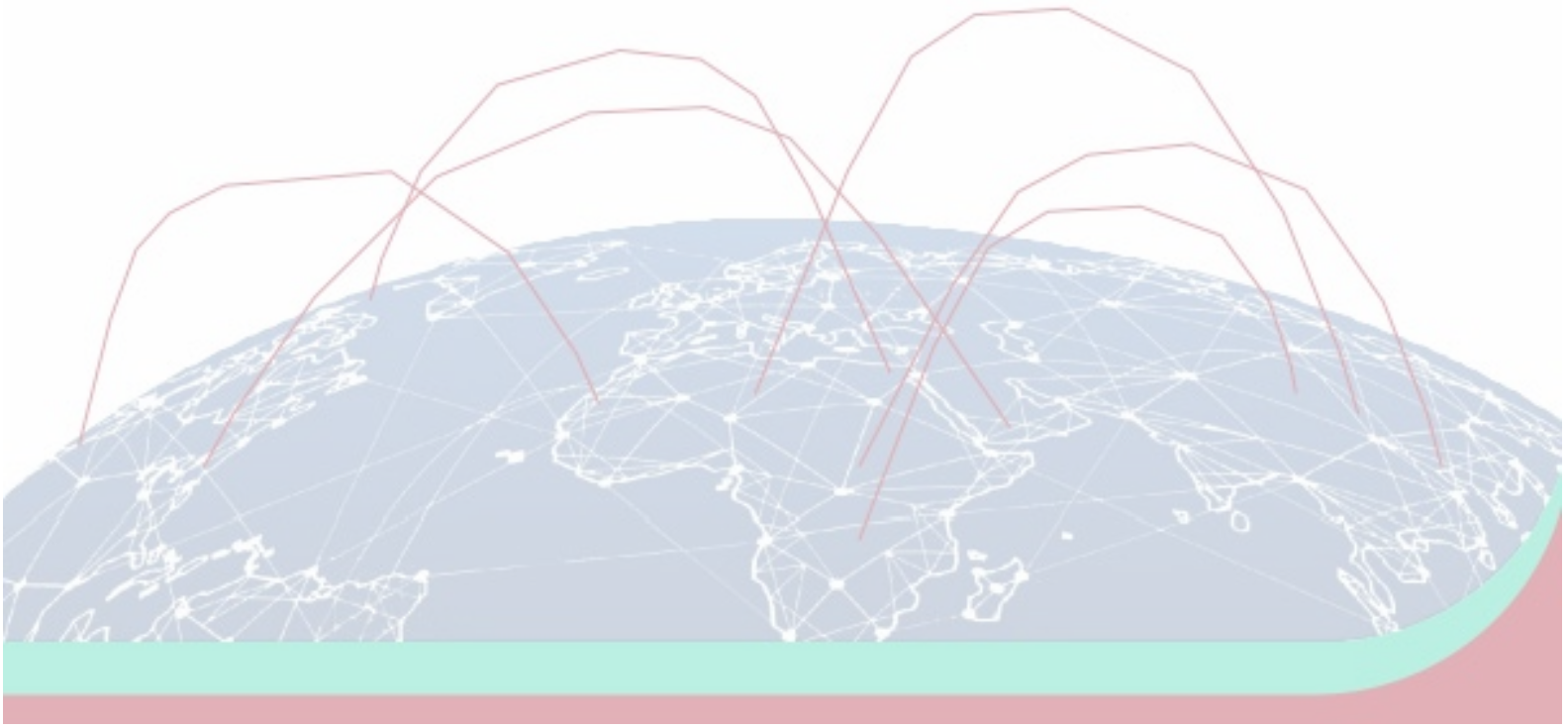
WG/GAML10/ST.REP/2

Eligibility criteria for reporting against SDG 4.1.1

A discussion paper – GAML 2023

Tenth meeting of the Global Alliance to Monitor Learning (GAML)

Paris, 6 - 7 December 2023



Eligibility criteria for reporting against SDG 4.1.1

A discussion paper – GAML 2023

Colin Watson

November 2023

ACER

Background and purpose

Sustainable Development Goal (SDG) 4 aims to ensure that, by 2030, "all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes." UNESCO Institute of Statistics (UIS) is the custodian of indicator 4.1.1, which concerns the proficiency indicator referring to three levels of schooling: early grades, end of primary, and end of lower secondary; and two subjects (reading and mathematics). The indicator reads as follows:

> "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 [MPL] in (i) reading and (ii) mathematics, by sex."

The MPLs were formulated by a group of experts at the Concensus Building Meeting held in September 2018, endorsed by the GAML Fifth Meeting in November 2018, and approved by the TCG Fifth Meeting in November 2018. The agreement established definition of the MPL and mapped the assessment programs suitable for the reporting and the proficieny level within each one compatible with that definition. Since then, the UIS and its partners have produced a number of methodological tools to measure learning in a comparable way, such as the Global proficiency framework for reading and mathematics, the Protocol for reporting SDG indicator 4.1.1, the Metadata document for SDG indicator 4.1.1, and a set of linking approches to MPLs (Rosetta Stone, Statistical linking – AMPL-, Policy Linking, Pairwaise Comparison Method), among others.

