unesco
Institute for Statistics

GLOBAL ALLIANCE TO MONITOR LEARNING

POLICY LINKING FOR MEASURING GLOBAL LEARNING OUTCOMES TOOLKIT

Linking Assessments to the Global Proficiency Framework

JANUARY 2023

UKaid
from the British people

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

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 specialized 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.

Published in 2023 by:
UNESCO Institute for Statistics
C.P 250 Succursale H
Montréal, Québec H3G 2K8
Canada

Email: uis.tcg@unesco.org
http://www.uis.unesco.org

Acknowledgements

The Policy-Linking methodology is a UNESCO Institute for Statistics (UIS) collaborative project. The Policy-Linking toolkit has been developed in partnership with the World Bank Group and Great Britain. The ideas and views expressed in this publication are those of the authors; they are not necessarily those of the agencies involved in developing the toolkit.

CONTENTS

TABLES

FIGURES

ACRONYMS

ACER	Australian Council for Educational Research
AE	Above Exceeds Minimum Proficiency
CBA	Curriculum-Based Assessments
COR	Contracting Officer's Representative
E3/ED	Bureau for Economic Growth, Education and Environment
EGMA	Early Grade Math Assessment
EGRA	Early Grade Reading Assessment
E_j	Exceeds Minimum Proficiency
FCDO	Foreign, Commonwealth and Development Office
GAML	Global Alliance to Monitor Learning
GPD	Global Proficiency Descriptor
GPF	Global Proficiency Framework
GPL	Global Proficiency Level
GRN	Global Reading Network
ICAN	International Common Assessment of Numeracy
JE	Just Exceeds Minimum Proficiency
JM	Just Meets Minimum Proficiency
JP	Just Partially Meets Minimum Proficiency
M_j	Meets Minimum Proficiency
MPL	Minimum Proficiency Level
MSI	Management Systems International
NAEP	National Assessment of Educational Progress
NFER	National Foundation for Educational Research
PAL	People's Action for Learning
PLT	Policy Linking Toolkit
PM_j	Partially Meets Minimum Proficiency
SDG	Sustainable Development Goal
SE	Standard Error
SEND	Special Educational Needs and Disabilities
TEP Centre	The Education Partnership Centre
UIS	UNESCO Institute for Statistics
UNESCO	United Nations Educational, Scientific and Cultural Organization
SME	Subject Matter Expert
USAID	U.S. Agency for International Development
USG	United States Government

GLOSSARY OF TERMS

Angoff method – A benchmark setting method in which panelists rate items by GPL and then average all panelists' ratings for each GPL to create a benchmark.

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

Breadth of Alignment – Sufficient coverage of the domains, constructs, and subconstructs in the 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.

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

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

Distractor – A set of plausible but incorrect answers 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.

Global Proficiency Level (GPL) – The four levels of proficiency or performance (below partially meets global minimum proficiency, partially meets global minimum proficiency, meets global minimum proficiency, and exceeds global minimum proficiency) that students can achieve for all targeted grade levels and subject areas. The meets global minimum proficiency level aligns with SDG 4.1.1, and the others allow countries to show progress toward all students meeting or exceeding that level.

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

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

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 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.

Normative information – The distribution of benchmarks set by panelists, with each panelist's location indicated by a code letter or number known only to them.

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.

Item difficulty statistics – Information on the empirical difficulty of items (i.e., percentage of learners getting an item correct), which gives panelists a rough idea of how their judgments about items compare to actual learner performance.

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

Stem – The question part of a multiple-choice item on an assessment.

Test-centered method – A family of benchmark-setting methods that make judgments based on a review of assessment material and scoring rubrics; the Angoff method is included in this category.

Timed assessment – In this toolkit, the term 'Timed Assessment' is used to describe an assessment that involves at least one item/task that requires learners to undertake an activity in a set time limit, with their score related to how much of the activity they complete in that time. For example, an assessment of oral language fluency where learners are asked to read as many words as possible in one minute.

Untimed assessment – In this toolkit, the term 'Untimed Assessment' is used to describe an assessment where there are no assessment items/tasks that require learners to undertake an activity in a set time limit, with their score related to how much of the activity they complete in that time. These will often be curriculum-based assessment (CBA) and they may have an overall time limit for the assessment, but not for individual items.

ACKNOWLEDGMENTS

This original draft of this toolkit was developed following workshops sponsored by the Office of Education in the Bureau for Economic Growth, Education and Environment (E3/ED) of the United States Agency for International Development (USAID) and the UNESCO Institute for Statistics (UIS). USAID and UIS – as well as other agencies including the World Bank Group (WBG), the U.K. Foreign, Commonwealth and Development Office (FCDO) (formerly the U.K. Department for International Development), and the Bill & Melinda Gates Foundation – have been extremely supportive of introducing and exploring policy linking as a method for comparing and aggregating results from learner assessments within and across countries.

The project team would like to thank Benjamin Sylla for his leadership on the development of the original toolkit as the USAID Contracting Officer's Representative (COR) of the Reading and Access Evaluation Project, as well as Dr. Saima Malik, Rebecca Rhodes, and Dr. Elena Walls of USAID E3/ED for their direction and guidance throughout the process of developing the original draft. Silvia Montoya, UIS Director, has been instrumental in providing organizational support. Jennifer Gerst of the Global Reading Network (GRN) played a key role in hosting workshops. We are highly appreciative of all contributions.

The project team would also like to thank the authors of the original toolkit: Dr. Abdullah Ferdous, Sean Kelly, and Dr. Jeff Davis of Management Systems International (MSI), who had support from Melissa Chiappetta (an independent contractor working with UIS, USAID, and the Bill & Melinda Gates Foundation who has also been helpful through her leadership of the Policy Linking Working Group); Norma Evans of Evans and Associates; Colin Watson of the U.K. Department for Education; and Carlos Fierros (NORC), Nathalie Liautaud and Ryan Aghabozorg (MSI).

Finally, the team would like to thank all participants in the processes of developing, piloting, and revising the toolkit and materials, with special thanks to the Ministries of Education and/or the specialized agency or department for administering national learning assessments in Bangladesh, Cambodia, Djibouti, Gambia, Ghana, India, Kenya, Lesotho, Morocco, Nepal, Nigeria, Rwanda, and Senegal, these the pilots were run by UIS, USAID and WBG. Also, to the People's Action for Learning (PAL) Network, The Education Partnership (TEP Centre), and Zizi Afrique, which supported a pilot of the International Common Assessment for Numeracy (ICAN). There has been substantial worldwide participation in policy linking activities, which we trust will continue in the future.

Second phase of piloting was in India, Cambodia, Nepal, Lesotho and Zambia are technically supported by CITO under UIS contract and coordinated by Melchior de Vries. Silvia Montoya and Shailendra Sigdel had coordinated from UIS to complete the second phase piloting.

Current version of the toolkit has been updated by ACER, under UIS contract based on the feedback from number of pilots, independent evaluations conducted by CITO and National Foundation for Educational Research (NFER), and lessons learnt activities with groups of experts involved in implementing policy linking across the world. ACER has a strong interest in supporting the use of a range of tools for setting benchmarks on national and international assessments.

CHAPTER I

CHAPTER I. INTRODUCTION TO POLICY LINKING

A. RATIONALE FOR POLICY LINKING

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.

As a result, this toolkit describes a non-statistical, judgmental method called policy linking for measuring global learning outcomes (policy linking for short), which has also been referred to as social moderation.[1] The UNESCO Institute for Statistics (UIS) has included policy linking in its list of acceptable methodologies for reporting on Sustainable Development Goal (SDG) 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 in (i) reading and (ii) mathematics, by sex.

Other donor organizations – including USAID, FCDO, the World Bank Group, the Bill & Melinda Gates Foundation, and UNICEF – have demonstrated interest in using or supporting the use of policy linking for setting benchmarks on national and international assessments, which would facilitate reporting on key global indicators related to reading and mathematics and also make it possible for countries to set learning targets for long-term improvement of learning outcomes.[2,3] Along with UIS, these agencies formed a working group to develop the policy linking method. An earlier version of this toolkit was used to pilot the policy linking method in three countries from October 2019 to March 2020, after which point it was revised – with contributions from the working group and from an independent evaluation organization (the National Foundation for Educational Research [NFER]) – for this current version. The current version of the toolkit was updated following further pilots in five countries between March 2021 and May 2022 and recommendations from the independent evaluation. The NFER evaluation of the method, funded by the Bill & Melinda Gates Foundation, is ongoing and will continue to inform future changes to the method.

---

[1] The policy linking approach was proposed in September 2017 at a meeting of the Global Alliance to Monitor Learning (GAML) and then again in August 2018 at a global workshop organized by USAID. In February 2019, USAID published a paper on policy linking, with technical support from Management Systems International (MSI). A group of 30 international subject matter experts (SMEs) produced the first Global Proficiency Framework (GPF) in April and May 2019 covering Grades 2 through 6. The first draft of the policy linking toolkit was produced in September 2019 to guide pilots. Another draft of the GPF was produced by an expanded group of SMEs in October 2020, concurrently with a revised version of the toolkit. The second draft GPF added Grade 1 and Grades 7 through 9. The current draft of the toolkit was produced in December 2022.

[2] The Bill & Melinda Gates Foundation commissioned an evaluation in 2019 aimed at empirically evaluating the acceptability of policy linking as a method for linking assessment results to SDG 4.1.1. The recommendations from pilot observations in the evaluation have been incorporated into this revision, though the full evaluation, including a validation study, remains outstanding. The foundation's support of the method is conditional on the results of this evaluation.

[3] A benchmark is a numeric threshold on an assessment that indicates a learner has met a proficiency level.

This toolkit was designed for policy linking using the Global Proficiency Framework (GPF) (available on Edulinks and UIS' website), which is described in detail below. The GPF is composed of internationally agreed upon expectations of the knowledge and/or skills minimally proficient learners should have (these statements of knowledge and/or skill(s) are sometimes called content standards) and how much of that they should be able to demonstrate (referred to in the GPF as global proficiency descriptors, sometimes called performance standards) that form a common scale for global reporting on learner outcomes in reading and mathematics in grades 1–9.[4,5] However, while the toolkit was developed to assist countries and regional and international assessment organizations with setting benchmarks for global reporting, it can also be used to set national benchmarks for national reporting on existing assessments. A country government may choose to set national and global benchmarks for the same assessment, and those benchmarks could be the same if the national frameworks are aligned with the GPF and the benchmarks are set using the same approach. However, some countries may choose to maintain their own national standards, separate from the global standards outlined in the GPF. Countries may do this for reasons such as choosing to teach knowledge and skills at different grade levels than those represented in the GPF or because they wish for their national standards to incorporate additional knowledge and skills not captured in the GPF. In such cases, countries might choose to set separate benchmarks for national reporting and global reporting.

This toolkit sets out essential standards for the quality of the assessment to be used for policy linking and the outcomes of the policy linking method. If these standards are not met, then the results will not be accepted for reporting against SDG 4.1.1. Where standards are not met (for example, where the assessment is not sufficiently aligned to the GPF or the sampling design does not produce outcomes appropriately representative of the desired population), there may still be a rationale for implementing the policy linking method. The toolkit will make clear the benefits of continuing with certain activities from a capacity building perspective.

B. AUDIENCE

This toolkit was created for use by country governments and assessment agencies (for multinational assessments) and their partners. 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.

C. OVERVIEW OF THE GLOBAL PROFICIENCY FRAMEWORK

The GPF was created to respond to the call set up by the Global Education Monitoring Report (GEMR), 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 policy linking method described in this toolkit requires this common set of global proficiency descriptors (sometimes called performance standards) by grade level and subject area to which countries can link their assessments for producing indicators and reporting to UIS. Using a standardized benchmarking approach, results from different countries and assessments that are linked to the GPF standards for their grade and subject can then be compared, aggregated, and tracked. For instance, all end of primary reading assessments can be linked to the grade 5 reading GPF, which then allows countries to produce the SDG 4.1.1(b) indicators based on the proficiency descriptors for comparing, aggregating, and tracking outcomes from those end of primary reading assessments.

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 children and youth to achieve, no matter where in the world they live.

---

[4] Authors have purposefully not used the term "content standards" in the GPF because 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.

[5] Authors have purposefully not used the term "performance standards" in the GPF because 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.

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 GPF defines, for primary school reading and mathematics, the global minimum proficiency level that learners are expected to demonstrate at the end of each grade (one through nine). 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) described in the GPF based on their knowledge of developmental progressions and the UIS's Global Content Framework, which was based on 73 curriculum and assessment frameworks from 25 countries for reading and 115 assessment frameworks from 53 countries for mathematics.[6,7] The GPF was also reviewed and informed by language experts and those with expertise in working with students with disabilities, and education in crisis and conflict affected areas. 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 existing country content standards and curricula and did not set higher expectations for learners.

An example from part of the grade three mathematics GPF is shown in **Table 1**. It has the domains, constructs, subconstructs, statements of knowledge and/or skills, and the GPDs for the top three out of four performance categories, called Global Proficiency Levels (GPLs). Note the lowest performance category, Below Meets Global Minimum Proficiency, does not need GPDs since it includes all learners who do not meet the expectations described in the Partially Meets Global Minimum Proficiency level.

Table 1: Grade 2 Mathematics Example from the GPF

Domain	Construct	Subconstruct	Knowledge or Skill (Content Standards)	Global Minimum Proficiency Levels and Descriptors (Performance Standards)		
				Partially Meets Global Minimum Proficiency	Meets Global Minimum Proficiency	Exceeds Global Minimum Proficiency
Number and operations	Whole numbers	Identify and count in whole numbers, and identify their relative magnitude	Count, read, and write whole numbers	Count in whole numbers up to 30.	Count in whole numbers up to 100.	Count backwards from 20.
				Read and write whole numbers up to 30 in words and in numerals.	Read and write whole numbers up to 100 in words and numerals.	N/A
			Compare and order whole numbers	Compare and order whole numbers up to 30.	Compare and order whole numbers up to 100.	N/A
			Skip count forwards or backwards	N/A	Skip count forwards by twos or tens.	Skip count backwards by tens.
		Represent whole numbers in equivalent ways	Determine or identify the equivalency between whole numbers represented as objects, pictures, and numerals	Identify and represent the equivalence between whole quantities up to 10 represented as objects, pictures, and numerals *(e.g., when given a picture of 10 objects and other pictures of various numbers of objects, select the picture that has the same number of objects; associate a numeral with the appropriate number of objects).*	Identify and represent the equivalence between whole quantities up to 30 represented as objects, pictures, and numerals *(e.g., when given a picture of 30 flowers, identify the picture that has the number of butterflies that would be needed for each flower to have a butterfly; given a picture of 19 shapes, draw 19 more shapes).*	Use place-value concepts for tens and ones *(e.g., compose or decompose a two-digit whole number using a number sentence such as 35 = 3 tens and 5 ones, 35 =30 + 5, or using number bonds; determine the value of a digit in the tens and ones place).*

[6] See the previous footnote for a chronology of the development of the GPF.

[7] See UNESCO (2018a, 2018b) in the references for its global content frameworks for reading and mathematics. Note that these frameworks are not by grade level and do not have descriptors by global proficiency level (GPL).

As **Table 1** shows, in order to define the content for each grade and subject, the GPF is organized hierarchically, i.e., from general to specific, with domains, constructs, and subconstructs. The statements of knowledge and/or skill(s) associated with the subconstructs demonstrate what learners need to know and be able to do by grade and subject.

