



Australian Government
Department of Foreign Affairs and Trade



Pairwise Comparison Method Toolkit

A toolkit for countries to measure global
learning outcomes

2024

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).

Pairwise Comparison Method Toolkit. A toolkit for countries to measure global learning outcomes.

UNESCO Institute for Statistics and Australian Council for Educational Research.

UNESCO

The constitution of the United Nations Educational, Scientific and Cultural Organization (UNESCO) was adopted by 20 countries at the London Conference in November 1945 and entered into effect on 4 November 1946.

The main objective of UNESCO is to contribute to peace and security in the world by promoting collaboration among nations through education, science, culture and communication in order to foster universal respect for justice, the rule of law, and the human rights and fundamental freedoms that are affirmed for the peoples of the world, without distinction of race, sex, language or religion, by the Charter of the United Nations.

To fulfil its mandate, UNESCO performs five principal functions: 1) prospective studies on education, science, culture and communication for tomorrow's world; 2) the advancement, transfer and sharing of knowledge through research, training and teaching activities; 3) standard-setting actions for the preparation and adoption of internal instruments and statutory recommendations; 4) expertise through technical cooperation to Member States for their development policies and projects; and 5) the exchange of specialised information.

UNESCO Institute for Statistics

The UNESCO Institute for Statistics (UIS) is the statistical office of UNESCO and is the UN depository for global statistics in the fields of education, science, technology and innovation, culture and communication. The UIS was established in 1999. It was created to improve UNESCO's statistical programme and to develop and deliver the timely, accurate and policy-relevant statistics needed in today's increasingly complex and rapidly changing social, political and economic environments.

Email: uis.tcg@unesco.org

<http://www.uis.unesco.org>

ACER

The Australian Council for Educational Research (ACER) is an independent, not-for-profit research organisation that has been improving learning for more than 90 years. ACER's mission is to create and promote research-based knowledge, products and services that can be used to improve learning across the life span.

As an official partner of UNESCO, ACER supports global efforts to meet the United Nations Sustainable Development Goals (SDGs) by 2030. In partnership with the Australian Government's Department of Foreign Affairs and Trade (DFAT) and the UNESCO UIS, ACER through the Global Education Monitoring (GEM) Centre has developed methods and tools for countries to monitor and report progress towards achieving SDG 4: education for all. ACER's methods and tools provide policymakers with an understanding of student learning levels in the local context. Being able to benchmark such education monitoring data against global standards provides an important additional perspective on the quality of education systems. This evidence is crucial for developing strategies to improve learning for all.

www.acer.org

www.acer.org/au/gem

Published in 2024 by the United Nations Educational, Scientific and Cultural Organization (UNESCO) Institute for Statistics, C.P 250 Succursale H, Montréal, Québec H3G 2K8, Canada and the Australian Council for Educational Research LTD (ACER), 19 Prospect Hill Road, Camberwell VIC 3124, Australia.

© UNESCO-UIS and ACER 2024

ISBN 978-1-74286-655-0



With the exception of any material protected by a trademark, and where otherwise noted, all material presented in this document is provided under a [Creative Commons Attribution NonCommercial NoDerivatives 4.0 International Licence](https://creativecommons.org/licenses/by-nc-nd/4.0/)

Recommended attribution

The United Nations Educational, Scientific and Cultural Organization (UNESCO) Institute for Statistics and the Australian Council for Educational Research must be attributed as the copyright holder of this publication. To request use outside this licence, email: permissions@acer.org

Recommended citation

UNESCO Institute for Statistics and Australian Council for Educational Research. (2024). *Pairwise Comparison Method Toolkit. A toolkit for countries to measure global learning outcomes*. <https://doi.org/10.37517/978-1-74286-655-0>

Acknowledgements

This publication has been funded by the Australian Government through the Department of Foreign Affairs and Trade, and the Australian Council for Educational Research Ltd. The ideas and opinions expressed in this publication are those of the authors and are not necessarily the views of the Australian Government or of UNESCO.

Contents

Acronyms	5
Glossary of terms	6
1. Introduction to the Pairwise Comparison Method	10
1.1. Foreword	10
1.2. Rationale for the PCM.....	10
1.3. Audience	11
1.4. Overview of the Minimum Proficiency Levels	11
1.5. Overview of the Learning Progression Scales.....	12
1.6. Overview of the Global Proficiency Framework	12
1.7. Relationship between the MPLs and GPF	13
1.8. Overview of the PCM	13
1.9. Advantages of the PCM	14
1.10. Project team	15
2. Self-assessment of the appropriateness of the assessment	16
2.1. Collation of evidence and issues for consideration.....	16
2.2. Criteria for PCM validity.....	16
2.3. Next steps	21
3. Preparing for the Pairwise Comparison Method workshop	23
3.1. Task 1 – Item selection.....	23
3.2. Task 2 – Organise logistics	24
3.3. Task 3 – Select and invite SME participants	25
3.4. Task 4 – Pre-workshop analysis	27
3.5. Task 5 – Technology check.....	28
4. Implementing the Pairwise Comparison Method workshop.....	30
4.1. Task 6 – Training SMEs	30
4.2. Task 7 – Undertaking the comparative judgements	31
4.3. Task 8 – Analysing the outcomes.....	32
4.4. Task 9 – Plenary session	32
4.5. Task 10 – Evaluation	32
5. Self-assessment of the PCM outcomes	33
5.1. Production of the technical documentation at the end of the PCM process.....	33
5.2. Criteria for self-assessment	34

5.3. Submit Evidence to UIS	35
Bibliography	37
Annex A – Minimum proficiency levels unpacked	39
Annex B – Self-assessment template report (Appropriateness of assessment).....	51
Annex C – Alignment rating form.....	54
Annex D – Workshop preparation checklist and activity planner	56
Annex E – Invitation letter template for workshop participants	59
Annex F – Participant demographic information capture form.....	61
Annex G – Sample agenda for PCM workshop	62
Annex H – Pre-workshop analysis	63
Annex I – Workshop facilitation slides	65
Annex J – Post-workshop analysis	75
Annex K – Workshop evaluation form	78
Annex L – Certification of appreciation template	80
Annex M – Self-assessment template report (PCM outcomes)	81
Annex N – Pairwise Comparison Method Report	82

Acronyms

ACER	Australian Council for Educational Research
DFAT	Department of Foreign Affairs and Trade
DIF	Differential Item Functioning
GEM Centre	Global Education Monitoring Centre
GPF	Global Proficiency Framework
IRT	Item Response Theory
ISSE	International Standards Setting Exercise
LPS	Learning Progression Scale
MPL	Minimum Proficiency Level
PCM	Pairwise Comparison Method
PLT	Policy Linking Toolkit
SDG	Sustainable Development Goal
SME	Subject Matter Expert
UIS	UNESCO Institute for Statistics
UNESCO	United Nations Educational, Scientific and Cultural Organization

Glossary of terms

Administration materials – Manuals relating to the administration of the tests and contextual instruments (otherwise known as field guidelines or field operations manuals) as well as important supporting documents such as student attendance forms (sometimes referred to as student tracking forms).

Assessment agency – The body tasked with the organisation of the assessment. It could be a standalone agency, or a team within an existing organisation like a university or the ministry of education.

Assessment design – The implementation plan for the whole assessment, including its purpose, the target population, the content to be tested, testing cycles, etc.

Assessment materials – Test forms, questionnaires, interviews, observation forms

Benchmark – The score on an assessment that delineates having met a proficiency level.

Bias – A systematic distortion of results that is based on factors unrelated to ability.

Blueprint – A description of how the test will be constructed, including the details of the proportion of items that will assess different learning domains and skills and the response formats. Is sometimes referred to as a table of specifications.

Breadth of Alignment – Sufficient coverage of the domains, constructs, and subconstructs in the MPL/GPF by at least one assessment item.

Classical Test Theory – A psychometric theory based on the view that an individual's observed score on a test is the sum of a true score component for the test taker and an independent random error component.

Confidence interval – An interval that specifies a range of values for a parameter estimate, based on a predefined confidence level, and calculated from one sample of the population. The confidence level (usually 95%) for an interval indicates the proportion of intervals, computed from all possible samples, that includes the true value of the parameter being estimated.

Content standards – What content learners are expected to know and be able to do as described in the GPF table on knowledge and skills.

Correlation – Indication of a relationship between two phenomena/variables.

Cycle (assessment) – All activities related to a single main survey assessment administration within a program with repeated administrations designed to assess learning over time.

Depth of Alignment – Sufficient coverage of the MPL/GPF by assessment items.

Desired target population – The population to which inferences from the survey outcomes will be made.

Differential item functioning – When the probability of answering an item correctly depends on the subpopulation the respondent belongs to rather than her/his ability level.

Distractor – A plausible but incorrect answer to the multiple-choice item on an assessment.

Global Proficiency Descriptor (GPD) – A detailed definition crafted by subject matter experts that clarifies how much of the content described under the statements of knowledge and/or skill(s) in the GPF a learner should be able to demonstrate within a subject at a grade level. These are sometimes called performance standards. Authors have purposefully not used that term, however, as countries have their own performance standards that may differ from global standards for important reasons. The set of GPDs included in the GPF are not meant to be prescriptive in nature but rather to facilitate measurement against SDG 4.1.1.

Impact data – The data that help participants understand the consequences of their judgments on the learner population that are subject to application of the benchmarks recommended by the participants.

Inter-rater consistency – An index that indicates participants' overall agreement or consensus across all possible pairs of participants.

Items – The questions or tasks used in an assessment.

Item difficulty – The difficulty of an item as hypothesised by test developers and confirmed by statistics.

Item discrimination – The ability of an item to differentiate amongst learners on the basis of their understanding of the material being tested, reported on a scale from -1 to +1.

Item facility – The probability of a test taker responding correctly to an item on a scale from 0 to 1.

Item pool – The total set of cognitive or contextual items written for an assessment.

Item Response Theory – A mathematical model of the functional relationship between performance on a test item, the test item’s characteristics, and the test taker’s standing on the construct being measured.

Item statistics – The data used to assess whether items are functioning as they should (e.g. percentage of participants who correctly answered the item and average ability of participants who correctly answered the item).

Logit – Log odd units. This unit is based on the logarithm of odds ratio of an event. The odds ratio is the probability for an event divided by the probability against an event. Logits have a mean of 0 and standard deviation of 1.

Marker – Scorers, markers, judges or coders are the people responsible for scoring the participant responses to items or tasks.

Mean – The arithmetic average.

Multiple-choice item – An item that presents several options as answers, from which the participant selects one.

Parameter – A characteristic that defines a population, such as its variability or its average. A characteristic that defines a sample is called a statistic.

Performance standards – How much of the content described in statements of knowledge and/or skill(s) (content standards) learners are expected to be able to demonstrate. See also the definition for Global Proficiency Descriptor above.

Policy linking for measuring global learning outcomes – A specific, non-statistical method that uses expert judgment to relate learners’ scores on different assessments to global minimum proficiency levels. Policy linking includes processes of alignment and matching between assessments and the GPF and benchmark setting.

Population – See ‘target population’

Psychometrics – Theory and methods of measuring psychological traits, such as mathematical ability or motivation to read.

Reliability – The consistency and accuracy of test and contextual measures and results over replications of the testing procedure (American Educational Research Association et al., 2014).

Scale – A numeric or substantive description of progress in learning.

Scorers – see markers.

Scoring – The process of classifying responses and allocating (usually numerical) codes to represent the various categories of response.

Scoring guide – The description of the scoring categories that are used to categorise and score a participant's answer.

Skills – The ways of thinking, or intellectual approaches, that develop as individuals become increasingly proficient in a learning domain (sometimes called 'processes', 'cognitive domains' or 'aspects').

Standard deviation – A numerical measure of how the data values are dispersed around the mean.

Standard error (SE) – A statistic that indicates the measurement error associated with a benchmark (participant judgment).

Statements of knowledge and/or skill(s) – What content learners are expected to know and be able to do for a specific grade and domain, construct, and subconstruct. The statements of knowledge and/or skill(s) are sometimes referred to as content standards. Authors have purposefully not used that term, however, as countries have their own content standards that may differ from global standards for important reasons. The statements of knowledge and/or skill(s) included in the GPF are not meant to be prescriptive in nature but rather to facilitate measurement against SDG 4.1.1.

Statistical linking – Methods that use common persons or common items to relate learners' scores on different assessments. Statistical linking methods include equating, calibration, moderation, and projection.

Validity – The extent to which the assessment instruments measure what they claim to be measuring for a specified population, and the extent to which interpretations made from the data analysis are correct and appropriate for the proposed use of the data (American Educational Research Association et al., 2014).

I. Introduction to the Pairwise Comparison Method

I.1. Foreword

This toolkit has been co-authored by the Global Education Monitoring (GEM) Centre at the Australian Council for Education Research (ACER) and the United Nations Educational, Scientific and Cultural Organization (UNESCO) Institute for Statistics (UIS). The GEM centre provides technical support to UIS, which has been mandated to monitor the progress of countries towards achieving Sustainable Development Goal 4 (SDG 4) in education to “ensure inclusive and equitable quality education and to promote lifelong learning opportunities for all” (United Nations, 2021). The GEM Centre sponsors and contributes to public goods and activities that facilitate education systems reporting against SDG 4 in a globally consistent way. Consistent and high-quality monitoring of student learning will help systems understand the strengths they have and the challenges they face. Moreover, it provides evidence to inform the development of policies and practice to improve student learning. This toolkit has been developed to help support countries to align their assessment with global standards and report against SDG 4.1.

UIS has developed a menu of options to enable countries to report against SDG 4.1.1, of which the Pairwise Comparison Method (PCM) for measuring global learning outcomes is one. Where appropriate, and to support consistency, some content of this PCM toolkit is based on one of the other options for countries, the Policy Linking Toolkit (PLT).

I.2. Rationale for the PCM

Sustainable Development Goal (SDG) 4 aims to ensure that, by 2030, “all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes.”

Indicator 4.1.1 concerns the proficiency indicator referring to three levels of schooling: end of lower primary, end of primary, and end of lower secondary; and two subjects (reading and mathematics). The indicator reads as follows:

“4.1.1 Proportion of children and young people: (a) in grades 2/3; (b) at the end of primary; and (c) at the end of lower secondary achieving at least a minimum proficiency level [MPL] in (i) reading and (ii) mathematics, by sex.”

While the number of countries engaging in learning outcome assessments has increased substantially over the past two decades, methods for comparing assessment results within and across countries, as well as aggregating those results for global reporting, have been lacking. Ministries of Education, regional assessment officers, international education donors, partners, and other stakeholders need a method for accurately

determining how learning outcomes compare between contexts in a country and across countries, and how countries and donors can report on progress in key subject areas such as reading and mathematics. This information is critical for identifying gaps in learning outcomes so that resources can be focused on the areas and populations most in need.

The main challenge with conducting global comparisons and aggregations of assessment results is that countries generally use different assessment tools with varying levels of difficulty. Linking the different assessments to a common scale addresses this problem. Linking can be done either statistically, using common items between assessments or having common learners take more than one assessment, or non-statistically, using expert judgments. Although statistical methods are often associated with higher levels of precision, they are not always practically possible or financially feasible and involve several methodological prerequisites.

This toolkit provides countries with a method to link their national assessment to global standards that combines expert judgement, through a pairwise comparison or comparative judgement process, with statistical linking using item response theory (IRT). Pairwise comparison methods exploit the finding that people are better at comparing two objects or examples of student work against each other, than at evaluating one object or piece of student work against criteria (Thurstone, 1927). Based on multiple comparative judgements, a rank order of tasks or examples of student work is generated. This rank order is based on all decisions made across judges, and results in relatively reliable scales.

Once the PCM has been implemented, countries will understand the relative difficulty of their national assessment to other national assessments. They will, therefore, be able to compare the proportion of their learners meeting minimum proficiency levels with other countries and report against SDG 4.1.1.

1.3. Audience

This toolkit was created for use by country governments and assessment agencies (for multinational assessments) and their partners. Given that a primary focus of the toolkit is helping facilitate country reporting on SDG 4.1.1, all toolkit users, including assessment agencies, should closely coordinate with the relevant country government(s), as it is governments that will ultimately report outcomes to SDG 4.1.1.

1.4. Overview of the Minimum Proficiency Levels

As a custodian agency for reporting against the Sustainable Development Goals in Education, UIS worked with international experts to achieve a consensus on expectations of learners at the three reporting stages of SDG 4.1.1. These descriptions of minimum expected performance were first published in the *Final Report of the Results of*

the Consensus Building Meeting on Proficiency Levels (Nitko, 2018). Work was commissioned and has continued since then on reviewing and refining the draft MPLs.

