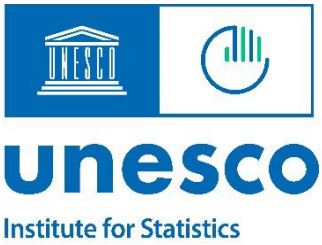

GAML/EDSC criteria for use of an assessment to report on SDG 4.1.1

Version 4
9 December 2024

Table of Contents

1. Background[1]

In the past few years, Montoya and Crouch have published blogs here, here, and here explaining how the market or ecosystem for assessments is dysfunctional: prices are untransparent, criteria that a good assessment should meet are not clear, which assessments are fit for what purpose, etc. These are all forms of information not easily accessible either to assessment organizations or to countries. As economists put it, it is a market rife with asymmetric information between producers, funders, and users. Some of this is difficult to avoid as it is a very technical field. But not all of the difficulties are inherently technical. This document contributes to the creation of a more efficient market or ecosystem in assessments, by setting out technical criteria that assessments ought to satisfy for reporting on SDG 4.1.1, and in general.

2. Introduction

The number of countries reporting on internationally comparable indicators on SDG 4.1.1 is not as high as desirable. For example, in the latest UIS data release available to the public online, only 37 countries report learning (using reading as a proxy) at the Grade 2 or 3 level, and 101 countries at the end-of-primary level at least once in the last six years. These numbers contrast sharply as compared to the 203 countries reporting primary school enrolment, indicating a mere 18% reporting at the lower primary level compared to reporting enrolment figures.[2] Perhaps, more importantly, the number of countries reporting is not increasing quickly enough. During 2013-2016, only 30 countries reported for SDG 4.1.1a, increasing marginally to 36 in the most recent three years. At this pace, it would take 35 years for the lower primary learning indicator, and 11 years for the end of primary learning indicator, to catch up to the enrolment reporting rates.

To some degree, this lack of reporting, especially for SDG 4.1.1a, is somewhat expected. Learning assessments for the end of primary and lower secondary have a relatively venerable history, whereas learning assessments suitable for SDG 4.1.1a are a much newer area of work. Furthermore, there seem to be stronger technical difficulties in measuring at the lower primary level. For example, at this level, language and orthography issues that are inherent to the process of learning to read (more so than mathematics) are not merely an artifact of the assessment methodology and tend to get in the way of the measurement of skill, or more accurately, get in the way of the use of the measurement of learning as a comparable proxy for school system quality. However, inherent and naturally given as this difficulty may be, it has had unfortunate consequences.

At the meeting of the UN-IAEG (Inter-agency and Expert Group on SDG Indicators) on 23 October 2023, the indicator for SDG 4.1.1a was "demoted" from a Tier I to a Tier II indicator due to lack of coverage.

[1] At the request of Silvia Montoya, Director of the UNESCO Institute for Statistics (UIS), this document was coordinated by Luis Crouch, UIS Governing Board Chair (First Vice Chair at the time this team was set up), with the kind collaboration of Abdullah Ferdous of AIR, and Kemran Mestan, Maurice Walker, and Colin Watson of ACER. This document has been revised to address comments received and in response to TAG advice and deliberation. This version of the document replaces all previous versions and will be used by UIS as the basis for making decisions about eligibility for global reporting.

[2] Using primary school enrolment as a simple benchmark of an indicator that is both relatively easy to report and is also relatively important.

The community of interest concerned with foundational learning, such as the Global Coalition for Foundational Learning, immediately expressed deep concern, due to the possible signalling that this "demotion" might imply to countries. The subtlety that the "demotion" is due to insufficient reporting rather than a lack of fundamental importance of the issue, is likely to be lost, with countries taking the demotion as a signal of lack of importance. As a result, no less than four blogs from opinion leaders in the sector were published within the two or three weeks after this decision, questioning the decision and/or proposing ways forward. One of them included many or most of the key global leaders of development agencies' education departments.[3] The IAEG decision did not close the door on reversing this decision. Specifically, the IAEG and opinion leaders, agreed on the need to increase reporting to at least 50% of countries *where the indicator is relevant* (according to the most current definition of Tier I).

On 6-7 December 2023, representatives and interested parties related to the Global Alliance to Monitor Learning, sponsored by UIS, met for the tenth time in its history, at a previously scheduled meeting in Paris. Naturally, given the change in status of SDG 4.1.1a, the issue of how to increase coverage received considerable attention, both formal and informal (sidebar conversations among key leaders). There was a common cause at the meeting to increase coverage, while also maintaining the necessity for methodological rigor. Key presentations on minimum criteria required to report, and on linking to agreed minimum proficiency levels, were made by consultants and advisors Abdullah Ferdous (AIR), Colin Watson (ACER), and Maurice Walker (ACER) at two important sessions of the meeting, available here. These presentations made specific suggestions on criteria that assessments would need to meet in order to report. At the same time, the fact that various assessments exist (such as UNICEF's FLM as part of the MICS surveys, various citizen-led assessments, and EGRA, to mention three examples) but were not being used for reporting, was noted. It was suggested that these could boost the count. (And this was the thrust of much of the blog commentary from stakeholders in Autumn 2023.) But it was noted by UIS at the meeting that these often did not meet standards that there were rigorous enough, or that the degree of rigor was unknown due to lack of standardized documentation. (See Appendix A for a discussion of this issue.) Thus, there were considerable discussion and requests from the floor, and from UIS itself, for further clarification and unification of criteria that could be compiled by UIS. Furthermore, the implications of the GAML recommendations were discussed and adopted at the 10th Meeting (virtual) of the Technical Cooperation Group (TCG) on SDG 4 Indicators – now known as the Education Data and Statistics Commission (EDSC - on 11 December 2023, via a presentation from UIS Director Montoya, available here.

This document seeks to clarify and lay out, in one single document, the state of play regarding the criteria that could allow an increase in reporting on SDG 4.1.1, while ensuring an acceptable standard of rigor. The document proceeds as follows:

- In section 3, the document sets out a general set of general considerations of a policy nature that, together with technical considerations, drive the criteria. These are an important preamble to the reporting acceptability criteria. They must be understood in order to then understand why

[3] Alicia Herbert, Foreign Commonwealth and Development Office (FCDO), United Kingdom; Robert Jenkins, Global Director, Education and Adolescent Development, UNICEF; Stefania Giannini, Assistant Director-General for Education, UNESCO; Allyson Wainer, Director of the Center for Education, USAID; Benjamin Piper, Bill and Melinda Gates Foundation; Luis Benveniste, Global Director for Education, The World Bank; and Jo Bourne, Chief Technical Officer, Global Partnership for Education.

the criteria read as they do.
- In section 4, the document lays out the technical criteria that assessments ought to comply with to be acceptable for reporting.

This document will remain as a draft document until it is discussed and approved at the upcoming meeting of the Education Data and Statistics Commission (EDSC) in February 2025.

3. General considerations

These considerations are provided to help guide planning and process development. They are important and should be considered in addition to the technical criteria described in section 4 below. However, the outcomes of considerations need not be documented to satisfy the technical requirements for reporting on SDG 4.1.1. Note that while the document specifies criteria for all of SDG 4.1.1, it occasionally uses 4.1.1a as a case in point, given the relative uncertainty over this indicator.

3.1. Country coordination

It will be up to any relevant country's authorities to decide whether they want to use any assessment for reporting on SDG4.1.1. The UIS should be notified of the intention to use an assessment for reporting SDG4.1.1 before the assessment is planned and data is collected. In particular, UIS must agree that the target population covered by the assessment is appropriate for reporting against the relevant SDG4.1.1 MPL for the national population. It is expected that no matter which assessment is used to collect data, all requirements described in Section 4 will be met. Therefore, any assessment organisation, provider, implementing partner or technical partner must be informed of the obligation to meet the technical requirements and provide the required supporting documentation. Any assessment, national, regional or international, may be used to report against SDG4.1.1 provided the requirements of the assessment or program, as described in Section 4, are met.