Expanding on the subcontracts, there are the GPDs, which describe how much of the content in the knowledge and skills learners need to demonstrate to be considered minimally proficient. Each of the GPLs is characterized by a definition – called a policy definition – that applies across grades and subjects. The four definitions – for the four performance categories, or GPLs – are provided below and also included in **Annex B**:

- **Below Partially Meets Global Minimum Proficiency:** Learners lack the basic knowledge and skills for their grade. As a result, they cannot complete the most basic tasks appropriate for their grade.
- **Partially Meets Global Minimum Proficiency:** Learners have partial knowledge and skills for their grade. As a result, they can partially complete basic tasks appropriate for their grade.
- **Meets Global Minimum Proficiency:** Learners have sufficient knowledge and skills for their grade. As a result, they can successfully complete basic tasks appropriate for their grade.
- **Exceeds Global Minimum Proficiency:** Learners have superior knowledge and skills for their grade. As a result, they can successfully complete complex tasks appropriate for their grade.

The Policy Linking Working Group developed the four levels through extensive consultation with national and international stakeholders. They are intended to allow countries to track and report progress over time, with the goal of an increasing percentage of learners moving from Below Partially Meets Global Minimum Proficiency to Partially Meets Global Minimum Proficiency and eventually Meets Global Minimum Proficiency or even Exceeds Global Minimum Proficiency.

The GPDs define what is expected of learners in the last three GPLs (there is no need for GPDs for the Below Partially Meets Global Minimum Proficiency level, as all learners who do not meet the benchmark for Partially Meets Global Minimum Proficiency will fall into this category) for grades one to nine in reading and mathematics. They describe how much content learners need to know and be able to do in relation to the statements of knowledge and/or skill(s) required by grade and subject. For example, in reading, the GPF says that a learner who meets global minimum proficiency in grade three should be able to identify the general topic in a grade three-level continuous text when the topic is prominent but not explicitly stated. In mathematics, a learner who meets global minimum proficiency in grade three should be able to compare and order whole numbers up to 1,000.

Note that policy linking is designed for use with the four GPLs. This provides information for reporting on some donor indicators, such as USAID's Foreign Assistance ("F") Indicators[8]. However, a country government/assessment agency can elect to use only the appropriate Meets GPL, which is sufficient for reporting on SDG 4.1.1 (see **Figure 1** for some criteria countries and assessment organizations may consider when deciding how many benchmarks they should set). However, as mentioned, setting benchmarks for the top three levels is encouraged, as it will allow countries and partners to better demonstrate progress over time toward meeting the requirements of SDG 4.1.1, though this does require an assessment of sufficient length (see the alignment criterion in **Chapter II**). Countries or partners reporting on USAID indicators will need to set benchmarks for the top three performance levels, since some of the "F" indicators measure improvement from one performance level to another.

[8] USAID's F indicators are worded as 'learners attain minimum grade-level proficiency in reading/mathematics' and are therefore related to the standards in the GPF.

Figure 1: Setting One versus Three Benchmarks

Three benchmarks are recommended because:

- They better facilitate tracking progress toward achieving the goals of SDG 4.1.1
- They are consistent with requirements for reporting against USAID Foreign Assistance Indicators
- They allow countries to better identify gaps in learning and target those in the most need

However, only one benchmark is necessary for reporting against SDG 4.1.1. It may make sense for countries/assessment agencies to set one benchmark if:

- Their assessments are short and unlikely to have a wide enough range in scores to facilitate multiple unique benchmarks
- They are not partnering with USAID
- They have other national assessment standards for which they also wish to set benchmarks for tracking need with their country

D. THE GLOBAL PROFICIENCY FRAMEWORK AND THE MINIMUM PROFICIENCY LEVELS

UIS developed minimum proficiency levels (MPLs) to support the reporting of SDG 4.1.1. These describe the basic knowledge in a domain, as measured through learning assessments, at a) end of lower primary (grades 2/3), b) end of primary and c) end of lower secondary. These 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.

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 math. 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 math. 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 math. The alignment with the GPF for reading and mathematics is closest to the 'Meets Global Minimum Proficiency' descriptions for Grade 8.

For reporting against SDG 4.1.1 (a), (b), and (c), countries will need to use the GPF at grades 2, 5, and 8 respectively for policy linking, regardless of the grade of the learners taking the assessment. This is to ensure that there is comparability across countries in the standard expected for SDG 4.1.1 reporting. Countries may choose to implement policy linking with the grade of the GPF that matches their assessment (for example, using the GPF at grade 4 for an assessment administered to grade 4 learners), but this would not be accepted by UIS for SDG 4.1.1 reporting.

E. OVERVIEW OF POLICY LINKING

Policy linking is a method that allows countries/assessment agencies to link their assessments to SDG 4.1.1 and determine the benchmarks on those assessments for meeting global minimum proficiency.[9] It brings together lead facilitators, content facilitators, panelists, and government official/assessment agency observers to complete this process. The roles and qualifications of each of these groups is presented in Chapter III.

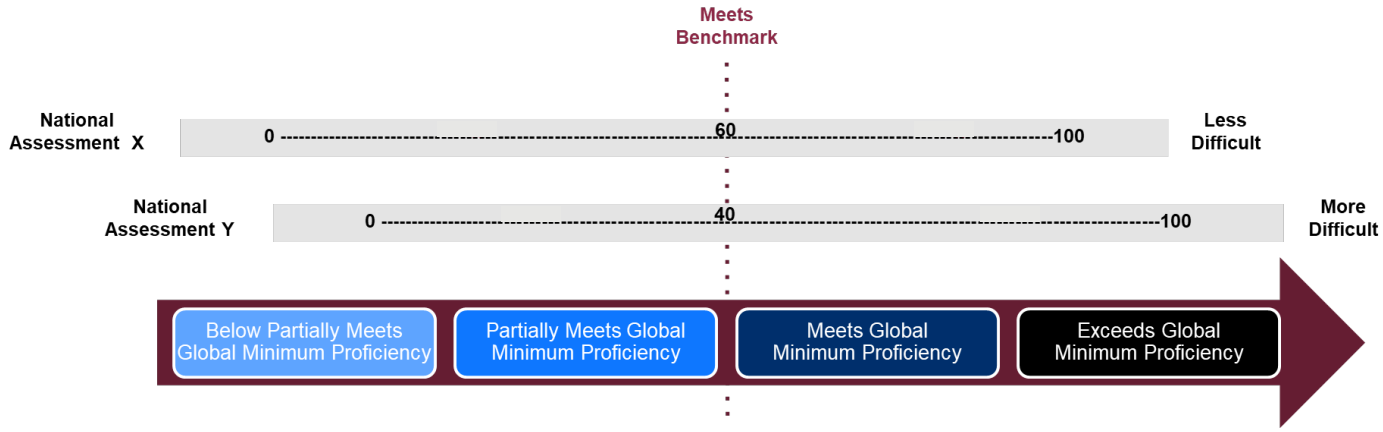
Since the GPF is used as a reference – or common criteria – for policy linking, these benchmarks represent the same standard of performance on those different assessments as defined by the GPDs, regardless of the difficulty or language

[9] The benchmarks on an assessment determine whether a learner is classified in a performance category or level; they are also known as cut scores, cut points, thresholds, or boundaries.

of the assessments, even though the benchmarks are set at different places (numeric scores) on the different assessments (unless the assessments are of equivalent difficulty).

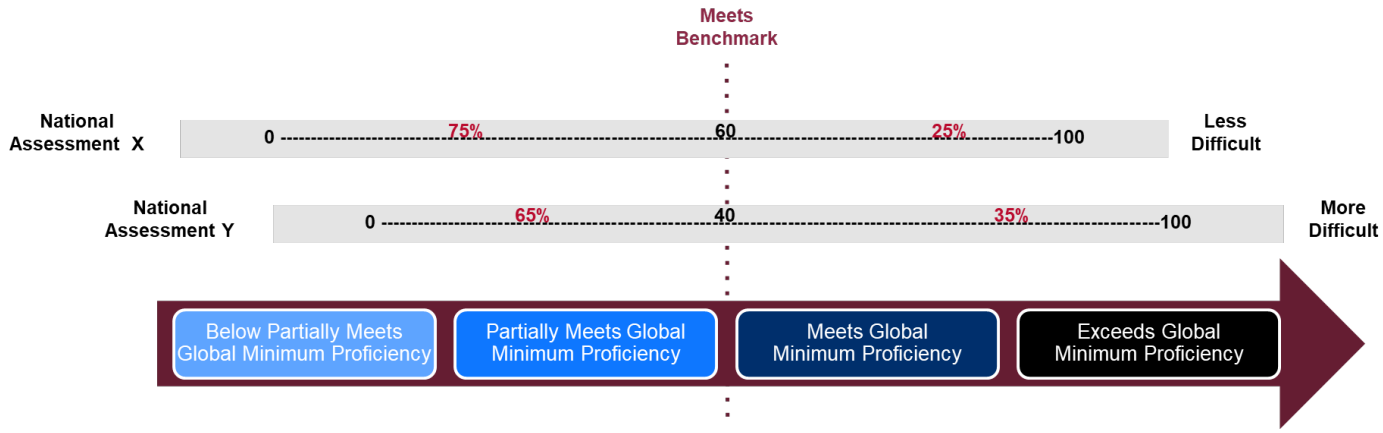
For instance, as **Figure 2** shows, two different assessments will most likely have different benchmarks for Meets Global Minimum Proficiency due to the unequal difficulty of those assessments. At a given grade and subject, less difficult assessments will have higher benchmarks and more difficult assessments will have lower benchmarks. In this example, Country X and Country Y have national assessments with scales of 0 (minimum) to 100 (maximum) points. They link their assessments to the GPF. National Assessment X – which is less difficult – has a Meets Global Minimum Proficiency benchmark of 60 points while National Assessment Y – which is more difficult – has a Meets Global Minimum Proficiency benchmark of 40 points. In theory, a learner with an ability level of just meeting global minimum proficiency who takes the two assessments would score 60 points on the less difficult assessment and 40 points on the more difficult assessment. As seen in the diagram below, the assessments vary in difficulty, but the GPF common scale remains constant, so benchmarks linked to the GPF are equivalent.

Figure 2: Example of Comparable Benchmarks on Various Assessments



By setting the benchmarks on different assessments based on the same descriptors in the GPF, the assessments are linked by their equivalent benchmarks, i.e., the benchmarks on each assessment that correspond to meeting global minimum proficiency. In this example, as shown in **Figure 3**, the process has determined that 25% of students in Country X and 35% of students in Country Y have met the minimum proficiency level.

Figure 3: Example of Comparable Benchmarks on Various Assessments



To set the benchmarks, policy linking uses an internationally recognized, standardized, test-centered, revised Angoff benchmarking procedure. The Angoff procedure requires groups of national SMEs, called panelists, to make judgments on the items used in the national assessments. The panelists include master teachers and curriculum experts from the

country (countries in the case of multinational assessments) who understand the performance of learners for specific grades and subjects. They follow the Angoff procedure to 1) examine the country/assessment agency's assessment instrument(s) in relation to the GPDs and 2) estimate how learners in each of the GPL categories would perform on the assessment. Planners and facilitators organize and conduct separate workshops by grade, subject, and language with different groups of panelists to set the equivalent benchmarks for those assessments.

F. POLICY LINKING STAGES

There are six stages to policy linking for measuring global learning outcomes that must be completed to facilitate global reporting, as shown in **Table 2**. Countries/assessment agencies and their partners must complete each of these stages for their results to be accepted for reporting against SDG 4.1.1 and USAID "F" indicators. This toolkit covers Stages 2, 3, 4 and 5. **Table 2** provides information on resources available to support the other stages.

Table 2: Policy Linking Stages

#	Policy Linking Stages	Purpose	Roles/Responsibilities	Resources (available on UIS website)
1	Initial engagement	For countries (or assessment agencies in coordination with relevant country governments) to determine whether, for a specific assessment, policy linking is the preferred approach, either at a national or regional/state level, 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, UIS, donors, and/or policy linking contractors. 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[10]
2	Self-assessment of appropriateness of assessment for policy linking	To determine whether assessment reliability, validity, and alignment with the GPF meet requirements for proceeding with policy linking for global reporting; and to determine the number of benchmarks to be set on the assessment depending on its length. (Where an assessment is not deemed appropriate for global reporting, activities will be proposed for capacity building)	Country governments/ assessment agencies with/without support of partners and UIS-approved independent observer	• Policy Linking Toolkit (**Chapter II**) • Self-assessment template report (**Annex C**)
3	Preparation for the policy linking workshop	To identify/confirm facilitators (if not done), invite panelists, prepare materials, and secure a venue	Country governments/ assessment agencies with/without support of partners	• Policy Linking Toolkit (**Chapter IV**) • Workshop Preparation Checklist (**Annex D**)

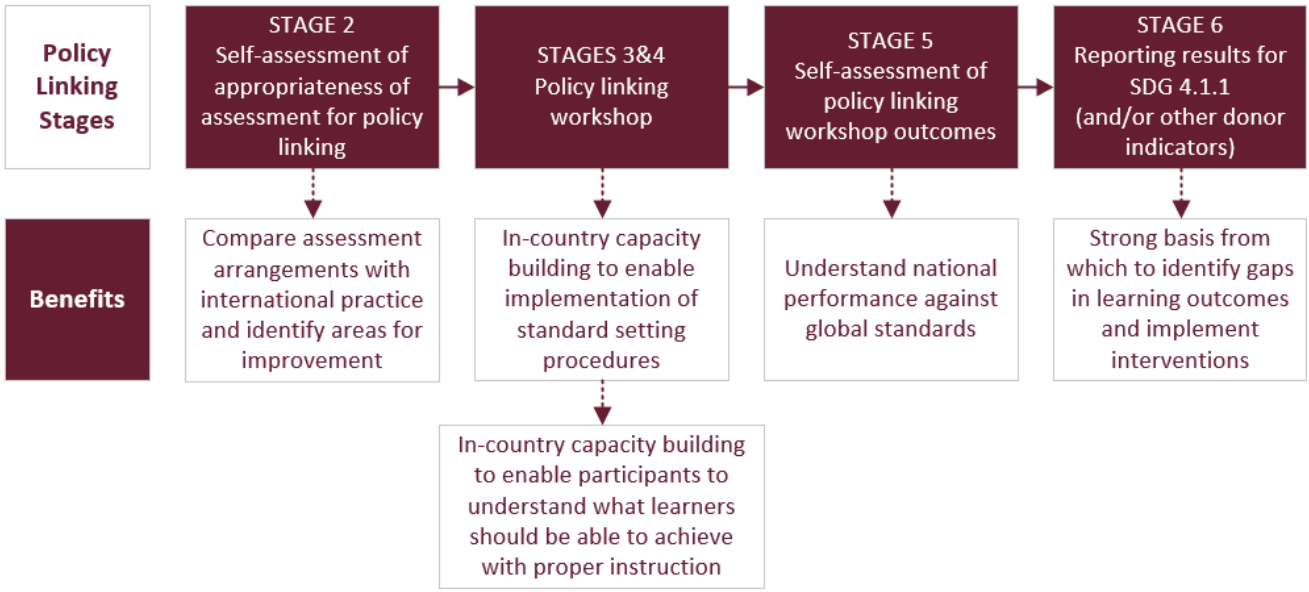
[10] https://tcg.uis.unesco.org/wp-content/uploads/sites/4/2022/08/Countrys-reporting-option-_Zambia_AAEA.Final_.pdf

#	Policy Linking Stages	Purpose	Roles/Responsibilities	Resources (available on UIS website)
4	Implementation of policy linking workshop (consisting of three steps - alignment, matching, and benchmarking)	To set benchmarks and document details regarding the process followed	Country governments/ assessment agencies with/without support of partners	• Policy Linking Toolkit (**Chapter V**)
5	Self-assessment of policy linking workshop outcomes	To determine whether workshop reliability and validity meet with criteria for global reporting	Country governments/ assessment agencies with/without support of partners and UIS-approved independent observer	• Policy Linking Toolkit (**Chapter VI**) • Self-assessment template report (**Annex V**)
6	Reporting results for SDG 4.1.1 (and/or other donor indicators)	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 (March 2022)[11] • Individual donor guidelines

G. USES AND BENEFITS OF POLICY LINKING

While the primary purpose of policy linking for measuring global learning outcomes is to link local, national, regional, and international assessments to global indicators, there are additional benefits of the process. For instance, as shown in **Figure 4**, the country government/assessment agency and its partners will self-assess their assessment against the quality criteria in this toolkit. This information might help inform improvements in country education systems. Finally, the results of the policy linking workshop should help countries identify the percentage and profile (assuming the country/assessment agency has collected demographic information on the assessment population) of learners in their country not meeting global minimum proficiency standards. Some countries use this information to conduct studies into why those gaps exist and how they might best address those.