[Annex A](#) contains the nutshell statement, expanded statement and domains, constructs and descriptors in reading and mathematics of the latest version of the MPL unpacked document. More information, including sample items to support interpretation, can be found in *Minimum Proficiency Levels Unpacked* (ACER, 2022b). Each of the three standards – Grades 2/3 (end of lower primary), end of primary and end of lower secondary – is described in terms of a single standard for reading and a single standard for mathematics.

1.5. Overview of the Learning Progression Scales

The PCM relies on the Learning Progression Scales (LPSs) for reading and mathematics, developed by the Australian Council for Educational Research (ACER)¹. Each of these scales is a robust, statistical ordering of items that was developed using a pairwise comparison method. The items were drawn from a range of different assessments. An International Standard Setting Exercise (ISSE) was carried out by ACER in 2022 to establish the MPL thresholds on the LPSs for reading and mathematics (ACER, 2022a). This ISSE used the Bookmark method to set the thresholds, involving participants with a diverse range of experience, background and skill, including sufficient geographical representation.

1.6. Overview of the Global Proficiency Framework

The Global Proficiency Framework (GPF) was created to respond to the call set up by the Global Education Monitoring Report, tasked with monitoring progress toward SDG 4, to create “shared definitions of what ‘relevant and effective learning outcomes’ are so that they can be comparative across countries and monitored globally.” The PCM described in this toolkit requires this common understanding of the constructs of reading and mathematics to ensure sufficient alignment between the assessment and the LPSs.

While countries define what knowledge and/or skills learners need to obtain in which grades based on their individual contexts and articulate that information through national standards, curricula, and assessments, the GPF defines the knowledge and skills that are important for all learners, no matter where in the world they live.

A team of more than 60 reading and mathematics subject matter experts (SMEs) from around the globe, all of whom have experience working in multiple countries and contexts, came together to create the GPF. The SMEs reached consensus on the statements of knowledge and/or skill(s) (sometimes called content standards) and the global performance descriptors (GPDs) (sometimes called performance standards)

¹ The development of the LPSs for its application in the global context was funded by the GEM Centre.

described in the GPF based on their knowledge of developmental progressions and the UIS's Global Content Framework. The Framework was based on 73 curriculum and assessment frameworks from 25 countries for reading and 115 assessment frameworks from 53 countries for mathematics. It was important that the GPF was grounded in the content framework and expert experience in diverse contexts to ensure the standards described within the document are aligned with and do not exceed existing country content standards and curricula.

1.7. Relationship between the MPLs and GPF

The MPLs were used in the development of the GPF, though since the MPLs are defined in terms of the stage of schooling and the GPF is defined in terms of the grade of a learner, there are some differences. In addition, the GPF provides much more detail on the descriptions of performance for each domain, construct and subconstruct. This means it is useful to use the GPF to understand whether assessments are aligned to the constructs of reading and mathematics as defined globally, despite these differences.

The end of lower primary MPLs (referred to in 4.1.1 as 'Grade 2/3') are described in terms of a single standard for reading and a single standard for mathematics. The alignment with the GPF for reading and mathematics is closest to the 'Meets Global Minimum Proficiency' descriptions for Grade 2.

The end of primary MPLs (4.1.1b) are also described in terms of a single standard for reading and a single standard for mathematics. The alignment with the GPF for reading and mathematics is closest to the 'Meets Global Minimum Proficiency' descriptions for Grade 5.

Finally, the end of lower secondary MPLs (4.1.1c) are described in terms of a single standard for reading and a single standard for mathematics. The alignment with the GPF for reading and mathematics is closest to the 'Meets Global Minimum Proficiency' descriptions for Grade 8.

1.8. Overview of the PCM

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 country's assessment and items that have already been located in relation to the LPS. In this way, the assessment items from the country are also located in relation to the LPS. This enables the MPL benchmarks set during the ISSE to be translated onto the assessment such that the proportion of learners meeting the MPL can be determined.

At present, the PCM can only be implemented with assessments in English, though it is hoped that future iterations of the process will enable it to be used with assessments in other languages.

Table 1 sets out the stages of the PCM that all countries must follow in order to ensure that the outcomes of the PCM are accepted for reporting against SDG 4.1.1.

Table 1: Stages of the Pairwise Comparison Method

#	PCM Stages	Purpose	Roles/Responsibilities	Resources
1	Initial engagement ²	For countries (or assessment agencies in coordination with relevant country governments) to determine whether, for a specific assessment, the PCM is the preferred approach to report against SDG 4.1.1.	Country governments/ assessment agencies may complete this stage themselves or they may request/receive support from their partners – for example, ACER, UIS and/or donors. It is critical that country governments own this process and are willing to provide the necessary information, reports and data to all involved at the appropriate time to support the work. Ownership of the process by country governments will also support capacity development, with a desired aim for them to be able to run future workshops on their own.	Reporting learning outcomes in basic education: Country's options for indicator 4.1.1 (UIS, 2022b)
2	Self-assessment of appropriateness of assessment for reporting against SDG 4.1.1	To determine whether assessment reliability, validity, and alignment with the MPL meets requirements for proceeding with the PCM for global reporting.	Country governments/ assessment agencies with/without support of partners.	PCM Toolkit (Chapter 2 and Annex B)
3	Preparing for the PCM workshop	To identify/confirm facilitators, invite participants and prepare materials.	Country governments/ assessment agencies with/without support of partners.	PCM Toolkit (Chapter 3)
4	Implementing the PCM workshop	To set benchmarks and document details regarding the process followed	Country governments/ assessment agencies with/without support of partners.	PCM Toolkit (Chapter 4)
5	Self-assessment of the PCM outcomes	To determine whether the PCM was implemented appropriately to meet with criteria for global reporting	Country governments/ assessment agencies with/without support of partners.	PCM Toolkit (Chapter 5 and Annex M)
6	Reporting results for SDG 4.1.1 ³	For a country to be counted in global reporting	Country governments with/without support of partners	Protocol for Reporting on SDG Global Indicator 4.1.1 (UIS, 2022a)

1.9. Advantages of the PCM

The PCM has a number of advantages over other methods that can be used to report against SDG 4.1.1:

- It is cheaper to run than statistical linking methods, since it does not require an additional administration of the assessment with either common items or

² This stage is outside the scope of this toolkit. Countries should contact UIS to undertake this stage.

³ Although some information is provided in section 5.3 of this toolkit, countries should contact UIS to discuss this stage if required.

common learners to enable statistical linking. It, therefore, provides a cost- and time-effective way for countries and development partners to align an assessment to global standards.

- The task for participants is relatively simple (determining which item in a pair of items is more difficult) and does not require them to internalise the standards in the MPLs/GPF or determine whether pupils at different standards would have, for example, a two-thirds chance of answering the item correctly, which can be difficult.
- The exercise can be run entirely remotely using a specifically designed platform to maintain consistency.
- As part of the PCM the assessment items are located on the LPSs, providing additional psychometric information for assessments that were only analysed using classical test theory.
- Countries can agree to contribute their assessment items to the global bank of items linked to the LPSs, building an invaluable resource for global education monitoring through the sharing of high-quality items.

1.10. Project team

There are a number of roles that make up the project team and staff need to be appointed early. The requirements for each role are provided below.

Project coordinator

Responsible for the management of the project, including project planning, recruitment of participants and logistics of the workshop, including the video-conferencing service. Ideally, they will have managed similar exercises previously, but should be able to use this toolkit to coordinate the project if not.

Facilitator

Responsible for leading the workshop by ensuring participants understand the PCM and what is expected. They must have expertise in the pairwise comparison method, strong organizational skills, excellent presentation skills, and experience with educators ranging from teachers to policymakers. They should be aware of challenges in the PCM and corrective measures that may be taken to address those challenges.

Data analyst

Responsible for analysing the data required to support the method and organising information for presentation to the participants. This role requires a background in statistics, computational and data visualization skills, and software skills (i.e., Excel or Google Sheets for the workshop data plus statistical software, such as Stata, SPSS, or R for the data).

2. Self-assessment of the appropriateness of the assessment

2.1. Collation of evidence and issues for consideration

The self-assessment activity will be led by the country, with support from the donor organization (if applicable) and implementing partner. It is essential that all evidence, including the assessment instrument itself, is shared with those involved in the self-assessment.

For assessments developed using Classical Test Theory (CTT), this will include operational information such as the assessment design, item facility and item discrimination.

For assessments developed using Item Response Theory (IRT), this will include operational information such as the type of IRT model used and the assessment design, item parameter estimates (difficulty, discrimination, fit, measurement error, etc.), student ability estimates and any other operational procedures in relation to reporting against the existing in-country standards (such the response probability adjustment when placing an item in a performance band).

To enable reporting against SDG 4.1.1, all evidence to support the judgements made as part of the self-assessment, excluding the assessment instrument itself, must be in the public domain. The assessment instrument itself must be kept secure until it is no longer being used for live administrations.

2.2. Criteria for PCM validity

There are five criteria for this first self-assessment to determine if the assessment is sufficiently valid for reporting against SDG 4.1.1. These criteria were agreed at the meeting of the Technical Cooperation Group on SDG 4 Indicators held on 11 December 2023. For each criterion, there are essential minimum requirements for an assessment to be self-assessed as sufficiently valid for SDG reporting. The five criteria relate to the following:

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

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

The project team will need to involve other experts to carry out the self-assessment, for example curriculum and assessment experts. They should record the outcomes of their self-assessment using the form in [Annex B](#).

Criterion 1 – Alignment

To self-assess against this criterion, the project team will need access to the following:

- The assessment instrument(s)
- The Alignment Rating Form in [Annex C](#)

Minimum requirements

To report against SDG 4.1.1, the assessment must be sufficiently aligned to the MPLs. This means that the items must assess content which is reflected in the definition of reading or mathematics in the MPL Unpacked document (depth) in [Annex A](#), and a sufficient range of the content of reading or mathematics should be assessed within the assessment (breadth).

The country will use the Frisbie alignment method described herein to complete the following three sub-steps using the Alignment Rating Form in [Annex C](#). To support this process, countries will need to use Table 3 from the GPF (UIS, 2023c and UIS, 2023d) for grade 2 for end of early primary, grade 5 for end of primary and grade 8 for end of lower secondary (depending on which of the MPLs are being set) because the GPF provides greater detail to support those making decisions than the MPL unpacked document and the two are sufficiently related.

1. For each assessment item, identify the knowledge and/or skill(s) that learners need to answer the item correctly.
2. Search through the GPF (Table 3) to find the domain, construct, subconstruct, and statement(s) of knowledge and/or skill(s) that align(s) with the knowledge and/or skills needed to answer the item correctly (for reading assessments, also examine the grade level of the text, using the criteria for assessing text complexity in Appendices A and B of the Reading GPF).
3. Use the alignment scale that follows to rate the level of alignment of the item.

Alignment Scale:

- **Complete alignment (C)** signifies that all content required to answer the item correctly is contained in the statement of knowledge and/or skill(s), i.e., if the learner answers the item correctly, it is because they only use the knowledge and/or skill(s) described in the statement.

- **Partial alignment (P)** signifies that part of the content required to answer the item correctly is contained in the statement of knowledge and/or skill(s), i.e., if the learner answers the item correctly, it is because they partially use the knowledge and/or skill(s) described in the statement.
- **No alignment (N)** signifies that no amount of the content required to answer the item correctly is contained in the statements of knowledge and/or skill(s), i.e., if the learner answers the item correctly, it is because they use knowledge and/or skill(s) that are different from those described in the GPF.

Once the alignment activity has taken place, the assessment will be considered to be aligned if it meets the following criteria for the appropriate MPL.

MPLa

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

MPLb

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

MPLc

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

algebra domains. The assessment must also cover 12 of the 21 subconstructs at grade 8 in the GPF.

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

Criterion 2 – Item review

To self-assess against this criterion, the project team will need access to the following:

- Assessment instrument, including scoring guidance
- Evidence from the development of the assessment instrument – this may be in the form of a technical report, outputs from data analysis, or from interviews with those responsible for developing the assessment instrument.
- Item statistics.

Minimum requirements

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

The qualitative review should consider whether:

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

The quantitative review should consider whether:

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

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

Criterion 3 – Sample

To self-assess against this criterion, the project team will need access to the following:

- Qualitative description of the population against which they wish to report against SDG 4.1.1

- Qualitative description of the cohort to whom the assessment was administered (if different)
- Information on sampling methodology. For example, if it is a stratified random sample, the project team should be provided with details of the strata (which should at least include district or other large administrative units).

Minimum requirements

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

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

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

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

Criterion 4 - Administration

To self-assess against this criterion, the project team will need access to the following:

- Administration materials
- Any reports on standardised implementation of the administration arrangements.

Minimum requirements

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

Administration guides must be reviewed for clarity and monitoring of the implementation must be undertaken. Any incidents of inappropriate administration,

identified through monitoring or reporting of concerns, should be recorded. Where significant incidents of inappropriate administration are recorded, relevant results should be excluded from the outcomes. This will require additional checks to confirm that this does not affect the representativeness of the sample.

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

Criterion 5 – Reliability

To self-assess against this criterion, the project team will need access to the following:

- Data from the most recent administration of the assessment
- Reliability statistics calculated from analysis of the data
- Details of the quality assurance arrangements for any human-scored items.

Minimum requirements

To report against SDG 4.1.1, the value of coefficient alpha (or equivalent reliability statistic) for the assessment must be greater than or equal to 0.7.

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

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

Overall self-assessment rating

To be eligible for reporting outcomes against SDG 4.1.1, countries must self-assess at the minimum requirement for each of the criteria.

2.3. Next steps

If countries self-assess as meeting the minimum requirements, they may continue to conduct the PCM. The outcomes of the PCM will be eligible for reporting against SDG 4.1.1 as long as the method is conducted in line with the guidance in this toolkit – this will be self-assessed at the end of the process as described in [Chapter 6](#).

Countries that self-assess as not meeting the minimum requirements may wish to consider improving their current assessment arrangements to increase compatibility with SDG 4.1.1 reporting requirements. How this is achieved will depend on which of the criteria the country did not meet.

If the assessment did not meet criterion 1 on alignment to the GPF, the country may wish to consider what changes would be required to their assessment framework or blueprint. For example, the country could include a wider variety of number items from the different subconstructs or including more reading comprehension items, to ensure the assessment is aligned.

If the assessment did not meet criterion 2 on item review, the country may wish to review their test development arrangements to ensure they align with international practice. For example, the country could implement processes as set out in the Standards for Educational and Psychological Testing (2014), which would include reviewing items before inclusion in an assessment.

If the assessment did not meet criterion 3 on sampling, the country may wish to review their sampling methodology to ensure it produces a nationally representative sample, for example by excluding fewer learners or ensuring regional coverage in the sample.

If the assessment did not meet criterion 4 on administration, the country may wish to review their administration guidance or provide more training to test administrators to ensure consistency.

If the assessment did not meet criterion 5, the country may wish to improve their quality assurance arrangements for human-scored items or work with a technical delivery partner to determine ways to improve the reliability of their assessment.

3. Preparing for the Pairwise Comparison Method workshop

This chapter explains the five tasks that need to be completed in order to prepare for the PCM workshop:

- Task 1 – Item selection
- Task 2 – Organise logistics
- Task 3 – Select and invite participants
- Task 4 – Construct and assign item pairs
- Task 5 – Technology check

3.1. Task 1 – Item selection

To make the PCM manageable, it is recommended that a maximum of 45 items are used in the process. The project team should consider whether there will be sufficient time for the participants to make all the required judgements when determining the number of items to use.

If the assessment is longer than 45 items, or a complex sampling design is used, then the project team will need to select which items are to be used.

When selecting the items, the project team should consider the following:

- **Content coverage** – the items selected should broadly reflect the domain, construct and subconstruct coverage of the whole test or the item pool. For example, if 60% of the test/item pool aligns with the ‘retrieve information’ construct in the reading comprehension domain, then 60% of the items selected for the PCM should also be aligned to ‘retrieve information’.
- **Item functioning** – the items selected should be the best functioning items in the assessment. For example, if there are items with negative discrimination, or that exhibit differential item functioning towards a particular gender, these should be removed.
- **Items classified as ‘no alignment’ during the alignment exercise** – the items selected should ideally all have at least partial fit with the GPF. It is helpful to remove non-aligned items to avoid needing to deal with them in the pairwise comparison exercise. For example, for a reading assessment, items assessing grammar, punctuation and spelling should be removed from the item pool.
- **Item difficulty** – when an assessment incorporates a rotation of items across different test forms (a matrix sampling of items approach) the selected set of items should be as evenly spaced as possible from the easiest to the most

challenging item on the underlying IRT scale. The most appropriate items for use in the PCM might not all be contained in a single existing test form. In contrast, the set of items selected for use in the PCM exercise will utilise all the assessment items to provide the best possible discrimination and coverage of knowledge and skills by items included in the PCM.

