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

Draft 1
31 January 2024

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 First Vice Chair, with the kind collaboration of Abdullah Ferdous of AIR, and Kemran Mestan, Maurice Walker, and Colin Watson of ACER.
Table of Contents

1. Introduction ........................................................................................................................................3
   1.1. Background ..............................................................................................................................3
   1.2. Process as context, status of this document ..............................................................................4
2. General principles and requirements ..............................................................................................5
3. Technical criteria that assessments must meet to be acceptable for reporting .......................8
4. References ........................................................................................................................................25
5. Annex ............................................................................................................................................26
1. Introduction

1.1. Background

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. This numbers contrast sharply as compared to the 203 countries reporting primary school enrollment, indicating a mere 18% reporting at the lower primary level compared to reporting enrollment figures. Perhaps, more importantly, the number of countries reporting is not increasing quickly enough. During 2013-2016, only 30 countries reported for SDG 4.1.1.a, 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 enrollment reporting rates.

To some degree, this lack of reporting, especially for SDG 4.1.1.a, is somewhat expected. Learning assessments for the end of primary and lower secondary have a relatively venerable history, whereas learning assessments suitable for SDG 4.1.1.a 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 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 4.1.1.a was “demoted” from a Tier I to a Tier II indicator due to lack of coverage. The community of interest concerned with foundational learning, such as the Global Coalition for Foundational Learning, immediately expressed deep concern, due to the possible signaling 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. 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 4.1.1.a, 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

---

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

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.
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. However, 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 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.a, while ensuring an acceptable standard of rigor. The document proceeds as follows:

- In the sub-section immediately following we place this document into the context of a process that we will follow in reviewing the document, reading comments from the community of interest and a Technical Advisory Group (TAG) to be appointed, and coming up with a final draft. The draft Terms of Reference for the TAG are attached as an Annex. The main thing to note about this TAG is that its purpose is to comment and advise on the criteria in this draft document, not to provide ongoing approval, or not, of specific assessments or assessment results as reported to UIS.
- In section 2 the document sets out a general set of principles and 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 the criteria read as they do.
- In section 3 the document lays out the technical criteria that assessments ought to comply with to be acceptable for reporting.

### 1.2. Process as context, status of this document

This document will remain as draft document until the TCG has discussed and recognized it as an elaboration or further specification of the outcome of the 11 December meeting. The process leading up to that point is as follows, and going forward (as per e-mail from Silvia Montoya to key foundational coalitions partners and advisors on 21 December 2023), in summarized form (with slight edits for sequential numbering):  

<table>
<thead>
<tr>
<th>Task 1.</th>
<th>Documentation regarding eligibility criteria for reporting on SDG 4.1.1.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1.a.</td>
<td>Draft operational Eligibility Criteria resulting from GAML meeting and TCG decisions (this document).</td>
</tr>
<tr>
<td>Task 1.c.</td>
<td>Document 1 shared with interested parties for comment. Note: this document at hand is Document 1.</td>
</tr>
<tr>
<td>Task 1.d.</td>
<td>UIS convenes the TAG to review the criteria contained in this document and provide recommendations to UIS.</td>
</tr>
<tr>
<td>Task 1.e.</td>
<td>UIS will produce a document for Criteria for Eligibility to</td>
</tr>
</tbody>
</table>

---

4 Some of these dates may change a little but the sequence is expected to stay the same.
5 For a description of the Terms of Reference of the TAG, see Annex 1.
be consulted with TCG.

| Task 1.f. | TCG is consulted and the results are communicated to the international community with an interest in foundational learning. | To be completed by 15 April 2024. |
| Task 2. | Data plan for indicator 4.1.1.a for TCG and IAEG-SDG. (Specification of template for contents of data plan to be shared by UIS by 10 April 2024). | To be completed 30 April 2024 |

**Note: highlighted issues for the TAG.** Given the need for speed in producing this document, and given that not all relevant issues were aired at the GAML meeting of 6-7 December 2023 and 11 December 2023 so that a “sense of the meeting” could develop, there are a few issues that are tagged for particular input from the TAG that will review this document, as set out in Task 1.d in the table above. Note that the TAG is not expected to determine the suitability of particular assessments for reporting; it is expected only to comment and advise on the criteria in this document. However, the TAG is expected to give advice on the following topics.

- Whether there need for an ongoing funded committee or team to assist the UIS in determining the acceptability of particular country assessment applications, by applying the criteria in this document, based on a checklist they would produce based on this document.

- Whether such a mechanism could be outsourced by the UIS to a team of experts, possibly an institution-based team of experts, with extensive background in assessment, versus a free-standing committee directly managed and coordinated by UIS.

- Whether it might be the case the methodologies of assessments such as UNICEF/MICS/FLM, EGRA/EGMA, PAL (and other similar ones that may come along, and/or recent reinventions of these tools) are sufficiently standardized and rigorously documented as per the criteria in this document, that any and all country applications could be accepted “in batch” on the strength of the standardization of the methodology.

