unesco
Institute for Statistics

25 YEARS

Education
Data and
Statistics
Commission

WG/GAML/11/1.4

Report on a Standard-Setting Exercise to Set the Minimum Proficiency Level for Indicator 4.1.1(a) in Mathematics Using AMPL

Global
Alliance
to Monitor
Learning

Table of Contents

1 Introduction 2

2 Methodology 3

2.1 Bookmark Method and adjustment 3

2.2 Procedure to set cut scores 4

2.3 Decision-decision making procedure 4

3 Expert panel 5

3.1 Selection criteria 5

3.2 Training 7

4 Expert session for standard setting pilot exercise using AMPL on mathematics 2/3 grades. 7

4.1 Description of expert session 7

4.2 Individual judgement process 10

4.3 Group discussion process and results 11

5 Discussion 16

6 Annex 19

6.1 Annex 1. Description of the expert panel members 19

6.2 Annex 2. Item reviewed in individual judgment and expert comments 20

6.3 Annex 3. Confidentiality agreement (Template) 22

DRAFT Report on a Standard-Setting Exercise to set the Minimum Proficiency Level for Indicator 4.1.1(a) in Mathematics Using AMPL

January 21st, 2025

1 INTRODUCTION

This report constitutes a draft version of the second deliverable of the project titled "Design and Implementation of a Procedure for Setting the Minimum Proficiency Level for SDG 4.1.1a", as outlined in contract number 4500526126. The final version of this report is due on May 2025. The project's overall objective is to develop and test a robust and transparent methodological approach to identify and set Minimum Proficiency Levels (MPLs) using different global educational assessments.

The Sustainable Development Goals (SDGs) set an agenda for advancing sustainable development across various domains by 2030. Within the field of education, SDG indicator 4.1.1 measures the proportion of children and young people achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, disaggregated by sex, at three key stages: (a) grades 2/3; (b) the end of primary education; and (c) the end of lower secondary education. This target seeks to increase the proportion of girls and boys who meet or exceed the MPL.

Measuring this indicator globally presents numerous challenges. These include ensuring comparability across diverse tests, addressing curriculum differences between countries, accommodating linguistic diversity that influences learning progression, and the absence of a universally comparable metric to standardize the substantive definition of the MPL.

To address these challenges, this project explores the use of a method for effectively measuring and monitoring MPLs in diverse contexts. This document reports on a pilot exercise conducted to set cut scores defining the MPL for indicator 4.1.1(a) in mathematics for grades 2/3, using data from the Australian Council for Educational Research's Assessment for Minimum Proficiency Level (AMPL). The exercise had two primary objectives: first, to align the substantive definition of the MPL with AMPL, and second, to identify and address challenges encountered during the process. By using AMPL as an example, this effort aims to develop a step-by-step guide that countries, as well as regional and subregional entities, can use to leverage their own large-scale assessments to locally set MPLs and estimate the proportion of students reaching SDG 4.1.1 target.

This documentation aims to convey the complexity of the process without oversimplifying it, while also shedding light on the decision-making processes and opportunity costs associated with various methodological approaches.

Finally, it is important to recognize that setting MPL cut scores using the method proposed in this document is a highly specialized procedure requiring the involvement of experts in measurement. Such experts should possess not only statistical expertise but also a deep understanding of how learning constructs are represented in tests and measured through various means. Having experience in front of a classroom is also crucial, as it enables experts to accurately judge the difficulty of assessment items and to have a practical understanding of the curriculum. Additionally, subject-matter experts in mathematics play a fundamental role in ensuring the validity and reliability of the outcomes. The group of experts must also be diverse, encompassing individuals with experience and knowledge of different geographical and contextual settings, to minimize potential biases and ensure the process is equitable and representative.

This document is organized into six main sections. Following this introduction, Section 2 details the methodology, explaining how the Bookmark Method was used and the adaptations made for this exercise. Section 3 describes the expert panel, including selection criteria and training procedures. Section 4 presents a detailed account of the cut-score setting process, covering both individual judgments and group discussions. Section 5 provides the results of the cut-score setting exercise, including detailed tables, graphs, and analysis of the proportions of students reaching the MPL. Finally, Section 6 discusses the main challenges encountered, limitations of the process, opportunities for improvement, and implications for monitoring SDG 4.1.1a. Additional details about the items reviewed and expert comments are included in the annexes.

2 METHODOLOGY

The content of this section describes the method used to implement the exercise of setting a cut score for indicator 4.1.1(a) for grades 2/3 in mathematics. It first explains the specific method used, then it describes the procedure to set the cut score, and finally it presents the decision-making procedure implemented.

2.1 BOOKMARK METHOD AND ADJUSTMENT

Although there are different methods to set standards that may be used as means to set the MPL, we will use an adaptation of the Bookmark Method. This method is one of the most widely used (Lewis, Mitzel, & Green, 1996). This method uses both statistical and substantive evidence as inputs for experts to set cut scores which represent the border between different levels of proficiency.

The Bookmark Method combines two perspectives. On the one hand, it is data-driven by providing the levels of difficulty of items in a test, as an initial input to empirically establish how students actually answered different questions. On the other hand, experts in the corresponding subject matter analyze substantively the items and, after organized discussions and deliberations, they have to reach a consensus on where the cut score must be set, establishing the border in which it is plausible to distinguish between students who meet or surpass the MPL and those who do not.

Two adaptations of the Bookmark Method are considered in this session. First, in contrast to the Bookmark Method, this session aims at setting only one cut score (the minimum proficiency level), instead of setting several cut scores which may distinguish different levels of achievement. The second adaptation refers to not repeating the discussion in case of disagreement, due to the pilot nature of this instance and the limited time to implement this exercise.

2.2 PROCEDURE TO SET CUT SCORES

The standard setting process must include the following steps: a) identify subject matter experts (6 to 12) who are led in the discussion by a measurement expert; b) analyze student responses using item difficulty and other indicators usually coming from an item response theory analysis; c) order the items of each test in ascending order of difficulty; d) define competency levels, which in this case is already defined by the MPL; e) experts read the items in ascending order and set a bookmark where they considered appropriate based on substantive judgment; f) calculate the threshold of the bookmarks based on the inputs from all experts; and, g) in case of disagreement, discuss results and perform a second round of bookmarking.