Alongside the items from the assessment, the project team will need to use the items that have been pre-selected from the relevant LPS for the PCM. The LPSs are organised into levels (14 for reading and 12 for mathematics) and the items selected have been drawn from levels that are appropriate for the MPL under consideration.

3.2. Task 2 – Organise logistics

The workshop for the PCM will take place remotely⁴ with a synchronous training session at the start (approx. 4 hours in total) and SMEs completing the comparative judgements asynchronously over the following period – completing the judgements within 48 hours is recommended to ensure the training is fresh for SMEs. Once the judgements are complete, SMEs should be brought together again remotely to be provided with the outcomes of the PCM and discuss their experiences.

A workshop preparation checklist and activity planner are provided in [Annex D](#) and a sample agenda is provided in [Annex G](#).

There are two systems required to undertake the PCM:

- The online system to enable participants to view items and make their comparative judgements⁵
- The video-conferencing service.

Online system for making judgements

The project team should choose the preferred online system for making judgements based on local availability and familiarity. Ideally, the system selected should allow for:

- easy uploading of items and metadata
- ability to control how pairs of items are presented (for example, the order and whether an item appears first or second on the screen)
- easy uploading of item pairs and assigning them to participants
- presenting of two items simultaneously with appropriate metadata (for example, keys for multiple choice items).

⁴ A face-to-face workshop for the PCM is possible, though this will increase the costs and logistical requirements.

⁵ It is possible to run a paper-based PCM, where SMEs are provided with packs of items to review and record their judgements on paper. This will increase the costs and logistical requirements and will have a significant environmental impact, so is not recommended.

- an intuitive way for participants to select which of the two items is more difficult
- an option for SMEs to go back and change the decision for a previous pair as they are working through allocated pair list
- appropriate response speed to ensure participants are not waiting for screens to refresh
- internal analysis of data to produce required outcomes or straightforward export of data for analysis in external software.

Video-conferencing service

The project team should make their choice of preferred video-conferencing service based on local availability and familiarity. Ideally, the service provided should allow for:

- presenting slides and sharing one's screen
- assigning participants to break-out groups
- recording the sessions (for SMEs who miss portions of the workshop due to technological issues to listen to after the sessions; if possible, find a platform that does not take long to process the recording so it can be released to SMEs quickly)
- muting everyone upon entry in the meeting
- typed chats
- raising one's hand to indicate a question or comment
- registration of participants to help track attendance (if the latter is not possible, administrative staff should be on hand to track changing attendance throughout each session – possibly noting who is there at the beginning, middle, and end; this allows facilitators to follow up with SMEs who missed significant portions of the workshop due to technological issues).

3.3. Task 3 – Select and invite SME participants

The SMEs are key to the workshop, as they are the ones who will actually make judgments on the difficulty of the items in the assessment compared to items that are already part of the LPS. The project team should plan separate workshops for each MPL, subject, and language of assessment for which the PCM will be used.

When selecting SMEs for a PCM workshop, the number of SMEs must be sufficiently large and representative. This is to provide reasonable assurance that the benchmarks 1) will be realistic, attainable, and unbiased and 2) would not vary greatly if the process were repeated with different SMEs. The SMEs must have strong content knowledge and teaching skills (reading or mathematics) to enable them to make the judgments required of them. They must also be perceived as experts in their field within their education system in order to foster the confidence of host governments in their decisions.

For each workshop, a group of 15 SMEs is a minimum and 20 SMEs is a maximum. A group of this size will ensure the process obtains a replicable outcome but is also practical and manageable. As shown in Error! Reference source not found., the SMEs in each workshop should be made up of at least 50 percent master classroom teachers/assessment markers and up to 50 percent non-teachers, preferably curriculum or assessment experts.

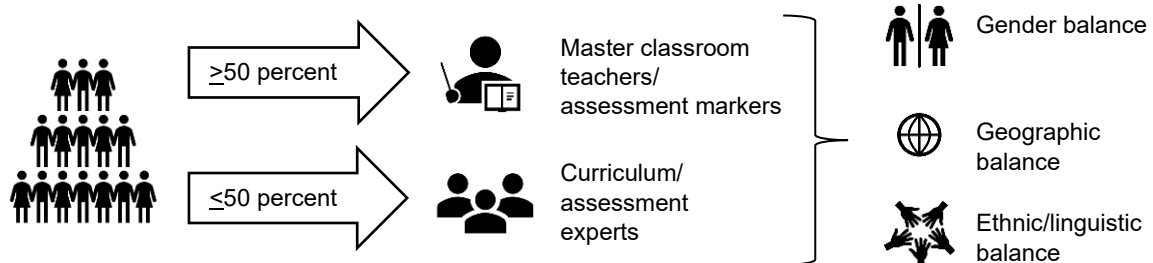


Figure 1: Composition of SMEs in each workshop

A typical workshop will include 6 teachers/markers and 6 curriculum/assessment experts as the SMEs. Qualifications for SMEs include the following:

- At least five years of teaching at or adjacent to the relevant grade level (teachers)
- At least five years of marking/scoring assessments (markers)
- At least five years of teaching experience (curriculum experts)
- At least five years of developing assessment items (assessment experts)
- Strong skills in the learning domain (reading or mathematics)
- Experience of interrogating system-level data
- Native skills in English
- Experience with a variety of learners at different proficiency levels
- Knowledge of the instructional system, including materials
- Teacher's college and/or university certification and licensing.

Aside from qualifications, representativeness for the SMEs for the workshop should be ensured through the following criteria:

- **Gender representation** – The SMEs must be selected to ensure a gender balance proportionate to the teaching profession in the country, both for the teachers and non-teachers.
- **Geographical representation** – The SMEs must be selected to ensure representation from regions, provinces, and/or states of the assessments.
- **Ethnic and/or linguistic representation** – The SMEs must have diversity that reflects the population.

- **Other representation** – Depending on its relevance to the context and specific learner populations for whom results will be reported, the composition of the SMEs might need to reflect other characteristics as well. These characteristics could include the following: assignment at private and public schools, experience with learners who have disabilities, background in accelerated learning programs, and location in crisis and conflict environments.
- **Representation for multinational assessments** – When the PCM workshop is seeking to link regional or international assessments to the GPF, it is important that SMEs represent multiple countries.

The project team should collaborate with the government, donor agency, implementing partner(s), and/or other stakeholders to determine the most appropriate way to recruit SMEs. This may be done through nominations by the Ministry of Education, assessment agency, or other government agency. The government, donor, partner, and facilitators should discuss how to apply the criteria in their context. It is important that the different parties agree to minimum requirements for the qualifications and representativeness criteria. A template letter to use to invite SMEs is provided in [Annex E](#).

SMEs' demographic information should be collected, aggregated, and submitted with the workshop outcomes using the form included in [Annex F](#). This form will give the project team sufficient data to address the degree to which the SME meet the criteria as part of the post-workshop self-assessment.

A list of SMEs and their contact details should be available sufficiently in advance of the workshop to ensure preparation activities can be appropriately conducted.

3.4. Task 4 – Pre-workshop analysis

Construct and assign item pairs

Once all of the items have been selected and the number of SMEs involved confirmed, the project team will construct the pairs of items that will be compared and assign them to the SME participants.

Since the items selected from those already aligned to the relevant LPS will be restricted to the levels that are appropriate for the MPL under consideration in the workshop and the items from the assessment should have been selected to have been of appropriate difficulty, it should be possible to randomly assign pairs of items. To maximise information from the exercise, though, it may be appropriate to avoid most of the trivial comparisons where a very easy item is paired with a very hard item.

It is recommended that each item is included in a comparison pair on average 50 times, with minimum exposure of 40. The maximum exposure will depend on the resource available to conduct the exercise in terms of number of available items and SME participants.

Further care must be taken to balance pair construction so that each item is compared to the pre-selected items from the relevant LPS and to items from its own assessments.

The position of an item in a pair should be randomised across allocated pairs. The pairs should then be randomly ordered with a condition that a single item should appear in up to three pairs sequence. This “chaining of item pairs” is done to reduce the cognitive load for the SMEs (see Pollitt & Crisp, 2004). Each SME should then be randomly assigned a set of pairs from the overall pair sequence.

Further details of how to construct the item pairs is provided in [Annex H](#).

The procedure for uploading items and pairs and assigning these to SMEs into the online system for making judgements will vary depending on the system selected. The project team should refer to the instructions provided by the online system provider.

3.5. Task 5 – Technology check

It is essential to carry out a technology check with participants in advance of the workshop, ideally with time to resolve issues for participants where connection issues may occur (i.e. at least a few days in advance of the workshop rather than at the start of the proposed workshop).

The technology check should take place with the equipment the participant will use to access the workshop. Ideally, participants should join via a laptop or PC, rather than a smartphone, so they can see slides (though joining by smartphone is acceptable if there are no other suitable options).

The project team may need to provide data cards to participants to ensure they have sufficient data to connect to the session(s) and should encourage participants to assess their service far in advance of the workshop in case they need to explore changing providers (if possible).

The project team should also set up a group chat on an agreed messaging service (e.g. WhatsApp or Telegram) in advance of the workshop to facilitate announcements, remind participants of sessions, and ensure ease of communication between workshop sessions when many participants do not have regular access to email communications.

The following activities should be undertaken during the technical test:

- **Connectivity** – do all participants have a suitable connection?
- **Audio** – do all participants have suitable microphones and speakers?
- **Platform** – are all participants familiar with the digital platform features (e.g., muting, raising hands, using chat functions, switching between breakout rooms etc.)?

- **Macros** – if the workshop intends to use digital forms containing macros, are these accessible to all participants?
- **Messaging app** – have all participants downloaded the chosen messaging app and can they access the group chat?

4. Implementing the Pairwise Comparison Method workshop

This chapter explains the six tasks that make up the workshop involving the SMEs:

- Task 6 – Training SMEs
- Task 7 – Undertaking the comparative judgements
- Task 8 – Analysing the outcomes
- Task 9 – Plenary session
- Task 10 – Evaluation

4.1. Task 6 – Training SMEs

The first part of the workshop involves training the SME participants. Training contains two sections:

- LPS – the facilitator provides information on the concepts behind the LPSs and the domain-relevant LPS is unpacked.
- PCM – the facilitator explains the PCM concept and the procedure that participants will follow, including use of the online system.

Presentation slides for reading and mathematics are available in [Annex I](#) and as separate files.

LPS training

The training is designed to provide SMEs with an understanding of the concept of a learning progression.

ACER's learning progressions are concerned with improvement in broad areas of learning. This improvement is essentially conceptualised as an increasing amount of a construct, e.g., proficiency in reading or mathematics. The quantitative representation of this increase is a continuous numerical scale, which in ACER's view is a central and necessary feature of a learning progression.

Within one of ACER's learning progressions, the numerical scale represents a long vertical range of improvement in the learning area. This long vertical range is divided into an arbitrary but convenient number of levels, typically no more than fourteen. The numerical scale underpins a structure of four layers of qualitative descriptions/illustrations of improvement in the learning area:

- The **Domain layer** gives 'big picture' information about the learning area.

This layer defines the learning area; explains why it is considered valuable for an individual and for society to develop understandings and skills in the learning area; and gives a broad outline of what improvement in the learning area looks like.

- The **Levels layer** describes levels in the development of understanding and skill in the learning area.

At this layer, improvement in the learning area is described for levels of attainment, corresponding to the divisions of the numerical scale. The description for each level comprises a 'nutshell' summary statement, and an elaboration of the understandings and skills that are typically associated with the level.

- The **Strands layer** describes levels in the development of understanding and skill for strands that make up the learning area.

At this layer, levels of attainment are described for the major 'threads' along which improvement within the learning area occurs. These threads are referred to as 'strands'.

- The **Illustrations layer** comprises examples of what might be observed in a student's behaviour or responses when particular understandings and skills are operationalised.

At this layer, the strand-level descriptions from the Strands layer are illustrated via exemplar tasks, assessment items, samples of student work and descriptions of student thinking and reasoning.

PCM training

This training is likely to focus primarily on the online system for making judgements, since the task that SMEs are asked to perform is relatively simple:

- When considering a pair of items, determine which item is more difficult.

The project team should refer to the instructions provided by the online system provider in order to develop suitable training.

4.2. Task 7 – Undertaking the comparative judgements

Once the SMEs have been trained, they can start to make their judgements. It is recommended that SMEs are encouraged to complete their judgements within 48 hours of the training, in order to ensure the information provided in the training is fresh in their minds. This period may be extended, though, if local circumstances dictate.

The online system will present SME participants with pairs of items from the assessment and the items already aligned to the LPS. SMEs will be required to answer the question:

Which of the two items presented is more difficult?

SMEs will continue to make judgements until they have made judgements about all the pairs that they have been assigned.

The project team will need to make themselves available during the period when SMEs are making their judgement (most likely via the chosen messaging app) to answer any questions that the SMEs have.

4.3. Task 8 – Analysing the outcomes

Once the pairwise judgements have been completed, they can be downloaded from the software and prepared for analysis. To produce an item difficulty scale for items included in the exercise, the Bradley-Terry-Luce (BTL) model is used (Bradley and Terry, 1952 and Luce, 1959)⁶. Differential item functioning (DIF) analyses can then be used to evaluate the stability of the relative difficulties of the items from the original scale to the LPS. Then, relevant cut scores (e.g., MPLs) can be transferred to the newly equated scale and subsequent secondary analyses (e.g., proportions of sample/population at described proficiency levels) can be done.

Further details of the analysis to conduct are provided in [Annex J](#).

4.4. Task 9 – Plenary session

Once the analysis has been completed, the workshop should be reconvened to share the outcomes with the SMEs. This plenary session will also provide opportunity for SMEs to share their views on the process.

4.5. Task 10 – Evaluation

The evaluation should be conducted at two stages in the process: following training and prior to SMEs starting to make their judgements; and once all comparative judgements have been made. The first evaluation enables issues to be identified and addressed and the second provides evidence of the participants views of the process to support the final self-assessment.

⁶ The BTL model is functionally equivalent to one-parameter IRT model and thus BTL item difficulty scales have the same properties as those developed in assessment programs from which the items used in the development of the LPS were sourced. Consequently, the BTL item locations can be interpreted in the same way and used to develop a set of described proficiency levels using the same approach as used in large-scale assessments

5. Self-assessment of the PCM outcomes

5.1. Production of the technical documentation at the end of the PCM process

At the end of the process, the project team should produce a report using the template in [Annex N](#). This report should detail the process followed, issues experienced and how these were resolved. In particular, to support the self-assessment of the PCM outcomes and SDG 4.1.1 reporting, the report must contain information on:

- The SME participant demographics
- The benchmark for the MPL on the assessment calculated following the pairwise comparison exercise
- The proportion of learners achieving the MPL
- The precision, accuracy, and consistency of the judgements
- The outcomes of the SME participants' evaluation.

This information should be presented at an impact analysis workshop, where the project team present the outcome of the pairwise exercise, statistical linking and impact analysis to relevant in-country stakeholders. This workshop should examine all the reliability indices, the magnitude of the statistical linking error and the impact analysis statistics.

The purpose of this workshop is to gather further evidence for the validity of the statistical linking including relevant to implemented process and procedures and outcome implications.

Participant demographics

The demographics of the SME participants will have been captured using the form in [Annex F](#). The project team will need to confirm that all SME participants met the requirements for participation and, as a group, were sufficiently representative. The requirements for SME participants and the requirements for representation are provided in [Chapter 3](#).

MPL benchmark

The analysis required to determine the MPL benchmark is provided in [Annex J](#). The project team will need to ensure appropriate quality assurance of all analysis undertaken.

Outcome data

The final benchmarks should be used to determine the proportion of learners achieving the MPL using the process described in [Annex J](#). The outcomes should be provided for

all learners and male and female learners separately.

Precision, Accuracy and Consistency

The analysis required to determine the precision, accuracy and consistency of the PCM outcomes is provided in [Annex J](#).

Evaluation

The evaluation ([Annex K](#)) contains a series of statements that participants record their level of agreement against a 5-point Likert scale (1 – strongly disagree to 5 – strongly agree) grouped into two sections:

- Training
- Making judgements.

For each statement, the mean average score for all participants (excluding outliers) should be calculated. For each section, the minimum and maximum of the scores for the statements should be noted. Where multiple workshops for different grades/subjects are being conducted simultaneously, evaluation results should be calculated separately for each workshop.

5.2. Criteria for self-assessment

The information documented above should be used to confirm that the outcomes of the process meet the requirements for reporting against SDG 4.1.1.