To date, however, GAML has not proposed the criteria for assessments in order to be eligible for reporting in terms of content coverage. Some criteria elements have been published in the policy linking toolkit (PLT) and the pairwise comparison method (PCM) document, but they have neither been endorsed by the GAML not approved by the TCG.

This paper proposes the eligibility criteria for assessments that wish to report against SDG 4.1.1. They are based on the criteria in the PLT and the PCM toolkit, which were developed following extensive consultation with groups of experts involved in implementing policy linking across the world.

Criteria overview

This paper proposes five criteria to determine if the assessment is sufficiently valid for reporting against SDG 4.1.1. These criteria were selected to ensure consideration of the quality of the assessment instrument and its implementation. For each criterion, there are essential minimum requirements for an assessment to be considered suitable for SDG reporting.

The five criteria relate to the following:

- **Criterion 1** – is the assessment sufficiently aligned to the MPL?
- **Criterion 2** – is there evidence that the items in the assessment have been reviewed qualitatively and quantitatively to determine their suitability for inclusion in the assessment?
- **Criterion 3** – is the sample of learners that took the assessment representative of the population against which the results will be reporting?
- **Criterion 4** – is there evidence that the assessment was administered in a standardised way?

- **Criterion 5** – are the outcomes of the assessment sufficiently reliable?

The evidence used to support the judgements against each of the criteria should ideally be in the public domain to facilitate a transparent process.

For criterion 1, several options are proposed for MPLa for discussion and decision.

Criterion 1 – Alignment

To be suitable for reporting against SDG 4.1.1, the assessment instrument must assess sufficiently and reliably similar knowledge and skills to those that are described in the relevant MPL in an appropriately comprehensive way.

This means that the assessment must contain a minimum of **20 items in total** (though it is likely to contain more) that assess the appropriate domains for the level of schooling in a sufficiently broad, and deep way. To determine breadth and depth, it is proposed to use the equivalent grades of the Global Proficiency Framework (GPF) that have been determined to link most closely to the MPLs (grade 2 for MPLa, grade 5 for MPLb and grade 8 for MPLc). This is because the domains, construct and subconstructs in the GPF provide a more detailed framework against which assessment instruments can be assessed.

Evidence for how the assessment aligns to the criterion must be made available to UIS.

MPLa

The issue for MPLa is that, depending on the country, learners may still be focused on developing foundational skills rather than the main constructs that are required for the indicator. In reading, this means that learners in some countries are still focusing on decoding and listening comprehension/comprehension of spoken or signed language to support their reading comprehension whereas in others, assessments focus solely on reading comprehension. In mathematics, this means that learners in some countries are only focussing on number and operations whereas in others, assessments include all mathematical domains.

This has led to complications in developing the eligibility criterion, with proposals for different levels of alignment, that have been determined to be confusing. The options presented below are intended to simplify the criterion, though each has advantages and disadvantages, which are discussed at the end.

Option 1

In this option, alignment is linked to the MPL, requiring all domains to be assessed:

- **Reading** – there should be a minimum of 10 score-points assessing *decoding*, 5 score-points assessing *listening comprehension*/comprehe*nsion of spoken or signed language* and 5 score-points assessing *reading comprehension*. The assessment must also cover 5 of the 9 subconstructs at grade 2 in the GPF.
- **Mathematics** – there should be a minimum of 10 score-points assessing *number and operations*, 5 score points assessing *measurement* and *geometry* and 2 score-points assessing *statistics and probability* and *algebra*. The assessment must also cover 7 of the 14 subconstructs at grade 2 in the GPF.

Option 2

In this option, alignment is linked to the minimum number of domains that would allow the most countries to report:

- **Reading** – there should be a minimum 10 score-points assessing *reading comprehension* and the assessment must cover both *reading comprehension* subconstructs at grade 2 in the GPF. The remaining items, to meet the minimum 20 items required can be drawn from any of the domains (*decoding*, *listening comprehension*/comprehe*nsion of spoken or signed language* or *reading comprehension*).
- **Mathematics** - there should be a minimum 10 score-points assessing *number and operations* and the assessment must cover all four *number and operations* subconstructs at grade 2 in the GPF. The remaining items, to meet the minimum 20 items required can be drawn from any of the domains (*number and operations*, *measurement*, *geometry*, *statistics and probability* or *algebra*).