Additionally, participating in a standard-setting exercise requires to ensure the confidentiality of the materials used in the process. For this reason, each participant signed a confidentially agreement. Furthermore, during the training session, the presentation stressed the commitment to the confidentiality of the materials, which were only used for the session and could neither be reproduced nor disseminated in any form.

2.3 DECISION-DECISION MAKING PROCEDURE

The decision-making procedure involves several steps, starting with the familiarization with definitions, item difficulties and their contents, continuing with the analysis of items to bookmark cut scores individually by experts, to finalizing with a group discussion to agree on a cut score. The following bullets describe the steps for the decision-making procedure:

- First, the substantive experts analyze the definitions of the Minimum Proficiency Level, starting with the general definition and ending with the specific definition of the MPL in mathematics for grades 2/3.
- Second, the team in charge of the organization of the session presents the list of items in a table ordered by difficulty in ascending order. This table only contains the item numbers and codes, as well as the difficulties.
- Third, the experts work individually to analyze the content of the test items. They "bookmark" the item (cut score) where they believe a test-taker just meets the threshold of the MPL. This is the item they believe represents the cut score of the MPL, according to the substantive definition of the MPL. This bookmark implies that a student who answers correctly that item meets the MPL, and those students who only answer correctly items below such item do not meet the MPL.
- Fourth, content experts submit their individual bookmarks to the organizers to prepare a comparative table including the bookmarks of the different experts which represents an input for the group discussion.
- Fifth, group discussion in which experts have to agree on a common cut score. During this session each expert presents her “bookmark”, explaining the rationale for the selection in relation to the substantive definition of the MPL. After the presentation of each expert, a mediator from the organizing team invites experts to re-visit the substantive definition of MPL and reconsider their initial bookmark in light of the presentation of the rest of the experts. This process is repeated several times until an agreement is reached (although in this pilot, due to time limitations, there were only two rounds of discussion and adaptation of bookmarks).
- Sixth, the organizing team collects the information on the cut score and registers the rationale of individual judgments and group discussions to substantiate the decision-making process with evidence on the process.

3 EXPERT PANEL

This section describes both the selection criteria and the training for the expert panel members who participated in the standard setting exercise and a brief description of the training the experts received.

3.1 SELECTION CRITERIA

The selection process for the expert panel involved two key steps:

a. **Call for Expressions of Interest (EoI)**

A public Call for EoIs was issued, providing a comprehensive overview of the project and outlining the responsibilities and contributions expected from panel members. The call included a detailed description of the project objectives, the importance of the expert panel in achieving those goals, and the specific timeline for their participation.

Additionally, the Expert Profile was outlined, specifying the essential qualifications and desirable attributes required for candidates to be considered.

Expert Profile (as included in the Call for EoI):

- **Essential Requirements:**
  - A minimum of a bachelor's degree in education and/or teaching.
  - In-depth knowledge of curriculum content and learning trajectories in reading and/or mathematics at the primary level.
  - Experience in curriculum analysis and learning trajectory evaluation.
  - Availability to participate in a training and a meeting in January 2025.
  - Fluency in English for effective communication with international colleagues.

- **Desirable Attributes:**
  - Experience teaching at the primary level.
  - Postgraduate studies in education, psychometrics, or other relevant areas.
  - Experience in setting cut-off points and describing achievement levels.
  - Familiarity with the SDG 2030 framework in education.
  - Involvement in curriculum development and assessment projects.
  - Proficiency in Spanish and/or French.

b. Evaluation of Expressions of Interest

Submitted EoIs were evaluated by the project team using a systematic approach to ensure that selected candidates possessed the necessary qualifications and experience. Two key criteria guided the selection process:

- **Alignment with the Requested Profile and Experience:** Candidates were assessed based on the extent to which their qualifications and expertise matched the essential requirements and desirable attributes outlined in the call.
- **Diversity of Expertise and Backgrounds:** Special attention was paid to ensuring a diverse panel, with members bringing a variety of professional and geographical perspectives, as well as experiences in different educational systems, curricula, and cultural contexts. This diversity was crucial to minimize potential biases and enhance the quality and applicability of the panel's outcomes.

This two-step procedure ensured that the selected experts were not only highly qualified but also represented a broad spectrum of experiences and insights necessary for setting the Minimum Proficiency Levels in an equitable and effective manner.

3.2 TRAINING

The experts participating in the adapted Bookmark Method exercise were trained in two venues. First, experts received an annotated agenda which included an executive explanation of the adapted Bookmark Method, its rationale, and its steps. Also, the document included the general and specific definitions of the Minimum Proficiency Level for indicator 4.1.1(a) on mathematics for grades 2/3. Finally, the document explained the nature of individual judgment and group discussion.

The second venue of training took place during the standard setting meeting on January 15th, 2025. During a session of 23 minutes, experts received training for the implementation of the standard setting exercise. The training included an explanation of the adapted Bookmark Method used in this session, highlighting its purpose, steps, and technical detail of its application–such as the definition of performance levels and the criteria for each level. Additionally, the training delved into how the method used both statistical information (item difficulties) and content analysis of items to bookmark an item which represents a cut score for defining the split between performance levels. Afterward, the training concentrated on the specifics of the specific exercise of defining a cut score for the Minimum Proficiency Level (MPL). This started with the description of the MPL, and the focus of the exercise on defining only one cut score to distinguish in which item test takers demonstrate to have achieved the MPL. The next step in the training session included a brief introduction to the large-scale Assessment for Minimum Proficiency Level (AMPL) for mathematics 2/3 grades designed to measure MPL. The training session ended with a question and answers moment before the start of the bookmark exercise.

4 Expert session for standard setting pilot exercise using AMPL on mathematics 2/3 GRADES

This section describes the expert session for standard setting, its implementation and decision-making process, and the results of the pilot exercise.

4.1 DESCRIPTION OF EXPERT SESSION

The pilot procedure with the expert panel was carried out on January 15, 2025, at 5:00 PM local time in London (GMT) via Zoom. Due to small connectivity issues for some of the experts, the session started at 5:10 PM. The panel was moderated by both principal investigators of the project, Andrés Sandoval and Ernesto Treviño, and assisted by PhD students Manuel Cheyre and Adam Coates. The final attending panel was composed by the following experts: Afzal Sayed, Artemio Cortez, Gail Coates, Israel Moreno, Lizzie Emelue, Natalia López, Nurullah Eryilmaz, Reywathi Arumal, and Yusuf Olaniyan.