Figure 4: Policy Linking Process and Benefits



[11] https://tcg.uis.unesco.org/wp-content/uploads/sites/4/2022/03/Protocol-for-Reporting-SDG-4.1.1.pdf

H. USING THE POLICY LINKING TOOLKIT

This policy linking toolkit is designed for project teams, most specifically workshop facilitators, and resource persons – i.e., government officials, assessment agency officers, donor representatives, and partners – who will be organizing, funding, and/or implementing the method in their country or region.[12] It has guidelines for implementing the method.

Chapter II describes the self-assessment process to enable project teams to confirm that the assessment is appropriate for policy linking. **Chapter III** includes details on the policy linking methodology. **Chapter IV** presents guidance on how to prepare for a policy linking workshop, including how to select facilitators and participants, what invitations should look like, what logistics need to be coordinated, what materials to prepare and how to prepare them, and how to train the content facilitators on leading sections of the workshop. **Chapter V** provides step-by-step guidance on how to implement a policy linking workshop. Finally, **Chapter VI** presents key considerations for documenting the outcomes of the policy linking workshop and presents details on the self-assessment process to enable project teams to confirm that the workshop outcomes are sufficiently valid.

The bibliography contains references on policy linking, benchmarking, and other psychometric issues. It includes the *Policy Linking Justification Paper* (2019), which provides background on the policy linking method, support for the method by international donors, and information on the importance of the method for measuring reading and mathematics outcomes globally.[13]

The annexes provide all the materials and forms needed for applying the policy linking procedures outlined in the toolkit. This includes, among other things, a sample workshop agenda, facilitation slide templates, alignment and item rating forms, a workshop evaluation template, formulas for calculating benchmarks and statistics, and an outline for a technical report.

Although all those involved in the implementation of policy linking should have a broad understanding of the end-to-end process, **Table 3** contains navigation guides that provide details of the most relevant sections for each of the roles of lead facilitator, content facilitator and data analyst that are described in **Chapter IV**, **section A**. To note, the lead facilitator, content facilitator and data analyst may be part of the project team, but this will depend on how the country wishes to organize themselves to deliver the project.

Table 3: Navigation Guides

Role	Most Relevant Main Sections	Most Relevant Annexes
Project Team	All	All
Lead Facilitator	**Chapter III** and **Chapter V**	**Annex G, Annex H, Annex I, Annex P, Annex R,** and **Annex S**
Content Facilitator	**Chapter III** and **Chapter V**	**Annex H, Annex I, Annex P, Annex R,** and **Annex S**
Data Analyst	**Chapter III**	**Annex G, Annex J, Annex N, Annex O,** and **Annex T**

[12] Ideally, the government's assessment, examination, or evaluation would use this toolkit and training to carry out the policy linking process with its own resources and expertise. However, in instances in which the government is not organizing the policy linking process independently, the responsible organization and project team must work closely with the government in planning and implementing the policy linking process to ensure buy-in and capacity building for future workshops.

[13] Management Systems International (2019). *Policy linking method: Linking assessments to a global standard*. U.S. Agency for International Development (USAID), Washington, D.C.

CHAPTER II

CHAPTER II. SELF-ASSESSMENT OF THE APPROPRIATENESS OF THE ASSESSMENT

A. 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, raw data and sampling design, 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. It will also include the score distribution to enable the development of impact data (see **Annex N**).

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

Countries should consider how the language of the assessment may affect the outcomes of policy linking where this is not the first language of the learners. It is likely that, in this case, the achievement of learners will be lower than it would have been had they been assessed in their first language. This does not affect the ability to carry out a successful policy linking workshop but will be reflected in reporting for SDG 4.1.1.

B. CRITERIA FOR POLICY LINKING 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. If an assessment does not meet these criteria, they may choose to continue with a policy linking workshop, but the results will not be accepted by UIS for global reporting.

For each criterion, there are essential minimum requirements for an assessment to be self-assessed as sufficiently valid for SDG reporting. In some cases, there are also desirable requirements, including descriptions of what 'good' and 'excellent' assessments would look like to support capacity building. Although these requirements are desired, the minimum requirements are the only ones essential for SDG reporting.

The five criteria relate to the following:

- **Criterion 1** – is the assessment sufficiently aligned to the GPF?
- **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 reporting?
- **Criterion 4** – is there evidence that the assessment was administered in a standardized way?
- **Criterion 5** – are the outcomes of the assessment sufficiently reliable?

The project team should record the outcomes of their self-assessment using the form in **Annex C**.

Criterion 1 – Alignment

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

- The assessment instrument
- The assessment framework or specification

- The curriculum framework.

There are three categories of alignment, though all are sufficient for SDG 4.1.1 reporting:

- **Minimal alignment** – The content of the assessment aligns with the minimum number of reading/mathematics skills in the GPF to be suitable for reporting against SDG 4.1.1, though the reporting will be qualified with a note to the level of alignment.
- **Additional alignment** – The content of the assessment aligns with more than the minimum number of reading/mathematics skills in the GPF to be suitable for reporting against SDG 4.1.1 but does not meet the requirements for strong alignment and will be qualified as such.
- **Strong alignment** – The content of the assessment aligns strongly with the reading/mathematics skills in the GPF and is, therefore, suitable for unqualified reporting against SDG 4.1.1.

MINIMUM REQUIREMENTS

To report against SDG 4.1.1, the assessment must be sufficiently aligned to the GPF. In addition to a minimum total score on the assessment, countries must determine whether there is sufficient depth (number of items that have at least a partial fit with at least one statement of knowledge and/or skill(s) from the GPF) and breadth (coverage of GPF domains, constructs, and subconstructs by at least one item with a partial fit) of alignment. This is determined using an alignment exercise. This exercise is repeated by panelists during the policy linking workshop, but it is the exercise conducted during self-assessment that takes precedence in determining the suitability of the assessment for reporting. Additional examples to support the alignment process are provided in **Chapter III**.

The project team, including appropriate content experts, will use the Frisbie alignment method described herein to complete the following three sub-steps using the same Alignment Rating Form that will be used by panelists during the workshop, which can be found in **Annex H**.

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 (using 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 Fit** (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 completely use the knowledge and/or skill(s) described in the statement.
- **Partial Fit** (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 Fit** (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 do not use knowledge and/or skill(s) described in the GPF.

Once all items have been rated, the overall alignment to the GPF is determined. The criteria for mathematics are presented in **Table 4** and those for reading are presented in **Table 5**. When summarizing results to the subconstruct level the project team should only consider the subconstructs with knowledge and/or skill(s) expected at the grade level for which alignment is being conducted (Grade 2 for SDG 4.1.1(a), Grade 5 for SDG 4.1.1(b) and Grade 8 for SDG 4.1.1(c)), i.e., those that have an "x" listed under the appropriate grade level column in GPF Table 3.

Table 4: Mathematics Assessment Alignment Criteria for Grades 1–9

Level of Alignment	Category	SDG 4.1.1 (a) GPF grade 2	SDG 4.1.1 (b) GPF grade 5	SDG 4.1.1 (c) GPF grade 8
Minimally Aligned	Test length	Minimum total score of 20 if setting only 'meets' level Minimum total score of 45 if setting 'partially meets', 'meets', and 'exceeds' levels		
	Domain (depth):	N (minimum 10 score-points)		
	Subconstructs (breadth):	Score-points covering at least 2 of the 4 N subconstructs	Score-points covering at least 5 of the 10 N subconstructs	Score-points covering at least 4 of the 8 N subconstructs
Additionally Aligned	Test length	Minimum total score of 20 if setting only 'meets' level Minimum total score of 45 if setting 'partially meets', 'meets', and 'exceeds' levels		
	Domain (depth):	N (minimum 10 score-points) and M and G (minimum 5 score-points)		
	Subconstructs (breadth):	Score-points covering at least 6 of the 11 N, M, and G subconstructs	Score-points covering at least 9 of the 17 N, M, and G subconstructs	Score-points covering at least 7 of the 14 N, M, and G subconstructs
Strongly Aligned	Test length	Minimum total score of 20 if setting only 'meets' level Minimum total score of 45 if setting 'partially meets', 'meets', and 'exceeds' levels		
	Domain (depth):	N (minimum 10 score-points) and M and G (minimum 5 score-points) and S and A (minimum 2 score-points)	N (minimum 10 score-points) and M and G (minimum 5 score-points) and S and A (minimum 5 score-points)	
	Subconstructs (breadth):	Score-points covering at least 7 of all 14 subconstructs	Score-points covering at least 12 of all 21 subconstructs	Score-points covering at least 12 of all 21 subconstructs

Key: N – Number and operations
M – Measurement
G – Geometry
S – Statistics and Probability
A – Algebra

Table 5: Reading Assessment Alignment Criteria for Grades 1–9

Level of Alignment	Category	SDG 4.1.1 (a) GPF grade 2	SDG 4.1.1 (b) GPF grade 5	SDG 4.1.1 (c) GPF grade 8
Minimally Aligned	Test length	Minimum total score of 20 if setting only 'meets' level Minimum total score of 45 if setting 'partially meets', 'meets', and 'exceeds' levels		
	Domain (depth):	D (minimum 10 score-points) C (minimum 5 score-points)	R (minimum 10 score-points)	R (minimum 20 score-points)
	Subconstructs (breadth):	Score-points covering at least 4 of the 7 D&C subconstructs	Score-points covering at least 4 of the 8 R subconstructs	Score-points covering at least 5 of the 10 R subconstructs
Additionally Aligned	Test length	Minimum total score of 20 if setting only 'meets' level Minimum total score of 45 if setting 'partially meets', 'meets', and 'exceeds' levels		
	Domain (depth):	D (minimum 10 score-points) R (minimum 5 score-points)	N/A	B1 (minimum 5 score-points) B2 (minimum 5 score-points)
	Subconstructs (breadth):	Score-points covering at least 3 of the 6 D&R subconstructs	N/A	Score-points covering at least 5 of the 10 R subconstructs
Strongly Aligned	Test length	Minimum total score of 20 if setting only 'meets' level Minimum total score of 45 if setting 'partially meets', 'meets', and 'exceeds' levels		
	Domain (depth):	R (minimum ten score-points)	B1 (minimum 5 score-points) B2 (minimum 5 score-points)	B1 (minimum 5 score-points) B2 (minimum 5 score-points) B3 (minimum 5 score-points)
	Subconstructs (breadth):	Score-points covering at least 1 of the 2 R subconstructs	Score-points covering at least 4 of the 8 R subconstructs	Score-points covering at least 5 of the 10 R subconstructs

Key: D – Decoding
C – Comprehension of spoken or signed language
R – Reading comprehension
B1 – Retrieve information
B2 – Interpret information
B3 – Reflect on information

DESIRABLE REQUIREMENTS

For any high-quality assessment, it is essential that there is a clear link between what is in the curriculum, what actually is taught, and what is assessed. If assessments do not align with the curriculum and what is taught in the classroom, students are unlikely to perform as well on the assessment.

Good Rating

To self-assess as 'good', the assessment must be at least additionally aligned (assessments that are minimally aligned cannot receive a 'good' rating) and there must be clear evidence that the assessment instrument meets the quantitative requirements of the specification or test blueprint in the assessment framework.

Excellent Rating

To self-assess as 'excellent', the assessment must be strongly aligned and there must be clear evidence that the assessment and curriculum frameworks are aligned.

Countries should have a curriculum framework[14] that includes details on domains, constructs, subconstructs, and skills[15] that are expected to be taught in classrooms by grade. Descriptors should be detailed enough to make it clear what should be taught. A similar alignment activity to that carried out with the GPF can be conducted between the country's curriculum framework and assessment instrument to determine if these are sufficiently aligned.

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 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).

[14] Where a country has a highly decentralized/federalized system in relation to curriculum arrangements, they should consider a small number of examples of local curricula that are considered broadly representative.

[15] It is not expected that all countries will make use of these terms within their curriculum frameworks, but rather that there is an attempt to detail the topics they expect to be taught.

DESIRABLE REQUIREMENTS

For any high-quality assessment, it is important to review the items being considered for inclusion in the assessment to ensure they are performing appropriately, including for subgroups of the population (for example, learners of different genders, those with special educational needs and disabilities (SEND), those of different ethnic or cultural backgrounds, those affected by crisis and conflict, those from rural and urban areas, and those living in poverty.)

Good Rating

There are no additional expectations for this criterion to support a good rating.

Excellent Rating

To self-assess as 'excellent', there must be evidence that items have been reviewed to ensure appropriateness for relevant subgroups of the population. This evidence will be qualitative and quantitative.

The assessment should appear free from bias. Items should not ask questions about foreign concepts or concepts familiar to only some cultural, ethnic, ability, socioeconomic, gender, or geographic groups. For instance, reading comprehension passages that discuss holidays that may be celebrated only by some groups or that discuss snow in countries where it only snows in certain parts of the country would be inappropriate. There should be evidence that experts have reviewed materials for such biases and removed them.

Countries should also demonstrate what, if any, test adaptations they have made for students with SEND.

Countries should also use appropriate statistical techniques to investigate differential item functioning for different subgroups of the population, taking action as appropriate.

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, countries should understand 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 sample of learners who took the assessment is representative of the population against which the results will be reported.[16]

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.

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

[16] It is accepted that for some countries, defining what 'nationally representative' means may be difficult given a lack of accurate sampling frame. In such cases, governments should consider how they have attempted to achieve an appropriate sample and identify any known limitations with their approach.

DESIRABLE REQUIREMENTS

For any high-quality assessment where a sampling approach is used, the sample size should be determined to be sufficiently powered to allow countries to capture changes in outcomes over time with appropriate confidence.

Good Rating

There are no additional expectations for this criterion to support a good rating.

Excellent Rating

To self-assess as 'excellent', the minimum detectable effect size should have been calculated and thought through ahead of finalizing sample size calculations to ensure that differences over time are detectable.

Criterion 4 - Administration

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

- Administration guidance materials
- Any reports on 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 are consistent, or length of time to administer the assessment is adhered to). Administration guides must be reviewed for clarity and any incidents of inappropriate administration 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.

DESIRABLE REQUIREMENTS

There are no additional desirable requirements for this criterion.

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[17] (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. This could include scoring of items with a pre-agreed score or double scoring of a sample of responses.

DESIRABLE REQUIREMENTS

For any high-quality assessment, it is essential that the assessment results have appropriate levels of reliability, meaning that if the test were given again to another sample of students with a different set of enumerators or test proctors,

[17] Also known as Cronbach's alpha

results would be similar. Assessment results that fluctuate significantly based on who is administering the assessment, who scored the assessment, or which specific students in a sample take the assessment are not reliable assessments.

Good Rating

To self-assess as 'good', the level of agreement between human-scorers and pre-agreed scores, or double-marked scores, should be over 80%.