As of the writing of this document, various assessments are being revised, with a view to satisfying the criteria in this document. Assessment submissions for reporting will naturally be evaluated on the basis of the newer version of the assessments.

3.2. Documentation

All evidence that is used to demonstrate that an assessment meets the criteria described in section 4 must be in the public domain and accessible to and via UIS. How this is managed will be determined at country level and will depend on a country's own document sharing approaches, e.g., through the Ministry of Education's website. It is possible that UIS may design a portal for this purpose in future, to facilitate ease of submission.

When submitting evidence to UIS for evaluation against the criteria, countries should provide a single document containing the following:

- An overall description of the assessment
- A written description for each relevant criterion of how it has been met or is planned (concretely, in detail) to be met
- Hyperlinks to (ideally) or extra files of relevant, publicly-available documents that provide the supporting evidence that each criterion has been met (with page/section numbers as appropriate)

Table 1 provides a proposed format to provide the evidence for each criterion.

Table 1: Submission of evidence against UIS eligibility criteria

Criterion reference	**Description of how criteria has been met**	**Hyperlink to relevant, publicly-available document**	**Page/ Section number**
1.1a (R)	The assessment contains 20 items, each worth 1 score-point, that are all aligned to the reading GPF	www.MoE/assessment_framework www.MoE/alignment_study_outcomes	Section 2, page 5 Section 3, page 10
1.2a (R)	The assessment contains 15 items, each worth 1 score-point, that assess the reading comprehension domain at grade 2 in the GPF	www.MoE/alignment_study_outcomes	Section 3, page 10
…	…	…	…

Copies of each of the relevant, publicly-available documents used to provide the supporting evidence (e.g., assessment frameworks, test development processes, technical analysis reports, etc.) should also be provided to UIS at the time of submission.

3.3. Requirements to apply to past and future assessment

The reporting requirements are absolute. There will be no diminishing of the technical criteria to apply retrospectively, however the degree of supporting documentation may be reduced in agreement with the UIS.

If an assessment program meets all the technical requirements, the same assessment or program may be used in the future to plan for reporting against SDG4.1.1. However, each iteration of the assessment must adhere to the technical requirements and documentation. Even if the assessment content remains unchanged over iterations, countries must consider maintaining the requirements around sampling and operations.

If the content of the assessment does change over successive iterations, countries must consider how the content requirements are met each time. Importantly, if content changes, then either:

- a psychometric link is required to ensure the MPL benchmarks can be located on empirical

scales across iterations (see section 4.7)
- an exercise to locate the MPL benchmarks will need to be conducted each iteration (see section 4.6)

These considerations should inform medium term assessment planning.

3.4. Sustainability plan

In considering the long-term nature of education monitoring, including reporting against global standards, each country or assessment program should develop a sustainability plan. The plan should express the desire to use the assessment again over time, and to have national capability in the use of the assessment, and similar assessments, built up. If an organization is providing technical support, there ought to be a plan to transfer as much capability to the country in question as possible or as desired by the country. It will help if the organization responsible for the assessment support has a track record of providing capacity building and transfer of capabilities.

3.5. Utility to the country

Related to the sustainability plan, ideally the assessment should be of great utility to the country, above and beyond global reporting. The assessment program should add value to policy dialogue, policy setting, and capacity building. The program could be considered for monitoring general policies or specific programs for improvements. Ideally, the assessments should not just report on outcomes but assist the countries to identify where they can do better on the outcomes on which they are reporting.

3.6. Humaneness and ethical testing

Assessments should adhere to principles of humaneness and ethical testing, as well as efficiency.

4. Technical requirements for reporting SDG 4.1.1

This section sets out in detail the criteria for assessments to be considered for reporting on SDG 4.1.1, with numerical values to the extent possible, and with an extensive illustration from AMPL-a.[4] [5] As will be noted, the criteria tend to be more specific for SDG 4.1.1a as this is the weakest of the SDG 4.1.1 indicators in terms of numbers of countries thus far reporting and methodological clarity. But the criteria hold for all of SDG 4.1.1. Most of these are elaborations and specifications of the issues discussed at the 6-7 December 2023 GAML meeting and at the 11 December 2023 TCG (now EDSC) meeting. The relevant documents from those meetings are here and here respectively. This second draft benefits also from feedback provided to UIS by the interested parties and above all by a meeting of a Technical Advisory Group on 4-6 March 2024 in London. Unless explicitly stated otherwise, all criterion guidelines and recommendations apply to both household and school-based assessments.

The information in each criterion is structured in the same way. The section begins with an elaboration of the criteria to aid understanding. The specific requirements for each criterion are then provided in a table. **Requirements in bold** are essential and must be met for an assessment to be considered acceptable for reporting. Where a requirement is not in bold, at this time these are recommendations and are not essential to meet the reporting requirements but are considered advisable for high-quality assessments and may be required at a later date. Following the table, there is a section detailing the documentation requirements to provide evidence that the criterion has been met. An exemplification of this has been provided using AMPL-a.

4.1. Criterion 1 – Alignment to the MPL and construct validity

[4] AMPL-a is part of the Assessments for Minimum Proficiency Levels (AMPL). The main aim of AMPL is to measure and analyse the reading and mathematics proficiency of students at the end of lower primary (SDG indicator 4.1.1a – 'AMPL-a') and at the end of primary school education (SDG indicator 4.1.1b – 'AMPL-b'). To date, AMPL was implemented in 13 countries with different populations, either as a stand-alone assessment, or integrated with a national or regional assessment. In 2024, UIS is supporting the implementation of AMPL in 11 countries, with a pilot in 7 countries and 10 national languages.

[5] There is no implication that any given assessment has to pass the same bar as the AMPL-a set for itself. This is used as a best practice example. For other examples of a good standard of documentation from the two assessments, ERCE and PASEC, that have been legacied into 4.1.1a, see the following links. For PASEC see the overall technical report here, and a typical country report here. The reader is invited to peruse the websites linked here to get a sense of how standardized the country reporting is. For ERCE, here is the background curricular analysis, here is the technical report on psychometric characteristics, assessment design, etc., and here is a typical country report. The reader may peruse the website links given to see how standardized the country reports are. As for general AMPL documentation that summarizes in just a few files the approach and shows good practice, see: a) On test design, here. On sampling, here. And on standard-setting and linking to the MPL, here.

The main purpose of this criterion is to determine whether the assessment is aligned to the MPL, using the GPF for reading and for mathematics as proxies to make the task of content alignment easier to manage.

A review of the MPL and the GPF has determined that the MPL for SDG 4.1.1a is most closely aligned to the description of "meets global minimum proficiency" for grade 2 in the GPF, the MPL for SDG 4.1.1b is most closely aligned to the description of "meets global minimum proficiency" for grade 5 in the GPF and the MPL for SDG 4.1.1c is most closely aligned to the description of "meets global minimum proficiency" for grade 8 in the GPF. In conventional terms, this criterion is based in the concept of "construct validity." To be aligned, the assessment should measure the value added of the skills from the grade previous to the targeted grade. For example, to report on SDG 4.1.1a in mathematics, an aligned measurement focuses on the skills that are specific to Grade 2 , such as add and subtract within 11 to 20 (whereas Grade 1 was 1 to 10). As noted, the GPF is used as an anchor in these criteria due to its greater specificity. If a country wishes to explain its submission in terms of the MPL and/or a slightly different grade level, a case can be made.