The duration of the panel was planned to last 3 hours and 20 minutes. As the session started a little later than stipulated, the timetable was adjusted to end as close to 20:00 as possible. Table 1 shows the proposed timetable for each section of the session contrasted with the actual duration of each.

Table 1. Timetable with proposed times for each segment of the session and actual times.

PHASE	**ALLOCATED TIME**	**ACTUAL TIME**
WELCOME AND PRESENTATIONS	10 minutes	10 minutes
METHOD DESCRIPTION AND EXPLANATION	20 minutes	23 minutes
INDIVIDUAL JUDGMENT OF MPL	70 minutes	77 minutes
BREAK	10 minutes	15 minutes
GROUP DISCUSSION	60 minutes	45 minutes
COMPARISONS WITH PREVIOUS MPL EXERCISES	20 minutes	5 minutes
CLOSING REMARKS	10 minutes	5 minutes
TOTAL	200 minutes	180 minutes

The first segment of the meeting consisted of a general greeting between the participants, in which everyone, including moderators and assistants, presented themselves by name and a short description of the relevant professional experience for the event. For instance, most of the experts related their experiences as teachers in primary schools in different parts of the world. This section lasted the 10 minutes that were allotted and allowed us to contextualize and provide the appropriate setting for the session.

The second stage of the panel was highly relevant, consisting of a thorough description and explanation of the methodology being carried out. The section started with a description of the context that makes such a procedure necessary. Next, a brief description of the key indicator was provided, as well as a context of what standard setting is and what is its purpose and relevance for international comparisons. Following, was a description of the Bookmark Method, a type of standard setting on which the piloted procedure is based. This allowed for a proper understanding of the rationale behind the exercise by the experts to provide them with an adequate mindset for the procedure. Then, the adjustments to the Bookmark method were presented. Lastly, a thorough description of the test being used for the exercise was carried out. At the end of this stage, participants were asked to share any questions or clarifications needed. Only one question was asked at this point, which consisted of clarifying if the experts were supposed to think about second-grade students or third-grade students, given that the presentation used the term “2/3 grade”. Participants were told that due to curricular differences, grades in each country may vary, but that they were supposed to base their analysis on the provided Minimum Proficiency Level definition, and not on their individual countries. The time dedicated to this section was only 3 minutes longer than planned, mainly due to the stressing of some key information relevant to the exercise and the Q&A.

The third segment of the session was the central part of the panel. It consisted of the individual work carried out by each expert. This phase started with a 17 minutes subsection describing in full detail both the general Minimum Proficiency Level definition, as well as the specific MPL definition on mathematics 2/3 grades provided by UNESCO for SDG 4 indicator 4.1.1(a).

The Minimum Proficiency Level (MPL) is the benchmark of basic knowledge in a domain (mathematics, reading, etc.) measured through learning assessments. The MPLs for reading and mathematics used to report on indicator 4.1.1 describe the basic knowledge and skills students must be able to demonstrate for specific grade levels. These benchmarks are based on an analysis of curriculum and assessment programs from around the world.

For the standard setting exercise, we first presented the general definition of the MPL for mathematics 2/3 grades, namely:

> *"Students demonstrate skills in number sense and computation, reading simple data displays, shape recognition and spatial orientation."*

Due to the general nature of the above-cited MPL definition, the information was complemented with the following more specific definition of the MPL, which is more informative for the item analysis for the Bookmark exercise:

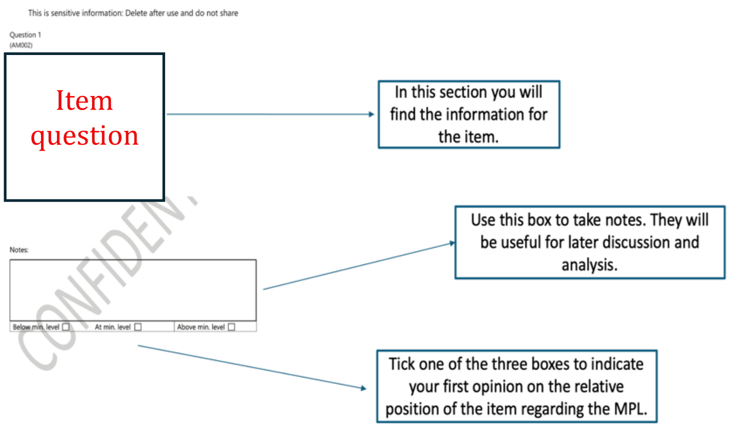
> *"Students recognize, read, write, order and compare whole numbers up to 100. They demonstrate computational skills involving the processes of addition, subtraction, doubling and halving for whole numbers within 20. They recognize and name familiar shapes and describe their basic attributes. They recognize time in days, weeks and months. They describe location in a space using simple language."*

After reviewing both MPL definitions, experts started the individual judgment exercise. In order to engage in this process, experts received the following instruction:

> *Review the ordered booklet of test items*
>
> > *Each expert has an individual PDF file that allows for note-taking and indicating, for each item, if it is below the cut point, at the point or above the cut point.*
> >
> > *Additionally, each expert is provided with a simplified worksheet that allows for the final bookmark, indicating the item that should represent the cut score.*
> >
> > *IMPORTANT:*
> >
> > > *Note-taking on key items will be relevant for the latter group discussion, as a means for reminding the justification for each decision.*
> > >
> > > *The definitive bookmark has to be placed on the worksheet before the group discussion.*

Then, the written instructions were also explained in the session supported by visual aids, as shown in Figure 1 and Figure 2.

Figure 1 Visual aid from a page of ordered item booklet demonstrating the sections of each page.



This is sensitive information: Delete after use and do not share

Question 1
(AM002)

Item question

Notes:

Below min. level □ At min. level □ Above min. level □

In this section you will find the information for the item.