Advantages and disadvantages

	Advantages	Disadvantages
Option 1: All domains	• The domain requirements align with the description in the MPL • Encourages countries where learners are still focused on foundational skills in reading to include all domains in their assessments	• Countries that assess with an assessment of reading comprehension only would not be able to report against SDG 4.1.1a • Countries that are focused only on foundational skills in mathematics (number and operations) would not be able to report against SDG 4.1.1a
Option 2: Minimal domains	• Maximises the number of assessments that can be used for reporting	• Assessments being used for reporting are likely to contain different domains outside the core elements required

MPLb

- **Reading** – the minimum 20 items must all relate to the *reading comprehension* domain. There should be 5 score-points assessing the *retrieve information* construct and 5 score-points assessing the *interpret information* construct from the GPF. The assessment should also cover 4 of the 8 *reading comprehension* subconstructs at grade 5 in the GPF.
- **Mathematics** – there should be a minimum of 10 score-points assessing *number and operations*, 5 score-points assessing *measurement* and *geometry* and 5 score-points assessing *statistics and probability* and *algebra*. The assessment must also cover 12 of the 21 subconstructs at grade 5 in the GPF.

MPLc

- **Reading** – the minimum 20 items must all relate to the *reading comprehension* domain. There should be 5 score-points assessing the *retrieve information* construct, 5 score-points assessing the *interpret information* construct and 5 score-points assessing the *reflect on information* construct from the GPF. The assessment should also cover 5 of the 10 *reading comprehension* subconstructs at grade 8 in the GPF.

- **Mathematics** – there should be a minimum of 10 score-points assessing *number and operations*, 5 score points assessing *measurement* and *geometry* and 5 score-points assessing *statistics and probability* and *algebra*. The assessment must also cover 12 of the 21 subconstructs at grade 8 in the GPF.

Criterion 2 – item review

To be suitable for reporting against SDG 4.1.1, there must be evidence that the items in the assessment have followed an appropriate test development process, and in particular, have been reviewed quantitatively and qualitatively to determine their suitability for inclusion in the assessment.

The qualitative review should consider whether:

- Each assessment item is considered appropriate by relevant experts for inclusion in the assessment
- The scoring guides are consistent with what the item is intended to measure.

The quantitative review should consider whether:

- Item difficulty (e.g., item facility (CTT) or item location on the scale (IRT)) is appropriate for the grade level
- Item discrimination (e.g., Discrimination Index for each item is generally greater than 0.2, with any exceptions rationalized or the distractors in a multiple-choice item should be negatively correlated with ability).

Details of the test development process followed, and evidence that suitable qualitative and quantitative reviews have been carried out should be in the public domain as part of a technical report.

Criterion 3 – sample

To be suitable for reporting against SDG 4.1.1, there must be evidence that the group of learners who took the assessment is representative of the population against which the results will be reported.

Where the assessment is administered to the whole cohort, the project team should consider whether there are any subgroups of the population that have been systematically excluded. For example, learners not in school, learners in conflict-affected areas, learners with special educational needs. Any systematic exclusions should be noted for reporting along with an estimate of the number of exclusions, and the exclusions as a proportion of the population.

Where the assessment is administered to a sample of the population, evidence must be provided to demonstrate the representativeness of the sample. The margin of error should be 5 percent or less at the 95 percent confidence level.

Details of the target population definition, population coverage, design effect, sampling frame development and the post sampling treatment of data to account for any issues identified in the achieved sample (for example weightings used to account for sampling bias) should be described in a technical report. This report must be made publicly available.

Criterion 4 – administration

To be suitable for reporting against SDG 4.1.1, there must be evidence that the assessment was administered in an appropriate and standardised way (for example, administration conditions were consistent, or length of time to administer the assessment was adhered to).

Administration guides must be reviewed for clarity and monitoring of the implementation must be undertaken. Any incidents of inappropriate administration, identified through monitoring or reporting of concerns, should be recorded. Where significant incidents of inappropriate administration are recorded, relevant results should be excluded from the outcomes. This will require additional checks to confirm that this does not affect the representativeness of the sample.

Documentation relating to administration should be in the public domain. Details of administrator training, quality assurance procedures and quality assurance outcomes should also be made available publicly.

Criterion 5 – reliability

To be suitable for reporting against SDG 4.1.1, the value of coefficient alpha/Cronbach's alpha (or equivalent reliability statistic) for the assessment must be greater than or equal to 0.7.

In addition, there must be evidence of appropriate quality assurance arrangements for any human-scored items. As a minimum, this quality assurance should take place during the training for those responsible for scoring the items. Ideally, however, such quality assurance should take place during the live administration. The method of quality assurance may be determined locally, though common procedures include scoring of items with a pre-agreed score to check that the scorer assigns the same score or double scoring of a sample of responses to check levels of agreement.

The approach to quality assurance must be documented and provided to UIS as a minimum, though publication is advised. UIS must also be provided with statistical outcomes from the quality assurance arrangements, for example agreement rates between scorers or with pre-agreed scores.