In order to be considered eligible for reporting against SDG 4.1.1, countries will need to be able to answer YES to all of the following criteria questions:

- Criterion 1 – Did all participants meet the requirements for participation? (YES / NO)
- Criterion 2 – Were the group of participants sufficiently representative in terms of the characteristics agreed by the country? (YES / NO)
- Criterion 3 – Were SMEs removed from analyses if their responses did not fit the model well? (YES / NO)
- Criterion 4 – Were items/SME participants considered for removal from analyses if they did not fit the model well, and was there a clear rationale for the ultimate decision? (YES / NO)
- Criterion 5 – Is the pairwise scale reliability index equal to or higher than 0.75? (YES / NO)
- Criterion 6 – Were items removed from analyses if they exhibited item DIF? (YES / NO)

- Criterion 7 – For the items from the assessment being linked, is the dis-attenuated correlation between the items original scale location and LPSs’ location equal to or higher than 0.75? (YES / NO)
- Criterion 8 – Was the average (mean) score for each section of the evaluation greater than or equal to 4? (YES / NO)
- Criterion 9 – Did the impact analysis workshop confirm the validity of the statistical linking exercises? (YES / NO)

5.3. Submit Evidence to UIS

Table 2 indicates the information countries will need to provide to UIS for SDG 4.1.1 reporting.

Table 2: Information required to report against SDG 4.1.1

Level	Indicator ID	Indicator description
SDG 4.1.1a	MATH.G2T3	Proportion of students in Grade 2 or 3 achieving at least a minimum proficiency level in mathematics, both sexes (%)
	MATH.G2T3.F	Proportion of students in Grade 2 or 3 achieving at least a minimum proficiency level in mathematics, female (%)
	MATH.G2T3.M	Proportion of students in Grade 2 or 3 achieving at least a minimum proficiency level in mathematics, male (%)
	READ.G2T3	Proportion of students in Grade 2 or 3 achieving at least a minimum proficiency level in reading, both sexes (%)
	READ.G2T3.F	Proportion of students in Grade 2 or 3 achieving at least a minimum proficiency level in reading, female (%)
	READ.G2T3.M	Proportion of students in Grade 2 or 3 achieving at least a minimum proficiency level in reading, male (%)
SDG 4.1.1b	MATH.PRIMARY	Proportion of students at the end of primary education achieving at least a minimum proficiency level in mathematics, both sexes (%)
	MATH.PRIMARY.F	Proportion of students at the end of primary education achieving at least a minimum proficiency level in mathematics, female (%)
	MATH.PRIMARY.M	Proportion of students at the end of primary education achieving at least a minimum proficiency level in mathematics, male (%)
	READ.PRIMARY	Proportion of students at the end of primary education achieving at least a minimum proficiency level in reading, both sexes (%)
	READ.PRIMARY.F	Proportion of students at the end of primary education achieving at least a minimum proficiency level in reading, female (%)
	READ.PRIMARY.M	Proportion of students at the end of primary education achieving at least a minimum proficiency level in reading, male (%)
SDG 4.1.1c	MATH.LOWERSEC	Proportion of students at the end of lower secondary education achieving at least a minimum proficiency level in mathematics, both sexes (%)
	MATH.LOWERSEC.F	Proportion of students at the end of lower secondary education achieving at least a minimum proficiency level in mathematics, female (%)
	MATH.LOWERSEC.M	Proportion of students at the end of lower secondary education achieving at least a minimum proficiency level in mathematics, male (%)
	READ.LOWERSEC	Proportion of students at the end of lower secondary education achieving at least a minimum proficiency level in reading, both sexes (%)
	READ.LOWERSEC.F	Proportion of students at the end of lower secondary education achieving at least a minimum proficiency level in reading, female (%)
	READ.LOWERSEC.M	Proportion of students at the end of lower secondary education achieving at least a minimum proficiency level in reading, male (%)

Countries will need to submit the following evidence to UIS to demonstrate that the information in Table 2 was generated in accordance with the requirements for policy linking:

- Self-assessment – appropriateness of assessment (see [Annex B](#))
- Self-assessment – workshop outcomes (see [Annex M](#))
- Policy linking workshop report (see [Annex N](#))

Bibliography

- ACER (2022a). International Standard Setting Exercise. WG/GAML/5.
https://tcg.uis.unesco.org/wp-content/uploads/sites/4/2022/11/WG_GAML_5_ISSE_ACER.pdf
- ACER (2022b). Minimum Proficiency Levels Unpacked. WG/GAML/4.
https://tcg.uis.unesco.org/wp-content/uploads/sites/4/2022/11/WG_GAML_4_MPLs-Unpacked_ACER.pdf
- Bradley, R., A. and Terry M., E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* 39 (3/4), 324–345
- Heldsinger, A.S. & Humphry, M.,S. (2013) Using calibrated exemplars in the teacher-assessment of writing: an empirical study. *Educational Research*, 55(3), 219-235.
- Huynh, H. & Meyer, P. (2010). Use of Robust z in Detecting Unstable Items in Item Response Theory Models. *Practical Assessment, Research & Evaluation*, 15(2).
<http://pareonline.net/getvn.asp?v=15&n=2>
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley
- Luce, R., D. (1959). Individual choice behaviour: a theoretical analysis. Wiley, New York
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York, NY: Springer, doi: 10.1007/978-1-4757-4310-4
- Nitko, A. (2018). Final Report of the Results of the Consensus Building Meeting on Proficiency Levels. Retrieved from <https://gaml.uis.unesco.org/wp-content/uploads/sites/2/2018/10/Final-Report-of-September-2018-Paris-Consensus-Meeting.pdf>
- Pollitt, A. & Crisp, V. (2004). *Could Comparative Judgements Of Script Quality Replace Traditional Marking And Improve The Validity Of Exam Questions?* Paper presented at the British Educational Research Association Annual Conference, UMIST, Manchester.
- Pollitt, A. (2012) The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice*, 19(3), 281-300.
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review* 34(4), 273-286
- UIS (2022a). Protocol for reporting on SDG global indicator 4.1.1.
<https://tcg.uis.unesco.org/wp-content/uploads/sites/4/2022/03/Protocol-for-Reporting-SDG-4.1.1.pdf>
- UIS (2022b). Reporting learning outcomes in basic education: Country's options for indicator 4.1.1. *38th Annual Conference for Educational Assessment*.
<https://tcg.uis.unesco.org/wp-content/uploads/sites/4/2022/08/Countrys-reporting-option- Zambia AAEA.Final .pdf>

UIS (2023c). Global Proficiency Framework for Reading.

<https://gaml.uis.unesco.org/policy-linking/#:~:text=To%20produce%20reliable%20benchmarks%20for%20international%20reporting%2C%20the,assessments%20are%20administered%20according%20to%20minimum%20quality%20standards.>

UIS (2023d). Global Proficiency Framework for Mathematics.

<https://gaml.uis.unesco.org/policy-linking/#:~:text=To%20produce%20reliable%20benchmarks%20for%20international%20reporting%2C%20the,assessments%20are%20administered%20according%20to%20minimum%20quality%20standards.>

Annex A – Minimum proficiency levels unpacked

Reading: End of lower primary (SDG 4.1.1a)

Nutshell statement

Students accurately read aloud and understand written words from familiar contexts. They retrieve explicit information from very short texts. When listening to slightly longer texts, they make simple inferences.

Expanded statement

In a short simple text of one or two sentences, students read aloud most words – including some unfamiliar words – accurately but slowly and often word by word. They identify the meaning of familiar words, including when they have common morphological changes, and also some unfamiliar words. They retrieve explicit information from a single sentence. When listening to longer texts, and looking at the illustrations, students retrieve explicit information about main events, ideas or characters and use that information to draw simple inferences.

Domains, constructs and descriptors

Decoding

- In a short and simple connected text of one or two sentences, decode most words, including some unfamiliar words with familiar sound–symbol patterns (applies to alphabetic and alpha-syllabic languages only).

Reading comprehension

Retrieving information

- Identify the meaning of familiar words in a sentence.
- Locate most pieces of explicit information within a sentence when the information is prominent and there is no or limited competing information.

Listening comprehension

Retrieving information

- In a longer text that is read aloud to them, identify key events, ideas and major characters.

Interpreting information

- In a longer text that is read aloud to them, make simple inferences and identify the meaning of key words that may be unfamiliar.

Reading: End of primary (SDG 4.1.1b)

Nutshell statement

Students independently and fluently read simple, short narrative and expository texts. They retrieve explicitly stated information. They interpret and give some explanation about the main and secondary ideas in different types of texts, and establish connections between main ideas in a text and their personal experiences.

Expanded statement

In a short, simple narrative or expository text, students read aloud at a pace and a level of accuracy and expression (prosody) that demonstrate understanding. They use previously taught morphological (word-level) and contextual (sentence- or text-level) clues to understand the meaning of familiar and unfamiliar words and to distinguish between the meanings of closely related words. When reading silently or aloud, they locate explicit information in a paragraph. They use that information to make inferences about behaviours, events or feelings. They identify the main and some secondary ideas in a text if they are prominently stated, and recognise common text types when the content and structure are obvious. They make basic connections between the text and their personal experience or knowledge.

Domains, constructs and descriptors

Decoding

- In a short, simple narrative or expository text, read at a pace and with a level of accuracy and expression (prosody) that meet minimum standards for fluency in the language of instruction.

Reading comprehension

Retrieving information

- Locate most pieces of explicit information when the information is prominent and found within a single paragraph containing limited competing information.

Interpreting information

- Use morphological or contextual clues to identify the meaning of most unfamiliar words, familiar words used in unfamiliar ways, different shades of meaning of closely related words, synonyms or basic figurative language.
- Establish the main idea of a text when it is prominent in the text.
- Make simple inferences by relating two or more prominent pieces of explicitly stated information, when there is little or no competing information, in order to identify behaviours, feelings, events and factual information.

Reflecting on information

- Establish basic connections between the key ideas in a text and personal knowledge and experience.
- Distinguish between text types (narrative and expository) and recognise some other common text types (for example, poetry, recipe, game instructions) when the content and structural clues are obvious.

Reading: End of lower secondary (SDG 4.1.1c)

Nutshell statement

Students retrieve and connect multiple pieces of related information across sections of texts to understand key ideas. They make straightforward inferences when there is some competing information. They reflect and draw conclusions in a variety of text types.

Expanded statement

In a range of continuous and non-continuous texts, including narrative, expository, descriptive, argumentative, instructional, and transactional texts, students locate multiple pieces of information across a text, including information in paratextual elements. They make straightforward inferences by drawing on prominent explicit and implicit information to summarise key ideas, and select evidence to support an interpretation. They reflect on texts in relation to personal experience and draw on general knowledge to identify if there is an obvious flaw in a text-based idea.

Domains, constructs and descriptors

Decoding

- In languages with large and complex sets of symbols, accurately decode most words.

Reading comprehension

Retrieving information

- Locate multiple pieces of related information that are dispersed throughout a text with familiar structures, when there is some similar information nearby.
- Locate paratextual information in continuous and non-continuous texts (for example, footnotes in continuous texts, legends in maps).

Interpreting information

- Connect pieces of related information across multiple sections of a text, including when ideas are well separated and there is competing information, in order to demonstrate
- understanding of less prominent ideas.

- Sequence events when there are overlapping timelines.
- Make inferences, drawing on obvious clues or prominent information, to summarise main ideas in paragraphs or across entire texts, when there is some competing information.
- Select evidence from a text, including obvious tone, to support an interpretation (for example, a simple comparison of two characters or two events).
- Apply information from the text to new examples (for example, classifying new items according to a described scheme).

Reflecting on information

- Recognise the implied audience of a text with a familiar format and content when there are multiple clues.
- Provide an example of how a text relates to personal experience.
- Draw on external knowledge to identify an obvious flaw in an idea or to make a prediction.
- Recognise different text types when they have familiar styles, language or text layouts.
- Distinguish between fact and opinion when the distinction is straightforward (for example, 'Evidence shows that ...' [fact] versus 'In my view, ...' [opinion]).
- Recognise the purpose of common print conventions, such as use of symbols and simple graphics.

Mathematics: End of lower primary (SDG 4.I.1a)

Nutshell statement

Students recognise, 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 recognise and name familiar shapes and describe their basic attributes. They recognise time in days, weeks and months. They describe location in a space using simple language.

Expanded statement

Students can read, write and compare whole numbers up to 100. They can add and subtract numbers within 20, double and halve whole numbers within 20, and solve application problems involving numbers within 20. Students can recognise simple shapes and their attributes and use these shapes to make other shapes. They can also measure and compare lengths of shapes and lines using non-standard units. They use calendars and recognise days in a week and months in a year. They can read simple data displays. They possess foundational knowledge of spatial orientation, and can appraise the relative size of real-world objects.

Domains, constructs and descriptors

Number and operations

Whole numbers

- Count, read, write, compare, and order whole numbers up to 100.
- Represent quantities up to 100 concretely, pictorially, and symbolically.
- Solve addition and subtraction problems within 20 that are presented concretely, pictorially, and symbolically.
- Divide a group of up to 20 objects into 2 equal sets.
- Solve simple real-world problems using addition and subtraction facts within 20.

Measurement

Length, weight, capacity, volume, area and perimeter

- Use non-standard units to measure and compare length and weight.

Time

- Tell time using a digital clock.
- Tell time using an analogue clock to the nearest hour.
- Recognise the number of days in a week and months in a year.
- Solve problems, including real-world problems, using a calendar (for example, given a calendar, answer the question: March 2 falls on which day of the week?).

Currency

- Count combinations of commonly used currency denominations.
- Combine commonly used currency denominations to make a specified amount.

Statistics and probability

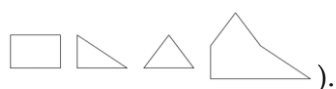
Data management

- Compare categories of simple data displays (that is, simple column graphs / bar graphs, tally charts, pictographs) with up to four categories and a single unit scale (for example, for a column graph showing favourite colours, make statements like: 'More children chose green than yellow, 'Blue was the most popular colour', 'Three more children chose blue than chose red').

Geometry

Spatial visualisations

- Compose/decompose a larger two-dimensional (2D) shape from a small number of given shapes without lines showing where the shapes go (for example, use the smaller shapes to make the larger shape:



Properties of shapes and figures

- Recognise and name shapes that are regular and irregular (for example, if shown an irregular triangle, recognise that it is a triangle; name a hexagon).
- Recognise and name straight and curved lines and attributes of shapes (for example, number of sides, number of corners).
- Recognise when a 2D shape has been rotated or reflected (for example, when shown a number of shapes, identify those that are the same, even when some are rotated or reflected).

Position and direction

- Interpret and use positional terms (for example, in front of, behind, opposite, between).
- Accurately use the terms left and right (for example, answer, 'Where is the teacher's desk?' 'To the [left] of the chalkboard.').

Algebra

Patterns

- Extend non-numerical repeating patterns, recognise repeating units, and identify a missing element (for example, $\bigcirc \square \square \bigcirc \square \square _ \square \square$).

Mathematics: End of primary (SDG 4.1.1b)

Nutshell statement

Students recognise, read, write, order and compare whole numbers within 100,000, unit fractions and their multiples. They add/subtract with whole numbers within 1,000 and multiply/divide with whole numbers within 100. Students can measure length, weight and capacity using standard units; read time on an analogue clock; calculate the perimeter of simple 2D shapes and the area of rectangles; and describe the attributes of familiar 2D and 3D shapes. They read, interpret and construct different types of data displays such as tables, column graphs and pictographs, and recognise, describe and extend number patterns. They can solve simple application problems.

Expanded statement

Students can add and subtract whole numbers within 1,000 and demonstrate fluency with multiplication facts up to 10×10 and related division facts; solve simple application problems with whole numbers using the four operations; identify simple equivalent fractions; compare and order unit fractions and fractions with related denominators; identify and represent quantities using decimal notation up to the tenths place; select and use a variety of tools to measure and compare length, weight and capacity/volume; read time to the minute on an analogue clock and calculate elapsed time in minutes within and across the hour; construct data displays with data arranged into categories and single or multi-unit scales; retrieve multiple pieces of information from data displays to solve problems; recognise and name 2D shapes and familiar 3D objects by their simple attributes such as number of faces, edges and vertices for 3D shapes and number of sides and corners for 2D shapes; describe and continue number patterns that increase or decrease by a constant value from any starting point; or that increase or decrease by a constant multiplier; and apply the concept of equivalence by finding a missing value in a number sentence.

Domains, constructs and descriptors

Number and operations

Whole numbers

- Read, write, compare, and order whole numbers up to 10,000.
- Skip count forwards and backwards using twos, fives, tens, hundreds, and thousands.
- Round whole numbers up to the nearest hundred and thousand.
- Add and subtract whole numbers within 1,000.
- Demonstrate fluency with multiplication facts up to 10×10 , and related division facts.