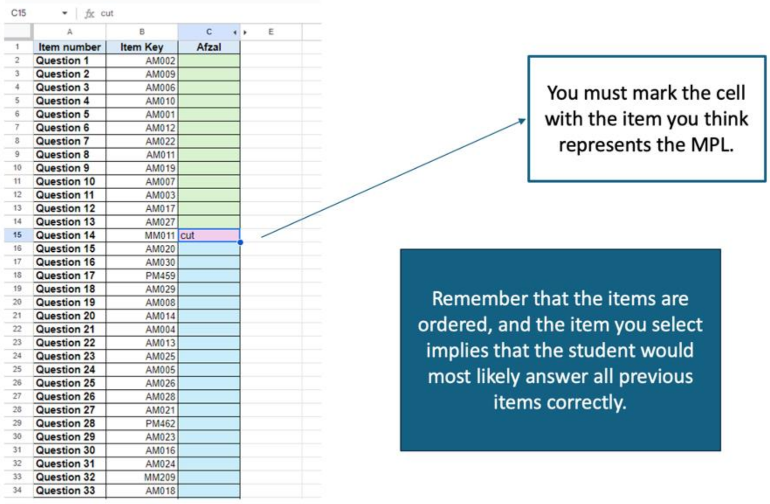
Use this box to take notes. They will be useful for later discussion and analysis.

Tick one of the three boxes to indicate your first opinion on the relative position of the item regarding the MPL.

4.2 Individual Judgement Process

After the full explanation of the procedure and answering questions, individuals were sent out to individual "virtual rooms" for 60 minutes. One of the assistants delivered the personalized material for each expert via email and went to each "virtual room" answering questions and clarifying instructions as needed on a case-by-case basis. After the hour elapsed, 10 minutes for a break was provided. The panel session resumed 15 minutes after the break was provided. During the break provided for the participants, the panel assistants compiled each individual cut score set by the experts.

Figure 2. Visual aid of the final bookmark worksheet demonstrating how to set the cut point after reviewing all items.



C15 fx cut

	A	B	C	E
1	Item number	Item Key	Afzal	
2	Question 1	AM002		
3	Question 2	AM009		
4	Question 3	AM006		
5	Question 4	AM010		
6	Question 5	AM001		
7	Question 6	AM012		
8	Question 7	AM022		
9	Question 8	AM011		
10	Question 9	AM019		
11	Question 10	AM007		
12	Question 11	AM003		
13	Question 12	AM017		
14	Question 13	AM027		
15	Question 14	MM011	cut	
16	Question 15	AM020		
17	Question 16	AM030		
18	Question 17	PM459		
19	Question 18	AM029		
20	Question 19	AM008		
21	Question 20	AM014		
22	Question 21	AM004		
23	Question 22	AM013		
24	Question 23	AM025		
25	Question 24	AM005		
26	Question 25	AM026		
27	Question 26	AM028		
28	Question 27	AM021		
29	Question 28	PM462		
30	Question 29	AM023		
31	Question 30	AM016		
32	Question 31	AM024		
33	Question 32	MM209		
34	Question 33	AM018		

You must mark the cell with the item you think represents the MPL.

Remember that the items are ordered, and the item you select implies that the student would most likely answer all previous items correctly.

During the individual judgment segment of the panel, the experts carefully reviewed each item in ascending order of difficulty and placed a bookmark on the item that they considered to be at the precise Minimum Proficiency Level. In other words, students at the minimum proficiency level should correctly answer the bookmarked item. It is important to note that the method assumes that, by answering the bookmarked item, students would answer the items of lower difficulty than that of the cut score correctly. Any student that either reaches or surpasses the cut score would be regarded as attaining the MPL.

4.3 Group discussion process and results

Individual cut scores presented in the same table served as input for the group discussion, where they saw the bookmarks from the different participants. This phase was set to last 60 minutes but was cut to 45 due to the saturation of information and to compensate for a slightly longer than-planned individual judgment section. During this part of the session, each participant gave their rationale for their cut score. The description started with the descending order from the participants that set the higher cut scores to the participants that set the lowest. During each individual description, participants and moderators were free to intervene and ask further questions to clarify. The purpose of this section was twofold. First, it aimed at aligning criteria and reaching a consensus regarding a cut score. At the end of the discussion, each participant was offered the chance to change their original cut score to a new one, to arrive at a consensus. Second, this section also aimed at collecting qualitative information regarding the rationale used by different experts to set their bookmarks, identifying strengths and weaknesses of the pilot exercise that will serve as inputs to improve the next standard setting sessions.

Next, the moderators showed the experts the official cut score generated as part of the design and implementation of AMPL. This allowed for a discussion regarding the conclusions reached by each individual. This section was set to last 20 minutes but was cut short to 5 due to time constraints.

Finally, the moderators lead a short 5-minute section for closing remarks and providing closure for the session. This section was also cut short by 10 minutes due to the time constraints already mentioned. In this stage, moderators highlighted key elements from the discussion and thanked all the experts. They also reminded them to send their responses and delete the confidential information after the meeting.

Table 2 presents all items in order of difficulty, just as they were presented to experts. Each expert response is registered in a separate column. The first bookmark placed by every expert is indicated with the term "Cut 1". The second bookmark changed after the group discussion, is indicated by the term "Cut 2". Experts that did not change their bookmark after discussion are indicated by the term "Cut 1-2".

Table 2 shows a summary of all bookmarks by individual experts, with their corresponding overall median and standard deviation. Results from expert number 4 are excluded from the analysis (albeit shown in

Table 2) because there is evidence that the expert used different criteria to set the bookmark (namely, the expert’s own experience in his/her country prevailed over the provided definition of Minimum Proficiency Level).

Table 2. Cut scores set by experts.

