-Moderation.pdf, dated 22 October 2018.

1. **Define Content Standards**: what students are expected to learn in reading and mathematics at the three education levels, i.e., grades 2/3, end of primary, and end of lower secondary.
2. **Determine Performance Levels**: number of categories and names for levels (e.g., levels separated by minimum proficiency in reading and mathematics at each education level).
3. **Develop Policy Definitions (PD) of Performance:** what students should demonstrate in each category, in generic terms and not by subject area, at each education level.
4. **Develop Performance Level Descriptors (PLDs):** what students should demonstrate in reading and mathematics (details of knowledge, skills, abilities) at each education level.
5. **Develop Proficiency Scale Maps**: how performance levels of various NAs and CNAs map to the UIS-PSs in reading and mathematics at each education level.
6. **Develop Socially-Moderated Performance Standards**: what students need to score on NAs and CNAs in reading and mathematics at each education level for placement into category.

Outputs 1-4 facilitate constructing UIS-PSs, and outputs 5-6 facilitate linking the UIS-PSs with NAs and CNAs. The steps and outputs are briefly described below.

3.1 Steps

Step 1: Define Content Standards. In order to develop stand-alone reporting scales for each of the three education levels in reading and mathematics (i.e., six scales), the first step is to define the content standards for each domain and for each grade span of K-3, 4-6, and 7-9 separately. The common content standards are predefined knowledge and skills that students are expected to learn in reading and mathematics by the end of grades 3, 6, and 9 across countries. The UNESCO's International Bureau of Education (IBE-UNESCO) and the UIS have collaboratively made significant progress in describing these content standards for each domain and grade. Based on the review and analysis of over 115 Mathematics and 76 Reading national assessment frameworks, and building on the learning cognitive theory, the global content framework (GCF) were developed for each domain. These GCFs are the basis of the content standards.

Step 2: Determine Performance Levels. The UIS built a preliminary Proficiency Scale (PS) for each domain based on the compiled cross-national (regional and international) assessment's[3] PLDs. The PS was reviewed by assessment experts from the regional and international assessment agencies[4] in a Consensus-Building meeting in Paris in September 2018. During that meeting, the group identified four performance levels and provided a name to each level.

[3] Assessments included EGRA, EGMA, MICS6, ASER, Uwezo, PASEC, PILNA, SACMEQ, TERCE, TIMSS, PIRLS, and PISA.

[4] Representatives from ASER, CONFEMEN, LLECE, IEA, SEAMAO, OECD, UNICEF, and Uwezo.

Step 3: Develop Policy Definitions (PD) of Performance. The group also agreed on a generic policy definition for each performance level. These definitions are not linked to content but are more general statements that assert policymakers' position on the desired level of performance.

Step 4: Develop Performance Level Descriptors (PLDs). Using existing cross-national assessments as a starting point, the UIS rank-ordered the PLDs from all cross-national assessments for each education level and each subject area (reading and mathematics)[5]. The full descriptions will express the knowledge and skills required to achieve the performance levels. They will be used to provide stakeholders with more information on what students at each performance level should know and be able to do, as well as what they need to know and be able to do to reach next performance level.

Steps 1-3 have already been completed. **Step 4** is still work-in-progress. Once completed it would be used as reference for linking

Step 5: Develop Proficiency Scale Maps. After performance levels of UIS-PSs for each grade and domain are determined from cross-national assessments, the next step is to link the UIS-PSs for each education level and subject area with corresponding NAs and CNAs for SDG 4.1.1 reporting. The different assessments can be linked through the PLDs. This process is called social moderation (SM) or policy linking (Buckendahl & Foley, 2015[6]).

Step 6: Develop Performance Standards. In an earlier step, assessment experts identified the performance level that is considered as the minimum proficiency. A socially moderated cut score for the NAs and CNAs will be established using a standard-setting method in order to link these NAs and CNAs with the UIS-PSs for each education level and subject area. Different standard-setting method could be used for estimating the cut score on the NAs and CNAs. The selection of the standard setting method mostly depends on item formats of the NAs or CNAs. In general, if the test contains only multiple-choice items, then the yes-no variation of Angoff method would be used. If the test contains both multiple and open-ended items, then the Bookmark or the Body of Work method would be used.