- Solve simple real-world problems using the four operations, with the unknown in different positions (addition and subtraction within 1,000 and multiplication problems using facts up to 10×10 and their associated division facts).

Fractions

- Identify simple equivalent fractions where one denominator is a multiple of another (for example, $\frac{1}{3} = \frac{2}{6}$).
- Compare and order unit fractions (for example, $\frac{1}{4}, \frac{1}{3}, \frac{1}{2}$) or fractions with different but related denominators (for example, $\frac{2}{3}, \frac{7}{12}, \frac{5}{6}$).

Decimals

- Identify and represent quantities using decimal notation (symbols) up to the tenths place (for example, identify that 0.8 is eight tenths).

Measurement

Length, weight, capacity, volume, area, and perimeter

- Select and use a variety of tools to measure and compare length, weight, and capacity/volume (to the nearest marked increment on the scale).
- Identify the relationship between the relative size of adjacent units within a familiar standard system of measurement for length, weight and capacity/volume (for example, identify the number of millimetres in a centimetre, the number of pints in a quart, the number of grams in a kilogram).
- Calculate the perimeter of a polygon.
- Solve problems, including real-world problems, involving the area of a rectangle.

Time

- Tell time using an analogue clock to the nearest minute.
- Solve problems, including real word problems, involving elapsed time in minutes across hours (for example, calculate the difference between 3:24 and 5:12 or the difference between 16:35 and 18:22), including problems involving schedules (that is, timetables, agendas, itineraries).

Statistics and probability

Data management

- Complete missing information in simple data displays using data arranged into categories, with a single or multi-unit scale, with some support provided (for example, labelled horizontal and/or vertical axes).
- Retrieve multiple pieces of information from data displays to solve problems (for example, calculate a total represented by multiple bars on a graph, compare two categories on the graph).

Geometry

Spatial visualisations

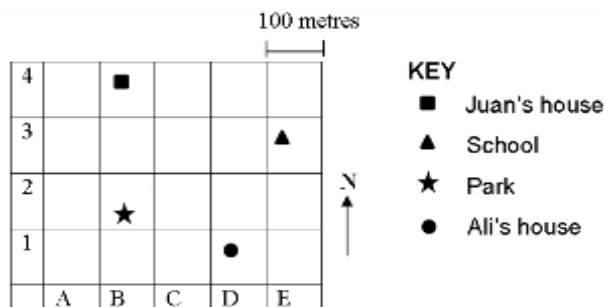
- Identify the net of a cube or specific faces on the net of a cube (for example, fold mentally to answer the question, 'Which of these is the net of a cube?'; 'Identify opposite faces on a net.').

Properties of shapes and figures

- Recognise and name 2D shapes and simple 3D objects by their attributes (that is, their lines and angle properties; for example, distinguishing between equilateral, isosceles and scalene triangles; describing the number of faces, edges and vertices of a rectangular prism).

Position and direction

- Follow more complex directions and/or give simple directions to a given location (for example, go straight, turn right at the corner with the tree, turn left at the next corner, keep going to the green house).
- Use different kinds of simple maps, such as alphanumeric maps, grid maps, or local equivalents, to give and follow two-step directions to a given location (for example, 'Using this map, if you are at the school, you walk 100 metres north, and turn left. What would you be facing?'; 'Which of these is closest to the distance between the park and Juan's house? 100 metres / 150 metres / 200 metres / 250 metres).



Algebra

Patterns

- Describe numerical patterns as increasing by a constant value but starting at a number that is not a multiple of the value of the pattern (for example, the pattern 5, 8, 11, 14 starts at 5 and goes up by 3).
- Describe numerical patterns that increase or decrease by a constant multiplier, and use this information to identify a missing element or extend the pattern (for example, describe that the pattern 2, 4, 8, 16 starts at 2 and doubles or that the pattern 20, 10, 5, 2.5 starts at 20 and halves; identify the missing element in the pattern 3, 6, __, 24, 48; write the next two numbers in the pattern 80, 40, 20, 10).

Relations and functions

- Demonstrate understanding of equivalence by finding a missing value in a number sentence using addition, subtraction, multiplication or division of numbers within 100 (for example, $23 + \underline{\quad} = 29$; $6 \times \underline{\quad} = 54$).

Mathematics: End of lower secondary (SDG 4.1.1c)

Nutshell statement

Students demonstrate skills in computation with fractions, decimals, rates, ratios, percentages and integers. They apply geometric relationships and formulae such as area, volume, Pythagoras' theorem, and the angle sum of a triangle. They interpret and construct a variety of data displays and calculate measures of central tendency. They make use of algebraic representations of linear relationships. They can use their mathematics knowledge to solve application problems.

Expanded statement

Students can apply the order of operations and solve simple problems involving fractions, decimals and whole numbers. They can apply geometric relationships and formulae (namely, area of a triangle, circumference and area of a circle, volume of a rectangular prism, Pythagoras' theorem, and angle sum of a triangle) to solve straightforward problems in simple contexts. They can interpret and construct a variety of data displays and calculate measures of central tendency. They can graph linear equations on a coordinate grid. They can solve equations in one variable and model context-based situations using simple algebraic representations. They can evaluate and calculate with simple algebraic expressions. They can use proportional reasoning to solve problems.

Domains, constructs and descriptors

Number knowledge and operations

Operations across number

- Evaluate numerical expressions requiring application of order of operations.
- Solve problems with fractions, decimals, and whole numbers.
- Identify and express percentages less than 1% and greater than 100% as fractions or mixed numbers and vice versa (for example, $124\% = 1\frac{24}{100}$; $0.2\% = \frac{2}{1000}$).
- Multiply and divide two decimal numbers and divide a whole number by a decimal.
- Solve real-world application problems involving the multiplication or division of two decimal numbers.

Fractions/decimals

- Compare and order positive and negative decimals and fractions (for example, place these numbers on a number line from -1 to $+1$: -0.4 , $+\frac{1}{2}$, $-\frac{4}{5}$, 0.25 , $-\frac{1}{3}$, $\frac{3}{4}$).

Exponents and roots

- Apply the laws of exponents.

Measurement

Length, weight, capacity, volume, area, and perimeter

- Make conversions of units of length and weight between different systems of measurement when the conversion factor is provided (for example, convert 12 cm to inches given 1 inch is 2.54 cm; convert pounds to kilograms given 1 pound is 0.45 kg).
- Solve problems, including real-world problems, involving the calculation of the volume of a rectangular prism (for example, calculate the volume in cubic centimetres of a box with a length of 10 cm, width of 10 cm, and height of 15 cm).
- Solve simple problems involving the area of triangles and the area and/or circumference of circles.

Statistics and probability

Data management

- Read, interpret and construct a variety of data displays, including two-way tables, line graphs, circle (pie) graphs, compound bar graphs.
- Calculate range and measures of central tendency (namely, mean, median and mode).

Chance and probability

- Compare probabilities of simple events.

Geometry

Properties of shapes and figures

- Classify angles in polygons.
- Recognise and name parts of the circle (namely, radius, diameter, circumference) and identify the relationship between the radius and diameter.
- Describe and implement 2D shape transformations (namely, reflection, rotation, translation, enlargement/reduction).
- Determine measurements in right triangles using Pythagoras' theorem.
- Use the angle sum of a triangle to solve problems (for example, determine the missing angle of a triangle where two angles are given).

Spatial visualisations

- Identify the net of a familiar 3D figure, such as a prism, cylinder, cone, or pyramid (for example, fold or unfold mentally to answer the question, ‘What figure does this make when folded?’; ‘What figure does this make when unfolded?’).

Position and direction

- Identify the outcomes of one or more transformations on a 2D object.
- Locate and plot points on a plane in all four quadrants of a Cartesian coordinate system.

Algebra

Patterns

- Describe, complete, and extend geometric and other non-linear sequences of numbers and objects.

Expressions

- Use expressions to represent problem situations with multiple variables (for example, ‘Akeelah bought 4 blouses for x dollars and a wristwatch for y dollars. Represent this as an expression.’).
- Evaluate and simplify exponential expressions using the laws of exponents (for example, evaluate $2x^3$ when $x = 7$; simplify $(3x^4)^2$).
- Multiply and divide linear monomials, and simplify linear expressions, by using the distributive property (for example, multiply $(3x)(5y)$; simplify $2x(3x + 4)$).

Relations and functions

- Solve linear equations in one variable.
- Represent context-based situations with expressions and equations in one or two variables.
- Interpret equations and their solutions in terms of context (for example, given an algebraic graph, such as a distance-time graph, interpret the slope as speed).
- Use formulas to solve context-based problems.
- Solve problems involving ratios, proportions, and percentages

Annex B – Self-assessment template report (Appropriateness of assessment)

Assessment Instrument	[Insert name of instrument]
Country	[Insert country where assessment instrument is administered]
SDG 4.1.1 level	End of lower primary / End of primary / End of lower secondary [delete as appropriate]
Subject	Mathematics / Reading [delete as appropriate]
Date of self-assessment	[Insert date on which self-assessment was undertaken]

Criterion 1 – Alignment

Assessors	[Insert names and organizations of those who undertook alignment exercise]
Level of alignment	Minimal / Additional / Strong [delete as appropriate]
Number of score-points in assessment instrument	[Insert number of score-points]
Number of score points per relevant domain	[Insert number of score-points per relevant domains for alignment level]
Number of subconstructs in relevant domains	[Insert number of subconstructs in relevant domains for alignment level]
Number of relevant subconstructs assessed	[Insert number of relevant subconstructs covered in assessment]
Percentage of relevant subconstructs assessed	[Insert percentage of relevant subconstructs assessed]

Criterion 1 rating: Insufficient / Sufficient [delete as appropriate]

Criterion 2 – Item Review

Assessors	[Insert names and organizations of those who undertook self-assessment]
Is there evidence that the items in the assessment have been reviewed quantitatively?	Yes / No [delete as appropriate]
Is there evidence that the items in the assessment have been reviewed qualitatively?	Yes / No [delete as appropriate]

Criterion 2 rating: Insufficient / Sufficient [delete as appropriate]

Criterion 3 – Sample

Assessors	[Insert names and organizations of those who undertook self-assessment]
Was the assessment administered to the whole cohort or a sample?	Whole cohort / sample [delete as appropriate]
Were any subgroups of the population systematically excluded from administration?	[Insert excluded subgroups of the population for reporting]
For sample – based assessments, is the margin of error 5 percent or less at the 95 percent confidence level?	Yes / No / Not Applicable [delete as appropriate]

Criterion 3 rating: Insufficient / Sufficient [delete as appropriate]

Criterion 4 – Administration

Assessors	[Insert names and organizations of those who undertook self-assessment]
Was the assessment instrument administered in an appropriate and standardized way?	Yes / No [delete as applicable]
Were administration guides clear on the administration process?	Yes / No [delete as applicable]
Were quality assurance procedures developed and implemented to identify and document incidents of inappropriate administration?	Yes / No [delete as applicable]
Were significant incidents of inappropriate administration recorded and relevant results excluded from the outcomes?	Yes / No [delete as applicable]
Did the exclusion of results from inappropriately administered assessments affect the representativeness of the sample?	Yes / No / Not Applicable [delete as appropriate]

Criterion 4 rating: Insufficient / Sufficient [delete as appropriate]

Criterion 5 – Reliability

Assessors	[Insert names and organizations of those who undertook self-assessment]
Is the value of coefficient alpha (or equivalent reliability statistic) for the assessment greater than or equal to 0.7?	Yes / No [delete as applicable]
Is there evidence of appropriate quality assurance arrangements for any human-scored items?	Yes / No / Not Applicable [delete as appropriate]

Criterion 5 rating: Insufficient / Sufficient [delete as appropriate]

Overall Self-Assessment Rating

Criteria	Insufficient	Sufficient
Criterion 1 – Alignment	<input type="checkbox"/>	<input type="checkbox"/>
Criterion 2 – Item review	<input type="checkbox"/>	<input type="checkbox"/>
Criterion 3 – Sample	<input type="checkbox"/>	<input type="checkbox"/>
Criterion 4 – Administration	<input type="checkbox"/>	<input type="checkbox"/>
Criterion 5 – Reliability	<input type="checkbox"/>	<input type="checkbox"/>
Overall Self-Assessment Rating	<input type="checkbox"/>	<input type="checkbox"/>

Annex C – Alignment rating form

To update this form, facilitators should check the total number of questions/items listed on the left and modify to fit the needs of the assessment being used. If using this form electronically, facilitators may wish to create conditional drop-down menus or autofill certain columns.

Table 3: Alignment Rating Form Template

Question	Domain	Construct reference	Subconstruct reference	Complete, Partial or No Alignment	These columns are only required where there is partial fit. You can use these to record any other domains, constructs, and subconstructs that relate to the item.			
					Domain	Construct reference	Subconstruct reference	Complete, Partial or No Alignment
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								

21								
22								
23								
24								
25								
26								
27								
28								
29								
30								
31								
32								
33								
34								
35								

Annex D – Workshop preparation checklist and activity planner

Table 4: Workshop Preparation Checklist

Activity	Owner	Deadline	✓
1. Select and contract project team			
a. Identify and contract facilitator			
b. Identify and contract data analyst			
c. Identify and contract project coordinator			
2. Prepare workshop logistics			
a. Identify and contract online system for making judgements			
b. Identify and contract video-conferencing service			
c. Identify phone card/data allowances and agree on amounts for participants and observers with government/ assessment agency and donor officials			
d. Identify method for receiving funds in country (if necessary); this might involve a wire or cash transfer			
e. Make cash/wire transfer, if needed			
f. Transfer funds to participants			
3. Select and invite SME participants			
a. Finalize teacher/marker participant list			
b. Finalize curriculum/assessment specialist participant list			
c. Finalize observer list			
d. Prepare and distribute invitations, with pre-workshop assessment instructions, to teacher/marker participants			
e. Prepare and distribute invitations, with pre-workshop assessment instructions, to specialist participants			
f. Prepare and distribute invitations for observers			
4. Prepare Materials			
a. Finalize and distribute the agenda			
b. Finalize and distribute the acronym list			
c. Finalize and distribute the glossary			
d. Assign and distribute participant IDs (if required)			
e. Extract and distribute the relevant grade/domain of Table 3 of the GPF			
f. Finalise and distribute evaluation forms			
g. Construct and assign item pairs			
i. Upload participant details into online system for making judgements			
h. Upload item pairs into online system for making judgements and assign			
j. Distribute login details for participants for the online system			
i. Produce impact data			
j. Finalise and distribute facilitation slides			
5. Technology check			
a. Organise session with SME participants to ensure they can use the chosen video-conferencing system and have suitable connectivity			

Table 5: Workshop Activity Planner

Number	Activity	Role/Responsibility
Four Weeks before the workshop		
1	Initiate contact with country	UIS/Donor Organization (DO)
2	Decide on which assessment and MPL(s) to use in PCM process	Country with UIS/DO and Delivery Partner (DP) support
3	Decide the timing of the workshop	Country with UIS/DO/DP support
4	Identify Facilitator	Country
5	Identify data analyst	Country
6	Identify project coordinator	Country
7	Identify SME participants (both teachers and content specialists), including collecting their contact information; ensure panel is representative	Country
8	Extract the relevant grade/domain of Table 3 of the GPF	DP
3-4 Weeks before the Workshop (the earlier the better)		
9	Draft agenda	DP
10	Provide feedback on draft agenda	Country
11	Finalise agenda	DP
12	Invite participants	Country, UIS/DO, or DP - depending on country's preference
3 Weeks before the Workshop		
13	Identify and invite any workshop observers - from other donors, Ministries, etc.	Country with UIS/DO/DP support
14	Identify other potential costs for the workshop, including phone/internet cards and materials during the workshop	Country
15	Submit budget to UIS/DO	Country
16	UIS/DO and DP complete NDAs	UIS/DO and DP
17	Send assessment instruments to UIS/DO and DP	Country
18	Identify and contract online system for making judgements	DP
19	Provide Ministry logo for certificates and determine who from the Ministry will sign	Country
20	Draft workshop slides and rating forms to send to UIS/DO for review	DP
2 Weeks before the Workshop		
21	Review workshop slides and rating forms and send feedback to DP	UIS/DO
22	Finalise MOU with country based on approved budget	UIS/DO
23	Draft certificates	DP
24	Finalise item rating forms and slides based on UIS/DO feedback	DP
25	Send data to UIS/DO and DO (if possible)	Country
26	Confirm participant participation	Country
27	Decide on video-conferencing service for workshop	Country
28	Transfer funds and/or phone/internet cards to participants	Country
1 Week before the Workshop		
29	Finalise the agenda (with any last-minute changes)	DP
30	Finalise the agenda, acronym list, glossary, assessment, GPF, rating forms, evaluation forms, slides with notes fields, certificates, and any other documents	DP
31	Distribute the agenda, acronym list, glossary, assessment, GPF, rating forms, evaluation forms, slides with notes fields, certificates, and any other documents	Country
32	Assign participant IDs	DP
33	Distribute participant IDs	Country

Number	Activity	Role/Responsibility
34	Construct and assign item pairs	DP
35	Upload item pairs into online system for making judgements	DP
36	Analyse data to produce impact data.	DP
37	Finalise facilitation slides and distribute	DP
A Few Days before the Workshop		
38	Remote platform testing with participants can access the platform and don't need technical support	All
Workshop begins		

Annex E – Invitation letter template for workshop participants

This annex includes a letter template for SME participants. All details that need to be filled in are included in brackets. The letter should be modified as needed to fit the context.