Item #	Item Key	Difficulty	E1	E2	E3	E4	E5	E6	E7	E8	E9
1	AM002	-4.11									
2	AM009	-3.56									
3	AM006	-3.43									
4	AM010	-3.27									
5	AM001	-3.22				Cut 1					
6	AM012	-3.08									
7	AM022	-3.02									
8	AM011	-2.97									
9	AM019	-2.54									
10	AM007	-2.33									
11	AM003	-2.27									
12	AM017	-2.26									
13	AM027	-2.19									
14	MM011	-2.17									
15	AM020	-2.11									
16	AM030	-2.09									
17	PM459	-2.07								Cut 1	
18	AM029	-2.01									
19	AM008	-1.97									
20	AM014	-1.82									
21	AM004	-1.78							Cut 1-2	Cut 2	
22	AM013	-1.77									
23	AM025	-1.76	**Official AMPL Cutscore**								
24	AM005	-1.6									
25	AM026	-1.52					Cut 2				
26	AM028	-1.42									
27	AM021	-1.4	Cut 2		Cut 1-2						
28	PM462	-1.4									Cut 1-2
29	AM023	-1.38									
30	AM016	-1.35									
31	AM024	-1.13									
32	MM209	-1.06	Cut 1				Cut 1				
33	AM018	-1.02									
34	MM029	-0.9		Cut 1-2				Cut 1-2			
35	MM125	-0.85									
36	MM004	-0.76									
37	MM208	-0.74									
38	MM022	-0.63									
39	MM212	-0.6									
40	PM468	-0.57				Cut 2					
41	PM422	-0.49									
42	PM445	-0.44									
43	PM942	-0.29									
44	MM060	-0.09									
45	PM469	-0.08									
46	MM016	-0.05									
47	MM211	0.16									
48	MM019	0.22									
49	MM089	0.28									
50	MM210	0.29									
51	PM449	0.3									
52	MM197	0.35									
53	AM015	0.39									
54	PM454	0.41									
55	MM104	0.61									
56	MM090	0.82									
57	MM191	1.17									
58	MM175	1.19									

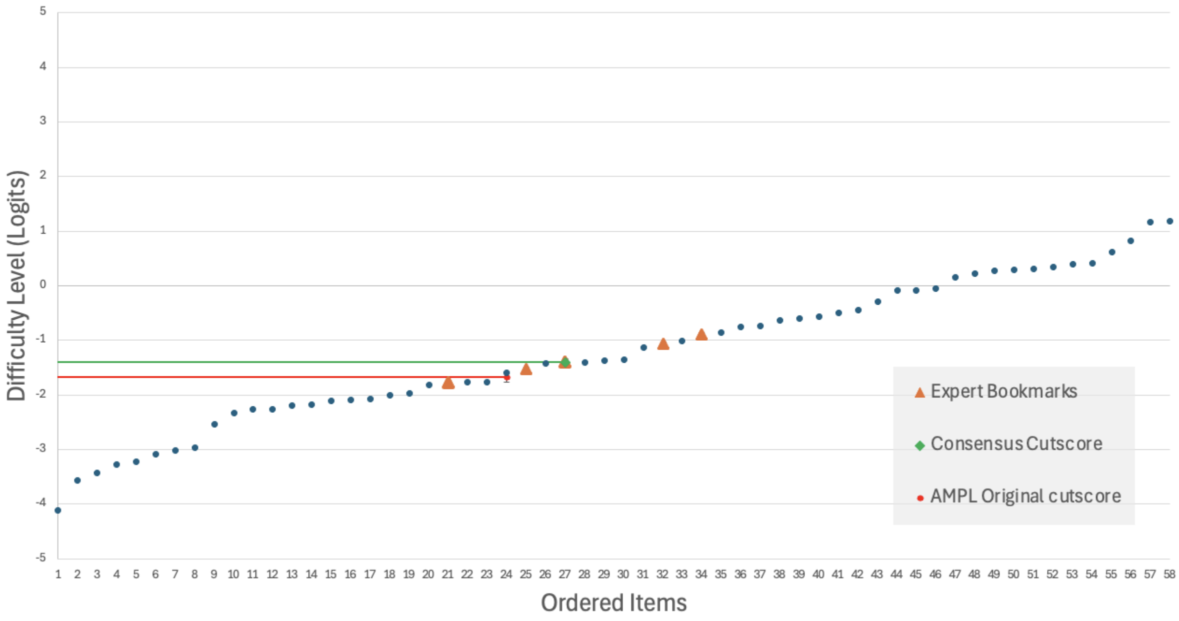
As can be seen in

Table 2, all experts set their cut scores below item 34 (in ascending order of difficulty) and above item 17. The median cut score at this stage was set at item 30. However, after the group discussion and the presentation of each expert's justification for their decision, 4 out of the 9 experts changed their bookmark position, most of them towards the center of the distribution. The new median after the discussion was set at item 27, with a range of 13.

These results suggest that the group discussion phase was effective in aligning some of the criteria used by the experts for setting their bookmark and getting closer to a consensus, by reducing the dispersion of bookmarked items. It is reasonable to expect that an increased number of group discussion sessions would reduce the dispersion of the bookmarks until a unanimous consensus is reached. However, such an exercise would require an increased number of hours, which was not feasible during this pilot due to the time limit to test the potential of this type of exercise as an alternative to set and measure the MPL in different contexts using different assessments.

A second analysis was conducted using only the difficulty levels of the items as measured in logit units. The mean difficulty of the bookmarked items after the individual judgment was -1.32 logits, a median of -1.23, and a standard deviation of 0.43 logits. After the group discussion, the new average difficulty of the bookmarked items decreased to -1.39 logits, with a median of -1,4 and a standard deviation of 0.34 logits. When the mean and the median of the item difficulty of the consensus cut score are transformed back to its equivalent item, the resulting item coincides with number 27 in the ordered item booklet. The final results of the procedure are shown in Graph 1.

Graph 1. Final Ordered Item Map with corrected bookmarks and final cut score.



Using the relative position of the ordered items and their difficulty levels yields item 27 as the definitive cuts-core that would represent the Minimum Proficiency Level as defined by UNESCO for its SDG 4 indicator 4.1.1(a) using AMPL. This means that any student with a Minimum Proficiency Level should be able to answer correctly up to item 27; all items above exceed the skills and knowledge defined as minimal.

Considering that the test has a total of 58 items, setting the cut score at item 27 implies that a student with a Minimum Proficiency Level would respond correctly to 46,5% of the assessment. As can be seen in graph 1, the official cut score for the assessment was set between items 23 and 24, indicating that a student with a Minimum Proficiency Level would have to answer 41,3% of the test correctly. The similarities between the results from the official standard setting procedure by the AMPL team and the present pilot strengthen the robustness of the proposed methodology.

Table 3. Summary of experts cut scores for the individual judgment exercise (Cut score 1) and after the discussion session (Cut score 2) by the relative ordered position of the bookmarked item and item difficulties of the items expressed in logits.