In reading assessments that are aimed at low-income or low/middle-income countries, or countries with low educational performance, and especially when the country is below benchmark for reading comprehension, the reading comprehension score itself may not be very informative. In those cases the country can be encouraged to include other subconstructs as specified in the MPL and GPF that can be considered precursors of the two chosen here, such as decoding, fluency, etc. These are likely to add to the in-country utility (utility being seen as important value in addition to reportability, as per Section 3.5) of the assessment for programming and policy, beyond reporting.

The following tables provide more information on the criterion for each MPL and subject.

- Table 2: Technical details for criterion 1 that assessments must meet to be acceptable for reporting for SDG 4.1.1a – Reading
- Table 3: Technical details for criterion 1 that assessments must meet to be acceptable for reporting for SDG 4.1.1a – Mathematics
- Table 4: Technical details for criterion 1 that assessments must meet to be acceptable for reporting for SDG 4.1.1b – Reading
- Table 5: Technical details for criterion 1 that assessments must meet to be acceptable for reporting for SDG 4.1.1b – Mathematics
- Table 6: Technical details for criterion 1 that assessments must meet to be acceptable for reporting for SDG 4.1.1c – Reading
- Table 7: Technical details for criterion 1 that assessments must meet to be acceptable for reporting for SDG 4.1.1c – Mathematics

Table 2: Technical details for criterion 1 that assessments must meet to be acceptable for reporting for SDG 4.1.1a – Reading

Ref.	Description of requirements
1.1a (R)	**Minimum 20 score-points aligned to the GPF in reading**
1.2a (R)	**Minimum 10 score-points assessing the reading comprehension domain in the GPF grade 2**
1.3a (R)	**The assessment must cover both reading comprehension subconstructs at grade 2 in the GPF (see Appendix B – Grade-level GPF subconstructs)**
1.4a (R)	**The remaining items can be drawn from any of the domains (decoding, listening comprehension or reading comprehension)**
1.5a (R)	For timed fluency tasks, students should be given sufficient time to read to the end of the text, but fluency should be tracked within one minute
1.6a (R)	For individually administered assessments, implementing a stop rule is recommended but it is also recommended to begin with easier items, potentially starting with a word list, to ensure the assessment is approachable
1.7a (R)	**Differences between the language of instruction, home language, and language of assessment must be noted and implications considered for interpretation of the outcomes**

Table 3: Technical details for criterion 1 that assessments must meet to be acceptable for reporting for SDG 4.1.1a – Mathematics

Ref.	Description of requirements
1.1a (M)	**Minimum 20 score-points aligned to the GPF in mathematics**
1.2a (M)	**Minimum 10 score-points assessing number and operations domain in the GPF at Grade 2.**
1.3a (M)	**The assessment must cover at least three out of the four number and operations subconstructs at grade 2 in the GPF (see Appendix B – Grade-level GPF subconstructs) as selected in 1.2a(M) above.**

Ref.	Description of requirements
1.4a (M)	**A minimum of 10 items must be included from any of the 4 non-number and operations domains (measurement, geometry, statistics and probability, and algebra). At the time of publishing of this document, these items will not be counted against the reporting requirement, pending more research on performance and item difficulty cut points, but must be reported.**
1.5a (M)	Within the non-number and operations domains (measurement, geometry, statistics and probability, and algebra), items should cover at least 3 out 4 of these domains
1.6a (M)	Within the non-number and operations domains (measurement, geometry, statistics and probability, and algebra), at least 5 out of 8 constructs should be represented (see Appendix B – Grade-level GPF subconstructs)
1.7a (M)	If there is an intention to report on individual domains, a minimum of 7 items per domain is required.
1.8a (M)	**Differences between the language of instruction, home language, and language of assessment must be noted and implications considered for interpretation of the outcomes**
1.9a (M)	The language(s) of instruction of the children being assessed should be used for the assessment.

Table 4: Technical details for criterion 1 that assessments must meet to be acceptable for reporting for SDG 4.1.1b – Reading

Ref.	Description of requirements
1.1b (R)	**Minimum 20 score-points assessing reading comprehension aligned to the GPF**
1.2b (R)	**As defined in the GPF, there should be a minimum of:** • **5 score-points assessing the retrieve information construct at grade 5** • **5 score-points assessing the interpret information construct at grade 5**
1.3b (R)	**The assessment should also cover 4 of the 8 reading comprehension subconstructs at grade 5 in the GPF (see Appendix B – Grade-level GPF subconstructs)**

Table 5: Technical details for criterion 1 that assessments must meet to be acceptable for reporting for SDG 4.1.1b – Mathematics

Ref.	Description of requirements
1.1b (M)	Minimum of 10 score-points assessing *number and operations* aligned to GPF
1.2b (M)	Minimum of 5 score-points assessing *measurement* and *geometry* aligned to the GPF
1.3b (M)	Minimum of 5 score-points assessing *statistics and probability* and *algebra* aligned to the GPF
1.4b (M)	The assessment must include 12 grade 5 items covering 12 of the 21 subconstructs at grade 5 in the GPF (see Appendix B – Grade-level GPF subconstructs)

Table 6: Technical details for criterion 1 that assessments must meet to be acceptable for reporting for SDG 4.1.1c – Reading

Ref.	Description of requirements
1.1c (R)	Minimum 20 score-points assessing reading comprehension aligned to the GPF
1.2c (R)	As defined in the GPF, there should be minimum of: • 4 score-points assessing the retrieve information construct at grade 8 • 4 score-points assessing the interpret information construct at grade 8 • 4 score-points assessing the reflect on information construct at grade 8
1.3c (R)	The assessment should cover 5 of the 10 reading comprehension subconstructs at grade 8 in the GPF (see Appendix B – Grade-level GPF subconstructs)

Table 7: Technical details for criterion 1 that assessments must meet to be acceptable for reporting for SDG 4.1.1c – Mathematics

Ref.	Description of requirements
1.1c (M)	**Minimum of 10 score-points assessing *number and operations* aligned to the GPF**
1.2c (M)	**Minimum of 5 score-points assessing *measurement* and *geometry* aligned to the GPF**
1.3c (M)	**Minimum of 5 score-points assessing *statistics and probability* and *algebra* aligned to the GPF**
1.4c (M)	**The assessment must include 12 grade 8 items covering 12 of the 21 subconstructs at grade 8 in the GPF (see Appendix B – Grade-level GPF subconstructs)**

4.1.1. Statement of documentation requirement

The country or its assessment advisors for the assessment being used for reporting should produce an assessment specification document that should include the information about the assessment as outlined here, all in one place. This should include information on the following: purpose(s) of the assessment; definitions of domain, constructs, subconstructs, and learning outcomes measured; definition of the assessed population; interpretations for the intended uses; definition of the content of the test; the item formats; time allowed for testing; directions for test takers; and scoring and reporting procedures.

The documentation should cover the issues and items in the relevant table for the indicator and subject,

4.1.2. Best practice examples (AMPL-a)

The AMPL-a reading assessments include listening decoding, decoding and listening comprehension items in addition to reading comprehension, as follows:

- Listening comprehension (Audio): 10 items
- Listening decoding (Audio): 5 items
- Decoding: 5 items
- Reading comprehension: 25 items

Mathematics: 30 items

Sources:
UIS & ACER (2023) *Study Design: AMPL-ab*.
UIS & ACER (2023) *Assessment Blueprint: AMPL-ab*.

4.2. Criterion 2 – Item content and quality

The main purpose of this criterion is to determine whether there is evidence that the items in the assessment have been reviewed qualitatively and quantitatively prior to their inclusion in the final assessment instrument. Most of the requirements in this criterion relate to activities that take place during the test development phase before the assessment is administered live. As a result, analysis is often conducted on data from field tests/trials. Where analysis needs to be confirmed with live data, this is clearly stated.