[Date]

Dear [Name],

Invitation to participate in the Pairwise Comparison Method

In pursuit of the Sustainable Development Goals on education (SDG 4.1.1), [Country/Regional or International Assessment] has decided to proceed with using a global reporting method called the Pairwise Comparison method (PCM). This method allows countries/assessment agencies to determine whether its learners are reaching global minimum proficiency in reading and mathematics, according to SDG 4.1.1.

Through the PCM, countries/assessment agencies will link their national assessments to a common global reporting scale using benchmarks. Setting the benchmarks requires judgments by panels of subject matter experts.

[Country/Assessment Agency] is planning to host virtual PCM Workshop on **[training date]** with a plenary session on **[plenary date]**. Participants will be required to undertake an online activity between these two dates. The Workshop will focus on linking [Assessment Name(s)] with SDG 4.1.1 for [end of lower primary/end of primary/end of lower secondary]. Participants will include master teachers, [markers/scores/raters/coders] of the assessment, curriculum experts and assessment experts, and they will be guided through a systematic process that involves reviewing pairs of assessment items and determining which is more difficult.

[Government Ministry/Assessment Agency] needs a total of [Number of SMEs] to participate in the workshop, including [X number from Location, with experience in Grade level, Subject, and Language of Assessment; Y number from . . .]. As such, [Government Ministry/Assessment Agency] would like to invite you to participate in the workshop.

Participation in the workshop will provide a valuable learning opportunity for the selected participants, who will gain an increased understanding of international standards for learner performance.

[Logistical details]

If you have questions or require further clarifications, please contact [Name] via phone [number]. Please kindly confirm your participation by [Date]. If you do decide to participate, we ask that you complete the pre-workshop activity detailed in the attachment to this letter ahead of the workshop. Your participation in this workshop is crucial and we look forward to you joining us.

Sincerely,

[Name and Title]

Annex F – Participant demographic information capture form

Facilitators should update this form to reflect the geographical distinctions (specifically, the region and district) that need to be tracked to ensure appropriate representativeness of the panel for the workshop and should add any other details needed for reporting. They may also want to create an electronic form that enables easier capture of the data.

Subject Group	<input type="checkbox"/> Reading <input type="checkbox"/> Mathematics
SDG 4.1.1 reporting level	<input type="checkbox"/> End of lower primary <input type="checkbox"/> End of primary <input type="checkbox"/> End of lower secondary
Name	
Occupation	
Region where you teach/work	
District where you teach/work	
Email	
Mobile number	
Gender	<input type="checkbox"/> Woman <input type="checkbox"/> Man <input type="checkbox"/> Non-binary / gender diverse <input type="checkbox"/> My gender identity isn't listed. I identify as: <hr style="width: 200px; margin-left: 0;"/> <input type="checkbox"/> Prefer not to say
Ethnicity	
Education level	
Years of experience	
Years of teaching/working with relevant school stage and subject	
Professional organisation/affiliation (e.g. school, ministry etc.)	
Prior Training(s) in Reading/Mathematics (answer only for the subject for which you are serving as a subject matter expert)	<input type="checkbox"/> Yes <input type="checkbox"/> No
Experience teaching learners with disabilities	<input type="checkbox"/> Yes <input type="checkbox"/> No
Experience working with conflict- and crisis-affected population	<input type="checkbox"/> Yes <input type="checkbox"/> No
First language	
Other languages spoken	
Language(s) Used for Classroom Instruction (for teachers only)	

Annex G – Sample agenda for PCM workshop

The following sample agenda may be adapted by the project team to reflect local decisions on timings etc.

Activity	Timing
Welcome and introductions	15 minutes
Overview slides: <ul style="list-style-type: none">• What is reading/mathematics?• What is a learning progression?• What is the pairwise comparison method?	30 minutes
Factors that influence reading/mathematics item difficulty	45 minutes
Practice items	30 minutes
Pairwise comparison exercise <ul style="list-style-type: none">• Number of pairs to be compared• Deadline for completion of work• Evaluation survey	15 minutes

Once the participants have completed their judgements, and the analysis has been completed, you may decide to bring the group back together for a short session to explain the outcomes of the exercise.

Annex H – Pre-workshop analysis

The main purpose of these analyses is to:

1. Construct the item pairs.
2. Allocate item pairs to different SME participants.

Although details are provided below to show the process, the selected pairwise software is expected to be configured to complete these actions as manual construction of pairs and pair sequences is laborious and error prone process. The logistics of preparing the software, e.g., uploading the items, the compilation and allocation of pairs will depend on software solution. The software should be able to provide report that shows.

- The total number of generated pairs.
- The number of times an item is included in the pairs.
- The number of item pairs allocated to each SME participant.

The software must have ability to accurately allocate item pairs and allow SMEs to record their decision.

I. Item pairs construction

The outcome of self-assessment regarding the Criterion 1: Alignment and Criterion 2: Item Review will inform the level of control of the item pair construction. If these activities indicate high level of alignment with the LPS, then a more structured approach in construction of pairs could be implemented. Alternatively, a random allocation of items across pairs is preferred.

Structured pairs construction

In this approach, pairs are constructed to limit the number of pairs where very easy items are compared to very hard items, as indicated by the location of the items on the LPS. The steps of the process are:

- Provisionally allocate the assessment items from the assessment being linked to the most appropriate LPS level - using qualitative judgment of two experts with excellent understanding of LPS levels (to note, these should not be experts who will be involved as SME participants).
- If the assessment being linked contains items that cover the whole range of LPS levels, then pairs should be constructed to exclude pairings where the difference in levels of the items is plus/minus four levels. That is, an item should only be compared to items that are up to four levels above or four levels below the level assigned to the item by the SMEs.

- If the assessment being linked contains a restricted range of LPS levels, then pairs should be constructed excluding pairings where the difference in levels of the items is plus/minus two levels.

Random pair of items

Where it is not possible to assign provisional LPS levels to the items in the assessment being linked, the random allocation approach should be used. Randomly pair items regardless of their estimated difficulty for the items from the assessment being linked or their LPS level for the pre-selected items that are already linked to the LPS.

Balance item position in the pair

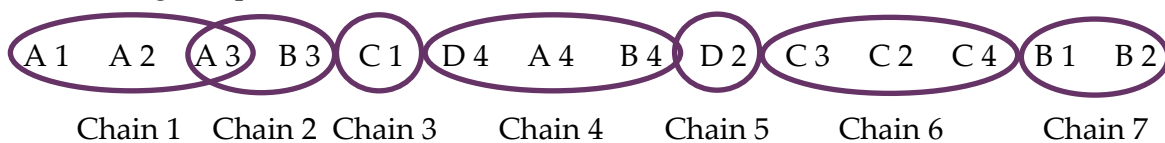
Once the pairs have been created, ensure that the position of item in a pair (i.e., whether it appears on the left or on the right side of a screen in the pairwise software selected) is balanced for each item across all the pairs in which it appears.

Item pair chaining

To reduce the cognitive load for SME participants, it is useful if consecutive pairs have one common item. To achieve this, the following steps should be followed:

- Place pairs that contain the same item in a sequence that has length of one, two, or three pairs.
- Vary the length of chain for each item (i.e. a chain of three pairs containing the same item should be followed by a chain with a length of one or two pairs)
- Start the chaining process from a random item and ensure the sequence of pairs is built at random to ensure that the complete list of pairs is randomly compiled.

Here is an example of a list of chained pairs for two sets of items where the first (A, B, C, D) is being compared to a second (1, 2, 3, 4)



2 Item pair allocation

To allocate pairs to SME participants, the following steps should be followed:

- Allocate SME participants a number from 1 to N at random (N = total number of SME participants), which will be their order ID.
- Split the pairs list to N approximately equal parts.
- Allocate the partitioned lists to SME participants using the randomly generated order ID.

Annex I – Workshop facilitation slides

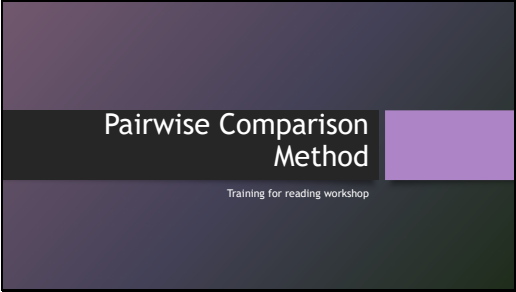
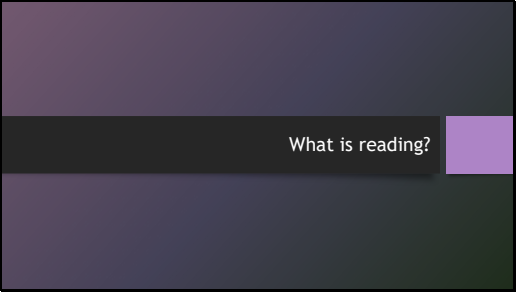
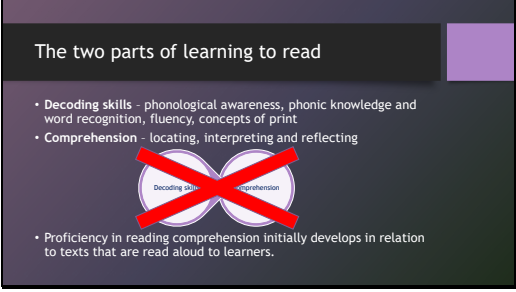
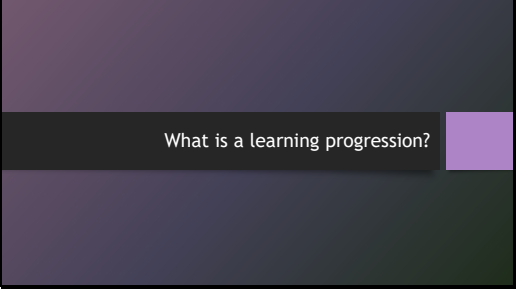
Facilitation slides have been developed for reading and mathematics. The slides contain notes to support facilitation. Facilitators should read through the slides and make any necessary adaptations prior to delivery.

The reading slides contain information about both reading comprehension and decoding. If the assessment you are using does not contain explicit decoding items, then these slides can be removed (slides 27 to 39 and 42).

The mathematics slides contain some tables that may be difficult to read when shared, depending on screen size. The project team should make separate versions available for participants who need them (slides 23 and 25).

The project team will also need to source pairs of practice items for the end of the presentation. This will give the participants the opportunity to practice making the difficulty judgement before they start the real exercise. These items should be similar in nature to those that are contained within the assessment being linked but they should not be items that will be used in the pairwise comparison exercise as this could bias the outcomes of the exercise.

Reading slides

<p>Slide 1</p>  <p>Pairwise Comparison Method</p> <p>Training for reading workshop</p>	<p>Slide 2</p>  <p>What is reading?</p>
<p>Slide 3</p>  <p>The two parts of learning to read</p> <ul style="list-style-type: none">• Decoding skills - phonological awareness, phonic knowledge and word recognition, fluency, concepts of print• Comprehension - locating, interpreting and reflecting <p>Proficiency in reading comprehension initially develops in relation to texts that are read aloud to learners.</p>	<p>Slide 4</p>  <p>What is a learning progression?</p>

Slide 5

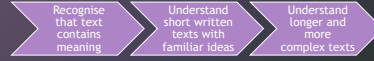
Locating and comparing tasks and learners

- We can imagine a continuum of reading development and there have been many attempts to systematically describe this (e.g., progress maps, reporting systems, learning progressions)
- Assessment aims to identify/compare the location of individuals on a growth continuum
- Assessment uses tasks appropriate to the expected location of the individuals assessed

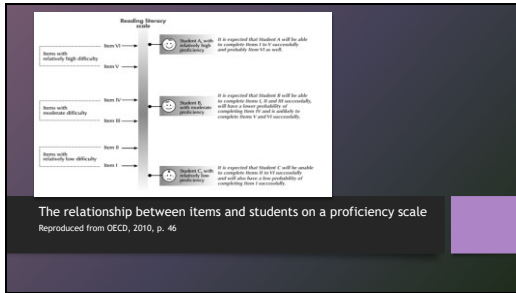
Slide 6

Using assessment tasks

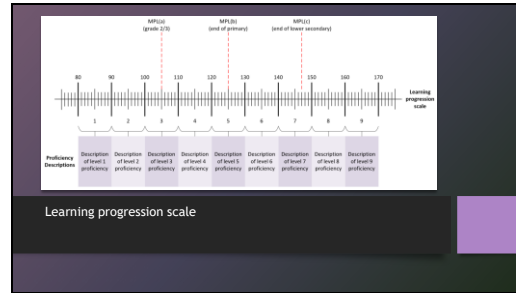
- We can use assessment tasks to determine where a student is at on this journey



Slide 7



Slide 8



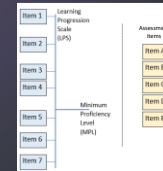
Slide 9

What is the pairwise comparison method?

Slide 10

Two sets of items

- One set mapped to a learning progressions scale
- The other set from the [insert name of assessment]



Slide 11

Make comparisons between the items

- Consider pairs of items at a time
- Which of the pair of items is more difficult?



Slide 12

One combined set of items

- Once all the comparisons are made, we can map all of the new items on the same scale as the old items
- We can also determine the cut score on the [insert name of assessment] that equates to the minimum proficiency level



Slide 13

Advantages of the approach

- Comparing two items is easier than evaluating one against some criteria
- Multiple comparative judgements → rank order of items
- There is no limit on the type of items that can be compared - multiple choice, short response, open ended, interactive, etc.
- The process is generally fast and efficient, allowing for comparison of a large number of items
- The process is robust and reliable in generating links with Minimum Proficiency Levels

Slide 14

Factors that influence reading item difficulty

Slide 15

Variables

1. Number of features and conditions: information to be located in the text
2. Proximity of pieces of required information to each other
3. Competing information in the stimulus and/or the distractors that the reader may mistakenly select, or may generate
4. Prominence of necessary textual information
5. Relationship between task and required information

Slide 16

Variables

6. Semantic match between task and text: matching task wording and the necessary information
7. Concreteness of information. The kind of information that readers must identify to complete a question
8. Familiarity of information needed to answer the question: close to the experience of the reader, or remote and unfamiliar?
9. Register of the text: implied relationship between the reader and the text; lexico-grammatical density
10. Extent to which information from outside the text is required to answer the question.

Slide 17

Ong mapu septat: diron nifit. The boy ran to school.

- | | |
|-------------------------------|---------------------------|
| 1. Ablah nit ong mapu septat? | 1. Where did the boy run? |
| a. nifit | a. school |
| b. diloco | b. shops |
| c. vilat | c. home |

- A very simple task**
- Students can select the **only word** in the options that appears in the text.
 - Students **do need letter/word matching skills**
 - No comprehension of the question or text required
 - No decoding

Slide 18

The boy ran to school. He was not happy about the rain. He ran fast because he was late. He was scared he would get into trouble.

Why did he run fast?

- a. He was happy.
- b. He was late.
- c. He was scared.

- A simple task:** All the options are words in the text.
- Students can match words from the question to the text and find adjacent information.
 - They need some knowledge of grammar.
 - They do not have to read or understand the meaning of the words.

Slide 19

Van is at school. He has new pencils.
Van draws a picture of a big tree with green leaves and red flowers.

Where is Van?

Match 'is' and 'Van' and copy 'at school'.

What colour are the flowers?

Match 'flowers' and copy 'red'

Slide 20

Orange and Cardamom Fruit Salad

Ingredients
4 oranges 1 tablespoon of honey
1/2 cup of raisins 1/2 a teaspoon of cardamom powder (a spice)

Instructions

- Peel 3 oranges, cut into slices and put in a bowl.
- Pick over the raisins to remove any stalks and add to the bowl.
- Put the juice of one orange into a saucepan with the cardamom and honey. Stir over a gentle heat for 5 minutes.
- Pour the hot sauce over the fruit in the bowl and mix gently.
- If you don't eat it immediately, keep it cool.