	Relative Ordered Position of Bookmarked Item		**Item Difficulty (Logits) of Bookmarked Item**	
Expert	**Cut score 1**	**Cut score 2**	**Cut score 1**	**Cut score 2**
E1	32	27	-1.06	-1.40
E2	34	34	-0.90	-0.90
E3	27	27	-1.40	-1.40
E4				
E5	32	25	-1.06	-1.52
E6	34	34	-0.90	-0.90
E7	21	21	-1.78	-1.78
E8	17	21	-2.07	-1.78
E9	28	28	-1.40	-1.40
Mean	28.13	27.13	-1.32	-1.39
Median	30	27	-1.23	-1.40
SD	6.26	4.99	0.43	0.34

Table 1 presents the proportion of students meeting or surpassing the MPL comparing the original exercise of AMPL with the pilot exercise reported in this document. First, it is important to briefly describe the differences in cut scores that led to the two estimations. In the case of MPL original proportion of students, the estimation is based on a cut score on item 23 with a difficulty of -1.76. The pilot exercise established a cut score on item 27 with a difficulty of -1.4. Therefore, there is a gap of 0.36 in item difficulties when comparing the two cut scores.

Second, as a consequence of the differential in cut scores, there are statistically significant differences between the original estimation of the proportion of students meeting or surpassing the MPL when contrasting the two exercises. A general trend shows that the pilot exercise led to an estimation of students meeting or surpassing the MPL, which is nearly 0.10 lower than the original estimation in each country – i.e. the results from the pilot exercise led to a "more difficult cut point". Furthermore, at the country level, all the differences are statistically significant. For example, in Gambia, the original exercise estimated a proportion of students meeting the MPL of 0.28, and that figure is 0.20 in the pilot exercise. In Kenya, the original estimate of the proportion of students is 0.88, while the estimate in the pilot exercise reaches 0.77. In the case of Lesotho, using the original cut score, the estimate of the proportion of students meeting the MPL is 0.83, while the proportion estimated in the pilot is 0.70. In Zambia, the estimation with the original cut score is 0.49, in contrast with the 0.36 estimated in the pilot exercise. As stated before, all the differences are statistically significant, as can be seen in the fact that the confidence intervals (ranges between upper and lower limit of the standard errors) do not cross when comparing the original with the pilot estimation.

Table 4. Proportion of students meeting or exceeding the MPL in the original AMPL assessment and proportion of students meeting or exceeding the MPL as set in the pilot exercise.

Country	**Proportion of students meeting MPL (original)**			**Proportion of students meeting MPL (pilot exercise)**		
	Estimate	**SE (lower limit)**	**SE (upper limit)**	**Estimate**	**SE (lower limit)**	**SE (upper limit)**
Gambia	0.28	0.24	0.31	0.20	0.17	0.23
Kenya	0.88	0.86	0.89	0.77	0.75	0.80
Lesotho	0.83	0.80	0.85	0.70	0.66	0.73
Zambia	0.49	0.46	0.51	0.36	0.34	0.39

Sources: For the original estimation Assessments for Minimum Proficiency Levels a and b (AMPL-ab). International Report, January 2024. https://ampl.uis.unesco.org/wp-content/uploads/sites/27/2024/02/International-Report__AMPLab_WEB.pdf, based on a cut score on item 23 with a difficulty of -1.76. For the pilot exercise own calculations based on a cut score on item 27 with a difficulty of -1.4

Besides the discussion on the estimations, the statistical analysis of items, and the content analysis of the items for the exercise of standard setting, it is important to consider the nature of the group discussion among experts. Such analysis provides important lessons on the elements–biases, previous experiences, and decoupling of the instructions–that may influence the individual judgment process and the discussion among experts when trying to agree on a cut score.

The basic strategy reported by the experts was consistent with the instructions provided, that is, carrying out a sequential analysis progressing from the easiest to the most difficult items (with one exception of one expert who started from the hardest to the easiest item). This strategy allowed them to identify a "cut-off" point where the complexity of questions no longer aligned with the described Minimum Proficiency Level. This strategy helped the experts ensure consistency with the provided Minimum Proficiency Level by checking case-by-case where questions began to exceed expected knowledge and skills.

Simultaneously, some experts considered each item's cognitive demands, distinguishing between items requiring basic knowledge and skills and those involving more abstract or analytical reasoning. Questions that introduced probability, multi-step problem-solving, or abstract spatial reasoning were often flagged as exceeding the Minimum Proficiency Level. This approach reflected an understanding of developmental appropriateness, ensuring that items beyond the cut score required skills that surpassed the described MPL and *vice versa*.

Experts frequently connected their judgments to their professional experiences with learners of similar ages, which was a double-edged sword. On the one hand, it is good that their experience helps experts identify the knowledge and skills required to correctly reply to a question. Their curricular expertise is key to the exercise. On the other hand, their experience sometimes biased them by distracting their attention from the MPL definition and using their memories of how students they taught of the same grade would have answered. In some cases, the experts even reported thinking of their own children as anchors for their reasoning. These strategies could have distanced the experts from the expected task at hand because the essential premise of the procedure was to set a bookmark based on a specific definition, not on an experiential basis. These individual experiences would most certainly be contaminated by cultural and curricular differences between countries, which deviates from the purpose of the exercise.

Upon hearing each other, 4 of the 9 experts switched their initial cut scores based on some of the arguments they heard from their colleagues. In some cases, the cut score was set to change as a "means to achieve consensus" more than a result of convincing. In other cases, the explicit rationale presented by some of the experts resonated and made them switch their original cut scores to a position closer to the median.

Some discussion was also carried out when the experts' context was provided as a rationale. For instance, in one case, an expert argued that a student in one of their classes in the country of origin could not have been able to recognize the complexity of the phrasing of most of the items, especially considering that the test would not have been in the local language. This situation led this expert to set an extremely low cut score. Some of the other experts set out to discuss if their own students would have been able to understand some of the items, but the moderators explicitly said that the exercise was supposed to be focused solely on the provided Minimum Proficiency Level definition and not on specific individual experiences or certain cultural contexts.

5 DISCUSSION

The pilot exercise implemented to set standards for Indicator 4.1.1(a) in mathematics for grades 2 and 3, using AMPL results, yielded important lessons. First, this process appears to be a promising approach to support the monitoring of Indicator 4.1.1 by using different national and international standardized tests. However, this type of exercise presents a significant tension.

On one hand, it offers a more feasible method for measuring Indicator 4.1.1 globally, given the lack of large-scale assessments that are fully aligned with this indicator and administered across all or most countries adhering to the SDGs. Additionally, it enables the contextualization of results by using existing national or international standardized tests that are already in place in various countries.