The following tables provide more information on the criterion. The same requirements apply to all indicators and domains.

- Table 8: Technical details for criterion 2 that assessments must meet to be acceptable for reporting for SDG 4.1.1 – qualitative review
- Table 9: Technical details for criterion 2 that assessments must meet to be acceptable for reporting for SDG 4.1.1 – quantitative review

Table 8: Technical details for criterion 2 that assessments must meet to be acceptable for reporting for SDG 4.1.1 – qualitative review

Ref.	Description of requirements
2.1	**Each assessment item must be considered appropriate by relevant experts for inclusion in the assessment**
2.2	The relevant experts for 2.1 should include nationals of the reporting country or similar countries.
2.3	**The assessment items must be developed under advice from subject matter experts (SMEs)**
2.4	**The assessment items must be thoroughly discussed and vetted by local experts to ensure suitability for the local context**
2.5	**The SMEs responsible for developing the items must be trained in item development principles and procedures**

Ref.	Description of requirements
2.6	**The items must be field tested on a representative sample of the learner population** Where the same assessment is administered in the same language across multiple countries, field testing may not be required in all countries. Caution is advised that considerable waste could result if issues that could disqualify the assessment and could have been identified in trialling, such as mistranslations leading to poor or biased item performance, were not identified until live administration.
2.7	**The scoring guides must be consistent with what the item is intended to measure**

Table 9: Technical details for criterion 2 that assessments must meet to be acceptable for reporting for SDG 4.1.1 – quantitative review

Ref.	Description of requirements
2.8	**Psychometric item analysis must be conducted on the field test data using classical test theory (CTT)**
2.9	Psychometric item analysis should be conducted on the field test data using item response theory (IRT) where possible
2.10	**Item difficulty (e.g., item facility (CTT) or item location on the ability scale (IRT)) must be reviewed following the field trial and determined to be appropriate for the grade level (given the MPL alignment requirements) before being included in the live assessment**
2.11	All assessment items should ideally have a difficulty level (facility or percent correct) between 0.20 and 0.90, with any exceptions being justified It is noted that this may not be the case where the population being assessed is largely working below the standard of the MPL and the assessment is correctly aligned to the MPL. The items should target the MPL to maximise information to calculate the estimate of the proportion of children or young people meeting the MPL.
2.12	**Item discrimination (e.g., discrimination index) for each item must be reviewed following the field trial and be determined to be appropriate before being included in the live assessment.**

Ref.	Description of requirements
2.13	All assessment items (with CTT) should ideally have an item-total correlation (or point biserial) value of at least 0.20, with any exceptions being justified

4.2.1. Statement of documentation requirement

The country or its assessment advisors for the assessment being used for reporting should produce a document that details the development process used for the assessment instrument that addresses each of the requirements in the tables above.

4.2.2. Best practice examples (AMPL-a)

Qualitative review

The UIS Global Item Bank was reviewed for suitable items for the AMPL-a tests in both English and French, using the following criteria:

- the items were suitable for students working at the level of lower primary
- the items were multiple- choice (or another closed item format)
- the items did not use a sentence fragment as the item stem (since this format can be difficult to translate)
- the items originated in either English or French, and
- (for reading) the item or stimulus did not rely heavily on language-specific features that would not translate well (e.g., a poem based on rhyming).
- No suitable items could be identified. Consequently, ACER developed new items in alignment with the "MPLs Unpacked" specifications for SDG4.1.1a or the GPF specifications for Grade 2.

Quantitative review

Psychometric quality assurance analysis of AMPL-a and AMPL-b items was undertaken. Analytical outputs include: 'Facility', 'Difficulty', Item-Rest', 'Delta', 'Threshold', 'Least Weighted MNSQ' and 'DIF Logits'. The analysis for reading items included response data from 21,994 students on 71 multiple-choice items and 1 constructed-response item.

Summary findings include:

- The mean score on the 72 items was 39.1 and the standard deviation was 14.9.
- The item with the highest item-rest correlation was Item 22 (ARM002) with a value of 0.59 and the item with the lowest item-rest correlation was

Item 43 (ARR021) with a value of 0.14.

- The analysis of mathematics items included response data from 21,941 students on 56 multiple-choice items, 1 constructed-response item and 1 partial-credit item.
- The mean score on the 58 items was 30.3 and the standard deviation was 13.7.
- The item with the highest item-rest correlation was Item 13 (AM013) with a value of 0.57 and the item with the lowest item-rest correlation was Item 36 (MM029) with a value of 0.06.

Sources:
ACER (2022). *Minimum Proficiency Levels Unpacked*.
UIS & ACER (2023) *Assessment Blueprint: AMPL-ab*.
UIS & ACER (2023) *Item Analysis Report - Reading: AMPL-ab*.
UIS & ACER (2023) *Item Analysis Report -mathematics: AMPL-ab*.

4.3. Criterion 3 – Population coverage and sampling

The main purpose of this criterion is to determine whether the learners that took the assessment are representative of the population against which the results will be reported. It applies to both sample-based assessments and census-based assessments, though the requirements for each are different.

It is to be noted that data on learning outcomes from sample-based household surveys that meet all of the requirements would be acceptable and encouraged.

The following tables provide more information on this criterion. The same requirements apply to all indicators and domains.

- Table 10: Technical details for criterion 3 that assessments must meet to be acceptable for reporting for SDG 4.1.1 – sample-based assessments
- Table 11: Technical details for criterion 3 that assessments must meet to be acceptable for reporting for SDG 4.1.1 – census-based assessments

Table 10: Technical details for criterion 3 that assessments must meet to be acceptable for reporting for SDG 4.1.1 – sample-based assessments

Ref.	Description of requirements
3.1 (S)	**The national desired target population to be reported upon in relation for achieving the MPL for the relevant SDG 4.1.1 indicator must be described**

Ref.	Description of requirements
3.2 (S)	**Any subgroup of the learner population that has been systematically excluded must be documented and justified.** It is strongly recommended that no more than 10% of the sampled population should be excluded from the reported results. Any purposeful exclusions must be explained and justified in writing and with data. Subgroups might include learners not in school, learners in conflict-affected areas, learners with special educational needs.
3.3 (S)	**The sample approach and design must be documented** This includes stratification, clustering, and description of how representation by gender, language of instruction, geography and socio-economic status is to be achieved
3.4 (S)	**The sample should be designed to achieve a 95% confidence interval of 5 percentage points for estimates of percentages of leaners meeting or exceeding each MPL.** **The calculation of this must be documented. This includes cluster effects and estimates of effective sample by gender.**
3.5 (S)	**For reporting SDG4.1.1a the national target population coverage of proportion of children in grades 2/3, by gender, must be documented**
3.6 (S)	**For reporting SDG4.1.1b the national target population coverage of proportion of children at the end of primary, by gender, must be documented**
3.7 (S)	**For reporting SDG4.1.1c the national target population coverage of young people at the end of lower secondary, by gender, must be documented**
3.8 (S)	**The sampling frame development must be documented and justified to demonstrate its suitability**
3.9 (S)	**Sampling weights must be calculated and applied in determining national estimates**

Ref.	Description of requirements
3.10 (S)	**An overall response rate of 70% (cluster x individual response) is required. Response rates by sample stage must be documented.** All surveys should plan revisits to clusters (schools, households) in advance of sampling to achieve this requirement.
3.11 (S)	**Substitution of non-responding sampled schools must be no more than 15% of the weighted population within all sampled schools.**

Table 11: Technical details for criterion 3 that assessments must meet to be acceptable for reporting for SDG 4.1.1 – census-based assessments