Excellent Rating

To self-assess as 'excellent', the country should report other measures of reliability on the assessment, for example, classification constancy, classification accuracy or inter-rater reliability, with levels that are consistent with international best practice.

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.

To self-assess as 'good', countries must additionally self-assess at the 'good' level for both criteria 1 and 5.

To self-assess as 'excellent', countries must self-assess at the 'excellent' level for all five criteria.

C. NEXT STEPS

If countries self-assess as meeting the minimum requirements (or as 'good' or 'excellent'), they may continue to conduct a policy linking workshop. The outcomes of the workshop will be eligible for reporting against SDG 4.1.1 as long as the workshop 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 VI**.

Countries that self-assess as not meeting the minimum requirements may choose to continue with a policy linking workshop for capacity building purposes, but the results will not be accepted by UIS for reporting against SDG 4.1.1. Alternatively, they may wish to consider whether it is possible to amend their current assessment arrangements to make them compatible with policy linking method. How this is achieved will depend on which of the criteria they did not meet.

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

If they did not meet criterion 2 on item review, they may wish to review their test development arrangements to ensure they align with international practice. For example, they 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 they do not meet criterion 3 on sampling, they 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 they do not meet criterion 4 on administration, they may wish to review their administration guidance or provide more training to test administrators to ensure consistency.

If they do not meet criterion 5, they 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.

CHAPTER III

CHAPTER III. THE POLICY LINKING METHOD

Once the assessment has been confirmed as appropriate for policy linking, the project team need to confirm which items will be used in the workshop. For assessments where the same instrument is administered to all students, the project team may decide to use all items. If the assessment is very long, has multiple forms or a complex sampling design, however, the project team will need to determine which items they will use based on the days available for the workshop and other criteria.

The Policy Linking Method begins with a thorough review of the main documents that provide the foundation for the workshop – the GPF and the assessment(s) being linked to the GPF and to SDG 4.1.1. Following this familiarization, facilitators lead panelists through three major tasks:

- **Task 1** – Check the content alignment between the assessment(s) and the GPF using a standardized procedure
- **Task 2** – Match the assessment items with the GPF, i.e., the GPLs and GPDs
- **Task 3** – Set three global benchmarks for each assessment using a standardized method (a modified version of the Angoff methodology) through two rounds of ratings[18]

Each of these activities is described in detail below in this chapter.

A. ITEM SELECTION

To make the policy linking workshop manageable, it is recommended that for untimed assessments, a maximum of 45 score-points are used in the process (which will mean 45 items if each item is worth one score-point, though fewer if there are polytomous items). As stated in the self-assessment section, this is sufficient to set the three GPLs ('meets', 'partially meets' and 'exceeds') for a single grade level. Fewer items may be selected if only the 'meets' GPL is being set through policy linking, though there must be a minimum of 20 score-points. For timed assessments, it will be possible to include a greater number of score-points depending on the nature of the task (for example, the number of score-points maybe related to the number of words read in a set time). The project team should consider whether there will be sufficient time in the workshop for the panelists to make all the required judgements when determining the number of items to use in a timed assessment.

If the assessment is longer than 45 score-points (for an untimed assessment), or a complex sampling design is used, then the project team will need to select which items are to be used. If the assessment is administered in multiple languages, items selected should be in the language in which the workshop will be delivered. Links to other languages should be made statistically, where possible, or by conducting separate workshops in each language of administration.

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 policy linking should also be aligned to 'retrieve information'.
- **GPL alignment** – the items selected should broadly reflect the different GPLs being set in the policy linking workshop. For example, if all three GPLs are being set in the workshop, then there should be 15 items aligned to each of the GPLs as identified through the self-assessed alignment task. If only the 'meets' GPL is being set, then the 20 items should mainly be aligned to the 'meets' GPL in the GPF, though some 'partial' and 'exceeds' items may also be included.

[18] Note that if during Stage 1, 2, or 3, the government decides that it only wishes to set a benchmark for the meets level or the government/assessment agency or 4.1.1 Review Panel decides the assessment is too short to accommodate three benchmarks at the three main GPLs, then panelists need only set one benchmark (rather than three) for each assessment.

- **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 are exhibiting differential item functioning towards a particular gender, these should be removed.
- **Items classified as 'no fit' during alignment exercise** – the items selected should ideally all have at least partial fit with the GPF. As the self-assessed alignment exercise has already determined that the assessment is sufficiently aligned to the GPF to support policy linking, it is helpful to remove non-aligned items to avoid needing to deal with them in the workshop. For example, where possible items assessing grammar, punctuation and spelling should be removed from the test/item pool for the workshop. The level of alignment to the GPF for reporting will use the outcome of the self-assessed alignment exercise that includes the items classified as 'no fit'.
- **Item difficulty (only applicable where IRT is used)** – 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 policy linking might not all be contained in a single existing test form. In contrast, the curated items set will utilize all the assessment items to provide the best possible discrimination and coverage of knowledge and skills by items included in the policy linking workshop.

B. FAMILIARIZATION

To successfully undertake the policy linking workshop, it is vital that panelists are familiar with the GPF and the assessment instrument. This familiarization can take place in advance of, or at the start of, the workshop depending on time availability.

For the GPF, familiarization should start by explaining the necessary terminology (GPL, GPD, domain, construct, subconstruct) and how the GPF is structured. Panelists should then review the GPF for the relevant grade (Grade 2 for SDG 4.1.1(a), Grade 5 for SDG 4.1.1(b) and Grade 8 for SDG 4.1.1(c)) to enable them to internalize the standards. Where possible, panelists should engage in active tasks to support their understanding of the GPF rather than just reading it.

For the assessment instrument, familiarization should explain how the assessment is structured and administered. Where possible, panelists should be given an opportunity to administer the assessment, or at least to take it themselves.

C. TASK 1 – ALIGNING THE ASSESSMENT TO THE GPF

Alignment is the process of checking if a certain item in an assessment used for policy linking corresponds with a subconstruct/statement of knowledge and/or skills in the GPF. It is important to distinguish the alignment activity in Task 1 from the alignment work conducted by the government/assessment agency as part of the self-assessment. The pre-workshop alignment exercise is intended to ensure there is sufficient alignment between the country/assessment agency's assessment and GPF to proceed with policy linking. In contrast, during the workshop the alignment activity is focused on further familiarizing the panelists with the GPF, in particular the knowledge and skills covered in it, and generating panelist ratings on the depth and breadth of the alignment between the assessments and the GPF. This understanding will aid panelists with the benchmarking process that occurs in Task 3, as it is the first step in narrowing in on which GPF expectations the assessment(s) measures. There are two steps in Task 1:

1. Panelists rate alignment between assessment being linked and the GPF
2. The workshop facilitators and data analyst summarize results of the alignment activity (roles and responsibilities are described in more detail in below)

Step 1 – Panelist Alignment Exercise

In Step 1, after being given instructions on the task and then working through some examples with the facilitators, panelists should work independently, going item-by-item using the Frisbie alignment method described herein to complete the following three sub-steps using the Alignment Rating Form, which can be found in **Annex H**.

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 (using 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 Fit (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 completely use the knowledge and/or skill(s) described in the statement.
- **Partial Fit (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 Fit (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 do not use knowledge and/or skill(s) described in the GPF.

Further details on the scale appear in **Figure 5** below.

Figure 5: Alignment Scale and Number of Statements of Knowledge and/or Skill(s) to Which an Item Aligns

If an item has a rating of **Complete Fit (C)** with a particular statement of knowledge and/or skill(s), the panelists should not match it with other statements of knowledge and/or skill(s), meaning it is aligned to only one statement in the GPF.

If an item has a rating of **Partial Fit (P)** with a particular statement of knowledge and/or skill(s), the panelists should generally match it to one or two other statements of knowledge and/or skill(s) in the GPF.

If an item has a rating of **No Fit (N)** with any statements of knowledge and/or skill(s), the panelists should not match it to any statements of knowledge and/or skill(s).

An example of a "complete fit" item follows in **Figure 6** with an items which asks a learner how eight hundred and seventy is written in standard form. In this example, the panelist identified that the knowledge or skill needed to answer this item correctly is the ability to read and write whole numbers up to 1,000. This skill is covered in the GPF under the "number knowledge" domain, "whole number" construct, and "identify and count in whole numbers" subconstruct. Finally, the panelist rated this alignment as a "complete fit" since all of the knowledge and/or skill(s) needed to correctly answer this item are contained in this single statement of knowledge and/or skill(s).

Figure 6: Example Alignment of an Item to the GPF with Complete Fit

1. How is eight hundred and seventy written in standard form?

A. 807

B. 870

C. 817

D. 871

Domain: Number and Operations

Construct: Whole Numbers

Subconstruct: Identify, count in, and identify the relative magnitude of whole numbers

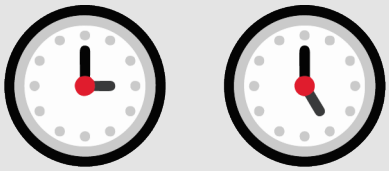
Knowledge or skill (content standard): Count, read, and write in whole numbers

Fit: To answer this item correctly, the learner needs to be able to identify and count whole numbers. Therefore, the item can be rated as "complete fit" with the statement of knowledge and/or skill(s) since it only requires the knowledge or skills from that single statement.

An example of a "partial fit" item follows in **Figure 7**. The panelist rated this item as a partial fit since to answer this item correctly, a learner would need knowledge or skills described by two different statements of knowledge and/or skill(s).

Figure 7: Example Alignment of an Item to the GPF with Partial Fit

What is the difference in time shown between these two clocks?



Domain: Measurement

Construct: Time

Subconstruct: Tell time AND solve problems involving time

Knowledge or skill (content standard): Tell time AND solve problems involving time

Fit: Partial fit since it requires the knowledge and/or skill(s) from two content standards.

An example of a 'no fit' item follows in **Figure 8**. The panelist rated this item as no fit since to answer the item correctly, a learner would need knowledge or skills that are no described in the GPF. Ideally, items with no fit to the GPF would have been removed in the item selection stage. If this is not possible, then instructions will be given for how to deal with non-fitting items in later stages of the process.

Figure 8: Example of Alignment of an item to the GPF with No Fit

Which of these sentences is punctuated correctly as a question?

A. Where is the cat!

B. Where is the cat.

C. Where is the cat?

D. Where is the cat:

Domain: Not applicable

Construct: Not applicable

Subconstruct: Not applicable

Knowledge or skill (content standard): Not applicable

Fit: No fit since the knowledge required to answer the item relates to punctuation which is not referenced in the GPF.

Step 2 – Facilitator Summary of Results

Once all panelists have completed their alignment task, the facilitators should summarize the results by taking an average of the number of items that the panelists aligned to each domain, construct, and subconstruct. Even though alignment occurs at the knowledge and/or skill level, the criteria for alignment are at the subconstruct level. As such, facilitators need to summarize results up to the subconstruct level. Both complete and partial fit items count toward alignment, but each item should only be counted once even if it is a partial fit (note: for items that have a partial fit, for summary purposes, facilitators should count the domain, construct, and subconstruct that they feel best describes the most important of the knowledge and/or skill(s) needed to answer the item correctly – this could be determined by using the outcomes of the original alignment exercise undertaken as part of the self-assessment process). As the purpose of the alignment activity is to further familiarize panelists with the GPF and the assessment, it is not essential for panelists

to agree either with each other or with the original alignment exercise at this stage. Assuming the original alignment exercise indicated sufficient alignment, the outcome of this exercise should provide additional confidence that policy linking is appropriate.

An example of summary results for a grade 3 assessment with 26 items appears below. To note, since policy linking for this assessment would be linked to grade 2 in the GPF for SDG 4.1.1(a) reporting, the domains, constructs and subconstructs shown relate only to those included in grade 2 (as shown in table 3 of the GPF).

Table 6: Example of Summary Alignment Results for a Grade 3 Assessment

Domain		Items
N	Number and operations	14
M	Measurement	7
G	Geometry	3
S	Statistics and probability	2
A	Algebra	0
Total		**26**

Construct		Items
N1	Whole numbers	14
M1	Length, weight, capacity, volume, area, and perimeter	3
M2	Time	4
M3	Currency	0
G1	Properties of shapes and figures	2
G2	Spatial visualizations	0
G3	Position and direction	1
S1	Data management	2
A1	Patterns	0
A3	Relations and functions	0
Total		**26**

Subconstruct		Items
N1.1	Identify and count in whole numbers, and identify their relative magnitude	4
N1.2	Represent whole numbers in equivalent ways	0
N1.3	Solve operations using whole numbers	8
N1.4	Solve real-world problems involving whole numbers	2
M1.1	Use non-standard and standard units to measure, compare, and order	3
M2.1	Tell time	2
M2.2	Solve problems involving time	2
M3.1	Use different currency units to create amounts	0
G1.1	Recognize and describe shapes and figures	2
G2.1	Compose and decompose shapes and figures	0
G3.1	Describe the position and directions of objects in space	1
S1.1	Retrieve and interpret data presented in displays	2
A1.1	Recognize, describe, extend, and generate patterns	0
A3.2	Demonstrate an understanding of equivalency	0
Total		**26**

Facilitators may choose to share the level of alignment of the assessment with the GPF in terms of the criteria used for self-assessment (see **Chapter II**), focusing on both the depth (number of items that have at least a partial fit with at least one statement of knowledge and/or skill(s) from the GPF) and breadth (coverage of GPF domains, constructs, and subconstructs by at least one item with a partial fit) of alignment. Where there is significant disagreement between the outcomes of the original alignment and the alignment by the panelists, facilitators will want to focus on this during the next task, where consensus is the aim.

From the criteria in **Chapter II**, it is clear that the example grade 3 assessment described in **Table 6** would be considered "strongly aligned" since it: 1) contains at least ten number items (14 total), at least five total measurement and geometry items (10 total), and at least two Probability and Statistics and Algebra items (2 total) and 2) has items covering at least 7 of the 14 subconstructs with knowledge and/or skills expected at grade 2 (9 out of 14 subconstructs are covered).

D. TASK 2 – MATCHING ASSESSMENT ITEMS WITH GPLS AND GPDS

Task 2 builds on the panelists' understanding of the assessment items and GPF gained through the alignment activity. Matching is the process of finding the specific Global Proficiency Descriptor (GPD) at the appropriate grade level (Grade 2 for SDG 4.1.1(a), Grade 5 for SDG 4.1.1(b) and Grade 8 for SDG 4.1.1(c)) in the GPF that corresponds with a certain item in an assessment used for policy linking. The purpose of Task 2 is to further narrow in on the expectations of learners measured by each assessment item. This will help panelists know which GPD (performance standard) they should be considering when rating whether or not a minimally proficient learner would answer the item correctly in the benchmarking process (Task 3). In this task, panelists are asked to take their alignment work to the next level by matching each item to the appropriate GPL and GPD in the GPF.[19] Where there have been differences of opinion in the alignment exercise, these will need to be discussed. Facilitators should refer to the original alignment exercise undertaken as part of the self-assessment to support the discussion. Panelists should work in groups to reach consensus on the answers to the following three questions for each assessment item:

1. **What knowledge and/or skill(s) are required to answer the items correctly?** Panelists can draw on their work on this during Task 1, compare responses, and reach consensus.
2. **What makes the item easy or difficult?** In this step, panelists should consider things such as: distractors (from multiple choice options), whether the language used to ask the question is language the learner is used to hearing in the classroom, whether the topic (for a reading passage) is likely to be familiar, and whether any images included in the item are likely to be familiar to the learner and similar or different to those presented in classroom materials. For instance, in the example provided in **Figure 9** below, the panelist might say that one thing that makes this item easy is that the question uses the same exact words as those used in the first sentence of the passage. One thing that might make it difficult would be if learners are not familiar with dogs because they do not exist in their context.
3. **What is the lowest GPL that is most appropriate for the item?** Panelists should read through the GPDs for each GPL at the grade level (and the lower grades) to determine what GPL(s) and GPD(s) is the best match at which grade level. They should select the lowest GPL that corresponds with the knowledge and/or skill(s) learners need to answer the item correctly. If the item aligns to more than one statement of knowledge and/or skill(s) (as determined in Task 1) and, thus, more than one GPD, the panelist should select the higher of the GPLs since a learner would not be able to answer the item without the knowledge and/or skill(s) described in that GPD. If the item is too difficult to match to the grade level for which benchmarks are being set, panelists should note that the item falls above the exceeds level. One important note for this step is that for reading assessments, panelists will often have to assess the grade level of the decoding, reading

[19] Note that if during Stage 1, 2, or 3, the government decides that it only wishes to set a benchmark for the meets level or the government/assessment agency or 4.1.1 Review Panel decides the assessment is too short to accommodate three benchmarks at the three main GPLs, then panelists need only match to the grade-level GPD rather than the GPL.

comprehension, or comprehension of spoken or signed language passage since many of the GPDs are the same from one grade to another with the only difference being the grade level of the passage. Appendices A and B of the Reading GPF have criteria and examples to help panelists make this assessment of the grade level of the passage.