The cut score separates below minimum proficiency level from achieved minimum proficient level for SDG 4.1.1 reporting. In other words, the students classified into the achieved minimum proficient performance levels of the UIS proficiency scale would demonstrate required knowledge and skills assessed on the NAs and the CNAs. The steps range from defining the content standards and determining performance levels to developing the socially moderated performance standards. These performance standards for each assessment would facilitate SDG 4.1.1 reporting in a relatively short amount of time with lower cost and no test security issues.

---

[5] The Proficiency Scale and the Performance Level Descriptors are still under validation.

[6] Buckendahl, C. W., & Foley, B. P. (2015, April). Policy linking as cut score moderation: Considerations for practice. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Strategy 2a. Statistical approach – Psychometric item-based linking[7]

Another option proposed by the Australian Council for Educational Research (ACER) was developed for reporting against SDG 4.1.1 and covers learning from the foundational level to mid-secondary, representing a range of difficulties and content. The reporting scales are designed to provide an overarching framework that combines key concepts and skills found to be important in cross-national assessments and countries' curricula. The scales describe learning progressions in reading and mathematics and can locate the distribution of learning observed.

3.2 Phases of development

Phase I of the development of the reporting scale (RS), which is the development of draft reporting scales, has been completed and the draft scales have undergone a broad review process. The conceptual frameworks for the scales are based on international and regional assessment frameworks and items—including PASEC, SACMEQ, LLECE, PILNA, TIMSS Numeracy, PIRLS Literacy and others—and assessments from a broad range of countries (e.g., ASER, Uwezo, and Afghanistan's MTEG).

The development included:

1. developing a conceptual framework based on assessment frameworks, curriculum documents, and the relevant learning domain literature;
2. an analysis of cognitive demands of the items and comparisons of item difficulty;
3. a qualitative validation by comparing with other existing reading and mathematics scales.

The review process involves gathering feedback on the domain and level descriptions, an example of skill illustrations, and overall coverage of key concepts in reading and mathematics.

Phase II focuses on equating existing assessments (e.g., international or regional assessments) with the RS using item-based equating, establishing a pool of calibrated items that can be included in an assessment (e.g., a national assessment) such that that assessment can be linked to the RS, and validating the RS.

Equating existing assessments with the RS will allow any country using one of the equated assessments to report against the RS directly and to understand how that assessment and its standards align with the RS. Further, the equating will establish a pool of calibrated items that can be embedded into an assessment that is not already equated to the RS and to determine how that assessment aligns with the RS.

The validation process will involve multiple linking exercises across 10-15 countries. In each country, sets of items from the involved assessment program will be selected and administered to one or more samples of

---

[7] Reference document in the Fifth Global Alliance to Monitoring Learning meeting, 17-18 October 2018, Hamburg, Germany – UIS Reporting Scales concept notes - https://www.acer.org/files/UIS_Reporting_Scales_concept_note.pdf, dated 22 October 2018.

children. Each sample represents the target population of interest. After all separate linking exercises are completed, all items that were included will together form a pool of calibrated items—this will be a central tool in the future use of the RS.

The item-based equating also provides data to empirically validate the RS since the draft scales were developed based on a conceptual empirical analysis of item difficulties and the equating would provide empirical validation at the country level. While less technically rigorous than test-based equating, or item-based equating, which relies on non-equivalent groups and common items, has advantages in terms of reducing the burden on countries and children and it is more flexible. The resulting pool of items can be used to link other assessments to the RS if it is chosen as the scale to report on SDG 4.

Strategy 2b. Statistical approach—Psychometric test-based linking[8]

A third option proposed by the International Association for the Evaluation of Educational Achievement (IEA) presents a strategy to establish a link between the results on regional assessments conducted at the primary level and the TIMSS and PIRLS International Benchmarks for numeracy and literacy. There are five regional assessment agencies planning reading and mathematics assessments at the end of primary schooling in 2018 or 2019:

- SACMEQ – Southern and Eastern Consortium for Monitoring Educational Quality
- PASEC – Programme for the Analysis of Educational Systems
- LLECE – Latin American Laboratory for the Assessment for the Quality of Education
- SEA-PLM – Southeast Asia Primary Learning Metrics
- PILNA – Pacific Island Literacy and Numeracy Assessment

The reading and mathematics assessments planned for 2018/19 provide a perfect opportunity to link these regional assessment results to IEA's TIMSS and PIRLS achievement scales. These regional assessments measure achievement at the sixth grade, except SEA-PLM, which tests the fifth grade. The content of the regional mathematics assessments aligns well with the TIMSS fourth grade assessments of numeracy and mathematics. Similarly, the content of the regional reading assessments aligns well with the PIRLS fourth grade assessment of literacy and reading comprehension.