Ref.	Description of requirements
3.1 (P)	**The national desired target population to be reported upon in relation for achieving the MPL for the relevant SDG 4.1.1 indicator must be described**
3.2 (P)	**Any subgroup of the learner population that has been systematically excluded must be documented and justified.** It is strongly recommended that no more than 10% of the sampled population should be excluded from the reported results. Any purposeful exclusions must be explained and justified in writing and with data. Subgroups might include learners not in school, learners in conflict-affected areas, learners with special educational needs.
3.3 (P)	**For reporting SDG4.1.1a the national target population coverage of proportion of children, by gender, in grades 2/3 must be documented**
3.4 (P)	**For reporting SDG4.1.1b the national target population coverage of Proportion of children, by gender, at the end of primary must be documented**
3.5 (P)	**For reporting SDG4.1.1c the national target population coverage of young people, by gender, at the end of lower secondary must be documented**

4.3.1. Statement of documentation requirement

Countries and their assessment advisors should produce a comprehensive technical report on sampling, which should encompass a detailed description of sample size calculation and the process of sample selection. This report is crucial for providing transparency and understanding of the methodology employed in obtaining national representative samples. The documentation should cover the issues and items in the relevant table above.

4.3.2. Best practice examples (AMPL-a)

The AMPL-a involved a two- stage clustered sample design. At the first stage schools were sampled. At the second stage, an intact class of students from those schools was sampled. Where the class size exceeded a certain practical number, a sub-sample of students from the sampled intact class was selected. A minimum of 150 schools and 4000 students were required to participate in AMPL-a in each population assessed. Details, including how robustness was assured, are available in the Sampling Framework Report and The Weighting and Sample Outcomes Approach Technical Report.

A nationally representative sample was drawn in each of the participating countries. Samples were stratified using the following strata:

- School type, sector, ownership or proprietor: e.g. private/public/religious School location: urban/regional
- Region: e.g. all the national counties or provinces School size: e.g. small and large schools
- Students may have been excluded on the grounds of having functional disabilities, or insufficient language proficiency. Schools might be excluded if they exclusively cater for students who would be excluded, as well as on the grounds of:
- Accessibility: e.g. too difficult to reach
- Size: e.g. too small
- Non-standard curriculum: e.g. has a special curriculum.
- The population definition and sample Designs, and the sample outcomes for each country can be found in two reports developed for each country.

Sources:
UIS & ACER (2023) *Sampling Framework: AMPL-ab*.
UIS & ACER (2023) *AMPL-ab Sample Information and Outcomes.* (1 report for each country)
UIS & ACER (2023) *Population Definition and Sample Design.* (1 report for each country)

4.4. Criterion 4 – Assessment administration and data custodianship

The main purpose of this criterion is to determine whether the assessment was administered in an appropriate and standardised way and that there is confidence in the data collection, quality assurance and data management/security activities.

The following table provides more information on this criterion. The same requirements apply to all indicators and domains.

- Table 12: Technical details for criterion 4 that assessments must meet to be acceptable for reporting for SDG 4.1.1 – administration and data custodianship

Table 12: Technical details for criterion 4 that assessments must meet to be acceptable for reporting for SDG 4.1.1 – administration and data custodianship

Ref.	Description of requirements
4.1	**A standardized test administration manual must be produced – the manual should be presented with sufficient clarity so that it is possible for others to replicate the administration conditions under which the reliability and validity are obtained**
4.2	Allowable variations of administration procedures should be clearly described in the administration manual
4.3	**Processes must be in place for selecting, training, and qualifying enumerators/administrators and supervisors/quality control officers and confirmation must be provided that those recruited met the required selection criteria**
4.4	**Training must be developed and conducted for enumerators/administrators and confirmation must be provided that any post-training qualifying conditions were met**
4.5	**Training must be developed and conducted for supervisors/quality control officers (QCO) and confirmation must be provided that any post-training qualifying conditions were met**
4.6	Where appropriate, dry run or practice sessions should be conducted for enumerators/administrators and supervisors/QCOs
4.7	A process should be in place which details how those responsible for administering or monitoring the assessment will be replaced and under what circumstances
4.9	**An explicitly stated data quality assurance plan must form part of the quality assurance plan**

Ref.	Description of requirements
4.8	**The proportion of the administrations being observed by supervisors/quality control officers must be determined and justified as part of a quality assurance plan**
4.10	**Any incidents of inappropriate administration, identified through monitoring or reporting of concerns, should be recorded**
4.11	**Where significant incidents of inappropriate administration are recorded, relevant results should be excluded from the outcomes and recorded as part of a quality assurance outcomes report** **This will require additional checks to confirm that this does not affect the representativeness of the sample**
4.12	Where required, and where it cannot be assumed, informed consent should be obtained from learners or the rationale for not doing so documented See standards of good practice here and here. The latter refers mostly to big data but is a good summary of the issues

4.4.1. Statement of documentation requirement

Countries and their assessment advisors should generate a detailed technical report on assessment administration and data custodianship, which should include a thorough account of the procedures for administering assessments and managing the data collected. This report is vital for ensuring transparency and comprehension of the methodologies used in administering assessments and safeguarding the integrity and confidentiality of the data. The documentation should cover the issues and items in the table above.

4.4.2. Best practice examples (AMPL-a)

Seventy-one standards were developed and applied to direct the assessment administration and data custodianship.

The standards for data collection and submission were developed according to three major goals: consistency, precision and generalizability of the data. The standards and the rationale for these standards are in the Technical Standards Report, and the explanation of how the standards were met is provided in a review of that Report.

Independent Quality Monitors were responsible for assessing the implementation of activities. Four standards relate to quality monitoring, including:

- The AMPL-a test administration is monitored using school visits by trained independent QMs.
- At least 5% school visits are conducted in each participating country to observe AMPL-a test administration sessions.
- AMPL-a test administration sessions that are the subject of the national QM visit are randomly selected.

Sixteen standards relate to the security, data management, data submission and archiving material. Data is managed and submitted via the ACER Maple software, which separates personal identification during data management whilst retaining it at the national centre upon data submission.

Five specific standards relate to test administrators, including:

- All AMPL-a assessment sessions follow the procedures as specified in the Test Administrator (TA) manual.
- TAs are trained in the field operations procedures outlined in the TA manual.
- Manuals were provided to support the adherence to the technical standards, as referred to in the source documents.

Sources:
UIS & ACER (2023) *Technical Standards: AMPL-ab*.
UIS & ACER (2023) *Technical Standards Review: AMPL-ab.*
UIS & ACER (2023) *Field Operations Manual: AMPL-ab.*
UIS & ACER (2023) *School Coordinator Manual: AMPL-ab*.
UIS & ACER (2023) *National Project Managers Manual: AMPL-ab*.
UIS & ACER (2023) *Test Administrators Manual: AMPL-ab*.

4.5. Criterion 5 – Reliability

The main purpose of this criterion is to determine whether the assessment is reliable for a specific administration. Informally, this means that any student taking the same test twice ought to score the same, and any assessor scoring the same student twice on the same test ought to score the same.

Though simple equating using common items or other methods may be possible in theory, countries and assessment organizations are advised to adopt a simple Item Response Theory (IRT) model to develop pre-calibrated item banks and utilize them for constructing multiple equivalent forms and their score conversion tables.

The following table provides more information on this criterion. The same requirements apply to all indicators and domains.

