Analysis Plan

Assessment of Minimum Proficiency Level (AMPLab)

May 2023, version 1.0

The Global Education Monitoring (GEM) Centre drives improvements in learning by supporting the monitoring of educational outcomes worldwide. The GEM Centre is a long-term partnership between the Australian Council for Educational Research (ACER) and the Australian Government's Department of Foreign Affairs and Trade (DFAT).

unesco Institute for Statistics

Bill & Melinda Gates foundation

ACER

Acknowledgments

This project, Development of an Assessment of Minimum Proficiency Level (AMPL) aligned to Global Proficiency Framework (GPF) Level 2, is referred to AMPLab. This UNESCO Institute for Statistics (UIS) project is funded by the Bill & Melinda Gates Foundation.

The Australian Council for Educational Research (ACER) is the technical partner for this project. Support is provided from ACER's Global Education Monitoring Centre (GEM Centre), an ACER initiative in partnership with the Australian government's Department of Foreign Affairs and Trade.

Contents

Introduction

As part of Sustainable Development Goal (SDG) 4, Indicator 4.1.1 aims to measure the "proportion of children and young people: (a) in grades 2/3; (b) at the end of primary; and (c) at the end of lower secondary achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex." To meet this goal, UIS has coordinated efforts to establish common reading and mathematics scales for all three points of Indicator 4.1.1, building on existing cross-national and national assessments. As a result of these efforts, two important points of consensus have been reached: the definition of the Minimum Proficiency Level (MPL) and the Global Proficiency Framework (GPF).

The overarching objective of the AMPLab project is to measure and analyze the performance of students at the end of lower and upper primary using an assessment that aligns with the GPF. This will:

- enable the collection of more informative data about where students are performing in terms of the MPLs at the end of lower and upper primary in reading and mathematics,
- produce baseline measures to set targets and compare learning gains/losses over time
- facilitate reporting on SDG 4.1.1
- aid the tracking of learning progress over time
- complement tools that had been already developed in 2021 in the Monitoring the Impacts on Learning Outcomes (MILO) study.

The benchmarks that will be used to quantify performance are:

- the proportion of students in the end of lower primary school, in participating countries, that meet the Minimum Proficiency Levels (MPL) referred to in SDG indicator 4.1.1(a) and described in ACER-GEM (2022).
- the proportion of students in the end of primary school, in each country, that meet the Minimum Proficiency Levels (MPL) referred to in SDG indicator 4.1.1(b) and described in ACER-GEM (2022).

The links to these benchmarks will be established in the AMPLab assessments as part of this study.

The purpose of this document is to provide a rationale for the analysis that will be undertaken and an outline of the method that will be used to analyse the data.

ASSESSMENT CONTENT

An outline of the selected domains and constructs covered by the AMPLab assessments appears in Table 1 and Table 2.

Table 1: Domains and constructs in the AMPLab assessments for SDG 4.1.1(a)

Learning areas	Reading	Mathematics
Domains	Listening Comprehension Decoding Reading Comprehension	Number and Operations Measurement and Geometry Statistics, Probability and Algebra
Constructs	Retrieving information Interpreting information Precision	Whole numbers Length, weight, capacity, volume, area, and perimeter Time Properties of shapes and figures Spatial visualisations Position and direction Data management Patterns

Table 2: Domains and constructs in the AMPLab assessments for SDG 4.1.1(b)

Learning areas	Reading	Mathematics
Domains	Reading comprehension	Number and operations Measurement Geometry Statistics and probability Algebra
Constructs	Retrieving information Interpreting information Reflecting on information	Whole numbers Fractions Decimals Integers Exponents and roots Operations across number Length, weight, capacity, volume, area and perimeter Time Currency Spatial visualizations Properties of shapes and figures Position and direction Data Management Chance and probability Patterns Expressions Relations and functions

The assessments follow an *Assessment Blueprint* that defines the coverage of the learning areas, domains and constructs as referenced in documentation of the MPLs. The assessment blueprint specifies the coverage of learning areas and the relative proportion of domains. The assessment items are drawn from the UIS Global Item Bank and include French and English source items.

Participating countries are provided with following paper-based assessment instruments that measure the attainment of SDG 4.1.1(a) and SDG 4.1.1(b) Minimum Proficiency Levels (MPL) in reading and mathematics in students at the end of lower and upper primary education.