On the other hand, from a strictly methodological perspective, it may not be possible to provide empirical evidence that the results are comparable across countries or assessments. This limitation arises because there is no common test administered universally, nor is there any equating across tests (e.g., shared items across different assessments that allow for results to be estimated on a common scale).

Considering that having a global common assessment or implementing psychometric equating across all available tests seems highly implausible to implement between 2025 and 2030, it is necessary to identify a more practical method for estimating the proportion of students meeting the MPL. The exercise proposed here may represent a viable alternative for estimating the indicator, acknowledging that it is a second-best solution with limitations compared to a more robust methodological approach, which does not currently exist and is unlikely to be feasible within the given timeframe.

The results of this initial exercise must not be interpreted as definitive evidence to proceed or discard with standard setting exercises for different large-scale assessments to set MPL and measure indicator 4.1.1. Although promising, as suggested below, this exercise must be improved to develop the standard setting completely following the Bookmark Method, devoting more time and providing clearer guidance during group discussions. Furthermore, it is necessary to complement this first pilot with repeated applications of the standard setting exercise using data from other large-scale assessments (international and/or national). These replications would allow us to better comprehend the strengths and limitations of such an approach to produce estimates that may be within reasonable ranges for substantive comparability. Furthermore, replicating this exercise for the same country (or group of countries) with both international and national assessments would open the opportunity to study the concurrency of results providing stronger evidence on the scope and limitations of this methodology.

In any case, monitoring indicator 4.1.1 is a delicate and challenging endeavor. It is delicate because it must be careful in applying any method to reduce incentives for countries to intentionally overestimate the proportion of students reaching the MPL. It is also challenging because it entails ensuring an adequate level of agreement among experts who review the item contents to set the cut score for the MPL. As it has been found in this pilot exercise, biases may interfere in standard settings due to differences in life experiences, such as cultural differences due to the place of birth, teaching in different contexts, interpretations of how education and assessment take place in the places they know, and even family experiences, among others.

Finally, it is important to highlight three lessons for improvement emerging from this pilot exercise in, at least, the following areas:

- Highlighting the necessity of sticking to the specific definition of MPL when assessing the items and bookmarking the cut score, avoiding the interference of personal experiences.
- Applying the Bookmark Method without adjustments, a situation that entails:
    - Conducting a more stringent moderation aimed at conducting the group towards reaching an agreement in light of both the definition of MPL and the item contents.
    - Devoting more time to the group discussion session, and being more proactive in revisiting both the definition of MPL and the items all the times necessary until reaching an agreement.

- Preparing the standard setting session in advance to ensure the participation of a wider range of experts is challenging due to time differences and the need to find experts from different parts of the world who speak a shared language (English in this case).
- Aiming at performing standard setting exercises in a group of countries using both national and international assessments to explore the level of concurrency in the results obtained using the proposed method.

6 ANNEX

The annex comprises three parts. First, it presents a brief description of the expert panel members. Secondly, it includes a description of the individual judgment process and the type of results by including a sample of the material and the type of comments that experts produced during this stage of the exercise. Finally, there is a link to the confidentiality agreement form.

6.1 ANNEX 1. DESCRIPTION OF THE EXPERT PANEL MEMBERS

The expert panel for this project comprises a diverse group of educators and researchers with extensive experience across primary, secondary, and higher education. Their expertise spans curriculum development, assessment, and educational policy from various international contexts. The panel members possess in-depth knowledge of curriculum content and learning trajectories in reading and mathematics, with several having direct teaching experience and substantial familiarity with international large-scale assessments. This diversity makes the panel well-suited for the project.

Below is a brief summary of each expert's background and experience:

- **Nurullah Eryilmaz** is a Senior Research Analyst at the International Association for the Evaluation of Educational Achievement (IEA) in Hamburg. He has a background as a mathematics teacher at the primary and secondary levels in Turkey. His expertise includes analyzing educational outcomes, evaluating curriculum content, and understanding learning trajectories in mathematics and reading. Additionally, he has extensive experience with international large-scale assessments and is familiar with the SDG 2030 framework in education.

- **Israel Moreno Salto** is a scholar at the Autonomous University of Baja, California, in northern Mexico, with a research focus on large-scale assessments and governance. He has teaching experience across lower secondary, primary, and preschool education, providing him with a comprehensive understanding of educational progression at various levels.

- **Artemio Cortez** is a Lecturer in education at the University of Bath. He has 6+ years of experience teaching primary education in Colima, Mexico, working with students aged 6 to 10, and has also taught secondary-level arts. His areas of expertise include educational leadership, policy, and practice, particularly in global south contexts.

- **Afzal Sayed Munna** is a Senior Lecturer and program manager at the University of Hull, London campus. He has extensive experience teaching at both the university level and in primary schools in the UK and Bangladesh. His academic background combines business leadership and education, providing a unique interdisciplinary perspective.

- **Lizzy Emelue** is a PhD student in educational pedagogy and leadership at the University of Bath. She has teaching experience in primary and secondary schools in Nigeria, Japan, and American international schools. Additionally, she has worked as an ESL and SEN teacher in Japan, enhancing her understanding of diverse educational contexts.

- **Reywathi Arumal** is a PhD student at the University of Bath, focusing on public education. She has 10+ years of teaching experience in primary and secondary schools in Malaysia and an additional 10 years of experience in the policy and research division at Malaysia's Ministry of Education.