Figure 9 provides an example taken from the Workshop Facilitation Slides included in **Annex G**. In this example item, learners are asked to read the following passage:

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

Learners are then asked to respond to the question, "Where is Van?" This question matches with the statement of knowledge or skill of retrieving a single piece of explicit information from a grade-level text by direct-word matching. The panelist has identified what makes this item easy or difficult in the top box of this example. Because the Reading GPF requires assessment of the passage's grade level (using GPF Appendix C), panelists must determine what level the passage is before identifying the GPL and GPD. In this example, the panelist has determined that the passage is a grade two-level passage. As a result, the item aligns to the Partially Meets Global Minimum Proficiency level at grade two.

Figure 9: Example of Matching Items to the GPLs and GPDs

Easy or difficult: One thing that makes the question easy is that it uses the same wording as the passage. Both contain the word, "is". Also, Van is a common name in this context.

Domain: Reading comprehension

Construct: Retrieve information

Subconstruct: Locate explicitly stated information

Passage grade level: Grade 2

Knowledge or skill: Retrieve a single piece of explicit information from a grade-level continuous text by direct- or close-word matching

GPL and GPD (performance standard):

Partially Meets: Retrieve a single piece of prominent, explicit information from a grade 2-level text by direct- or close-word matching when the information required is adjacent to the matched word and there is no competing information.

Meets: Retrieve a single piece of explicit information from a grade 2-level text by direct- or close-word matching when the information required is adjacent to the matched word and there is limited competing information.

Exceeds: Retrieve a single piece of explicit information from a grade 2-level text by direct- or close-word matching when there is limited competing information.

When completing this matching process, facilitators ask panelists to focus on matching to the GPDs that match with the items. Panelists should record their group's responses to the three questions posed in this task directly next to each item on their test booklet/assessment instrument.

Figure 10: Matching items identified as 'No fit' with the GPF

For items that have been determined to have 'no fit' with the GPF, and where it has not been possible to remove the items from the process, the discussion should focus on trying to determine whether the item is most appropriate for learners at the partially meets, meets or exceeds performance standard. To do this, panelists should follow these steps:

- Imagine a group of learners who are best described by the GPDs in the 'partially meets' level.
- Using their experience of teaching such learners, determine whether this is an appropriate item for those learners and if they would be likely to answer the item correctly.
- If the item is determined to be appropriate for learners at the partially meets level, this can be recorded.
- If not, then the process should be repeated for the 'meets' level and then the 'exceeds' level, if required.
- If the item is determined to be too difficult for the grade, then it should be recorded as above the exceeds level.

E. TASK 3 – THE ANGOFF METHOD FOR SETTING BENCHMARKS

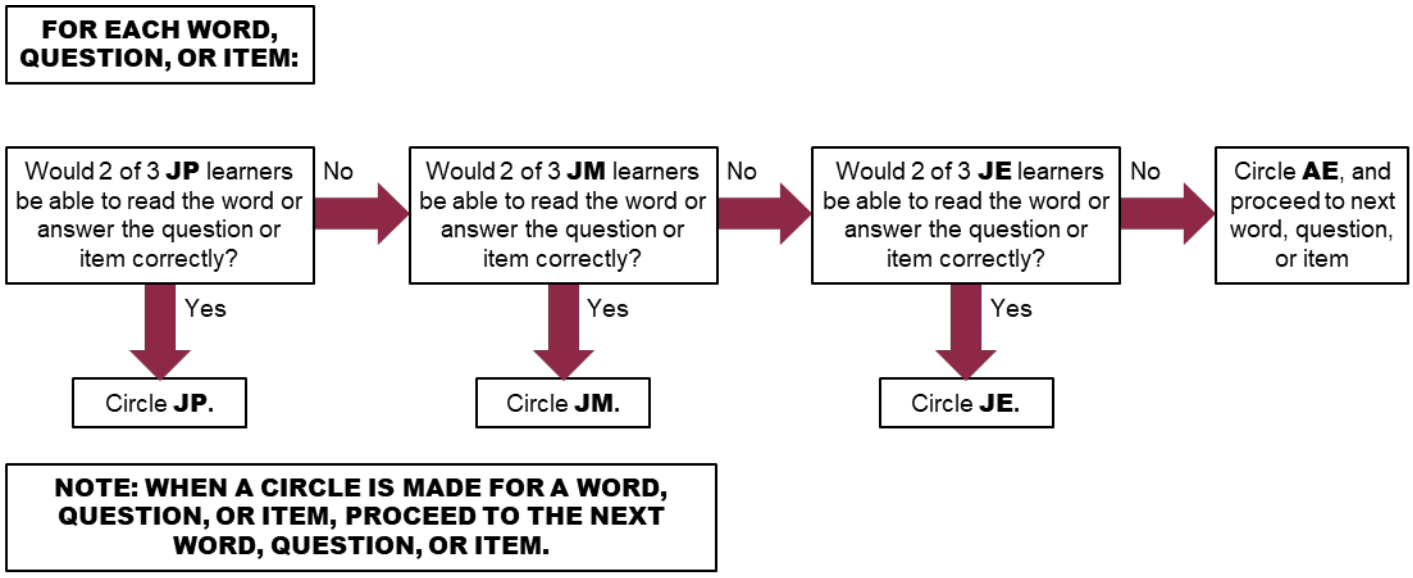
Task 3 is the most important task in the Policy Linking Workshop, as this is where panelists set benchmarks by making their judgements of how learners whose reading or mathematics abilities correspond with the knowledge and/or skill(s) aligned to each item in Task 1 and how the GPDs matched with each item in Task 2 would perform on each item. Task 3 relies on the Angoff method for setting benchmarks. The Angoff method is a test-centered method that is appropriate for the various kinds of assessments administered in different countries. With the Yes-No Angoff method, the panelists should use an item rating form (see **Annex I**) to rate each of the items on the assessment instruments, using the following four steps:

- **Step 1:** Identify or conceptualize three minimally-proficient learners at each GPL described by the GPF (Grade 2 for SDG 4.1.1(a), Grade 5 for SDG 4.1.1(b) and Grade 8 for SDG 4.1.1(c)).[20] Minimally proficient learners are those who perform at or just slightly above the GPDs that describe the GPL. Estimate how these learners would perform on each of the assessment items. These learners are called Just Partially Meets (JP), Just Meets (JM), and Just Exceeds (JE) learners.
- **Step 2:** Proceed item-by-item by reviewing the item and identifying the knowledge and/or skill(s) required to answer it correctly. The idea is to focus on the item content in relation to the statement(s) of knowledge and/or skill(s) in the GPF. Consider what makes the item easy or difficult (e.g., the wording of the item stem and the strength of the incorrect options or distractors) and what kind of errors may be possible or reasonable (Note: panelists should have recorded this information on their test booklet/assessment instrument during Task 2).
- **Step 3:** Select the lowest GPL, with the associated GPD, for the knowledge and/or skill(s) needed to answer the item correctly (panelists should have recorded this information on their test booklet/assessment instrument during Task 2). Where the item is classified as 'no fit' (and these haven't been removed from the set of items being used), panelists should use the discussion from the matching exercise to determine the lowest GPL that would answer the item correctly, even though it is not linked to the GPF.
- **Step 4:** Based on an understanding of Steps 1–3, follow the procedure shown in the flowchart in **Figure 11** below, which allows the panelists to rate each item to estimate whether learners in the different GPLs at the relevant grade level would answer each item correctly (yes or no) (note: **Figure 11** is only relevant when setting three benchmarks. When one benchmark is being set, facilitators can simplify this graphic to show only the JM and Above Meets (AM), instead of AE, levels). The flowchart has three decision points that must be considered to make the item ratings. These decision points correspond with the expectations for JP, JM, and JE learners described in the GPF. If a panelist does not believe that a JE learner (a learner who meets the expectations depicted in the Exceeds Global Minimum Proficiency Descriptor for the grade level and subconstruct) would correctly answer an item on an assessment, the panelist will circle AE, for Above Exceeds. In making a yes or no judgement at the three decision points, panelists must also consider the criteria depicted below that describe being "reasonably sure" and estimating how learners at each GPL/decision point *would* perform on an actual assessment in real life given assessment conditions, not how the GPF says they *should* perform. This means they will consider learners who meet the expectations of the appropriate GPL and GPD and determine if they are reasonably sure that those learners would answer the item correctly.[21]

[20] If, during Stage 1, 2, or 3, the government decides that it only wishes to set a benchmark for the meets level or the self-assessment determines that the assessment is too short to accommodate three benchmarks at the three main GPLs, then panelists need only conceptualize learners at the meets or JM level.

[21] For timed assessments, the rating process involves five steps, rather than four. Before panelists proceed to Step 2, they will first need to estimate how many items JP, JM, and JE learners will likely attempt (not get correct, but attempt) within the time limit. Then, in Step 4 (which is actually Step 5 for timed assessments) the panelist will only rate those items that they determined learners at that performance level would attempt. See Slide 119 of the timed assessments slide deck for more details.

Figure 11: Item Rating Process for Yes-No Angoff Modification



In completing Step 4, panelists should make their item ratings based on a consideration of four expectations, i.e., chances of whether the identified/conceptualized minimally proficient learners (as described in the GPF) would answer each item correctly:

- Probably not ("no")
- Somewhat possible ("no")
- Reasonably sure OR ≥ 67 percent chance OR two out of three learners ("yes")
- Absolutely positive ("yes")

To answer yes, panelists must be either reasonably sure or absolutely positive that a minimally proficient learner would answer the item correctly. Panelists should also be asked to base their ratings on "would" rather than "should" to set realistic expectations. Definitions of "would" and "should" follow:

- "Should" refers to performance-based only according to the GPDs
- "Would" is influenced by assessment constraints, e.g., difficulty of an item for a particular learner, testing conditions, learner anxiety, and random errors.

Important note for timed assessments: During the rating process, panelists working with a timed assessment will need to follow two steps:

1. Consider how many items a learner would attempt within the allotted time
2. Then determine whether or not the learner would have correctly responded to each item (following the typical steps for Task 3 described in **Figure 11**).

Important note for reading assessments: When panelists consider whether minimally proficient learners would correctly answer an item, they also need to consider the grade level of the word or passage the item references. For words, this consideration should be based on country expectations for words to be taught in a specific grade level, given all of the differences in languages across countries. For passages, panelists will need to consider the criteria for determining the grade level/text complexity of a passage, included in Appendices A and C of the Reading GPF. Details about how panelists should consider rating items based on their assessment of the grade level/text complexity of a passage are included in **Figure 12**.

Figure 12: Grade-Level/Text Complexity of Reading Passages

Overview – The GPDs in the reading GPF rely heavily on the assumption that the assessment being linked includes words and passages that are grade-appropriate. However, this is not always the case. Some assessments include passages from multiple grade levels purposefully so that results can help educators understand at what grade level learners are performing. Other assessments are used for more than one grade level of learners to examine improvement across grades. Also, as discussed above, assessments differ significantly in their level of difficulty. For this reason, it is critical that panelists working to link reading assessments work to determine the grade level of the words/passages in the assessment. For words, panelists will need information on what words are taught in the relevant grade level in that country – likely taken from national content or performance standards. For the passages, panelists should use the Appendices in the GPF to determine complexity.

Determining grade level/text complexity – For passages read to or signed for learners (ones that align with the Comprehension of Spoken or Signed Language domain), panelists should review the criteria included in Appendix A of the Reading GPF. For passages decoded by the learners (ones that align with the Decoding and/or Reading Comprehension domains), panelists should review the criteria included in Appendix B of the Reading GPF.

When the grade level of the word/passage is appropriate – If panelists assess the grade level of the word/passage to be appropriately aligned with the grade level for which the assessment is being linked, they can interpret the GPDs exactly as they are written.

When the grade level of the word/passage is too low – If panelists assess the grade level of the word/passage to be too low or easy for the grade level for which the assessment is being linked, they should assume that a minimally proficient learner might be able to do more than what is listed in the appropriate performance-level GPD. How much more depends on how easy the word/passage is (e.g., is it from the grade below or two or three grades below?).

When the grade level of the word/passage is too high – If panelists assess the grade level of the word/passage to be too high/difficult for the grade level for which the assessment is being linked, they should assume that a minimally proficient learner will likely not be able to do everything listed in the appropriate performance-level GPD. How much less depends on how easy the word/passage is.

The panelists should go through two rounds of ratings on two different days, with an in-depth discussion occurring between the two rounds. Literature suggests that having panelists rate items twice, through two separate rounds, works to improve the quality of ratings as well as the standard error of benchmarks (SE) and inter-rater reliability (See **Annex J** for details on how to calculate these and **Chapter V** for more details on when/why these are calculated), which have to be considered as part of the self-assessment process at the end of the workshop to inform whether the results of the policy linking workshop meet with the reliability and validity requirements to be accepted by UIS and other donors for global reporting.

During the discussion that occurs between Round 1 and 2 ratings, facilitators should present panelists with:

- **A summary of their ratings** as well as how their individual ratings compare with other panelist ratings. They should also lead panelists through discussions about items where there was considerable disagreement in the yes-no ratings.
- **Information on item difficulty** (guidance on how to generate this data is included in **Chapter V**), which helps panelists examine their own decisions on the difficulty of items.
- **Impact data** on the percentage of learners that would fall into each of the GPLs based on the most recent iteration of the assessment (guidance on how to generate this data is included in **Chapter V**), which helps panelists have an idea of the impact of their ratings and benchmarks.

Panelists should record their responses during each round on the same item rating form. An example of the form – with six items – is shown in **Table 7**.

Table 7: Item Rating Form for Use with Yes-No Angoff Modification

Item no.	Round 1 Individual and Independent Predictions				Round 2 Individual and Independent Predictions			
1	JP	JM	JE	AE	JP	JM	JE	AE

Item no.	Round 1 Individual and Independent Predictions				Round 2 Individual and Independent Predictions			
2	JP	JM	JE	AE	JP	JM	JE	AE
3	JP	JM	JE	AE	JP	JM	JE	AE
4	JP	JM	JE	AE	JP	JM	JE	AE
5	JP	JM	JE	AE	JP	JM	JE	AE
6	JP	JM	JE	AE	JP	JM	JE	AE

The panelists should submit their forms to the facilitators at the end of each round, and the facilitators will summarize the number of yes responses by GPL to yield an individual panelist's benchmark. The facilitators should then average the individual panelists' benchmarks to determine the panel's recommended benchmarks. The bullet points below show how the panelists' ratings are used to create benchmarks, both for each panelist and for the entire panel.