AMPL-A at the end of lower primary stage

AMPL-A assesses students at the end of lower primary stage on the Minimum Proficiency Levels corresponding to 4.1.1(a). Each booklet contains three clusters of items including:

- One cluster of 10 listening comprehension and 5 decoding items delivered via audio with students answering in their booklets
- One cluster of 25 reading items and 5 decoding items, paper-based
- One cluster of 30 mathematics items, paper-based

There are two AMPL-A booklets, to be rotated across students. The listening comprehension/decoding cluster appears at the beginning of both booklets. Each booklet contains both the reading/decoding and mathematics clusters, but the ordering of the clusters is reversed across the two booklets.

Table 3: AMPL-A Assessment: test design

Booklet	Part 1	Part 2	Part 3
AMPL-A Booklet 7	Listening Comp Decoding	Mathematics	Reading Decoding
AMPL-A Booklet 8	Listening Comp Decoding	Reading Decoding	Mathematics

AMPL-B at the end of upper primary stage

AMPL-B assesses students at the end of upper primary stage on the Minimum Proficiency Levels corresponding to 4.1.1(b). Each booklet contains two clusters of items including:

- One cluster of 32 reading items, paper-based
- One cluster of 30 mathematics items, paper-based

There are two AMPL-B booklets, to be rotated across students. Each booklet contains both the reading and mathematics clusters, but the ordering of the clusters is reversed across the two booklets.

Table 4: AMPL-B Assessment: test design

Booklet	Part 1	Part 2
AMPL-B Booklet 1	Mathematics	Reading
AMPL-B Booklet 2	Reading	Mathematics

AMPL-A+B at the end of upper primary stage

AMPL-A+B assesses students at the end of upper primary stage on the Minimum Proficiency Levels corresponding to both 4.1.1(a) and 4.1.1(b).

The audio-based cluster of 10 listening comprehension and 5 decoding items appears at the beginning of each booklet. Booklets 3 and 4 rotate the AMPL-B reading and mathematics clusters. Booklets 7 and 8 rotate the AMPL-A reading/decoding clusters and the mathematics clusters. Booklets 5 and 6 contain reading and mathematics clusters which are comprised of a mix of AMPL-A and AMPL-B items.

Table 5: AMPL-A+B Assessment test design

	Part 1 Audio	Part 2	Part 3
AMPL-A+B Booklet 3	Listen Comp(a) 10 items Decoding(a) 5 items	Maths(b) 30 items	Reading(b) 32 items
AMPL-A+B Booklet 4	Listen Comp(a) 10 items Decoding(a) 5 items	Read(b) 32 items	Maths(b) 30 items
AMPL-A+B Booklet 5	Listen Comp(a) 10 items Decoding(a) 5 items	Maths(a) 15 items Maths(b) 15 items	Read(a) 15 items Read(b) 15 items
AMPL-A+B Booklet 6	Listen Comp(a) 10 items Decoding(a) 5 items	Read(a) 15 items Read(b) 15 items	Maths(a) 15 items Maths(b) 15 items
AMPL-A+B Booklet 7	Listen Comp(a) 10 items Decoding(a) 5 items	Maths(a) 30 items	Decode(a) 5 items Read(a) 25 items
AMPL-A+B Booklet 8	Listen Comp(a) 10 items Decoding(a) 5 items	Decode(a) 5 items Read(a) 25 items	Maths(a) 30 items

Contextual Information

In addition, to characterise reading and mathematics performance, contextual information gathered alongside the assessments at the student and school levels.

Two questionnaires constructed, each focusing on a different level: student and school. The student-level questionnaire is completed by the students undertaking the assessments and the school-level questionnaire is completed by school principals.

The student questionnaire gathers information on student characteristics, household resources, home support and resources, and student nutrition and sanitation. The school questionnaire gathers information on the characteristics of the school principal, school characteristics, school facilities and resources, and teachers and students.

Sample design

The sample is designed to yield an effective sample size of 400 students. As the sample design involves a first stage of sampling schools and a second stage of sampling classes, it is expected that a sample of at least 150 schools and between 3500 and 5000 students will be drawn. More precise estimates of school and sample size are made on a country-by-country basis once the effects of clustering students within schools are explored during the sample design phase.

Analytic strategy

Some variables included in AMPLab can be measured directly by asking questions for example about characteristics of a student (e.g. gender, age), whereas others cannot (e.g. mathematical ability, socioeconomic status). Descriptive statistical methods are used to review the functioning of items that measured directly.

In the case of attributes that cannot be measured directly, it is necessary to assume there is an unobserved latent trait that can be indicated by a finite set of (manifest) items. This approach of latent measurement acknowledges that items can be of varying quality and together they can be stronger or weaker measures of underlying latent trait of interest. It is, therefore, important to consider the quality of the items used in a study such as AMPLab where the outcomes being measured are unobserved latent traits.