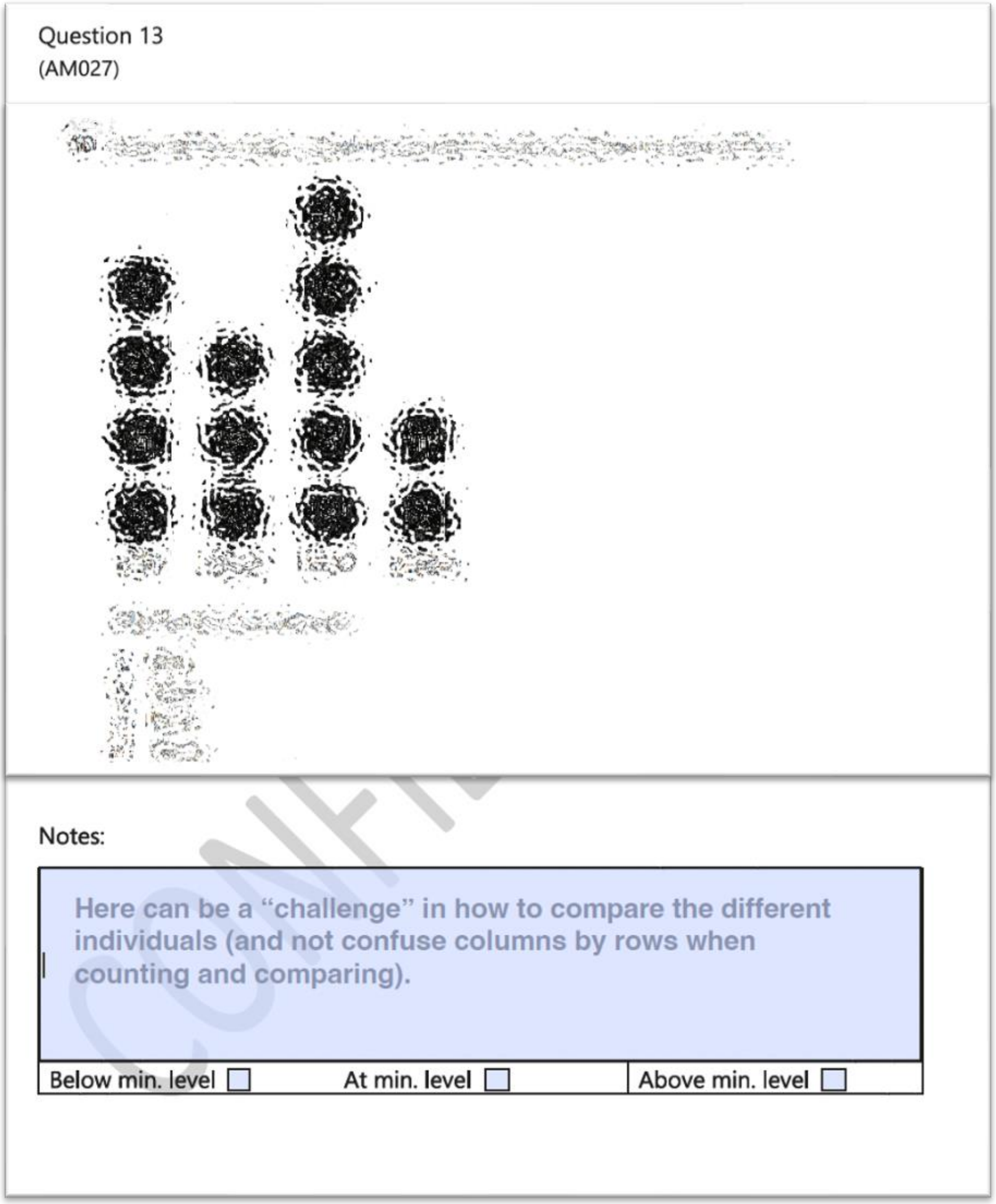
- **Gail Coates** has 14+ years of experience teaching in primary schools in the UK, primarily working with Year 3 and Year 4 students. Her extensive classroom experience offers valuable insights into curriculum content and student learning trajectories.

- **Yusuf Olaniyan** is a PhD student in Education at the University of Bath. He has 5+ years of teaching experience in primary education in Nigeria, specifically with Year 2 and Year 3 students, as well as additional experience in secondary and higher education. His research interests include higher education, labour markets, policy sociology, and global inequalities.

All experts meet the essential qualifications required for the project, including a minimum of a bachelor's degree in education and/or teaching, in-depth knowledge of curriculum content, and experience in curriculum analysis. Additionally, many panel members bring desirable attributes such as postgraduate studies, experience in setting cut-off points, familiarity with the SDG framework, involvement in curriculum development, and proficiency in other languages.

6.2 Annex 2. Item reviewed in individual judgment and expert comments

As part of the pilot exercise, experts were asked to analyze test items and determine the point at which a student would meet the minimum proficiency level (MPL). The experts reviewed each item in ascending order of difficulty and placed a bookmark on the item they considered to be at the precise MPL. This involved assessing whether the item aligned with the provided definition of the MPL and considering the cognitive demands of each question. Each expert was given an individual PDF file that allowed for note-taking on key items, which would be relevant for the later group discussion. The aim of this section is to provide a transparent record of each expert's rationale when setting their cut score, and to provide insight into the specific items that were the focus of discussion among the experts. This analysis will also highlight how the experts' backgrounds and experiences helped to inform their judgments, in the context of the detailed definition of the Minimum Proficiency Level. Note that the comments and items are blurred to maintain confidentiality and ensure the integrity of the assessment.



Question 13
(AM027)

Notes:

Here can be a “challenge” in how to compare the different individuals (and not confuse columns by rows when counting and comparing).

Below min. level □ At min. level □ Above min. level □

Figure 3. Sample of an item and comments.

6.3 Annex 3. Confidentiality agreement (Template)

Confidentiality Agreement for the use of AMPL and ERCE test materials

The University of Bath at Claverton Down, Bath, BA2 7AY, United Kingdom, hereafter referred to as "UoB", demands that all materials and data received from the ACER's Assessment for Minimum Proficiency Levels (AMPL) and the UNESCO's Regional Comparative (ERCE). Explanatory Study are kept strictly confidential.

All national and international experts, including subcontracted staff (if applicable), must understand and obey these confidentiality rules and practices and, regardless of their organisational affiliation, sign this confidentiality agreement in which they agree that they will not:

1. make any public disclosure or publication of any interim or final AMPL and ERCE materials, including, but not limited to, responses, data, instruments, items, documents, communication materials, analysis, reports, infographics, and videos.
2. use, disclose or publish any personally identifiable information.
3. have ownership or any other intellectual property of the subject data and any related documentation or accompanying software which at all times shall be and remain the sole and exclusive property of ACER and/or UNESCO.
4. publish or otherwise release research results based on the meetings and/or information provided for this project.

I agree to the above terms.

Name: ____

Job title: ____

Role/participant: International expert ____

Organization: ____

Full official address: ____

____

Date: ____

Signature: ____

Please fill, sign, and return to the UoB at Adam Coates (ac3615@bath.ac.uk).