Consider the text complexity and the skill demands of the item.
An unfamiliar or more complex text, such as a recipe with some unfamiliar vocabulary, may have some easy items and some harder items.

Slide 21

Orange and Cardamom Fruit Salad

Ingredients

4 oranges 1 tablespoon of honey
1/2 cup of raisins 1/2 a teaspoon of cardamom powder (a spice)

Instructions

- Peel 3 oranges, cut into slices and put in a bowl.
- Pick over the raisins to remove any stalks and add to the bowl.
- Put the juice of one orange into a saucepan with the cardamom and honey. Stir over a gentle heat for 2 minutes.
- Pour the hot sauce over the fruit in the bowl and mix gently.
- If you don't eat it immediately, keep it cool.

1. Where do the instructions tell you to put the raisins?
Answer: in a bowl

2. The ingredients list says 4 oranges but only 3 oranges are peeled and sliced. What is the other orange used for?
Answer: juice

3. What can you learn from this text?
a. how to be safe in the kitchen
b. how to cool hot food
c. how to cut fruit
d. how to make a dessert
Answer: d

Slide 22

Dwarf Lantern Shark
Are you afraid of sharks?

Some sharks are harmless. The Dwarf Lantern Shark cannot hurt you. You might think sharks are large but this one is not. It is so small you can hold it in one hand.

Another unusual thing about Dwarf Lantern Sharks is that they glow in the dark. They live at the bottom of very deep oceans. There is no light where they live. They make their own light.

- The text is about an unfamiliar topic which increases difficulty.
- It is short and uses simple familiar, ideas, and simple language which reduces difficulty.
- Consider these elements as well as the difficulty of the tasks.

Slide 23

Dwarf Lantern Shark
Are you afraid of sharks?

Some sharks are harmless. The Dwarf Lantern Shark cannot hurt you. You might think sharks are large but this one is not. It is so small you can hold it in one hand.

Another unusual thing about Dwarf Lantern Sharks is that they glow in the dark. They live at the bottom of very deep oceans. There is no light where they live. They make their own light.

1. Which part of the ocean do Dwarf Lantern Sharks live in?
Answer: at the bottom of very deep oceans / deep part

2. Why does the Dwarf Lantern Shark need to glow in the dark?
Answer: Because there is no light where they live

3. 'Some sharks are harmless.'
What does 'harmless' mean?
a. safe b. light
c. large d. dangerous
Answer: safe

Slide 24

	Alphabetistan	Vietnam	Philippines	Hanoi
Climate	arid to semi-arid; freezing winters and hot summers	tropical in south; monsoonal in north	usually hot and humid	subtropical in south; cool summers and severe winters in north
Geography	landlocked and mountainous	the fertile Mekong river delta covers a large part of south western Vietnam	made up of 7,107 islands	landlocked; contains eight of the world's 10 highest peaks
Main crops	wheat, fruits, nuts, wool, sheepskins	paddy rice, coffee, rubber, cotton; fish	sugarcane, coconuts, rice	rice, corn, wheat, sugarcane, milk
Typical exports (goods sold to other countries)	fruits and nuts, carpets, saffron	crude oil, marine products, rice, coffee, rubber, garments	electronic equipment, transport equipment, garments	carpets, clothing, leather goods
Wildlife	the Marco Polo sheep: it has the longest horns of any sheep	the saola (a kind of antelope): one of the world's rarest mammals	the Philippine Eagle: the largest eagle in the world	the one-horned rhinoceros: the world's fourth largest land mammal

Which country exports rice?

Slide 25

	Alphabetistan	Vietnam	Philippines	Hanoi
Climate	arid to semi-arid; freezing winters and hot summers	tropical in south; monsoonal in north	usually hot and humid	subtropical in south; cool summers and severe winters in north
Geography	landlocked and mountainous	the fertile Mekong river delta covers a large part of south western Vietnam	made up of 7,107 islands	landlocked; contains eight of the world's 10 highest peaks
Main crops	wheat, fruits, nuts, wool, sheepskins	paddy rice, coffee, rubber, cotton; fish	sugarcane, coconuts, rice	rice, corn, wheat, sugarcane, milk
Typical exports (goods sold to other countries)	fruits and nuts, carpets, saffron	crude oil, marine products, rice, coffee, rubber, garments	electronic equipment, transport equipment, garments	carpets, clothing, leather goods
Wildlife	the Marco Polo sheep: it has the longest horns of any sheep	the saola (a kind of antelope): one of the world's rarest mammals	the Philippine Eagle: the largest eagle in the world	the one-horned rhinoceros: the world's fourth largest land mammal

What do all the kinds of wildlife in the table have in common?
A. They are large. B. They are horned.
C. They are unusual. D. They are endangered.

Slide 26

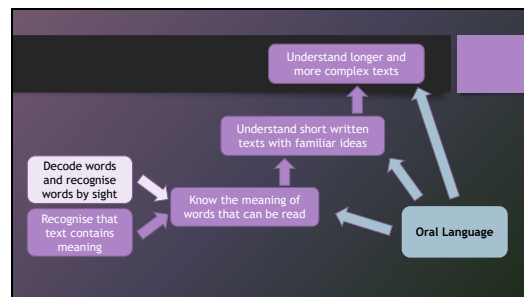
	Alphabetistan	Vietnam	Philippines	Hanoi
Climate	arid to semi-arid; freezing winters and hot summers	tropical in south; monsoonal in north	usually hot and humid	subtropical in south; cool summers and severe winters in north
Geography	landlocked and mountainous	the fertile Mekong river delta covers a large part of south western Vietnam	made up of 7,107 islands	landlocked; contains eight of the world's 10 highest peaks
Main crops	wheat, fruits, nuts, wool, sheepskins	paddy rice, coffee, rubber, cotton; fish	sugarcane, coconuts, rice	rice, corn, wheat, sugarcane, milk
Typical exports (goods sold to other countries)	fruits and nuts, carpets, saffron	crude oil, marine products, rice, coffee, rubber, garments	electronic equipment, transport equipment, garments	carpets, clothing, leather goods
Wildlife	the Marco Polo sheep: it has the longest horns of any sheep	the saola (a kind of antelope): one of the world's rarest mammals	the Philippine Eagle: the largest eagle in the world	the one-horned rhinoceros: the world's fourth largest land mammal

Maria says the typical exports show that Vietnam is the most successful country. Do you agree or disagree with Maria?
Circle one. Agree Disagree
Use evidence from the text to give a reason for your choice

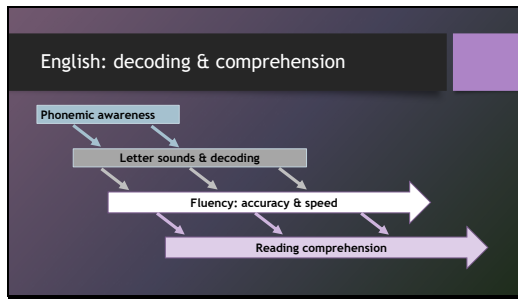
Slide 27

Factors that influence decoding difficulty

Slide 28



Slide 29



Slide 30

- Factors that affect English decoding difficulty
- Ease of differentiating phonemes (sounds /s/ and /m/ easier than /t/ and /p/)
 - Frequency & consistency of letter-sound relationship
 - Ease of letter shape recognition (e.g. b, p, d, q may be confused)
 - Phonetic simplicity of word
 - Word length
 - Frequency of exposure to word (likelihood it is recognised by sight)
 - Competing information (words with letters in common with target word)
 - Number of words to decode

Slide 31

Task 1: Point to and name the upper-case 'M'. Ask the student to find another letter with the same name. Indicate and name lower case 'd' and ask the student to find another letter with the same name.

Home Made Dinner

Task 2: What is the first sound in table? _____

Task 3: What is the last sound in dog? _____

Task 4: Tell me the names and sounds of these letters.

E n h R L i O s A p

Student names at least 8 letters correctly. _____

Task 5: Say these words.

is to me you for one mum was

Reads at least 4 words correctly.

Slide 32

Words or sentences easier?

Read aloud:

A cat sat on a mat.

This is my friend.

blaup maip thworp

Easier task - for decoding OR for sight words

Slightly harder but familiar sight words

Harder task - decoding requires knowledge of more complex phonics

- Nonsense words so not known by sight

Slide 33

Task 9: Ask the student to point to the word 'playing' (audio prompt).

The boy in the park was playing football.

Task 10: It is a hot day so Jana is wearing a hat.

What is Jana wearing?

A. a hat
B. a shirt
C. a top

What Skills are required?

Slide 34

- Factors that affect reading comprehension difficulty for very simple texts and tasks
- Prominence of information (e.g. in first sentence)
 - Similarity between key words in the question and the text
 - Proximity of information to matched words
 - Competing information (in the text and/or in multiple-choice options)
 - Extent to which knowing word meaning is required
 - Need to interpret the text e.g. infer meaning, make links across the text
 - Decoding factors:
 - Ease of word recognition e.g. simple decoding OR known sight words
 - Number & familiarity of words to be read

Slide 35

The student is asked to read the word themselves and select the matching picture.

Task 1

hat

Students only need to know:
The first letter-sound in the written word and the first phoneme in the spoken word.

Task 2

sock

pig

table

map

Slide 36

Task 3:

dog

class

dress

bread

Task 4:

Task 4: What is the woman doing?

A. cooking
B. playing
C. sleeping
D. reading

Task 5: What can you eat?

A. car
B. pen
C. egg
D. book

Slide 37

Joe rides his bike to the market.
Joe wants to sell some yams and buy some beans.

Task 6: What does Joe ride to the market?
A. his car
B. his bike
C. the bus

Task 7: What does Joe want to sell?
A. his bike
B. some yams
C. some beans

Word matching is sufficient.
Options are not in the text.

Word matching to adjacent text is sufficient.
Options are in the text.

Slide 38

Dana

Dana is making dinner. She needs some eggs.
She goes to the shops. Dana has eggs for dinner.

Task 13: Why does Dana go to the shops?
Answer: to get eggs / get things for dinner

Task 14: What does Dana do with the eggs?
A. eats them B. sells them C. Gives them away

Task 15: How do you think Dana feels at the end?
A. cross B. happy C. sad

Task 16: What was Dana like in this story?
A. lazy B. silly C. busy

Slide 39

Dana

Dana is making dinner. She needs some eggs.
She goes to the shops. Dana has eggs for dinner.

Task 8: What does Dana do with the eggs?
A. eats them B. sells them C. Gives them away

Task 9: How do you think Dana feels at the end?
A. cross B. happy C. sad

Task 10: What was Dana like in this story?
A. lazy B. silly C. busy

Slide 40

Practice examples

Slide 41

Factors contributing to reading difficulty

1. Number of features and conditions
2. Proximity of pieces of required information
3. Competing information (in task or text)
4. Prominence of necessary textual information
5. Relationship between task and required information
6. Semantic match between task and text
7. Concreteness of information
8. Familiarity of information needed to answer the question
9. Register of the text
10. Reference to information from outside the text

Slide 42

Factors that affect English decoding difficulty

1. Ease of distinguishing phonemes (spoken sounds)
2. Ease of distinguishing letter shapes
3. Consistency of letter-sound relationships
4. Phonetic simplicity of word
5. Word length
6. Frequency of exposure to word (likelihood it is recognised by sight)
7. Competing information (words with letters in common with target word)
8. Number of words to decode

Mathematics slides

Slide 1

Pairwise Comparison Method

Training for mathematics workshop

Slide 2

What is mathematics?

Slide 3

Global proficiency framework

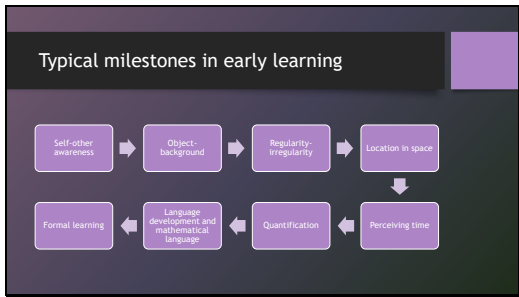
The global proficiency framework, developed by international experts, identifies five domains:

- Number and operations
- Measurement
- Geometry
- Statistics and probability
- Algebra

Slide 4

What is a learning progression?

Slide 5



Slide 6

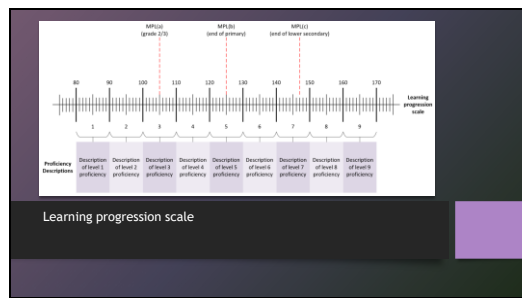
Locating and comparing tasks and learners

- We can imagine a continuum of mathematics development and there have been many attempts to systematically describe this (e.g., progress maps, reporting systems, learning progressions)
- Assessment aims to identify/compare the location of individuals on a growth continuum
- Assessment uses tasks appropriate to the expected location of the individuals assessed

Slide 7

The relationship between items and students on a proficiency scale
Reproduced from OECD, 2010, p. 46

Slide 8



Slide 9

What is the pairwise comparison method?

Slide 10


Two sets of items

- One set mapped to a learning progressions scale
- The other set from the [insert name of assessment]

Slide 11

Make comparisons between the items


- Consider pairs of items at a time
- Which of the pair of items is more difficult?



Slide 12

One combined set of items

- Once all the comparisons are made, we can map all of the new items on the same scale as the old items
- We can also determine the cut score on the [insert name of assessment] that equates to the minimum proficiency level



Slide 13

Advantages of the approach

- Comparing two items is easier than evaluating one against some criteria
- Multiple comparative judgements → rank order of items
- There is no limit on the type of items that can be compared - multiple choice, short response, open ended, interactive, etc.
- The process is generally fast and efficient, allowing for comparison of a large number of items
- The process is robust and reliable in generating links with Minimum Proficiency Levels

Slide 14

Factors that influence mathematics item difficulty

Slide 15

Variables

- Decoding information (text/image complexity)
- Devising strategies (mental manipulation and recall, mathematical content knowledge, logical connections)
- Solving
- Checking or interpreting the solution

Slide 16

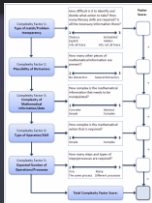
Item complexity schema

- Developed by the Adult Literacy and Lifeskills (ALL) Numeracy Expert Group
- Used in Programme for the International Assessment of Adult Competencies (PIAAC)
- Initially developed for adult learners but has been successfully applied more generally to learners across Primary and Secondary levels

Slide 17

Item complexity schema

- Type of match/problem transparency
- Plausibility of distractors
- Complexity of mathematical information/data
- Type of operation/skill
- Expected number of operations/processes



Slide 18

Item complexity schema

- Type of match/problem transparency
- Plausibility of distractors
- Complexity of mathematical information/data
- Type of operation/skill
- Expected number of operations/processes

Literacy aspects

Mathematical aspects

Slide 19

Complexity factor 1: Type of match/problem transparency


How difficult is it to identify and decide what action to take? How many literacy skills are required? Is all the necessary information there?





Score 1	Score 2	Score 3
<p>In the question and the stimulus, the information, activity or operation required:</p> <ul style="list-style-type: none"> is clearly apparent and explicit and all required information is provided and where minimal translation or interpretation is required is specified in little or no text, using simple, familiar and non-formal language/symbols, familiar objects and/or photographs or other clear, simple visualizations is about locating obvious information or relationships only closed question - not open ended 	<p>In the question and the stimulus, the information, activity or operation required:</p> <ul style="list-style-type: none"> is given using clear, simple sentences and representations including some formal language/ symbols and/or visualizations where some translation or interpretation is required is located within a number of sources within the text/activity may need to bring to the problem simple information or knowledge from outside the problem fairly closed question 	<p>In the question and the stimulus, the information, activity or operation required:</p> <ul style="list-style-type: none"> is embedded in text including more technical or formal language/ representations where considerable translation or interpretation is required may need to be derived or estimated from a number of sources within or outside the text/activity the information or action required is not explicit or specified or necessary information or knowledge is missing, so outside information or knowledge needs to be brought in more complex, open-ended task

Slide 20

Example item

Select the net that makes this cube.



A.  B.  C.  D. 

Slide 21

Complexity factor 2: Plausibility of distractors


How many other pieces of mathematical information are present?





Score 1	Score 2	Score 3
<ul style="list-style-type: none"> no other mathematical information is present apart from that requested no distractors 	<ul style="list-style-type: none"> there is some other mathematical information in the task that could be a distractor the mathematical information given or requested can occur in more than one place 	<ul style="list-style-type: none"> a range of other irrelevant information appears mathematical information given or requested appears in several places

Slide 22

Example item

Select the net that makes this cube.