**2. General principles and requirements**

These considerations and requirements are above and beyond the technical criteria described in section 3 below. However, they are not less important. They are listed separately because they pertain more to process than to technical requirements. This section also covers one or two issues that were only very lightly discussed during the GAML meeting of 6-7 December and on which there was, therefore, no “sense of the meeting”. These are posed as less binding criteria or are even posed as issues on which to seek further advice from consulting experts, the community, and the TAG.

UIS regards this set of principles as largely non-negotiable, and expects the TAG to understand them, whereas the criteria in section 3 are more open to discussion, interpretation, and advice from the TAG. There are two reasons why these principles are seen as relatively non-negotiable. First, we see these principles as following directly in spirit and often in words, the sense of the meeting at the GAML meeting on 6-7 December and the TCG meeting on 11 December. The principles are seen as necessary in order to be consequent with these meetings and decisions. Second, if we do not hold these principles firmly, we risk having to go back to the beginning of all the discussions, and delaying implementation further.
Retrofitting of assessments. Some of the Grade 2/3 assessments that have been proposed for reporting on SDG 4.1.1.a. 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 retro-fit these assessments for the purpose of global reporting is a difficult task, as their very purpose, originally, was something quite different from the current retro-fitted purpose of reporting. And to do it in a rush, given the change in status of the indicator 4.1.1.a., is even more difficult. There is 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 retro-fit 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 4.1.1.a. The criteria in this document, particularly in Section 3 below, 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.

Country interest and coordination. It will be up to any relevant country's authorities to decide whether they want to use a certain assessment for reporting on SDG 4.1.1. This interest should be expressed formally to UIS by the country authorities. The authorities may also specifically ask that a certain assessment (or its application in any given year) is not to be used. To prevent excessive lobbying of countries by assessment organizations or bilateral donors, it is expected that the reporting by any country, and the decision as to which assessment is used, will be coordinated by UIS. At the same time, if a country chooses to report according to an assessment, it is expected that the suppliers of that assessment assist the country in lining up the documentation, especially in cases where application country by country is not completely standardized. Assessment organizations are expected to budget for this work, which can also help build capacity.

Documentation. A dossier or set of files documenting the technical criteria described below should be made available by the country in question or its assessment advisors for reporting purposes, to UIS and to the public, in one single, simple, well-organized online portal. The dossier may consist of more than one file, but the files need to be well organized and easy to find, with hyper links between files offered where relevant. The contents of that file or dossier need to include documentation that shows how all the criteria in this section and the next have been met or plan (concretely, in detail) to be met, including purpose(s) of the assessment; definitions of domain, constructs, subconstructs, and learning outcomes measured; define the examined population; interpretations for the intended uses; define the content of the test; the item formats; time allowed for testing; directions for test takers; and scoring and reporting procedures. The dossier or file may include links to other files.

Sustainability plan. The reporting should go beyond reporting, and ought to contain a sustainability plan that expresses the country's desire to use the assessment again over time, and to have its capability in the use of the assessment, and similar assessments, built up by the organization providing the assessment support. That is, 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 in fact has a track record of providing capacity building and transfer of capabilities.

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, for policy dialogue, policy setting, capacity building, monitoring, etc., either of general policies or of specific improvement programs. Ideally, the
assessments should not just report but help the countries do better on the skills on which they are reporting. The file or dossier should contain and explanation of how this utility has been generated or will be generated.

**Comparability over time.** To be useful for reporting, but also to the countries themselves, the assessments must be comparable over time, which means that techniques needed to equate their difficulty over time must have been used or plan to be used. The assessment must be susceptible, technically, of being equated over time. Acceptable techniques for guaranteeing comparability over time are discussed in Table 1, row 5. Note that these need to have been documented in the manner described often in this document for the rest of the technical criteria.

**Criteria to apply to past data as well as future data.** In an ideal world we want all criteria to apply to forthcoming assessment applications as well as to previous ones. For one thing, it would make little sense to report past data that are not comparable to future data, in terms of the basic nature of, the difficulty of the assessment (and thus the equating assessment versions), if applied at various points in time. Similarly, to ensure proper comparability, confidence intervals for the assessment, and other reliability considerations, ought to meet a similar bar for the past as for the future. Otherwise one runs the danger of creating unjustified despondency if the indicator seems to be going down, or unjustified optimism if it appears to be going up, at best, and a loss of respect for the measurement at worst.

**Consistency with an efficient ecosystem or market for assessments.** In the past few years, Montoya and Crouch have published blogs [here](#), [here](#), [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, and in general.