- Calculate totals for the initial and final benchmarks for each panelist:
  - Partially Meets = Total of each "yes" in the JP column of the rating form
  - Meets = Total of each "yes" in the JP and JM columns of the rating form
  - Exceeds = Total of each "yes" in the JP, JM, and JE columns of the rating form

- Calculate averages for the initial and final global benchmarks for the panel:
  - Partially Meets = Average of the "partially meets" benchmarks across all panelists
  - Meets = Average of the "meets" benchmarks across all panelists
  - Exceeds = Average of the "exceeds" benchmarks across all panelists

Since the panel's initial and final benchmarks are calculated by taking the averages of the panelists' benchmarks, the benchmarks will almost always have fractional values, i.e., not whole numbers. When this happens, the **benchmarks should always be rounded down** to the next score point, even if this goes against typical mathematical rounding rules. The reason is that the benchmarks designate minimum proficiency levels, and the advantage should be given to the learner (following the principle of "do no harm").

The calculation of the final benchmarks and presentation of the results by the lead facilitators and the data analyst completes the policy linking workshop. Details for calculating the benchmarks are included in **Annex T**. Details for preparing for the workshop are presented in **Chapter IV** below, and facilitator notes for implementing this methodology in an in-person or remote workshop are included in **Chapter V**.

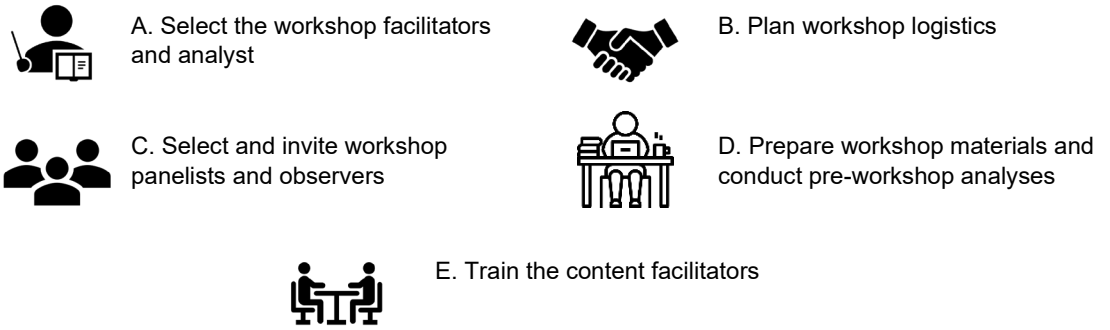
CHAPTER IV

CHAPTER IV. PREPARING FOR THE POLICY LINKING WORKSHOP

Government officials/assessment agency officers and donor representatives, if relevant, should have met during Stage 1: Initial Engagement to reach agreement on whether to conduct policy linking for global reporting and which assessment(s) they will link to global standards through this process. Resources for Stage 1 are linked in **Table 3**. One key goal of Stage 1 is ensuring government buy-in and ownership of the process as well as engagement throughout planning and preparation – with the intention that if the government is not implementing the workshop on its own, following the workshop, it should have the capacity to repeat a similar workshop to set additional benchmarks on different assessments in future years if necessary.

In this stage (Stage 4: Preparation for the Policy Linking Workshop), the project team – composed of the team of government or partner facilitators and logisticians designated to conduct the workshop – should carry out the five activities shown in **Figure 13**. A detailed checklist of technical and logistical preparations used by the project team, in conjunction with the government officials and donor representatives, is in **Annex D**.

Figure 13: Activities to Prepare for the Policy Linking Workshop



A. SELECT WORKSHOP FACILITATORS AND ANALYST

The project team will select facilitators and a data analyst for the workshop based on these criteria:

Lead facilitator(s) – Responsible for leading the workshop by ensuring panelists understand the policy linking method and what is expected. They must have expertise in policy linking and benchmarking, strong organizational skills, excellent presentation skills, and experience with educators ranging from teachers to policymakers. They should be aware of challenges in the policy linking process and corrective measures that may be taken to address those challenges.

Content facilitators – Responsible for helping the panelists interpret and understand the GPF and the assessment content, based on an understanding of local language and context. There is one facilitator for each assessment, i.e., by subject, grade, and language. They must be able to learn quickly since they will not usually have had previous experience with policy linking or benchmarking. The content facilitators must have experience in the theories and techniques of educational measurement, group facilitation skills, and experience in the content area (reading and/or mathematics) and context. They should understand curriculum and content standards, and how they are implemented by teachers in the classroom in the context where the assessment(s) was implemented. They must be fluent in the language of assessment.

Data analyst – Responsible for analyzing the data from the workshop and organizing information for presentation to the panelists. The analyst could be one of the lead facilitators who has the requisite skills, if that person has enough time during the workshop, though having a dedicated data analyst is recommended. 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).

Note that it is recommended that recruitment efforts also cover a **national workshop coordinator** and a **national logistician**. Also note that facilitators may be selected in Stage 1 as well to help coordinate the government/assessment agency's collation of documents in Stage 2.

B. PLAN WORKSHOP LOGISTICS

USE ANNEX D, ANNEX E AND ANNEX F

There are three main options for hosting a policy linking workshop. These are set out below along with their advantages and disadvantages. It is recommended, where possible, that policy linking workshops be held with the facilitators and panelists gathering in person, though successful workshops have been held using remote and hybrid approaches.

Table 8: Options for hosting a policy linking workshop

Option	Description	Advantages	Disadvantages
In Person	Facilitators and panelists attending the same location for the workshop	• Enables greater opportunity for informal discussion between panelists • Enables facilitators to provide more targeted support to panelists as required	• Cost, which may restrict the geographical representativeness of the panelists • May be impacted by travel restrictions
Remote	Facilitators and panelists attending the workshop from their own homes/offices	• Enables potential for greater geographical representation (though see associated disadvantage related to technology implications) • Cost	• Requires careful management to ensure high-quality discussions and support for panelists • May have technology implications, particularly in areas with poor connectivity
Hybrid	Panelists and content facilitators gathering in person, in country (as one group or as regional groups) and the lead facilitators attending remotely (assuming lead facilitators are internationally based)	• Enables greater opportunity for informal discussion between panelists • Enables content facilitators to support panelists	• Requires careful management by lead facilitators to ensure high-quality discussions and support for panelists • May be impacted by travel restrictions

The project team should work with relevant government and partner stakeholders to select the appropriate option based on the context, participants' safety, and budget. If it is possible for at least some participants to attend the workshop in person, the project team will need to work with the government to select an appropriate venue for this activity. If it is not possible to gather in person, the project team and government should agree on an appropriate digital platform (ensuring appropriate licenses are purchased to enable access to all relevant features of the platform). They should also agree and plan for other logistics, such as whether workshop interpretation and/or material translation is necessary; whether they will cover the costs of panelist transportation, hotel, and per diem costs or phone/internet cards; whether they provide food during the workshop; whether they will send out the assessment or a sample of it to panelists in advance; etc. A workshop preparation checklist and a workshop activity planner are provided in **Annex D** and **Annex E**. A budget estimation template is provided in **Annex F** to help countries estimate the costs of the workshop.

In addition to the digital platform on which to host a remote or hybrid workshop, the project team will also need to consider the following:

- **Connectivity** – this may require the purchase of Internet data cards to ensure panelists have sufficient data to connect to the workshop sessions with a sufficiently stable connection.
- **Hardware (panelists)** – it is strongly recommended that panelists join the workshop using a computer or laptop with a sufficiently large screen. This will enable them to more clearly follow the presentations and any

other materials that are shared on screen. Where this is not possible, and panelists will join by mobile phones, it will be essential to provide hardcopy versions of all documents, including the presentation slides.

- **Hardware (room)** – in a hybrid workshop, it is important to ensure a screen is available so that all panelists can see the lead facilitators, with appropriate cameras and microphones so that the lead facilitators can see and hear what is happening in the workshop room.
- **Messaging app** – during remote and hybrid workshops it is important for panelists to be able to message facilitators. A secure messaging app (e.g., WhatsApp, Telegram) should be selected for this purpose.

Finally, in addition to general logistics, during this activity, the project team should agree with the government(s) about ways in which they will continue the engagement with the country government(s)/assessment agency that started prior to the workshop (in Stage 1). This engagement should ideally continue throughout the workshop and after its conclusion. The goal with engagement of the country government/assessment agency is to actively give key representatives a role in the preparations and execution of the workshop, which will build capacity and permit governments and assessment agencies to conduct future workshops as needed.

C. SELECT AND INVITE WORKSHOP PANELISTS

Selecting Panelists

USE ANNEX M

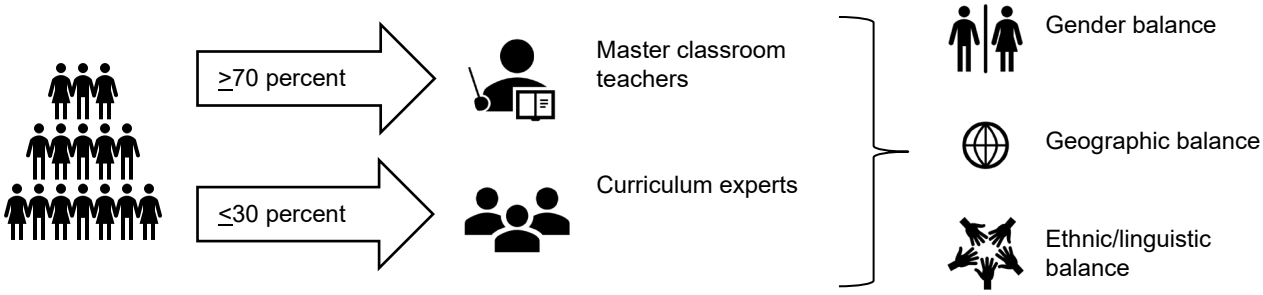
The panelists are key to the workshop, as they are the ones who will actually make judgments on the link between the assessment(s) and the GPF and then set benchmarks on the assessment(s) based on that link. The project team should plan separate panels for each grade, subject, and language of assessment used for policy linking. If multiple assessments are included in a single workshop, e.g., grade three reading and grade three mathematics, there will be plenary sessions for training, discussion, and presentation, but each panel will have separate group activities to check the alignment with the GPF, match the items with the GPLs and GPDs, and set the benchmarks.

When selecting a panel (or panels) for a policy linking workshop, the number of panelists 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 panelists. The panelists must have strong content knowledge and teaching skills (reading or math). They must be qualified to make the judgments required of them to set the benchmarks. The panelists must 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 assessment, a group of 15 panelists is a minimum and 20 panelists is a maximum. A group of this size will ensure the process obtains a replicable outcome but is also practical and manageable.[22] As shown in **Figure 14**, the panel should be made up of at least 70 percent master classroom teachers and up to 30 percent non-teachers, preferably curriculum experts.

[22] See Livingston & Zieky, 1982; Norcini, Shea, & Grasso, 1991; Mehrens & Popham, 1992; Hurtz & Hertz, 1999 for literature on the panel's size and the panelists' characteristics and qualifications.

Figure 14: Composition of Panelists



A typical panel composition is 12 teachers and 3 curriculum experts. Qualifications for panelists include the following:

- At least five years of teaching at or adjacent to the relevant grade level (teachers)
- At least five years of teaching experience (curriculum experts)
- Strong skills in the subject area (reading or math)
- Native skills in the language of instruction and assessment
- 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 panels should be ensured through the following criteria:

- **Gender representation –** The panelists 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 panelists must be selected to ensure representation from regions, provinces, and/or states of the assessments.
- **Ethnic and/or linguistic representation –** The panelists must have diversity that reflects the population as well as the language of assessments.
- **Other representation –** Depending on its relevance to the context and specific learner populations for whom results will be reported, the composition of the teachers and non-teachers 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 policy linking workshop is seeking to link regional or international assessments to the GPF, it is important that panelists represent multiple countries or that separate workshops are held for each country and then results compared to determine final benchmarks. Facilitators should reach out the 4.1.1 Review Panel for more details on appropriate representation with regional/international assessments.

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 panelists. This may be done through nominations by the Ministry of Education, assessment unit, 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. Final panelist demographics should be collected, aggregated, and submitted with the workshop outcomes using the form included in **Annex M**. Note: facilitators may want to send this form electronically to invited panelists ahead of the workshop to confirm representativeness of the panel, or facilitators may print this and collect it from panelists during the workshop. This form will give the project team sufficient data to address the degree to which the panelists meet the criteria as part of the post-workshop self-assessment.

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

Inviting Panelists and Observers and the Pre-workshop Activity

USE ANNEX K AND ANNEX L

Panelists should be invited well in advance of the workshop; at least three to four weeks is recommended. **Annex K** and **Annex L** include draft invitation letters for observers (e.g., government/assessment agency representatives, donors, other international or local donors/partners who may be interested in conducting a future policy linking workshop or understanding the process for more general purposes) and panelists respectively. The invitation letters should include basic information on the workshop and logistics, i.e., objectives, expectations, dates, transportation, lodging, meals, and per diems.

If at all possible, the invitations should include the full assessment tool(s) that will be linked to global standards with instructions on how it should be administered to learners ahead of the workshop. If the government has security concerns related to releasing the assessment, a sample of assessment questions can be used, as described in the following bullet points. However, using a sample assessment rather than the full assessment is not the preference, as it will not give panelists insight into reasonable benchmarks. See **Figure 15** for more information on assessment security.

- For individually administered timed assessments, such as early grade reading or mathematics assessments (EGRAs or EGMAs), the sample assessments will include subtasks from reading or mathematics, as appropriate.
- For group or individually administered untimed assessments, such as most curriculum-based assessments (CBAs), the sample assessments will include items from reading or mathematics, as appropriate.

During the workshop, the panelists will receive additional training and practical experience administering and scoring the assessments.

Given the length of the document, it may also be helpful for panelists to receive the GPF at this stage. However, the project team will need to decide how best to introduce the GPF, since panelists may be confused if the grade of the assessment is different from the grade of the GPF used for policy linking.

Figure 15: Assessment Security Considerations

Reasons for assessment security – To avoid teachers teaching to the test or learners cheating on tests, it is important to maintain the security of assessment instruments.

Which tests should be kept secure – Security is most critical for CBAs, especially those administered to all learners in a particular grade nationwide. Security among assessments that are administered only to a sample of learners and/or that change regularly (e.g., every year) is less important. However, security protocols should be left up to the government/assessment agency.

Security protocols for policy linking workshops – Assessment security protocols will vary depending on government and/or assessment agency preferences. However, the following security protocols are often used with CBAs:

- **Pre-workshop activity –** If the assessment is implemented with a census of learners or is not changed regularly, the government/assessment agency may wish to only send out a sample of questions from the assessment or a sample of similar assessment items.
- **Workshop protocols –** The assessments may not be included in panelist packets but might instead be handed out with panelist ID numbers (see **Section D** of this chapter for more on panelist ID numbers and packet preparation) listed on the top at the beginning of each day or for each activity in which the assessment is needed and then collected at the end of the day or activity.

D. MATERIALS AND ANALYSES

USE ANNEX A, ANNEX G, ANNEX H, ANNEX I, ANNEX M, ANNEX N, ANNEX O, ANNEX P, ANNEX Q, ANNEX R AND ANNEX S

All materials and analyses needed for the workshop are listed below in a series of three lists, organized by materials that need to be obtained from the government or regional/international assessment agency, analyses that need to be conducted using these materials in advance of the workshop, and materials that need to be created/adapted. Use of each of these materials in the workshop is also referenced in the following chapters and sections.

Most of these should have been obtained when conducting the self-assessment activity; thus, if the facilitators were involved in that stage, they should already have access to all except the starred items below (which they will need to request).