The overarching idea is to construct a concordance table that translates between the scores on each of the regional assessments in mathematics and reading and scores on TIMSS and PIRLS, respectively. The concordance table

---

[8] Reference document in the Fifth Global Alliance to Monitoring Learning meeting – IEA's Rosetta Stone: Measuring global progress toward the SDG for quality education by linking regional assessment results to TIMSS and PIRLS international benchmarks of achievement - http://gaml.uis.unesco.org/wp-content/uploads/sites/2/2018/10/4.1.1_24_IEA's-Rosetta-Stone-Proposal.pdf, dated 22 October 2018.

provides a translation from the countries' regional assessment results to the TIMSS and PIRLS achievement scales or a link between regional assessments and the TIMSS and PIRLS achievement scales. The countries participating in the regional assessments can use the translations to determine what percentage of their students could be expected to reach the TIMSS and PIRLS International Benchmarks.

This involves IEA (the agency that develops TIMSS and PIRLS) to work with the study centers for each of the five regional assessments. The proposal is to have a subset of countries (3-5) from each regional assessment administer selected booklets of TIMSS and PIRLS achievement items at the same time as their upcoming regional assessments. Depending on the level of mathematics and reading achievement in a region, the booklets can be tailored to contain primarily items assessing TIMSS Numeracy and PIRLS Literacy. The same students should take the regional mathematics and reading assessments and then the TIMSS and PIRLS booklets, preferably on the following day. The combined data across the 3-5 countries will provide scores on both the regional assessment and TIMSS and PIRLS for approximately 15,000 students from the region, which can be used to construct the concordance tables for numeracy and literacy achievement. Since the concordance tables provide a projected TIMSS or PIRLS score for all possible regional assessment scores, it will be possible to determine the regional assessment scores equivalent to each of the TIMSS and PIRLS International Benchmarks. For each country participating in a regional assessment, progress toward an International Benchmark can be estimated by the percentage of students reaching the regional assessment score equivalent to the International Benchmark. For example, a country may want to determine the percentage of students reaching the Low International Benchmark. Hypothetically, if the concordance table showed that a regional assessment score of 562 in reading was equivalent to 400 on the PIRLS reading scale, then all students in the country reaching 562 could be considered to have reached the Low International Benchmark. Although based on data from the 3-5 countries that participate in the linking study, the concordance table and the benchmark equivalent scores can be applied in all the countries in the regional assessment (whether they participated in the linking study or not).

Although this option is the most statistically rigorous, since it involves the same students taking the same test, it does have its own disadvantage in term of cost and complexity. However, it has more political supports than item-based approach.

Strategy 2c. Statistical approach – alignment through statistical modeling[9]

In this option, UIS explores the database and method that Altinok (2017) has developed. This method of harmonizing data based on countries who participated in more than one cross-national assessment is suggested as a way to validate results based on psychometric approach mentioned above. The methodology builds on literature that aims to produce comparable estimates of cognitive skills across countries and over time.

---

[9] Reference document in the Fifth Global Alliance to Monitoring Learning meeting – Mind the gap: Proposal for a standardised measure for SDG 4 – Education 2030 agenda - http://uis.unesco.org/sites/default/files/documents/unesco-infopaper-sdg_data_gaps-01.pdf dated 22 October 2018.

International and regional assessments differ in the definition of what students should know and skills tested. This method built on the methodology and database presented in Altinok, Angrist, and Patrinos (2017)[10] to create comparable estimates across various international and regional assessments. The idea is to use countries who took part in several assessments and use them as anchored countries. In this approach, Altinok et al. (2017) links international assessments such as PISA, TIMSS, PIRLS, and their precursors, as well as regional student achievement tests (RSATs), such as MLA, LLECE, SACMEQ, or PASEC3. This enables a database on the quality of student achievement for the largest set of countries to date, and the largest number of developing countries. This method allows the largest globally comparable panel database of cognitive achievement, including 163 countries and regions, 32 of which are from Sub-Saharan Africa, over the last 50 years (1965-2015).