*It may be that some ambiguity or difficulties need to remain* and further decisions need to be made over time, in real time. It may not be possible to offer quantitative numerical benchmarks, in this document, that are clear and very simple and unambiguous (such as that the alpha coefficient must be above a certain threshold, or the sample must be of a certain size) on every single criterion in this note. Some ongoing committee or small team of experts will be needed on an ongoing basis to provide UIS with advice on whether a certain assessment meets the criteria.

One area that seems destined to remain fraught with the need for human judgment on a case by case basis is the issue of how to link to, or benchmark to, the Global Proficiency Framework (*GPF*) or the Minimum Proficiency Levels (*MPL*). This appears to require some judgment specific to each preferred assessment at least, and perhaps specific to each country. Certainly, that would be the case if, for example, a country chooses to use its own assessment, whether of a standard pencil-and-paper type or a one-on-one assessment. But, also broadly in order to prevent a sense from the community that the process is excessively top-down.

**National assessments.** Related to the point immediately above, the issue of using countries’ own national assessments to report on SDG 4.1.1. did not receive much discussion at the 6-7 December 2023 GAML meeting or the 11 December 2023 TCG meeting and hence there is no “sense of the meeting.” All the criteria stated in this document would presumably apply to national assessments. However, it would make sense to set out a process, as recommended at the end of section 1.2 above, on how UIS can decide which assessments to accept. The TAG is asked to make a recommendation in this issue.
3. Technical criteria that assessments must meet to be acceptable for reporting

This section sets out in detail the criteria assessments to be considered for reporting on SDG 4.1.1, with numerical values to the extent possible, and with an extensive illustration from AMPLa. As will be noted, the criteria tend to be more specific for SDG 4.1.1.a as this is the weakest of the SDG 4.1.1 indicators. 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 meeting. The relevant documents are here and here respectively.

<table>
<thead>
<tr>
<th>Criterion Area</th>
<th>Elaboration</th>
<th>Criterion threshold numerical value as per GAML</th>
<th>Notes, explanations, extensions</th>
<th>Best practice examples (AMPLa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Alignment and validity</td>
<td>Is the assessment aligned to the MPL and GPF?</td>
<td><strong>Reading</strong> - minimum 10 score-points assessing reading comprehension and the assessment must cover both reading comprehension sub-constructs at grade 2 in the GPF (see right). The remaining items can be drawn from any of the domains (decoding, listening comprehension or reading comprehension). <strong>Mathematics</strong> - minimum 10 score-points assessing number and operations and</td>
<td>In conventional terms, this criterion is based in the concept of “validity” but also possibly “utility.” Wording from the GPF for Grade 2 for reading comprehension: 1.1 Recognize the meaning of common grade-level words. R1.2 Retrieve explicit information in a grade-level text by direct- or close-word</td>
<td>The AMPL-a reading assessments include decoding and listening comprehension items in addition to reading comprehension, as follows: Listening comprehension (Audio): 10 items Decoding (Audio): 5 items Reading comprehension: 25 items</td>
</tr>
</tbody>
</table>

---

6 There is no implication that any given assessment has to pass the same bar as the AMPLa set for itself. This is used as a best practice example.
7 In Draft 1 of this document no links or bibliographical references are provided for most rows of the matrix, except for examples from AMPLa and in row 6 of the matrix, and elsewhere if the point being made does not refer to standard and easily available literature. Full references could be provided in a subsequent or final draft, if there is a demand from the community.
8 For convenience and to save space, AMPLa is used. AMPLa is part of the AMPL family of assessments. The main aim of the Assessments for Minimum Proficiency Levels ‘a’ and ‘b’ (AMPL-ab), is to measure and analyze the reading and mathematics proficiency of students at the end of lower primary (SDG indicator 4.1.1a) and at the end of primary school education (SDG indicator 4.1.1b). Four countries participated in the international AMPL-ab study: The Gambia (Grade 3), Kenya (Grade 6), Lesotho (Grade 7) and Zambia (Grade 4 & Grade 7).
Table 1. Technical criteria that assessments must meet to be acceptable for reporting

<table>
<thead>
<tr>
<th>Criterion Area</th>
<th>Elaboration</th>
<th>Criterion threshold numerical value as per GAML</th>
<th>Notes, explanations, extensions</th>
<th>Best practice examples (AMPLa)⁸</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>the assessment must cover all four <strong>number and operations</strong> sub-constructs at grade 2 in the <a href="#">GPF</a>. The remaining items can be drawn from any of the domains (<strong>number and operations, measurement, geometry, statistics and probability or algebra</strong>). 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 as noted in Section 2: purpose(s) of the assessment; definitions of domain, constructs, subconstructs, and learning outcomes measured; define the examinee population; interpretations for the intended uses; define the content of the test; the item formats; time allowed for testing; directions for test takers; and scoring and reporting procedures.</td>
<td>matching. Wording from the GPF for Grade 2 for mathematics, 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. In reading assessments that are aimed at LI or LMI countries, or countries with low educational performance, and especially when the country is below benchmark for reading comprehension (see row 6 of this matrix), the reading comprehension score itself will not be very informative. In those cases the country can be encouraged to include other sub-constructs as specified in the <a href="#">MPL</a> and <a href="#">GPF</a> that can be considered precursors of the two chosen here. Sub-constructs such as decoding, accuracy of reading, fluency, etc., would be suitable. These are likely to add to the</td>
<td>Decoding: 5 items Mathematics: 30 items UIS &amp; ACER (2023) <em>Study Design: AMPLab</em>. UIS &amp; ACER (2023) <em>Assessment Blueprint: AMPLab</em>.</td>
<td></td>
</tr>
</tbody>
</table>
### Table 1. Technical criteria that assessments must meet to be acceptable for reporting