A.  B.  C.  D. 

Slide 23

Complexity factor 3: Complexity of mathematical information/data


How complex is the mathematical information that needs to be manipulated?





Score 1	Score 2	Score 3
<p>Context</p> <ul style="list-style-type: none"> Based on common, real life activities <p>Text</p> <ul style="list-style-type: none"> A combination of clearly stated and clear, brief mathematical information, diagrams and visualizations related to the mathematics of length, area, volume, mass, etc. <p>Activities</p> <ul style="list-style-type: none"> Large whole numbers including addition, subtraction, multiplication, division, etc. <p>Tables and other representations</p> <ul style="list-style-type: none"> Tables with whole numbers and simple mathematical relationships and patterns <p>Measurement/units</p> <ul style="list-style-type: none"> Simple standard measures for length, weight, volume, etc. <p>Diagrams</p> <ul style="list-style-type: none"> Simple 2D shapes, nets, etc. <p>Formulas</p> <ul style="list-style-type: none"> Simple formulas for area, perimeter, etc. 	<p>Context</p> <ul style="list-style-type: none"> Based on real life activities, but less obvious <p>Text</p> <ul style="list-style-type: none"> A combination of stated and clear, brief mathematical information, diagrams and visualizations related to the mathematics of length, area, volume, mass, etc. <p>Activities</p> <ul style="list-style-type: none"> Large whole numbers including addition, subtraction, multiplication, division, etc. <p>Tables and other representations</p> <ul style="list-style-type: none"> Tables with whole numbers and simple mathematical relationships and patterns <p>Measurement/units</p> <ul style="list-style-type: none"> Simple standard measures for length, weight, volume, etc. <p>Diagrams</p> <ul style="list-style-type: none"> Simple 2D shapes, nets, etc. <p>Formulas</p> <ul style="list-style-type: none"> Simple formulas for area, perimeter, etc. 	<p>Context</p> <ul style="list-style-type: none"> Based on abstract or unfamiliar activities <p>Text</p> <ul style="list-style-type: none"> A combination of stated and clear, brief mathematical information, diagrams and visualizations related to the mathematics of length, area, volume, mass, etc. <p>Activities</p> <ul style="list-style-type: none"> Large whole numbers including addition, subtraction, multiplication, division, etc. <p>Tables and other representations</p> <ul style="list-style-type: none"> Tables with whole numbers and simple mathematical relationships and patterns <p>Measurement/units</p> <ul style="list-style-type: none"> Simple standard measures for length, weight, volume, etc. <p>Diagrams</p> <ul style="list-style-type: none"> Simple 2D shapes, nets, etc. <p>Formulas</p> <ul style="list-style-type: none"> Simple formulas for area, perimeter, etc.

Slide 24

Example item

Select the net that makes this cube.



A.  B.  C.  D. 

Slide 25

Complexity factor 3: Complexity of mathematical information/data

How complex is the mathematical information that needs to be manipulated?

Score 1	Score 2	Score 3
<p>Measures/dimension/space</p> <ul style="list-style-type: none"> everyday standard measures for length, weight, volume including common fraction and decimal units common 3D shapes and their representation via diagrams, nets or photos common types of maps or plans with visual scale indicators 	<p>Measures/dimension/space</p> <ul style="list-style-type: none"> everyday standard measures for length, weight, volume including common fraction and decimal units more complex 2D and 3D shapes, or a combination of 2 shapes, and their representation via diagrams, nets, including geometric properties area and volume formulae common types of maps or plans with ratio type scales 	<p>Measures/dimension/space</p> <ul style="list-style-type: none"> all kinds of measurement scales complex shapes or combinations of shapes

Slide 26


Complexity factor 4: Complexity of Type of operation/skill





How complex is the mathematical actions that is required?

Score 1	Score 2	Score 3
<p>Operations/skills</p> <ul style="list-style-type: none"> Simple mathematical operations (addition, subtraction, multiplication, division, etc.) <p>Text</p> <ul style="list-style-type: none"> Simple mathematical operations (addition, subtraction, multiplication, division, etc.) <p>Tables and other representations</p> <ul style="list-style-type: none"> Simple tables with whole numbers and simple mathematical relationships and patterns <p>Measurement/units</p> <ul style="list-style-type: none"> Simple standard measures for length, weight, volume, etc. <p>Diagrams</p> <ul style="list-style-type: none"> Simple 2D shapes, nets, etc. <p>Formulas</p> <ul style="list-style-type: none"> Simple formulas for area, perimeter, etc. 	<p>Operations/skills</p> <ul style="list-style-type: none"> Simple mathematical operations (addition, subtraction, multiplication, division, etc.) <p>Text</p> <ul style="list-style-type: none"> Simple mathematical operations (addition, subtraction, multiplication, division, etc.) <p>Tables and other representations</p> <ul style="list-style-type: none"> Simple tables with whole numbers and simple mathematical relationships and patterns <p>Measurement/units</p> <ul style="list-style-type: none"> Simple standard measures for length, weight, volume, etc. <p>Diagrams</p> <ul style="list-style-type: none"> Simple 2D shapes, nets, etc. <p>Formulas</p> <ul style="list-style-type: none"> Simple formulas for area, perimeter, etc. 	<p>Operations/skills</p> <ul style="list-style-type: none"> Complex mathematical operations (addition, subtraction, multiplication, division, etc.) <p>Text</p> <ul style="list-style-type: none"> Complex mathematical operations (addition, subtraction, multiplication, division, etc.) <p>Tables and other representations</p> <ul style="list-style-type: none"> Complex tables with whole numbers and simple mathematical relationships and patterns <p>Measurement/units</p> <ul style="list-style-type: none"> Complex standard measures for length, weight, volume, etc. <p>Diagrams</p> <ul style="list-style-type: none"> Complex 2D shapes, nets, etc. <p>Formulas</p> <ul style="list-style-type: none"> Complex formulas for area, perimeter, etc.

Slide 27

Example item

Select the net that makes this cube. 

A.  B.  C.  D. 

Slide 28

Complexity factor 4: Complexity of Type of operation/skill

How complex is the mathematical actions that is required?

Measure/shape properties	Score 1	Score 2	Score 3
<ul style="list-style-type: none"> - visualizing/representing, comparing and describing 2D and 3D shapes, objects or objects - knowing common straight forward measures and personal measures - naming, comparing common 2D shapes - comparing whole unit measurements 	<ul style="list-style-type: none"> - visualizing/representing, comparing and describing 2D and 3D shapes, objects or objects - geometric patterns or relationships, including simple nets - estimating, making and interpreting standard measurements using common measuring instruments and scales 	<ul style="list-style-type: none"> - using angle properties and symmetry to describe shapes or objects - transposing shapes (rotations/reflectors) - understanding relationship between length/area - estimating, making and interpreting non-standard measurements - converting between standard measurement units within the same system - interpolating values on scales 	<ul style="list-style-type: none"> - understanding more formal geometric representations and relationships e.g. parallel lines and angle relationships/properties - understanding relationships between area/volume - converting between non-standard measurement units within the same system - converting between measurements across different systems

Slide 29


Complexity factor 5: Expected number of operations/processes





How many steps and types of steps/processes are required?

Score 1	Score 2	Score 3
<ul style="list-style-type: none"> • one operation, action or process 	<ul style="list-style-type: none"> • application of two or three steps, the same or similar operation, action or process Note: repeating the same sequence of operations/processes only counts once 	<ul style="list-style-type: none"> • integration of several steps covering more than one different operation, action or process

Slide 30

Example item


Select the net that makes this cube. 

A.  B.  C.  D. 

Slide 31

Total complexity score

Total = 8



Slide 32

Which item is more difficult

1. Use your own expertise and experience
2. Item complexity schema
3. Curriculum / learning progression consideration

Slide 33

Practice examples

Annex J – Post-workshop analysis

Pairwise judgements can be downloaded from the software once completed. Then, appropriate data cleaning and preparation activities can be undertaken. The format and structure of the resulting dataset will depend on the requirements of the IRT software being used for the pairwise comparison analysis. The structure typically includes details of each judgement, including judge ID, item pairs, and actual judgement (i.e., which item is more difficult).

This dataset can then be shared with the psychometrician who is undertaking the IRT analyses. This analysis will have to cover the following processes.

1. Scaling of the pairwise judgments
2. Evaluation of the pairwise robustness and relationship between ordering of items on the assessment original scale and the LPS
3. Statistical linking of assessment and LPSs and placement of MPL benchmarks on the assessment scale
4. Calculation of the proportion of learners achieving meeting each of the relevant MPLs

Adjusting Correlations for Reliability

1. Scaling of the pairwise judgments

Undertake a free calibration of the pairwise comparison model using the Bradley-Terry-Luce (BTL) Model (see Pollitt, 2012). This will involve:

- Estimating parameters for each item included in the model such that item difficulty is placed on the pairwise comparison scale
- Producing infit and outfit statistics for each item
- Producing infit and outfit statistics for each judge
- Removal of misfitting items and judges and rerunning of the pairwise comparison analysis

2. Evaluation of the pairwise robustness and reliability

The IRT software that you have selected to use should produce a reliability estimate equivalent to the KR20 index. The standard interpretation of this index with a value of 0.75 indicates the sufficiently reliability of item difficulty parameters and values of 0.90, indicating a high level of scaling outcome reliability.

Given that item parameters used in the pairwise correlation linking relate to the latent construct, it is necessary to adjust the correlation between the original assessment parameters and those obtained in the comparative judgment exercise. The first step is to calculate the empirical reliability of the IRT parameters and then use these indices to adjust that correlation. Operationally, this means calculating a dis-attenuated correlation between the original and comparative judgment item difficulty parameters.

IRT empirical reliability index

Reliability indices will need to be calculated for the source item IRT item difficulty parameters and those obtained from the comparative judgment exercises. These two sets of items are considered to be equivalent.

The reliability here is defined as the ratio of item difficulty true variance to the observed item difficulty variance under the true-score model (Lord & Novick, 1968), which can be expressed as empirical reliability (BTL model item parameters):

$$reliability_X = \frac{VAR^{\wedge}(\theta^{\wedge})}{VAR^{\wedge}(\theta^{\wedge}) + SE^{\wedge}(\theta^{\wedge})^2}$$

Where θ is item difficulty and SE item parameter standard error. $SE^{\wedge}(\theta^{\wedge})^2$ is obtained by averaging across the N standard errors of the respective θ^{\wedge}_i terms and squaring.

When calculating empirical reliability for the original assessment, item parameters for items included in the comparative judgement should be used if not all items from the original tests are included in the exercise. Similarly, only BTL model parameters for items from the initial assessment should be used when calculating the reliability of the comparative judgment outcomes.

The dis-attenuated correlation can be calculated using the following formula:

$$R_{xy} = \frac{r_{XY}}{\sqrt{reliability_X * reliability_Y}}$$

Where r_{XY} is the Person correlation coefficient between the original item difficulty parameters from the original assessment scale and those from the comparative judgement scale. Reliability X and Y are empirical reliability estimates for these scales, respectively.

The dis-attenuated reliability of at least 0.75 should be required to move to the statical linking of MPL parameters using the assessment items included in the pairwise exercise.

3. Statistical Linking

The following stages should be followed when conducting the statistical linking:

- Anchored calibration: The BTL model should be run where the pre-selected items that are aligned to the LPS and their parameters are used to anchor the calibration of the rest of the items used in the comparative judgment exercise.
- Differential item functioning (DIF) analyses: The purpose of this analysis is to identify items from the original assessment that might show evidence of differential ordering in the comparative judgment exercise and extract a set of well-functioning items to conduct the statistical linking between learning progression and the original assessment scales. The suggested procedure is the Robust z procedure (Huynh & Meyer, 2010). This procedure applies only to the items from the assessment being linked to the LPS.
- Placing MPL benchmark on the assessment scale for the assessment being linked: The mean equation where a constant adjustment to item parameters is based on the mean difference between item parameters anchored on the LPS and those from the assessment being linked to the LPS. The purpose of the exercise is to calculate the linking adjustment needed to place the MPL benchmark from the LPS on the original assessment scale. To calculate the linking error, see the description of the mean-mean linking procedure in Kolen and Brennan (2004).

4. Calculation of the proportion of learners achieving meeting each of the relevant MPLs

The data from the last relevant testing cycle for the assessment being linked to the LPS should then be reanalysed to calculate the percentage of students at and above the relevant MPL. The procedure will depend on the student ability estimates used in the original assessment reporting. The confidence intervals of these estimates should be calculated using the same methodology used during the original assessment reporting.

Annex K – Workshop evaluation form

Stage I: Training

This evaluation should take place following training and prior to the start of the pairwise comparison exercise.

Please read the following statements carefully and place a mark in that category indicating your level of agreement.

Statement	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I understand the purpose of the Learning Progression Scale (LPS)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
The descriptions in the LPS were clear and easy to understand	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
The practical exercise using the LPS was useful to improve my understanding	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I understand the process I need to follow to complete the PCM	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I understand how my judgements will contribute to the overall MPL benchmark	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I feel confident in using the system to record my judgements	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I felt able to ask questions during the training to clarify my understanding	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Please use the space below to record any comments you would like to make about the training. In particular, please provide any suggestions for improvements to the training for the future that would have aided your understanding.

Stage 2: Judgements

This evaluation should take place once all comparative judgements have been made.

Please read the following statements carefully and place a mark in that category indicating your level of agreement.

Statement	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I feel confident in the decisions I made during the exercise	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Where items were similar in difficulty, I used the training to support my decision	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I had sufficient time to make my judgements between training and the deadline	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
The system for recording my judgements was straightforward to use	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Please use the space below to record any comments you would like to make about the exercise. In particular, please provide any suggestions for improvements to the system for the future that would have improved your experience.

Annex L – Certification of appreciation template



Annex M – Self-assessment template report (PCM outcomes)

Assessment Instrument	[Insert name of instrument]
Country	[Insert country where assessment instrument is administered]
SDG 4.1.1 level	End of lower primary / End of primary / End of lower secondary [delete as appropriate]
Subject	Mathematics / Reading [delete as appropriate]
Date of self-assessment	[Insert date on which self-assessment was undertaken]

Criterion Number	Criterion	Delete as appropriate
1	Did all participants meet the requirements for participation?	Yes / No
2	Were the group of participants sufficiently representative in terms of the characteristics agreed by the country?	Yes / No
3	Were SMEs removed from analyses if their responses did not fit the model well?	Yes / No
4	Were items/SME participants considered for removal from analyses if they did not fit the model well, and was there a clear rationale for the ultimate decision?	Yes / No
5	Is the pairwise scale reliability index equal to or higher than 0.75?	Yes / No
6	Were items removed from analyses if they exhibited item DIF?	Yes / No
7	For the items from the assessment being linked, is the dis-attenuated correlation between the items original scale location and LPSs' location equal to or higher than 0.75?	Yes / No
8	Was the average (mean) score for each section of the evaluation greater than or equal to 4?	Yes / No
9	Did the impact analysis workshop confirm the validity of the statistical linking exercises?	Yes / No

Overall self-assessment rating

Did the PCM Workshop meet all 9 Self-Assessment Criteria?	Yes / No [delete as appropriate]
-----------------------------------------------------------	-------------------------------------

Annex N – Pairwise Comparison Method Report

In order to submit evidence for reporting against SDG 4.1.1, the project team must produce a report on the process and outcomes of the PCM. The following headings may be helpful in developing such a report.

1. Executive Summary
2. Overview to the Assessment
 - a. Introduction
 - b. Purpose of the Assessment
 - c. Design of the Assessment
 - d. Sampling and Test Administration
 - e. Scoring
3. Self-Assessment Results (appropriateness of assessment)
 - a. Criterion 1: Alignment
 - b. Criterion 2: Item Review
 - c. Criterion 3: Sample
 - d. Criterion 4: Administration
 - e. Criterion 5: Reliability
4. Preparation for the Pairwise Comparison Method
 - a. Item selection
 - b. Logistics
 - c. Selection and Description of Subject Matter Expert participants
 - d. Construction of item pairs and assignment to Subject Matter Experts
 - e. Technology check issues and resolutions
5. Implementation of the Pairwise Comparison Method
 - a. Training
 - b. Making judgments
 - c. Plenary session
6. Outcomes of the Pairwise Comparison Method
 - a. Analysis of the outcomes
 - b. Results
 - c. Precision, Accuracy and Consistency
7. Evaluation of Pairwise Comparison Method
 - a. Training
 - b. Post judgements
8. Self-Assessment Results (PCM outcomes)
9. Conclusions and Recommendations
10. References
11. Annexes
 - a. Forms used during the process
 - b. Other Relevant Documents and Data