- Table 13: Technical details for criterion 5 that assessments must meet to be acceptable for reporting for SDG 4.1.1 – reliability at any given point in time

Table 13: Technical details for criterion 5 that assessments must meet to be acceptable for reporting for SDG 4.1.1 – reliability at any given point in time

Ref.	Description of requirements
5.1	**An item analysis should be conducted on the data from live administration to examine aspects such as difficulty, discrimination, and differential item functioning (DIF). IRT methods of obtaining this information are generally recommended however equivalent methods under CTT are also permissible with documented justification and plan to implement IRT in future analyses**
5.2	**The live assessment must have a reliability coefficient (Cronbach's alpha) of at least 0.80 (or equivalent using a different measure of reliability)**
5.3	**If an assessment is used for a range of ages, the live administration must have a reliability coefficient (Cronbach's alpha) of at least 0.80 for the participants representing the national target population for reporting against SDG4.1.1**
5.4	**If an assessment contains constructed response (CR) and/or oral assessments with any type of performance-based items, the enumerators or those who score the assessment must have an inter-rater reliability (IRR) of at least 0.80 in the live administration**
5.5	**For oral one-on-one assessments, reported inter-rater reliability for the live administration must be greater than a kappa coefficient of 0.8**
5.6	**Items with weak reliability must be carefully considered and excluded from the analysis of the live data or included only with appropriate justification**
5.7	**Item DIF for gender, and other important factors in the countries in question, must be used to determine item inclusion and exclusion from the live item analysis using IRT or classical equivalent**
5.8	Relevant reliability coefficients should be included in the technical report for each total score, sub score, or combination of scores intended for interpretation

Ref.	Description of requirements
5.9	**In cases where a stop rule is applied, assign a zero score (impute zero) for subsequent items, and distinguish these item responses clearly from:** • **item not administered for other reasons** • **Item response missing** • **Item response provided but incorrect**

4.5.1. Statement of documentation requirement

Countries and their assessment advisors should create a detailed technical report on CTT and/or IRT-based item analysis and reliability, which must include a comprehensive explanation of the measures taken to ensure consistency and accuracy in the assessment process. This report is essential to offer clarity and insight into the methods used to guarantee the reliability of the national assessments.

The documentation should cover the issues and items in the table above.

4.5.2. Best practice examples (AMPL-a)

The reliability for each of the reading and mathematics scales in the AMPL is calculated from a unidimensional model for each construct. The reliability for the reading construct is provided on line 209 of the ACER Con Quest output file. Weighted EAP/PV reliability: 0.906

The reliability for the mathematics construct is provided on line 206 of the ACER Con Quest output file. Weighted EAP/PV reliability: 0.898

AMPL-a technical Standard
1.6 notes that participating countries should aim for a sample size that achieves 95% confidence interval widths within ±5% for student percentage estimates, and within 0.1 of a standard deviation around an estimated mean. All AMPL estimates of mean percentage of students at or above the MPL at the country level achieved this precision. This is documented through the provision of standard errors on these statistics in Table D1 and D4 of the international report.

A small number of items were excluded from the analysis due to weak items statistics. The final item statistics report does not include the excluded items.

Item DIFF (i.e. differential item functioning) for gender, was used to analyse item inclusion and exclusion using IRT. The DIF results for each item can be observed in the Item Analysis reports.

Sources:
UIS (in press). Assessment of Minimum Proficiency Level (AMPL-ab): International Report, UNESCO Institute for Statistics, ACER.
ACER (2023) ConQuest output file: CINT_R_itm_formreg(1).shw
ACER (2023). ConQuest output file INT_M_itm_formreg(1).shw
UIS & ACER (2023) Item Analysis Report -mathematics: AMPL-ab.

4.6. Criterion 6 – Benchmark-based linking to the MPL

The main purpose of this criterion is to determine how the assessment is linked to the MPL. That is, what constitutes evidence of minimum proficiency in the results obtained, noting the dependency in terms of the criteria for validity and alignment in criterion 1.

There are two approved ways of linking an assessment to the MPL standards:

- Statistical linking between the assessment and another assessment, or group of assessment items, that have already been calibrated to the MPL.
- A pairwise comparison method exercise using the items from the assessment and a group of items that have already been calibrated to the MPL and that will be available for the ones carrying out the exercise.

UIS will provide further information and clarification about any additional methods that are sufficiently rigorous to directly linking to the MPL and the operational dimensions on their implementation in the next GAML/EDSC (former TCG) meeting.

UIS acknowledges that, at present, this means that more work is needed to link assessments in all languages to the MPL, as there are limited languages for which MPL linking has taken place and the pairwise comparison approach can only currently be undertaken in English. UIS is working with partners to improve this situation, including through translation of tools such as AMPL, to enable a wider range of languages to be reported.

The following tables provide more information on this criterion. The same requirements apply to all indicators and domains.

- Table 14: Technical details for criterion 6 that assessments must meet to be acceptable for reporting for SDG 4. 1.1 – benchmark-based linking to the MPL

- Table 15: Technical details for criterion 6 that assessments must meet to be acceptable for reporting for SDG 4.1.1 - benchmark-based linking to the MPL through pairwise comparison

Table 14: Technical details for criterion 6 that assessments must meet to be acceptable for reporting for SDG 4. 1.1 – benchmark-based linking to the MPL through statistical linking

Ref.	Description of requirements
6.1 (SL)	**An appropriate process to statistically link the results of an assessment to the MPL must be implemented** There are likely to be two main ways to efficiently statistically link an assessment to the MPL: • Including sufficient items within the assessment that have previously been statistically linked to the relevant MPL (common-item design) • Administering the assessment alongside another assessment that has been linked to the relevant MPL (common-person design), this could be using a national assessment from another country that has already been accepted for reporting There may be other methods to link directly to the MPL, information on which will be provided by UIS. We note the above as likely to most efficient and easiest. When using a common-item design, IRT must be used to calibrate the new assessment on the same scale as the previously linked assessment using the common items. When using a common-person design, IRT would still be strongly recommended, though CTT methods may also be used. Once an assessment has been statistically aligned to the MPL, it can be re-used in subsequent iterations as long as there are no changes.
6.2 (SL)	**The statistical link must be established with a broadly representative sample of learners, with a sufficient sample to establish the link psychometrically** The appropriate sample size depends on the linking model and test design. Psychometric advice is needed to determine sample size. As a guideline, 500 participants assessed in both old and new material is likely to be necessary to establish the statistical link.

Ref.	Description of requirements
6.3 (SL)	**When using a common-item design, the statistical link must be established using a design that contains sufficient items from the previously-linked assessment to provide a sufficient anchoring between the two assessments** When selecting items form the previously-linked assessment, care should be taken to select items from a range of the subconstructs covered in the assessment.

Table 15: Technical details for criterion 6 that assessments must meet to be acceptable for reporting for SDG 4.1.1 - benchmark-based linking to the MPL through pairwise comparison

Ref.	Description of requirements[6]
6.1 (PCM)	**All participants must meet the requirements for participation as described in the PCM toolkit**
6.2 (PCM)	**The group of participants must be sufficiently representative in terms of the characteristics agreed by the country**
6.3 (PCM)	**Any participant whose responses did not fit the model well, must be removed from the analysis**
6.4 (PCM)	**Any items that did not fit the model well, must be considered for removal from analyses and a clear rationale for the decision to remove, or not, must be provided**
6.5 (PCM)	**The pairwise scale reliability index must be equal to or higher than 0.75**
6.6 (PCM)	**Items must be removed from analyses if they exhibited item DIF**
6.7 (PCM)	**For the items from the assessment being linked, the dis-attenuated correlation between the items original scale location and LPSs' location must be equal to or higher than 0.75**
6.8 (PCM)	**The average (mean) score for each section of the evaluation must be greater than or equal to 4**

[6] Full details of the Pairwise Comparison Method (PCM), including support to demonstrate alignment to the criterion, can be found in the PCM toolkit – Global Alliance to Monitor Learning (unesco.org)

Ref.	Description of requirements[6]
6.9 (PCM)	**The impact analysis workshop must confirm the validity of the statistical linking exercises**

4.6.1. Statement of documentation requirement

Countries or assessment organizations assisting them should generate a comprehensive standard setting report. This report should outline the selection, training, and qualifications of panellists, the implementation of benchmarking methods, and include both quantitative and qualitative evidence to support the benchmarks.