<table>
<thead>
<tr>
<th>Criterion Area</th>
<th>Elaboration</th>
<th>Criterion threshold numerical value as per GAML</th>
<th>Notes, explanations, extensions</th>
<th>Best practice examples (AMPLa)</th>
</tr>
</thead>
</table>
| 2. Item content and quality | Is there evidence that the items in the assessment? Does the item review process include empirical item analyses and expert judges? The qualitative review should consider | | in-country utility (utility being seen as important value in addition to reportability, as per Section 2) of the assessment for programming and policy, beyond reporting. If necessary, equivalences between these precursor skills and reading comprehension can be used, because the benchmark is reading comprehension, but one may be able to link comprehension to a precursor skill using a method such as the IRT method by Ferdous (2023). In that case, for example, mean fit statistics should be around 1.0, and standardized fit statistics ought to be between -1.9 and 1.9 (in terms of z scores) as suggested in the literature e.g. [here](#). Note that this approach would make a decision on a conjunctive versus compensatory model moot. If the assessment fulfills these characteristics and those in the other rows of this table, it could be considered satisfactory. | 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: |
### Table 1. Technical criteria that assessments must meet to be acceptable for reporting

<table>
<thead>
<tr>
<th>Criterion Area</th>
<th>Elaboration</th>
<th>Criterion threshold numerical value as per GAML</th>
<th>Notes, explanations, extensions</th>
<th>Best practice examples (AMPLa)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>have been reviewed qualitatively and quantitatively</td>
<td>whether: Each assessment item is considered appropriate by relevant experts for inclusion in the assessment. The expert teams ought to include nationals of the reporting country or similar countries. Were the assessment items developed by subject matter experts (SMEs)? Have the items been thoroughly discussed with and vetted by local experts? Are the SMEs trained in item development principles and procedures? Are the items field tested on a representative sample of the examinee population? The scoring guides are consistent with what the item is intended to measure.</td>
<td>this but ideally the item review should be explicit about these issues. As noted, there ought to be enough items on reading comprehension as per above. Items relating to decoding, fluency, accuracy, etc., may need slightly different analyses.</td>
<td>i) the items were suitable for students working at the level of lower primary ii) the items were multiple-choice (or another closed item format) iii) the items did not use a sentence fragment as the item stem (since this format can be difficult to translate) iv) the items originated in either English or French, and v) (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.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Quantitative review**

Psychometric quality assurance analysis of AMPL-a and AMPL-b items was undertaken. Analytical outputs include: ‘Facility’,
### Table 1. Technical criteria that assessments must meet to be acceptable for reporting

<table>
<thead>
<tr>
<th>Criterion Area</th>
<th>Elaboration</th>
<th>Notes, explanations, extensions</th>
<th>Best practice examples (AMPLa)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rationalized or the distractors in a multiple-choice item should be negatively correlated with ability.</td>
<td>‘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.</td>
<td>‘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.</td>
</tr>
</tbody>
</table>

Countries or their assessment advisors should produce a test development report documenting the procedure used to develop, review, and select items from the item pool. It should also include the qualifications, relevant experience, and demographic characteristics of the expert judges who reviewed the items.

<table>
<thead>
<tr>
<th></th>
<th>Is psychometric item analysis conducted on the field test data using classical test theory (CTT)?</th>
<th></th>
<th>Summary findings include:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Do all items have a difficulty level between 0.20 and 0.90 (with CTT it’d be % correct)</td>
<td></td>
<td>The mean score on the 72 items was 39.1 and the standard deviation was 14.9.</td>
</tr>
<tr>
<td></td>
<td>Do all items have an item-total correlation (or point biserial) value of at least 0.20?</td>
<td></td>
<td>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.</td>
</tr>
</tbody>
</table>

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 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.

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 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.