Some materials will need to be translated into the language in which the workshop will be conducted for the facilitators (panelists should be fluent in the language of the assessment). These are indicated with a (†). The decision regarding the languages for which to provide a translation will be determined by the country. It may not be possible for the documentation to be translated into the first languages for all panelists, depending on the number of languages spoken, but all panelists should be sufficiently fluent in at least one of the languages into which materials are being translated. Some materials are required to be provided to panelists in hardcopy. These are indicated with a (‡)

In order to obtain these materials, governments may require the development and signing of a non-disclosure agreement.

Materials That Need to be Obtained

- Assessment specifications
- All assessment instruments used in the assessment (†) (‡)
- Full set of assessment data files
- Answer keys and scoring rubrics
- Country standards on fluency/pace for decoding and grade-level text (if available and if countries are linking a reading assessment)*
- Technical report, including results from the most recent implementation of the assessment
- Sample assessment(s), created based on the full assessment (if necessary for security purposes, as described in **Section C**)* (‡)

Most of these documents/data will be used for the analysis that must occur before the workshop, which is described in detail below.

Analysis That Should be Conducted

Facilitators should calculate/prepare information on the following before the workshop using the assessment, data file, answer key, and scoring rubrics (if appropriate):

- **Item difficulty –** See **Annex N** for details on how to calculate these statistics using the data from the most recent assessment results.
- **Data distributions –** See **Annex O** for details on how to prepare these data. The data distributions will show the number and percentage of learners who took the assessment that achieved every possible score on the assessment. While these data can be prepared ahead of the workshop, they are not needed until Day 4, when they will form the basis of the impact information analysis between Angoff rating rounds 1 and 2 (what percentage of learners would meet each of the GPLs based on the initial panelist ratings/benchmarks and the data from the most recent iteration of the assessment).

This analysis will inform Round 2 of Task 3 Angoff ratings.

Materials and Data That Should be Created/Adapted

The project team/workshop facilitators should create (or adapt from the templates/examples provided in this toolkit) the following documents:

- **Workshop agenda** – A template is included in **Annex P**, which will need to be adapted as described below. (†)
- **Panelist IDs** – Need to be assigned on the first day of the workshop and should be confidential between the panelist and the project team.
- **Daily attendance sheet** – Needs to be created and tracked during the workshop to ensure each panelist has received all necessary training.
- **Panelist demographic information** – Form is included in **Annex M** but may need to be updated depending on criteria for representativeness of panelists.
- **Relevant grade/subject GPDs** – for workshops linking to SDG 4.1.1(a), panelists should be provided with grades 1, 2 and 3 from the GPF. For workshops linking to SDG 4.1.1(b), panelists should be provided with grades 4, 5 and 6 from the GPF. For workshops linking to SDG 4.1.1(c), panelists should be provided with grades 7, 8 and 9 from the GPF. (Facilitators will need to cut the GPF back to only the meets GPLs if benchmarks are only being set for one GPL.) Also, if the assessment is a reading assessment, the relevant appendices should also be included in the file so that panelists have criteria for assessing the grade level of a reading passage, as well as example items for the relevant grade levels. (†) (‡)
- **Facilitation slides** (workshop) – Details on how to locate the slide templates are included in **Annex G** for both timed and untimed assessments, but facilitators will need to adapt these; instructions on how to do so are included in the template. (†) (‡)
- **Facilitation slides** (content facilitator training) – Detail of which slides to use from the main facilitation slides is included in **Annex S**. (†)
- **Alignment rating forms and item rating forms** – Details on how to create these forms and examples are included in **Annex H** for the alignment form and **Annex I** for the item rating form.
- **Workshop evaluation forms** – A draft is included in **Annex R**. The project team may wish to add questions to the form and/or turn it into a daily evaluation form. (†)
- **Workshop feedback data** (Note that these cannot be created until after the Round 1 panelist ratings and then Round 2 ratings; instructions for how to generate this data are included in **Annex O**).

Details for how to create or adapt these materials and data, except the attendance sheet, which should be intuitive, are included below:

WORKSHOP AGENDA

The agenda for the workshop should remain similar, with the same tasks, regardless of whether an in-person, hybrid or remote workshop is planned (thought the remote agenda should allow for shorted days and more breaks over a longer period of time due to the fatigue of online meetings). Regardless of the type of workshop, there will always, though, be the need for some flexibility should tasks take longer than planned or should panelists finish more quickly. There are certain activities that should take place at the end of a day/session, to allow this flexibility, and to provide facilitators to prepare for the next session.

Table 9 provides an overview of the tasks in the workshop, with further details in **Chapter V** – to note: the activities within a task may take place over consecutive days for an in-person workshop or over several sessions for remote workshops. The template workshop agenda (**Annex P**) provides time allocations, and facilitation requirements, for the activities within each task. Example agendas for in-person and remote workshops are provided in **Annex Q**.

The structure of the tasks should remain constant for all workshops, though there may need to be slight modifications on the time allocations depending on logistics and other context-specific issues. Facilitators should review the agenda,

adjust the dates, adjust times for breaks (based on local norms), add in any necessary speeches from government officials, assessment agency officers, donors, etc., and then send to the government/assessment agency and its partners for their review before finalizing. The recommendation for in-person workshops is that they should take place for 8 hours per day, over 6 days. The recommendation for remote workshops is that the sessions take place in approximately 2–4 hours sessions, spread out over a longer period of time of two weeks to one month. The latter time period is to allow panelists to review the GPF and practice administering the assessment ahead of the workshop as recommended in the "Inviting panelists and the pre-workshop activity" subsection above.

Table 9: Brief Description of the Workshop Sessions

Tasks	Descriptions
Opening	This task welcomes panelists to the workshop and provides time for introductions and will usually take place in a single session. Dignitaries from the host country/ies, including the government(s), assessment agency (if relevant), and donor agency (if relevant), are invited to address the workshop. The workshop coordinator reviews logistics. The lead facilitators present the agenda, objectives, and a high-level summary of the method (acknowledging that this will be new for most panelists and should not get into detail that might be confusing to panelists if introduced too early).
Familiarization	The focus of this task is on introducing and carefully reviewing the GPF and assessment instrument(s) ahead of the activities where these documents will be used. There will be two sessions for this task, which may take place on the same day or separately (indeed, this task can take place before the main workshop if time is limited or to provide more time for panelists to understand the documents). For the session on the assessment instrument, facilitators may choose to have the panelists administer the assessment to one another for practice (timed assessments) or take the assessment themselves (untimed assessments). At the end of this task, the panelists complete the first part of the evaluation.
Alignment (Task 1)	The lead facilitators train the panelists on the alignment task. The content facilitators lead the Task 1 activity on aligning the assessments with the GPF, which is an individual and independent activity. The lead facilitators present the alignment results. At the end of this task, the panelists complete the next part of the evaluation.
Matching (Task 2)	The lead facilitators train the panelists on the matching task. The content facilitators lead the Task 2 activity on matching the assessments with the GPDs/GPLs, which is a group activity. At the end of this task, the panelists complete the next part of the evaluation.
Benchmarking (Task 3)	The lead facilitators present an overview on global benchmarking. The lead facilitators train the panelists on the Angoff method. The content facilitators lead the first Task 3 activity with Angoff practice. The content facilitators lead the second Task 3 activity with Angoff Round 1. The lead facilitators analyze the Round 1 results (this activity will need to take place overnight/between sessions) The lead facilitators present the Round 1 results. The content facilitators lead the third Task 3 activity with Angoff Round 2. The lead facilitators present the Round 2 results. At the end of this task, the panelists complete the final part of the evaluation.
Close	Dignitaries from the host country/ies, including the government(s), assessment agency (if relevant), and donor agency (if relevant), are invited to close the workshop.

PANELIST IDS

Panelists should be assigned unique and confidential (between the project team and panelist) IDs ahead of the workshop. They will use these to identify themselves on their ratings forms so facilitators can follow up with panelists who do not seem to be understanding concepts and so that anonymous panelist ratings (normative information) can be presented to panelists between Round 1 and 2 ratings and after Round 2, as described in more detail below. Every panelist should know what their ID number is. It might be included on a slip of paper in their folders or written on the inside of the folder somewhere.

DAILY ATTENDANCE SHEET

It is important to take attendance each day of the workshop so that facilitators know which panelists have missed sessions and can follow up with those panelists, as needed, to make sure they understand what they need to do.

PANELIST DEMOGRAPHIC INFORMATION

It is important to collect all of the information included in the form in **Annex M** to ensure that panelists are representative of the population being assessed. This information must also be reported to the 4.1.1 Review Panel along with details on the population being assessed and the teachers of that population. For instance, the 4.1.1 Review Panel will want details on the percentage of grade X teachers in the area of assessment that are male versus female in order to check gender representation of teachers. The form included in **Annex M** may need to be updated so that it asks the appropriate questions about geographic demographics. For instance, some countries don't have regions or districts but instead states or municipalities. The form can either be sent to panelists in advance of the workshop or passed out and collected during the workshop.

RELEVANT GRADE/SUBJECT GPDS

The GPF is available on Edulinks and UIS' website. However, it is not necessary to present panelists with the entire GPF. Instead, facilitators can create a modified version that only has the relevant grades (Grade 2 for SDG 4.1.1(a), Grade 5 for SDG 4.1.1(b) and Grade 8 for SDG 4.1.1(c)) and the grade below. Facilitators will take panelists through a careful review of these tables during the workshop.

The GPF Knowledge or Skills table, Table 3, and Table 5, which includes the GPDs for each of the GPLs, are the most useful for workshops focused on setting three benchmarks – one for each of the GPLs. Workshops focused on only setting one benchmark should use GPF Tables 3 and 4. In both cases, panelists will use Table 3 for Task 1 – Alignment. Depending on the number of benchmarks that will be set, they will then use either Table 4 (for one benchmark) or Table 5 (for three benchmarks) for Task 3 – Rating. GPF Table 1 defines each GPL and is a useful reference for panelists if they cannot remember a specific GPL. Table 2 illustrates the domains, constructs, and subconstructs across grade levels and provides a useful summary for policymakers and panelists.

Note again that if the assessment is a reading assessment, the relevant GPF Appendices should also be included in the file so that panelists have criteria for assessing the grade level of a reading passage for the grade level being linked and one above and one below as well as example items for the relevant grade levels.

Facilitators, with the government, should consider whether the two tables, at a minimum, may need to be translated if the language of assessment is not English (see **Figure 16** for details), but facilitators should not make any other changes to the content or language of the GPF.

Figure 16: Translation of the GPF

Translation firms or individual translators may assist with the translation, but translation should be led by content experts. It is critical that the meaning of each term is translated fully and accurately and that translation of examples for reading includes changing the examples, as needed, to ensure they are still appropriate for the grade level (since the length and complexity of the words may change in translation). The project team should also consider a backward translation into English to validate the translation into another language.

Finally, over time, there will be translations of the GPDs (and even the entire GPF) into many languages, some of which may be used in multiple countries with the same languages. Even with those translations, the individual countries should carefully read the translated GPDs and make any necessary modifications based on local language usage.

FACILITATION SLIDES (MAIN WORKSHOP AND CONTENT FACILITATOR TRAINING)

The facilitators will present the slides during Days 1 to 6 of the workshop (for in-person workshops) or through a series of eight workshop sessions (for remote workshops). The slides are included in **Annex G** and include details on

the: 1) agenda, objectives, and method, 2) how to introduce the GPF and the assessment, 3) alignment, 4) matching, 5) benchmarking, and 6) evaluation. Note that the slides required some adaptation depending on the nature of the assessment.

The project team should consult with the government and other key stakeholders to determine whether the facilitation slides need to be translated into the language of assessment or another international language. If the slides are not translated into local languages, then the content facilitators or translators can interpret as needed.

The slides for content facilitator training (**Annex S**) are a subset of the slides for the main workshop.

ALIGNMENT AND ITEM RATING FORMS

There are two types of rating forms. The project team will adapt the forms to match with the assessment instrument and relevant parts of the GPF.

- **Alignment rating forms** (**Annex H**) – These will be used for the panelists' ratings of the alignment between the assessments and the GPF.
- **Item rating forms** (**Annex I**) – These will be used for the panelists' ratings of each assessment item in relation to the GPLs and GPDs.

The annexes include example alignment and item rating forms from timed assessments and untimed assessments. The forms will need to be adapted from one assessment to another depending on the assessment format (e.g., number of domains and constructs), question type(s) (e.g., multiple choice or single word), and scoring (e.g., dichotomous or polytomous). The alignment rating form was created with ease of use in mind, but the project team may wish to update it to make it more dynamic, with drop-down menus and automatically generated totals. Several options and examples of item rating forms are included in **Annex I** with details on how to choose and adapt the forms.

WORKSHOP EVALUATION FORMS

At minimum, panelists should fill out an evaluation at the end of the workshop; however, ideally panelists should complete the relevant section of the evaluation form at the end of each task, to check in on knowledge acquisition, areas that may need further clarity, facilitation techniques that are working/not working, etc. **Annex R** includes the minimum evaluation questions that should be asked of panelists by the end of the workshop. It is designed to capture their views on the policy linking process and support the self-assessment in **Chapter VI**. The form consists of Likert-type scales and open-ended questions on the panelists' satisfaction with the orientation, training, and process. The results will provide evidence of the panelists' confidence in their judgments, as well as seek additional comments on the policy linking experience.

If the project team conducts a daily evaluation that identifies issues that require retraining, they should administer the evaluation again following the additional training to confirm that panelists are now content. This will be required to support the self-assessment in **Chapter VI**.

If the project team opts not to include a daily evaluation, the lead facilitators and content facilitators should at a minimum consider conducting verbal check-ins with the panelists at the end of each day to discuss the proceedings and possible adaptations, e.g., more interpretation of the presentations into local language, a need to review the steps of a task, etc.

WORKSHOP FEEDBACK DATA

Workshop feedback data include normative information on panelist ratings and impact data. (These analyses will take place during the workshop, not before). Instructions on how to generate these statistics and feedback charts are included in **Annex O**. The data analyst will need to calculate the statistics, graphics, and charts using panelist rating data from Round 1. As such, this will need to be done between the relevant sessions of the workshop. The same

process will need to be completed following Round 2 ratings, and sufficient time must be built into the agenda for this to take place.

WORKSHOP PACKETS

Once all documents are created or adapted and data is generated, the project team will need to print the following documents to be included in each of the panelists' packets (and mailed or delivered to the panelists in the case of remote workshops):

- Agenda
- Panelist ID (can be written in small numbers on the inside of the folder or printed on a piece of paper included in the folder)
- Glossary of terms (can be printed from the one included at the beginning of this document)
- Acronym list (can be printed from the one included at the beginning of this document)
- Relevant grade/subject GPDs from the GPF (Grade 2 for SDG 4.1.1(a), Grade 5 for SDG 4.1.1(b) and Grade 8 for SDG 4.1.1(c))
- Assessment instrument (should only be included if assessment security protocols allow for it; see **Figure 15** for details on assessment security)
- Slides (printed in notes format)
- Alignment rating form
- Item rating form

E. TRAIN CONTENT FACILITATORS

The lead facilitators will need to conduct a training session for the content facilitators, who are not likely to be familiar with the policy linking methodology. A content facilitator training slide template is available in **Annex S**. The training should include an overview of the agenda for the workshop; a detailed discussion of the GPF; a review of the assessment(s); and practice alignment, matching, and benchmarking exercises. These practice exercises are vital to ensure content facilitator engagement in the process rather than relying on passive knowledge transfer.