The documentation should cover the issues and items in the table above.

4.6.2. Best practice examples (AMPL-a)

The AMPL was linked to the MPL via three methods:

- standard setting
- pairwise comparison
- psychometric linking.

Standard setting

The MPL cut scores for reading and mathematics for SDG 4.1.1a, b and c were established on the Learning Progressions Scale (LPS) with an international standard setting exercise (ISSE) undertaken in 2022. The bookmark standard setting method was applied, which uses an Ordered Item Booklet. This consists of items ordered by difficulty. The easiest item is presented first, and the most difficult item is presented last. Sixty participants were asked to make judgements about the placement of bookmarks about the same set of items.

Pairwise comparison

The pairwise comparison method was used to equate the LPS with the AMPL scale for both reading and mathematics. Thirty-three judges were trained to independently judge the difficulty of items, by comparing a pair of items. The judgements formed a dataset that technical experts from ACER analysed to locate AMPL items on the LPS scale, providing validation of the benchmarks comparison method are in Appendix A of the AMPL-ab International Report.

Psychometric linking

The assessment data was psychometrically scaled, using a two-dimensional model to produce estimates for mathematics and reading proficiency; details of this scaling are provided in a Technical Note. The proportion of students above the MPLs for SDG 4.1.1a and SDG 4.1.1b were estimated. These estimates were made by determining the number of students above each of 2 benchmarks (the MPLs for SDG 4.1.1a and b) on the reading and mathematics scales.

Sources:
ACER (2022). *International Standard Setting Exercise*
UIS & ACER (2023) *Scaling AMPL-ab Items: Technical Note*
UIS (in press). *Assessment of Minimum Proficiency Level (AMPL-ab): International Report*, UNESCO Institute for Statistics, ACER.

4.7. Criterion 7 – Maintaining standards over time

The main purpose of this criterion is to determine whether the assessment standards are appropriately maintained over time. This is essential to ensure that any increase or decrease in scores reflects real changes in learner knowledge or skills, not a shift in assessment difficulty.

Though simple equating using common items or other methods may be possible in theory, countries and assessment organizations are advised to adopt a simple Item Response Theory (IRT) model to maintain standards, for example by developing pre-calibrated item banks and utilizing them for constructing multiple equivalent forms and their score conversion tables.

In the first year of submission for an assessment to be used for SDG reporting, countries need to provide details of the plan for maintaining standards over time, but they do not need to have met all criteria at this stage. This criterion will become essential when countries submit results for a second time, to show that have carried our standards maintenance processes appropriately where required (i.e. when a new version of the assessment instrument is used, rather than reusing the previously-approved assessment instrument).

The following table provides more information on this criterion. The same requirements apply to all indicators and domains.

- Table 16: Technical details for criterion 7 that assessments must meet to be acceptable for reporting for SDG 4.1.1 – reliability or comparability over time

Table 16: Technical details for criterion 7 that assessments must meet to be acceptable for reporting for SDG 4.1.1 – reliability or comparability over time

Ref.	Description of requirements

7.1	**Items that are not in the public domain may be used in multiple live test administrations**
7.2	**Conversely, items that have been released to the public domain cannot be used in future live test administrations**
7.3	**A process must be in place to ensure that the assessment is comparable or equated over time – the approach should involve either a common-item or the common-person assessment design**
7.4	**If a common-item design is employed for linking, the results of a delta analysis should be presented, offering evidence regarding the stability of common items over time** **It is essential to specify which items were common and which items were accepted (i.e., item parameters are not statistically significantly different between the administrations) after the delta analysis for linking purposes**
7.5	**In the case of a common-person design (or concordance), a concordance table should be generated using all student data within a 95% confidence interval.**

4.7.1. Statement of documentation requirement

Countries and their assessment advisors should create a detailed technical report on the proposed approach to maintaining standards, using a CTT or IRT-based approach. This report is essential to offer clarity and insight into the methods used to guarantee the maintenance of standards. At the point of submission, the procedure may not have been carried out (since the maintenance of standards only take place at the time of the second round of administration), but it is important that there is a confirmed plan for how this will be achieved.

The documentation should cover the issues and items in the table above.

4.7.2. Best practice examples (AMPL-a)

AMPL items are kept secure and confidential. This is ensured by the respective technical standard on security of material, which is also highlighted in the operational manuals.
AMPL uses a common-item design, where the same items are used over time. IRT scaling is applied, ensuring the reading and mathematics scales are stable over time. New items can be added and linked to the scale over time, ensuring ongoing stability and sustainability of the AMPL scales.

Sources:
UIS & ACER (2023) *Technical Standards: AMPL-ab*.
UIS & ACER (2023) *Field Operations Manual: AMPL-ab.*
UIS & ACER (2023) *School Coordinator Manual: AMPL-ab*.
UIS & ACER (2023) *National Project Managers Manual: AMPL-ab*.
UIS & ACER (2023) *Test Administrators Manual: AMPL-ab*.
UIS & ACER (2023) *Study Design: AMPL-ab*.
UIS & ACER (2023) *Assessment Blueprint: AMPL-ab*.

5. References

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thondike (Ed.) Educational Measurement (2nd ed.). Washington, DC.: American Council on Education.

Chang, L. (1999). Judgmental item analysis of the Nedelsky and Angoff standard-setting methods. Applied Measurement in Education, 12(2), 151-165.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37- 46.

Ferdous, A. & Plake, B. (2007). A mathematical formulation for computing inter-panelist inconsistency for Body of Work, Bookmark, and Yes/No Variation of Angoff methods. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Chicago, IL.

Ferdous, A. (2023). "Estimating Reading Comprehension Benchmarks using Oral Reading Fluency Benchmarks Set Through Policy Linking." CIES Conference, Washington, DC, February 2023.

Ferdous, A. & Muller, E. (2024 forthcoming). "Estimating Reading Comprehension Benchmarks Using Oral Reading Fluency Benchmarks Set through Policy Linking Method." Journal of Educational Assessment in Africa. Available at https://drive.google.com/drive/fold- ers/1ejU9mHU1CD4es6--C-NJ2fhrV9_daRJb.

Impara, J. C., & Plake, B. S. (1997). Standard Setting: An Alternative Approach. *Journal of Educational Measurement*, *34*(4), 353–366. http://www.jstor.org/stable/1435114

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. Journal of Educational Measurement, 32, 179-197.

Montoya, S. & Crouch, L. The learning assessment market: pointers for countries – part 1. Blog available at: https://world-education-blog.org/2019/04/26/the-learning-assessment-market-pointers-for-countries-part-1/

Montoya, S. & Crouch, L. The learning assessment market: pointers for countries – part 2. Blog available at: https://world-education-blog.org/2019/05/20/the-learning-assessment-market-pointers-for-countries-part-2/

Montoya, S. & Crouch, L. Compare, align, track: The foundational learning data challenge. Blog available at: https://world-education-blog.org/2023/09/13/compare-align-track-the-foundational-learning-data-challenge/

Plake, B., Ferdous, A., & Buckendahl, C. (2005). Setting multiple performance standards using the yes/no method: An alternative item mapping method. Paper presented to the meeting of the National Council on Measurement in Education (NCME), Montreal, Canada.