Summary findings include:
Table 1. Technical criteria that assessments must meet to be acceptable for reporting

<table>
<thead>
<tr>
<th>Criterion Area</th>
<th>Elaboration</th>
<th>Criterion threshold numerical value as per GAML</th>
<th>Notes, explanations, extensions</th>
<th>Best practice examples (AMPLa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Sample representativeness and sampling rigor</td>
<td>Is the sample of learners that took the assessment representative of the population against which the results will be reporting?</td>
<td>Inclusion of the specification and statistical justification of the sample size. Sample size robustness to Type 1 and Type 2 errors should be indicated. Documentation of minimum power 0.8 and minimum significance level 0.05. Explanation of the sample approach and design (stratification, clustering, etc.). Documentation of design effect to be included. Where the assessment is administered</td>
<td>Samples that are created ex-post from regional or special purpose samples could be accepted on a case-by-case basis and on the basis of thorough numerical analysis. These analysis include Montecarlo and other simulation procedures, and examination of the characteristics of the samples used in terms of correlation with known variables for a national sample. The predicted error from aggregation, in all of at least 5 correlations, must be less than a 5% confidence interval.</td>
<td>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). <em>Minimum Proficiency Levels Unpacked</em> UIS &amp; ACER (2023) <em>Assessment Blueprint: AMPLab.</em> UIS &amp; ACER (2023) <em>Item Analysis Report - Reading: AMPLab.</em> UIS &amp; ACER (2023) <em>Item Analysis Report - mathematics: AMPLab.</em> The AMPL-ab 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-ab in each population assessed.</td>
</tr>
<tr>
<td>Criterion Area</td>
<td>Elaboration</td>
<td>Notes, explanations, extensions</td>
<td>Best practice examples (AMPLa) noexcept</td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td>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. Evidence of correction for design effect should be included. 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. Was the assessed population defined? Does the country have an acceptable</td>
<td>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:</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 1. Technical criteria that assessments must meet to be acceptable for reporting

<table>
<thead>
<tr>
<th>Criterion Area</th>
<th>Elaboration</th>
<th>Criterion threshold numerical value as per GAML</th>
<th>Notes, explanations, extensions</th>
<th>Best practice examples (AMPLa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sampling frame? Is the assessment administered to the whole cohort? Is there any subgroup of the examinee population systematically excluded? Explain. Is the sample size adequate (based on statistical power analysis) for national-level estimates, disaggregated by gender? Is the margin of error considered 5% or less (at a 95% confidence level)? What is the design effect used in the sample size calculation when the cluster sampling method is used? What is the intra-class correlation (ICC) considered for sample size calculation? Are sampling weights calculated and accounted for in national estimates? If a country has multiple official languages of instructions (LOIs), are reading assessments conducted in all LOIs? For reading, are national level estimates computed after appropriately weighted assessment results conducted on all LOIs?</td>
<td></td>
<td>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 &amp; ACER (2023) Sampling Framework: AMPLab. UIS &amp; ACER (2023) UIS &amp; ACER (2023) Sampling Framework: AMPLab. UIS &amp; ACER (2023). AMPLab Sample Information and Outcomes. (1 report for each country) UIS &amp; ACER (2023) Population Definition and Sample Design. (1 report for each country)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 1. Technical criteria that assessments must meet to be acceptable for reporting

<table>
<thead>
<tr>
<th>Criterion Area</th>
<th>Elaboration</th>
<th>Criterion threshold numerical value as per GAML</th>
<th>Notes, explanations, extensions</th>
<th>Best practice examples (AMPLa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. Assessment administration and data custodianship</td>
<td>To be suitable for reporting against SDG 4.1.1, there must be evidence that the assessment was administered in an appropriate and standardized way</td>
<td>Has a standardized test administration manual been produced? Is multiple-day training conducted for enumerators? Is training conducted for supervisors or quality control officers (QCO)? Has there been any dry run or practice session conducted for enumerators and QCOs? Do enumerators meet the required selection criteria (e.g., knowledge, skills, and abilities observed during training and dry runs)? Do the enumerators have adequate time to administer the assessment?</td>
<td>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...</td>
<td></td>
</tr>
</tbody>
</table>
Table 1. Technical criteria that assessments must meet to be acceptable for reporting