[10] Altinok, N., Angrist, N., & Patrinos, H. (2017). *A Global Data Set on educational Quality (1965-2015).* Educational Global Practice, World Bank Group.

Summary of each approach: pros and cons

Strategy 1. Non-statistical approach—Policy linking: pedagogically informed recalibration of existing data

The approach involves using the proposed proficiency framework that describes the range of competencies that children/youth have at each level to locate proficiency levels from alternative assessment programmes based on the Performance Level Descriptors (PLDs). This approach of linking is guided by experts' judgement. This proposal would allow the expansion of coverage in terms of educational systems reporting for SDG 4. For instance, coverage at the primary level would double, in terms of the population-weighted world, if national assessments were included. Furthermore, it is more cost-effective as compared to the statistical psychometric approach. However, the error base on expert judgement cannot be compared to those sampled students who took the test.

Strategy 2. The statistical approach

2.a. Item-based linking: Psychometrically informed recalibration based on common items

This approach implies the use of common items in different assessment programmes. One version has been proposed by the Australian Council for Educational Research (ACER) as part of an overall proposal of progression in learning, but options are not exhausted.[11] However, this has proven to face technical, political, and other difficulties in implementation. As it is costly to conduct phase II equating and validating to generate calibrated item pool, and regional assessment agencies are not willing to share their items for the construction of its RS.

2.b. Test-based linking: Recalibration by running a parallel test on a representative sample of students

The IEA outlines the concordance table solution that deals only with the primary level and allows two assessments, one international and one regional, to be expressed on the same scale. Concretely, the proposal states that sub-samples of students in three to five countries per regional programme would write not just the regional tests, but also IEA's test. This would produce a "concordance table" based on psychometric modelling.[12] The table is not the reporting scale, but it facilitates SDG reporting by expressing a larger number of countries in the same scale. The outcomes of this still need to be put on a UIS-PS in order to report for 4.1.1. It is considerably costly, but it builds capacity in country to implement large-scale assessment.

2.c. Statistical alignment: Recalibration of existing data

This approach relies largely on statistical adjustments taking advantage of the fact that some countries, referred to as "doubloon countries," participate in more than one cross-national programme. Using several such overlaps allowed for the identification of roughly comparable proficiency thresholds. It could serve as a validation but it is unlikely to have political buy-in.

---

[11] Note that the reference scale is built from items coming from various assessments.

[12] For countries the option is to either participate in a regional programme or in a global programme (something that might be difficult or not possible if the region does not have a regional initiative).

Conclusion

Each methodological approach put forth carries different technical complexities, financial costs, and burdens on national authorities. Any measurement system must be sensitive to detect small changes. An all-or-nothing plan relying on the ideal approach is not pragmatic. It is more realistic to prioritise the measurement in a staggering hybrid approach, moving from easy to more complex and reliable methodology over time, and from weaker to better data collection systems.

The efforts described examine the different options for reporting. These options should be taken more as complementary routes than as mutually exclusive options in order to minimize risk, if some of the approaches prove to be too costly, the margin of error too high, politically unfeasible, or a combination of all these. The strategies help each other to build a sustainable reporting strategy, where it is easier to see steps between Strategy 1 and Strategy 2a and complementarity between Strategy 2b and Strategy 1, such as the concordance table results, which need to be expressed in a proficiency framework. Strategy 2c has a potential use as a check to compare statistics based on national assessments.[13] Given the variation in learning assessments regarding content, construct, and methodology used for data collection and reporting, this reporting framework shows three approaches of linking assessments that could produce comparable data and improve coverage of reporting for SDG 4.

[13] A third strategy could be a new test that everybody takes for reporting using a common comparable tool, but this is neither politically feasible nor cost-efficient so it has not been pursued.

The Center for Advanced Studies in Global Education
P.O. Box 871611, Tempe, AZ 85287-1611
CASGE@asu.edu
(480) 727-5346
education.asu.edu/casge

@asucasge

Mary Lou Fulton Teachers College
Arizona State University

Center for Advanced Studies in Global Education