UNESCO Institute for Statistics. Data browser available at: https://databrowser.uis.unesco.org/

6. Appendix A - Issues arising regarding the suitability of various newer assessments

As discussed briefly in the main text, much of the commentary on the "demotion" of indicator SDG 4.1.1a related to the fact that it was a shame that one could not make use of the fact that there was much measurement, in many countries, from some "newer" assessments such as EGRA, UNICEF's FLM, and various flavours of citizen-led assessments. In principle these could add a lot of countries to the list submitting data through ERCE and PASEC. However, it was also discussed at the GAML meeting in December 2023 that some of the Grade 2/3 assessments that have been proposed for reporting on SDG 4.1.1a. were not originally designed for the purpose of global reporting. In fact, comparability was distinctly and explicitly discouraged in some cases.

They were originally designed to underwrite policy dialogue, to track pilot projects, and for research purposes. Furthermore, some of them were not centralized and standardized. In fact, relatively free use was actively encouraged, with little assertion of intellectual property, and with little centralized control, by anyone, including the originators. This was done to encourage measurement in an accessible manner. However, the implication is that to retrofit these assessments for the purpose of global reporting would be a difficult task, as their very purpose, originally, was something quite different from the current retrofitted purpose of reporting. And to do it in a rush, given the change in status of the indicator 4.1.1a., is even more difficult.

It was noted that there would be a danger of losing credibility not only for these assessments but for the goal itself, if the community of interest on these issues proposes a retrofit that is excessively non-rigorous or inelegant. On the other hand, these assessments have been useful in programmatic design and implementation, and there is some documentation sustaining this claim though not as extensive and centralized as that which exists for the assessments that have already been accepted for reporting, such as ERCE and PASEC. So, it seems worthwhile to try to see how they can be useful, but with new rigor and centralized documentation, for reporting on SDG 4.1.1a.

The criteria in this document, particularly in Section 4, aim to make it possible to have more reporting, while maintaining a level of rigor and documentation that is needed for reporting purposes that are, as noted, quite new, and after-the-fact, to these assessments. In the end, various meetings of the TAG discussed in the main text, and the list of criteria in this document, makes it clear that it would be impossible to use these past assessments, even if one could make up ex-post documentation and carry out psychometric analysis of reliability, as few if any meet the relevance or alignment criteria (Criterion 1) especially around the number of needed score points in comprehension in reading, and in both numeracy and mathematics. In the end, while in principle it may be possible to find an assessment in this "newer" category that fits the criteria in this document, it seems worthwhile to focus one's efforts on the future and not spend too much effort retrofitting or justifying usage of assessments that were not designed to meet the demands of reporting.

7. Appendix B – Grade-level GPF subconstructs

The following sections detail the subconstructs in the relevant grades of the GPF. In reading, these are separated into the constructs of reading comprehension. In mathematics, these are separated into the domains.

7.1. Grade 2 reading comprehension

7.1.1. Retrieve information

R1.1 Recognize the meaning of common grade-level words
R1.2 Retrieve explicit information in a grade-level text by direct- or close-word matching

7.2. Grade 2 mathematics

7.2.1. Number and operations

N1.1 Identify and count in whole numbers, and identify their relative magnitude
N1.2 Represent whole numbers in equivalent ways
N1.3 Solve operations using whole numbers
N1.4 Solve real-world problems involving whole numbers

7.2.2. Measurement

M1.1 Use non-standard and standard units to measure, compare, and order
M2.1 Tell time
M2.2 Solve problems involving time
M3.1 Use different currency units to create amounts

7.2.3. Geometry

G1.1 Recognize and describe shapes and figures
G2.1 Compose and decompose shapes and figures
G3.1 Describe the position and direction of objects in space

7.2.4. Statistics and Probability

S1.1 Retrieve and interpret data presented in displays

7.2.5. Algebra

A1.1 Recognize, describe, extend, and generate patterns
A3.2 Demonstrate an understanding of equivalency

7.3. Grade 5 reading comprehension

7.3.1. Retrieve information

R1.1 Recognize the meaning of common grade-level words
R1.2 Retrieve explicit information in a grade-level text by direct- or close-word matching
R1.3 Retrieve explicit information in a grade-level text by synonymous word matching

7.3.2. Interpret information

R2.1 Identify the meaning of unknown words and expressions in a grade-level text
R2.2 Make inferences in a grade-level text
R2.3 Identify the main and secondary ideas in a grade-level text

7.3.3. Reflect on information

R3.1 Identify the purpose and audience of a text
R3.2 Evaluate a text with justification

7.4. Grade 5 mathematics

7.4.1. Number and operations

N1.1 Identify and count in whole numbers, and identify their relative magnitude
N1.2 Represent whole numbers in equivalent ways
N1.3 Solve operations using whole numbers
N1.4 Solve real-world problems involving whole numbers
N2.1 Identify and represent fractions using objects, pictures, and symbols, and identify relative magnitude
N2.2 Solve operations using fractions
N2.3 Solve real-world problems involving fractions
N3.1 Identify and represent decimals using objects, pictures, and symbols, and identify relative magnitude
N3.2 Represent decimals in equivalent ways (including fractions and percentages)
N3.3 Solve operations using decimals

7.4.2. Measurement

M1.1 Use non-standard and standard units to measure, compare, and order
M1.2 Solve problems involving measurement
M2.1 Tell time
M2.2 Solve problems involving time

7.4.3. Geometry

G1.1 Recognize and describe shapes and figures
G2.1 Compose and decompose shapes and figures
G3.1 Describe the position and direction of objects in space

7.4.4. Statistics and Probability

S1.1 Retrieve and interpret data presented in displays
S2.1 Describe the likelihood of events in different ways

7.4.5. Algebra

A1.1 Recognize, describe, extend, and generate patterns
A3.2 Demonstrate an understanding of equivalency

7.5. Grade 8 reading comprehension

7.5.1. Retrieve information

R1.1 Recognize the meaning of common grade-level words
R1.2 Retrieve explicit information in a grade-level text by direct- or close-word matching
R1.3 Retrieve explicit information in a grade-level text by synonymous word matching

7.5.2. Interpret information

R2.1 Identify the meaning of unknown words and expressions in a grade-level text
R2.2 Make inferences in a grade-level text
R2.3 Identify the main and secondary ideas in a grade-level text

7.5.3. Reflect on information

R3.1 Identify the purpose and audience of a text
R3.2 Evaluate a text with justification
R3.3 Evaluate the status of claims made in a text
R3.4 Evaluate the effectiveness of a text

7.6. Grade 8 mathematics

7.6.1. Number and operations

N3.2 Represent decimals in equivalent ways (including fractions and percentages)
N3.3 Solve operations using decimals
N3.4 Solve real-world problems involving decimals
N4.2 Solve operations using integers
N4.3 Solve real-world problems involving integers
N5.1 Identify and represent quantities using exponents and roots, and identify the relative magnitude
N5.2 Solve operations involving exponents and roots
N6.1 Solve operations involving integers, fractions, decimals, percentages, and exponents

7.6.2. Measurement

M1.1 Use non-standard and standard units to measure, compare, and order
M1.2 Solve problems involving measurement
M2.2 Solve problems involving time

7.6.3. Geometry

G1.1 Recognize and describe shapes and figures
G2.1 Compose and decompose shapes and figures
G3.1 Describe the position and direction of objects in space

7.6.4. Statistics and Probability

S1.1 Retrieve and interpret data presented in displays
S1.2 Calculate and interpret central tendency
S2.1 Describe the likelihood of events in different ways
S2.2 Identify permutations and combinations

7.6.5. Algebra

A2.1 Evaluate, model, and compute with expressions
A3.1 Solve problems involving variation (ratio, proportion, and percentage)
A3.3 Solve equations and inequalities