<table>
<thead>
<tr>
<th>Criterion Area</th>
<th>Elaboration</th>
<th>Notes, explanations, extensions</th>
<th>Best practice examples (AMPLa)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>What proportion of the actual sample has been observed by supervisors or quality control officers? The directions for test administration 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. Allowable variations of administration procedures should be clearly described. Moreover, the process for selecting, training, and qualifying enumerators and quality control officers should be specified by the test developer. 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. Protocol for field supervision, in writing, just exist and be adequate. Informed consent was used.</td>
<td>implementation of activities. Four standards relate to quality monitoring, including: The AMPLab test administration is monitored using school visits by trained independent QMs. At least 5% school visits are conducted in each participating country to observe AMPLab test administration sessions. AMPLab 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 center upon data submission. Five specific standards relate to test administrators, including: All AMPLab assessment sessions</td>
<td></td>
</tr>
<tr>
<td>Criterion Area</td>
<td>Elaboration</td>
<td>Notes, explanations, extensions</td>
<td>Best practice examples (AMPLa)²</td>
</tr>
<tr>
<td>----------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-----------------------------------</td>
</tr>
<tr>
<td></td>
<td>Privacy, encryption, and anonymization procedures were used. 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 of pre-pilot and pilot and corrections made on that basis, must be provided. Details of administrator training, quality assurance procedures and quality assurance outcomes should also be made available publicly.</td>
<td>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 &amp; ACER (2023) Technical Standards: AMPLab. UIS &amp; ACER (2023) UIS &amp; ACER (2023) Technical Standards Review: AMPLab. UIS &amp; ACER (2023) UIS &amp; ACER (2023) Field Operations Manual: AMPLab. UIS &amp; ACER (2023) UIS &amp; ACER (2023) School Coordinator Manual: AMPLab. UIS &amp; ACER (2023) UIS &amp; ACER (2023) National Project Managers Manual: AMPLab. UIS &amp; ACER (2023) UIS &amp; ACER (2023) Test Administrators Manual: AMPLab.</td>
<td></td>
</tr>
</tbody>
</table>
5. Reliability

**Reliability at any given point in time**

Does the assessment have a reliability coefficient (alpha) of at least 0.70? (Yes/No)

If an assessment is used for a range of ages (e.g., MICS-FLM), does the assessment have a reliability coefficient (alpha) of at least 0.70 for 7-9 years old, who attend grade 2 in formal schooling (Standard 2.12)?

If an assessment contains constructed response (CR) and/or oral assessments with any type of performance-based items, do enumerators or those who score the assessment have an inter-rater reliability (IRR) of at least 0.80? For oral one-on-one assessments, reported inter-rater reliability must be greater than a kappa coefficient of 0.7. Confidence interval on the proportion at or above the minimum must be reported, documented, and be equal to less than 0.05.

Items with weak reliability were considered and excluded or included only with great justification.

Item DIF for gender, and other important factors in the countries in question was used to analyze item inclusion and exclusion using IRT or classical equivalent.

Countries or the assessment

The assessments must be reliable at any given point in time. Informally, 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.

The assessment must also be reliable over time, in that any increase or decrease in scores must reflect improved or worsened student 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 develop pre-calibrated item banks and utilize them for constructing multiple equivalent forms and their score conversion tables.

The policy linking method (whether for one or more benchmarks) establishes benchmarks on a raw score scale (e.g., if a test consists of 15 reading comprehension items, each valued at 1 point, then the raw score scale for the reading comprehension test ranges from 0 to 15). Subsequently, these benchmarks are converted into Item Response Theory (IRT)-based theta values, representing students' true ability in reading comprehension. These theta values for each of the reading and mathematics scales in the AMPla+b is calculated from a unidimensional model for each construct. The reliability for the reading construct is provided on line 209 of the ACER ConQuest output file. Weighted EAP/PV reliability: 0.906

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

AMPLab 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 AMPlab 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
Table 1. Technical criteria that assessments must meet to be acceptable for reporting

<table>
<thead>
<tr>
<th>Criterion Area</th>
<th>Elaboration</th>
<th>Criterion threshold numerical value as per GAML</th>
<th>Notes, explanations, extensions</th>
<th>Best practice examples (AMPLa)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>organizations assisting them, are advised to calculate and include relevant reliability coefficients in the technical report for each total score, sub score, or combination of scores intended for interpretation.</td>
<td>benchmarks remain constant throughout the lifespan of the assessment program, serving as a reference for measuring students' progress in reading comprehension across various assessments over time.</td>
<td>the excluded items.</td>
<td>the excluded items.</td>
</tr>
<tr>
<td></td>
<td><em>Reliability or comparability over time</em></td>
<td></td>
<td></td>
<td>Item DIFF (i.e. differential item functioning) for gender, was used to analyze item inclusion and exclusion using IRT. The DIF results for each item can be observed in the Item Analysis reports.</td>
</tr>
<tr>
<td></td>
<td>The bases for judging the assessment to be comparable or equated over time must be documented.</td>
<td></td>
<td>Sources:</td>
<td>UIS (in press). <em>Assessment of Minimum Proficiency Level (AMPLab): International Report</em>, UNESCO Institute for Statistics, ACER.</td>
</tr>
<tr>
<td></td>
<td>The approaches should involve either a common-item or the common-person assessment design. 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.</td>
<td></td>
<td>ACER (2023) <em>ConQuest output file: CINT_R_itm_formreg(1).shw</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>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.</td>
<td></td>
<td>ACER (2023). <em>ConQuest output file INT_M_itm_formreg(1).shw</em></td>
<td></td>
</tr>
</tbody>
</table>
6. Linking to the MPL