Similar to MILO project, it is proposed to use Item Response Theory (IRT) for the psychometric analysis. The IRT model to be used is the One Parameter Logistic (1PL) model, that considers the probability a student responds correctly to an item as a function of the student's unobserved latent ability within that domain and the "difficulty" of the item.

Before using these models to yield ability estimates (that is, measures of each student on each of the two outcome domains, reading and mathematics) it is important to ensure that the set of manifest items are indeed good indicators of the underlying latent trait, and that the set of items fit together to form a reliable measure. It is also important in a cross-national study to ensure there is no bias in the items across different country or language groups.

For each domain 1PL models will be fitted and consideration will be given to the quality of the individual items, as well as the sets of items used together. The analysis will lead to the recommendation of the item treatment for the student proficiency estimation.

To represent student proficiency distributions, it is proposed that the plausible values approach be used. Plausible values are intermediate values that enable secondary data analysts utilising the application of replication techniques to compute correct standard errors, taking into account both measurement errors and sampling errors.

Psychometric analysis

Calibrations of items included in the cognitive assessments will be first performed separately by domain and country and then internationally using all available countries' data. The outcomes of the national calibrations will be used to review scaling properties of test items at the national level and make decisions about how to treat individual items in each country. When reviewing the national calibrations, particular attention will be paid to the fit of the items to the scaling model, item discrimination, item-by-country interactions, and the gender DIF.

For each domain, the following analysis will be undertaken:

Item level analysis:

Review of item statistics and scoring categories

This is a data cleaning and quality assurance step whereby a review of percentages for correct, incorrect and missing responses will be undertaken. Different types of missing responses will be considered (e.g., not reached, omitted).

Review of item fit statistics

The goodness of fit for individual items (e.g., departure from assumptions of model) can be determined by calculating a mean square (MNSQ, sometimes further broken down by infit and outfit) statistic. This residual-based item fit indicates the extent to which each item fits the item response model. A value of 1 indicates the best possible item fit to the Rasch model, whereas values above 1 show an item discrimination which is lower than expected, and values below 1 an item discrimination which is higher than expected. It is generally recommended that analysts and researchers interpret residual-based statistics with caution and in conjunction with other indicators of item fit. A typical rule of thumb is to identify items with a MNSQ statistic outside a range between 0.8 to 1.2. This should be interpreted along other indicators included averages of latent trait estimates within item categories and classical item statistics such as item-total correlations and point-biserial correlations. Graphical displays, item characteristic curves (ICC), can also be useful.

Differential Item Functioning (DIF)

DIF is variations in the difficulty of items between sub-groups within a sample after controlling for latent ability. For example, it should not be true that because an item is completed in one language it should be harder or easier for students otherwise estimated to be of the same ability who are completing the item in another language.

Statistical tests are undertaken to measure the magnitude of DIF, and usually calculated for subgroups by gender, country, and language group. Where unacceptable DIF is observed, then the item parameter (difficulty) can be "freed" so that it is not assumed to be equally difficult across (some or all) groups and therefore will not bias the estimate of abilities within one specific group.

Latent trait-level analysis

Targeting

A convenient property of IRT approaches is that item difficulty and student ability are reported on the same latent scale. The degree to which the items are well targeted to the ability of the student affects the reliability of the scale and the accuracy of reported statistics. This will be reviewed visually through the presentation of item-thresholds maps (Wright maps), and through the porting of test information functions (targeting).

Reliability

Reliability is a measure of the theoretical consistency of the assessments. It shows how similar would the estimated abilities be if the same assessment was given to the same student multiple times (without the interference of memory effects). Indices range from 0 (unreliable) to 1 (perfectly reliable). The coefficients to measure reliability are estimates of the above analogy and calculated by measuring the internal consistency of a test or the consistency of repeatedly estimated ability estimates.

Latent variance and correlations

The estimates of variance of the latent trait provide information about the range of abilities being assessed and the spread of observed abilities along the estimated scales. The analysis of the relationships among the domains will be useful as a measure of predictive validity of latent traits.

For each domain, the following analysis will be undertaken.

Equating and final item calibration

International AMPL-B item parameters obtained from the item calibration step will be compared to the MILO item parameters to verify usage of a fixed-item-parameter scaling approach. In this approach item parameters will be fixed to their values established in analysis of MILO data. AMPL-A item parameters will be determined by joints calibration all clusters in booklets with AMPL-B item parameters anchored so that the scaled results can be aligned to the AMPL scale reported in MILO.