<table>
<thead>
<tr>
<th>How does the assessment link to the MPL? That is, what constitutes evidence of minimum proficiency in the results obtained, and in terms of the criteria for validity and alignment, in row 1 of this matrix.</th>
</tr>
</thead>
<tbody>
<tr>
<td>This criterion in the matrix is harder to set, in terms of numerical threshold values and so on, than the others. There are a few reasons for this. First, this topic was not as thoroughly discussed at the 6-7 December GAML meeting or the 11 December TCG as the topics above. Second, there are many choices here, driven simply by the fact that there is no linking methodology provided by the psychometrics profession that dominates all others on every possible concept and on which there is consensus. While that is also not the case for the criteria above, it seems to be more nearly the case for those. Third, which method is best will therefore depend a lot on the type of assessment in question: one method may be best for the more standard assessments based on straightforward items, others may be more suitable for the one-on-one assessments. For now, the criteria will remain general. The AMPLa example to the right serves as a best-practice scenario and exemplifies various methods that can be used. Given the above, the following can be said.</td>
</tr>
</tbody>
</table>
for the standard written assessments typically for SDG 4.1.1.b and 4.1.1.c are well-known and the links have been established. Similar methods for similar 4.1.1.a assessments are valid and have been accepted. For unconventional or newer 4.1.1.a assessments, the following criteria apply.

Several methods can be used: policy-linking method (Angoff, 1971; Plake, Ferdous, & Buckendahl, 2005; Impara & Plake, 1997), a pairwise comparison method or other psychometric methods, if explained.

*Policy linking method*  

Do all panelists meet the requirements for participation?  

Are the group of panelists sufficiently representative in terms of the characteristics agreed by the country?  

Are all outliers removed before calculating the final benchmarks?  

Are benchmarks only set for GPLS that don’t exhibit floor or ceiling effects?  

Is the inter-rater consistency statistic kappa greater than or equal to 0.7 (Cohen, 1960; Ferdous & Plake, 2007)?  

For other linking or standard-setting methods, there may be equivalent statistics and they should be reported on.

analyzed to locate AMPL items on the LPS scale, providing validation of the cut-points. Details of the Pairwise 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 (MPLa and MPLb) on the reading and mathematics scales.

Sources:  
UIS & ACER (2023) *Scaling AMPLab Items: Technical Note*  
Is the intra-rater consistency statistic greater than or equal to 0.7 (Chang, 1999)? (This refers to whether the judgment is consistent with the measured difficulty level of the item.) For other linking or standard-setting methods, there may be equivalent statistics and they should be reported on.

Has the standard error for each benchmark been calculated and reviewed to be determined as appropriate? For other linking or standard-setting methods, there may be equivalent statistics and they should be reported on.

Has the confidence interval for each benchmark been calculated and reviewed to be determined as appropriate?

Is the mean average score for each section of the evaluation greater than or equal to 4?

Is the mean average score for the overall evaluation greater than or equal to 3?

Do actual classifications of examinees agree with those that would be made of their true scores greater than or equal to 0.7 (Livingston & Lewis, 1995)?

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

Sources:
4. References


5. Annex

Draft (as of 31 January 2024) of the Technical Advisory Group (TAG) Terms of Reference: SDG 4.1.1 Criteria Operational Implementation

Overview
The United Nations Educational, Scientific, and Cultural Organization (UNESCO) Institute of Statistics (UIS) is convening a technical advisory group (TAG) to provide recommendations that will contribute to finalizing the eligibility criteria for reporting assessment-based data on minimum proficiency levels for Sustainable Development Goal (SDG) 4.1.1 with some emphasis on 4.1.1.a, given its recent downgrade from a Tier I to a Tier II indicator: proportion of children and young people in grades 2/3 achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex.

Background
The Global Alliance to Monitor Learning (GAML) is the working group of the Technical Cooperation Group on SDG 4 Indicators (TCG) that focuses on learning data and aims at improving learning outcomes by supporting national strategies for learning assessments and developing internationally comparable indicators and methodological tools to measure progress towards key targets of Sustainable Development Goal 4 (SDG 4). An outcome of the fifth GAML meeting in Hamburg, Germany in 2018 was the notion of converging criteria around indicator 4.1.1a. It was proposed as a process of "social moderation" by MSI as contracted by UIS. The Consensus Building Meeting on Proficiency Levels in Paris in September 2018 confirmed the decisions and allowed the inclusion of the indicator in the framework based on it. Since that time, UIS has intellectually and financially supported it through external funding and the UNESCO Regular Program budget. Still, the 4.1.1.a community lacks consensus on a clear set of criteria (validity/alignment, item quality content and review, sampling, administration and data custody, and reliability) for reporting assessment-based data on minimum proficiency levels.

The need for these criteria is now critical. In December 2023, indicator 4.1.1a was downgraded from a Tier 1 to a Tier 2 indicator in the SDG framework as documented by the IAEG1. The Inter-agency and Expert Group on SDG Indicators (IAEG-SDGs) plans to eliminate all Tier 2 indicators in 2025. To maintain this significant indicator, it must be returned to Tier 1 status, and to do so, minimum reporting coverage must be achieved by 2025. Confirming the criteria is a critical and necessary step to make use of existing and forthcoming assessments conducted by the global education community to expand the coverage and reporting.