Generation of Plausible Values

The imputation methodology usually referred to as plausible values (PVs) will be used to select likely proficiencies for students that attained each score (von Davier, Gonzalez & Mislevy, 2009). Five plausible values (PVs) per domain will be generating using a two-dimensional model with conditioning.

The conditioning variables will be prepared using procedures based on those used in the United States National Assessment of Educational Progress (NAEP) (Beaton, 1987) and in TIMSS (Macaskill, Adams and Wu, 1998). The steps involved in this process are as follows:

Step 1. Some background variables where available (e.g. gender, school type, etc) will be prepared to be directly used as conditioning variables.

Step 2. Each variable in the student questionnaire will be dummy coded.

Step 3. For each country, a principal components analysis of the dummy-coded variables will be performed, and component scores will be produced for each student (a sufficient number of components to account for 99 per cent of the variance in the original variables).

Step 4. The item-response model will be fit to each national data set and the national population parameters will be estimated using item parameters anchored at their international location estimated AMPLab and conditioning variables derived from the national principal components analysis in Step 3 and the background variables in Step 1.

Step 5. Five vectors of plausible values will be drawn for each domain.

Standard setting exercise

To enable robust and valid reporting of student achievement against the MPL requirements, a systematic approach will be taken to establish cut-scores that correspond to the end of the lower primary MPL requirements for each AMPL domain (reading and mathematics). Note that the cut-scores that correspond to the end of primary MPL requirements have already been established (through MILO).

To set benchmarks, a standard setting exercise will be conducted in cooperation with officials and subject matter experts from all AMPL participating countries. The Pairwise Comparison Method (PCM) will be used. The PCM allows countries to determine the benchmark on their assessment for meeting global minimum proficiency. This is achieved by subject matter experts (SMEs) undertaking a pairwise comparison exercise using items from the AMPL assessment and items that have already been located in relation to the Learning Progression Scale, that have previously been established through an International Standard Setting Exercise (ISSE). This enables the MPL benchmarks set during the ISSE to be translated onto the AMPL assessment such that the proportion of learners meeting the MPL can be determined.

Using the benchmarks set in the standards setting exercise, the proportion of students meeting or exceeding the MPLs for SDG 4.1.1a and SDG 4.1.1b will be estimated using plausible values.

Reporting

The reporting will clearly articulate the outcomes of the study, in a way that is clear to policy and practitioner stakeholders.

ACER will provide a final report on the performance of students at the end of lower and upper primary school, with a focus on gender

The final report will include:

- A brief description of the study purpose and design, the target population, country samples and response rates
- A statistical summary of the scale score distribution by country and by gender, for each population assessed
- A statistical summary of the proportion of students reaching/exceeding SDG 4.1.1.a Minimum Proficiency Level cut-point by country and by gender, for each population assessed
- A statistical summary of the proportion of students reaching/exceeding SDG 4.1.1.b Minimum Proficiency Level cut-point by country and by gender, for each population assessed
- Descriptive statistics of contextual factors at the student and school levels, for each population assessed
- Inferential statistics of associations of contextual factors and achievement, aggregated at different levels, for each population assessed
- Individual country summaries with a focus on gender.

The final report will be provided by ACER to the UIS for dissemination.

References

ACER & UIS. (2017). *Principles of good practice in learning assessment*. http://uis.unesco.org/sites/default/files/documents/principles-goodpractice-learning-assessments-2017-en.pdf

Australian Council for Educational Research (ACER). (2022). *Minimum Proficiency Levels: Described, unpacked and illustrated.* Version 3.

Beaton, A.E. (1987). Implementing the new design: The NAEP 1983-84 Technical Report. (Report No. 15-TR-20). Princeton, NJ: Educational Testing Service.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Nielsen & Lydiche.

von Davier, M., Gonzalez, E. & Mislevy, R. (2009). What are plausible values and why are they useful? In *IERI Monograph Series* Volume 2, 9-36.

Macaskill, G., Adams, R.J. and Wu, M.L. (1998). Scaling methodology and procedures for the mathematics and science literacy, advanced mathematics and physics scales. In: M. Martin and D.L. Kelly (eds.) Third International Mathematics and Science Study, Technical Report Volume 3: Implementation and Analysis. Chestnut Hill, MA: Center for the Study of Testing, Evaluation and Educational Policy, Boston College

Adams, R.J, Wu, M.L, Macaskill, G, Haldan, S, Cloney, D, Berezner, A (2023). ACER ConQuest: Generalised Item Response Modelling Software [Computer software]. Version 5. Camberwell, Victoria: Australian Council for Educational Research.