The recent GAML meeting 6-7 December 2023 showed a way to move the process forward2. UIS has articulated a series of next steps to fill this gap in an open and public process, where arguments will be made through written pieces with recommendations, sharing documentation/data/analysis, and consensus reached in an open room. These steps are presented in the text box below. Note that the dates may change a little, but the sequence of steps is likely to be firm.

---

2 https://tcg.uis.unesco.org/10th-meeting-of-the-tcg/
**Task 1: Operational implementation process of current Eligibility Criteria resulting from GAML meeting and TCG decisions**

**Expected Output:**
- Document 1 summarizing the state of the recommendations regarding reporting in a clear operational tables and decision tree.
  - UIS will produce a table ("Criteria for Eligibility") like the one used at the end of the GAML meeting, used in various presentations, with the five key criteria (validity/alignment, item quality content and review, sampling, administration and data custody, and reliability) and the description of all the needed documentation/data sharing that is a non-negotiable condition of reporting.

**Task 1 will be executed as follows:**

1. **Define a drafting group to prepare Document 1 (to be completed by 7 January 2024).** This small group of up to 4 experts will be coordinated by UIS will put a forward draft for consultation for the discussion based on past data and research inputs.
2. **Establish a technical advisory group (TAG) (to be completed by 31 January 2024).** The TAG will receive the draft of Document 1, the feedback from interested parties as stated below, and will elaborate a final set of recommendations later.
3. **Document 1 shared with interested parties for comment (to be completed by 15 February 2024).** This step implies the circulation of Document 1 as in point 1.a for a 2 weeks’ period of analysis and comment. This will allow a final and public set of comments on the Criteria for Eligibility to be brought to UIS attention. Comments from the interested parties to be publicly posted in a platform to be shared by UIS.
4. **UIS convenes the TAG (to be completed by 28 February 2024).** This group will make a final set of recommendations based on analysis and looking at documentation on past data, research, and feedback. The TAG will have been provided with Document 1 and the comments received, and any extra documentation on the reasoning behind the recommendations in Document 1 as well as references to background and related documents.
5. **UIS will produce a document for Criteria for Eligibility to be consulted with TCG (to be completed by 15 March 2024).** UIS will collate recommendations from TAG and produce a final recommendation for accepting any assessment-based data on minimum proficiency levels.
6. **TCG is consulted and decision is communicated to the international community (to be completed by 15 April 2024).**

**Task 2: Data plan for indicator 4.1.1.a for TCG and IAEG-SDG**

**Expected Output:**
- Document 2 will provide a description of the feasible data coverage increase based on Document 1 and the agencies/assessment program documentation submission and the work of the TAG.

**Task 2 will be executed as follows:**

1. **UIS will share a template for the contents of data plans no later than 10 April 2024.**
2. **Submission of each assessment party’s data plans to UIS (to be completed by 30 April 2024)**
   - Data plan for future and current coverage based on Document 1 definitions, with the accompanying supporting documentation and requirements, including a self-assessment tool filled out by the agency proposing a set of countries that would report according to any given assessment.
   - No submission will be considered unless there is all the documentation, the microdata is available in a public site and there is an explicit country agreement approving the data to be disseminated for SDG reporting.
Purpose of this Group
The purpose of the TAG during the period January - April 2024, is to serve in a policy advisory capacity in the UIS-led process articulated above.

Responsibilities
- Task 1: The TAG will review the draft UIS table ("Criteria for Eligibility", Document 1 in the text box above) with the five key criteria (validity/alignment, item quality content and review, sampling, administration and data custody, and reliability) plus an additional criterion on rigor of the method for linking to the Minimum Proficiency Levels (MPL) and Global Proficiency Framework (GPF) and the description of all the needed documentation/data sharing that is a non-negotiable condition of reporting.
- Task 2: The TAG will review all comments from interested parties to UIS.
- Task 3: The TAG will review any extra documentation on the reasoning behind the recommendations as well as references to background and related documents.
- Task 4: The TAG will convene and provide a set of final recommendations based on items reviewed in Tasks 1-3.

Membership
The TAG is comprised of a small, regionally balanced group of subject matter and psychometric experts from a representative set of countries.

Level of Effort
Time (level of effort) requirements for TAG members for the four tasks listed above is approximately 5-8 days. Determination of level of effort for any subsequent tasks is pending.

Compensation
TAG members will be compensated through an honorarium for participation in the four steps above during the period January-April 2024. In addition, all travel and per diem costs associates with any in person TAG meeting will be covered.

Conflicts of Interest
TAG members ought not to be associated in any financial or similar manner to providers of assessment or advisory services whose nature would be driven by the TAG’s recommendations.