The training should also include a discussion of lead and content facilitator roles and responsibilities, managing group dynamics (particularly when observers with higher status are present, to avoid them unduly influencing panelist decisions) and should provide details on the do's and don'ts of facilitating discussions during and following completion of each of the tasks as shown in **Table 10** (the same rules apply to answering panelist questions and facilitating practice ratings).

The main point of the training will be to ensure the content facilitators are keenly familiar with the GPF and the assessment, as they will need to help the panelists interpret both, and to cover the three tasks – alignment, matching, and benchmarking. The lead and content facilitators are responsible for communicating the policy linking procedures to the panelists, while the content facilitators are responsible for reinforcing the overall training with the panelists during group work. Both facilitators must know how to answer panelist questions and facilitate appropriate discussions.

If there is sufficient time between the training and the workshop, it may be helpful to undertake a rehearsal of the relevant sections of the workshop with the content facilitators, with lead facilitators acting as panelists, to ensure understanding.

F. TECHNICAL TEST

In hybrid and remote workshops, it is essential to carry out a technical test with all locations and participants in advance of the workshop. This will allow for troubleshooting and enable time for back-up solutions to be implemented where issues cannot be resolved and ensure the start of the main workshop is not disrupted.

The following activities should be undertaken during the technical test:

- **Connectivity** – do all locations and participants have a suitable connection?
- **Audio** – do all locations and 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 panelists?
- **Messaging app** – have all participants downloaded the chosen messaging app and can they access the group chat?

Table 10: Discussion Purpose, Do's, and Don'ts by Task

Task	Discussion Purpose	Do's	Don'ts
Task 1 – Assessment and GPF alignment (panelists work independently)	To ensure panelists understood the task, find out what challenges they faced and also determine if there are any items that do not fit with the GPF and, thus, do not need to be rated.	• Make sure all panelists have the opportunity to speak, share their ratings, and ask questions. • Make sure all panelists are considering each of the alignment steps and that their explanations of how they selected "no fit," "partial fit," or "complete fit" make sense and demonstrate understanding of the concepts. • Explore disagreements between panelists' alignment with statements(s) of knowledge and/or skill(s) and fit by asking panelists on both sides to volunteer explanations of why they rated the way they did.	• Tell a panelist or imply that a panelist has incorrectly aligned an item. • Tell a panelist or imply that a panelist has selected the wrong level of fit. • Single out individual panelists to ask them why they aligned X item to X statement(s) of knowledge and/or skill(s).
Task 2 – Matching the assessment items with the GPLs and GPDs (panelists work together in groups)	To ensure panelists understood the task, find out what challenges they faced, make sure they considered what makes an item easy/difficult and also ensure the group has reached consensus on the GPL and GPDs that align with each item.	• Make sure all panelists have the opportunity to speak, provide opinions on whether they agree or disagree with the group consensus, and ask questions. • Make sure all panelists are considering each of the matching steps and that their explanations are clear and in line with the methodology with regards to how they selected the lowest GPL at which learners should have the knowledge and/or skill(s) to answer an item. • Bring up additional points that could make an item easy or difficult that panelists didn't identify.	• Tell panelists or imply that panelists have incorrectly matched an item to a GPL/GPD or that their points about what makes an item easy/difficult are wrong.
Task 3, Round 1 – Rating the items using the Angoff method (panelists work independently)	To ensure panelists understood the task, ask them to explain why they rated an item the way they did. Their explanation should reference the GPD and the questions of "would" and "reasonably sure." And, give the panelists an opportunity to talk about disagreements on ratings, as this might inform some panelists' Round 2 rating decisions.	• Make sure all panelists have the opportunity to speak, provide explanations of how they rated the items and why, and ask questions. • Make sure all panelists are considering each of the rating steps and that their explanations of why they rated an item the way they did reference the GPDs, their conceptualization of learners at each of the GPLs, things that make the item easy/difficult, and whether they are "reasonably sure." • Identify items where panelists disagreed, and ask volunteer panelists who rated no to explain why and vice-versa. • Encourage panelists to consider the item difficulty and impact data and decide if that affects their Round 2 judgements.	• Tell panelists or imply that panelists have incorrectly rated an item. • Single out individual panelists to ask them why they rated X item as the way they did (Note - panelist ratings are supposed to be confidential, which is why they are presented to the group by panelist number rather than name). • Imply that because item difficulty data show learners found an item difficult that it should be rated as "no." It is possible that many learners who took the assessment simply were not meeting the requirements of the GPLs.
Task 3, Round 2 – Rating the items using the Angoff method (panelists work independently)	Get panelist reactions to their final benchmarks and the impact data.	• Make sure everyone has the opportunity to speak and ask questions.	• Make unsubstantiated claims about how the government/regional or international assessment agency will use the benchmarks.

CHAPTER V

CHAPTER V. IMPLEMENTING THE POLICY LINKING WORKSHOP

While **Chapter III** provides an explanation of the methodology used in the policy linking workshop, this chapter provides guidance and tips for facilitators on how to lead the workshop and when to do what.

Where the language spoken by the lead facilitators is not the first language of the panelists, it will be important to ensure interpreters are present to support facilitation through simultaneous translation. This is important even where the panelists have a basic understanding of the language of the lead facilitator. Given the technical nature of the concepts involved in policy linking, and the level of discussion required, it will be easier for panelists to express their ideas in their preferred language. Lead facilitators may need some training in how to deliver the workshop with simultaneous translation if this is not something they have done before.

As described in **Chapter III** and **Chapter IV**, facilitators will lead presentations and activities over a number of sessions. During that time, they will introduce the workshop methodology, the GPF, and the assessment during familiarization, and then proceed to leading the panelists through the three main policy linking tasks:

- **Task 1:** Check the content alignment between the assessments and the GPF using a standardized procedure
- **Task 2:** Match the assessment items with the GPF, i.e., the GPLs and GPDs
- **Task 3:** Set three global benchmarks for each assessment using a standardized method (a modified version of the Angoff Procedure)[23]

The template in **Annex P** provides timings for each activity during the workshop. The project team will need to organize these into days/sessions depending on the type of workshop (in-person, hybrid or remote). Example agendas for an in-person and remote workshop can be found in **Annex Q**. **Table 11** sets out the activities for the workshop, linked to the 20 sets of slides in **Annex G** that are also organized by activity and can be used to create day-by-day or session-by-session slides depending on the agenda created by the project team. Although timings may vary slightly to accommodate local requirements, the order of the activities must remain constant, and the project teams must ensure there is sufficient time to enable panelists to fully understand what is expected of them and to carry out the tasks. The project team will also need to ensure there is sufficient time in the agenda for the facilitators to undertake necessary activities between relevant session (for example, collating data and producing data impact slides).

The presentations are led in plenary by the lead facilitators, and the activities are led in groups (panels) by the content facilitators. Calculations of benchmarks and indicators should be conducted by the lead facilitators and the data analyst. Lead facilitators and content facilitators should hold check-in discussions or administer short evaluations with the panelists at the end of each day/session (more details are included in the "**workshop evaluation form**" subsection). Regardless of what is decided for the regular check-ins/evaluations, panelists must complete a written evaluation on all activities by the end of the workshop for reporting purposes.

Facilitators should meet at the end of the day/session to prepare for the next day/session in case there are any changes required to the plans as a result of lessons learnt.

Table 11: Summary of Tasks and Activities for the Policy Linking Workshop

Task	Presentation	Activity
Opening	1	Welcome and introductions
		Address by government(s) representatives, assessment agency (if relevant), and donor organization (if relevant)

[23] Note that if during Stage 1, 2, or 3, the government decides that it only wish to set a benchmark for the "meets" level or the government/assessment agency or 4.1.1 Review Panel decides the assessment is too short to accommodate three benchmarks at the three main GPLs, then panelists need only set one benchmark (rather than three) for each assessment.

Task	Presentation	Activity
	2	Overview of agenda, objectives, and high-level overview of method
Familiarization	3	Familiarization with GPF
	4	Familiarization with assessment instrument
Task 1 – Alignment	5	Train panelists on the alignment exercise (including practice items)
	6	Panelists undertake alignment activity (independent activity)
	7	Presentation and discussion on the alignment results
Task 2 – Matching	8	Train panelists on the matching task
	9	Panelists undertake matching activity (group activity)
	10	Presentation and discussion on the matching results
Task 3 – Benchmarking	11	Overview of global standards and benchmarking approach
	12	Train panelists on Angoff method
	13	Panelists undertake Angoff method with practice items
	14	Round 1
	15	Presentation and discussion of Round 1 results and impact data
	16	Presentation on Angoff Round 2
	17	Round 2
	18	Presentation of Round 2 results
Evaluation	19	Workshop evaluation
Closing	20	Closing remarks and presentation of certificates
Documentation	After the workshop	Production of the technical documentation

Information on each of the above presentations (1–20) is provided below, along with tips for the facilitators.

A. OPENING

1. Welcome and Introductions

MATERIALS: FACILITATION SLIDES (PRESENTATION 1), PANELIST WORKSHOP PACKETS (SEE CHAPTER III, SECTION D)

In this presentation, you will introduce yourself and provide opening remarks. You should invite government officials and any donor education officials, if relevant, to make opening remarks. The implementing partner may also make remarks if a project is co-sponsoring the workshop. The workshop participants and the project team will introduce themselves. You will identify workshop materials found in the panelists' workshop packets. You will discuss logistics of the workshop, which will vary depending on the type of workshop (in-person, remote or hybrid) but may include information pertaining to the venue, plenary and breakout rooms, lodging, meals, per diem, transportation, technology, and methods of communication.

Figure 17: Tips for Facilitators on Opening Presentation

Government officials/assessment agency officers, donor education officials, and implementing partners should be provided about 10 minutes each for their remarks. As each panelist introduces themselves to the group, you may ask them to share their name, location, and position. Following the overview presentation, allow about 10 minutes for questions and answers. Assure participants that the formal introductions are just an overview and that the following sessions will dive more deeply into each of the topics mentioned.

2. Overview of Agenda, Objectives, and High-level Overview of Method

MATERIALS: FACILITATION SLIDES (PRESENTATION 2), PANELIST WORKSHOP PACKETS

In this presentation, you will provide an overview of the workshop agenda to the participants, background information on the policy linking method, the SDG 4.1.1 indicators, the USAID "F" indicators (where relevant), and the GPF. You will explain briefly the need for benchmarks that will determine global minimum proficiency on assessments. You will explain the three policy linking tasks: 1) check the alignment, 2) match the assessment items with the proficiency levels and descriptors, and 3) set the global benchmarks using a standardized method.

Figure 18: Tips for Facilitators on Background Presentation

When introducing the GPF and PLT, provide context for the workshop by giving a brief background and describing future activities. Use the graphic with the GPF scale, including the four proficiency levels and three benchmarks. Explain that the objective of the workshop is to set the benchmarks. The benchmarks will be used for comparing assessment results across countries, aggregating assessment results for global reporting, and tracking progress over time. Tell the panelists that more information will be provided during each session.

B. FAMILIARIZATION

Where possible, it is ideal to undertake the familiarization exercises in advance of the workshop to give panelists time to internalize the information prior to making use of it in the workshop. However, this may not be possible, particularly for in-person workshops.

3. Familiarization with GPF

MATERIALS: FACILITATION SLIDES (PRESENTATION 3), RELEVANT GRADE/SUBJECT EXTRACTS FROM THE GPF

In this presentation, you will introduce the GPF, including introducing each of the domains, constructs, subconstructs, statements of knowledge and/or skill(s), and GPLs and GPDs. You will provide background information on the development of the GPF and walk through all of the GPDs for the relevant grade level (Grade 2 for SDG 4.1.1(a), Grade 5 for SDG 4.1.1(b) and Grade 8 for SDG 4.1.1(c)). You will discuss confusing terms and ask panelists to give examples of items that might be used to measure the performance standard described in the GPD. Where the grade of the assessment differs from the grade used for policy linking, you should also provide the grade of the assessment from the GPF as this is likely to be needed later for the matching exercise.

Figure 19: Tips for Facilitators on Presentation of the GPF

Where familiarization takes place in the workshop, partition the information on a 'when needed' basis (for example, up to and including the knowledge or skill statements before the Alignment task, and the GPD and GPL between the Alignment task and the Matching task) to avoid overload on day one.

Make sure you spend enough time reviewing each of the key terms and the GPDs to ensure panelist understanding. You may wish to have content facilitators translate some terms into the local language to ensure everyone has the same understanding. Also, take time to pause when reviewing each GPD to engage panelists in a discussion about that GPD and what types of assessment items they might envision could be used to measure it. Make sure it is clear that when you talk about meeting global minimum proficiency in the workshop, you are talking about learners who have the skills defined in the GPF.

It is important to ensure this session is engaging for panelists, rather than just listening to facilitators. The Community of Practice for Policy Linking has shared materials, including activities for panelists, to support familiarization, which may be useful.

4. Familiarization with Assessment Instrument

MATERIALS: FACILITATION SLIDES (PRESENTATION 4), ASSESSMENT INSTRUMENT
(Note: you will need to create additional slides for this presentation; the recommendation is one slide per assessment item or pair or items)

In this presentation, you will introduce the assessment instrument, describe how it is administered, how it is scored, and what the sample population looked like for the last iteration of the assessment (e.g., what area/populations was it representative of). You will walk through each of the items in the assessment and make sure panelists understand each one. You may also have the panelists administer the assessment to one another (for individually administered assessments) or take the assessment themselves (for group-administered assessments) to ensure further understanding.

Figure 20: Tips for Facilitators on the Assessment Presentation

This presentation should be led by someone with a full understanding of the assessment. The session should be interactive and not just showing the assessment to panelists.

Make sure you spend enough time on each assessment item to ensure the panelists understand the item, how it is administered, and what some common stumbling blocks might be. When reviewing the pre-workshop activity, make sure panelists selected learners to assess based on those they knew had the knowledge and/or skills described in the GPF for a particular grade and GPL. If so, those learners' scores may prove especially helpful for panelists in setting benchmarks. If panelists were unable to assess learners who meet the GPF definitions for partially meets, meets, or exceeds global minimum proficiency, the scores of the learners they did assess are less important, and they should instead just use the findings from that activity to inform their understanding of item difficulty and test administration procedures. Take plenty of time for questions and discussion about the assessment.

C. TASK 1 - ALIGNMENT

5. Train Panelists on the Alignment Exercise (including practice items)

MATERIALS: FACILITATION SLIDES (PRESENTATION 5), PANELIST WORKSHOP PACKETS (SPECIFICALLY ANNEX H)

In this presentation, you will revisit the GPF, specifically, the subconstructs and the statements of knowledge and/or skill(s). You will describe the three-step process panelists will engage in to check the alignment of the assessments with the statements of knowledge and/or skill(s) described by the GPF (see the section on **Task 1** in Chapter III and Table 3 of the GPF) and the process the facilitators will use to summarize results. You will explain the three levels of alignment or fit – complete, partial, and no fit – with both complete and partial counting towards alignment. You may explain the standardized method for determining the level of breadth and depth of alignment between the assessment(s) and the GPF. You will walk the participants through some sample items to ensure they understand the task. There are sample reading and mathematics items included in **presentation 5** that you can use for this purpose, or you can select/develop your own. Note that sample items should not be too similar to the actual assessment items that panelists will rate, as this may bias ratings, but it is helpful if they cover similar subconstructs. Finally, you may share the alignment criteria listed in **Table 4** and **Table 5**, though this is not required as the overall reported alignment will be taken from the exercise carried out as part of self-assessment.

Figure 21: Tips for Facilitators on the Alignment Presentation

When describing the alignment activity, remind panelists that the GPF was developed as a global set of knowledge and skills and related GPDs that was drawn from consensus global content. Make sure that the panelists know the difference between the statements of knowledge and/or skill(s) and the GPDs (content and performance standards). Go carefully through the examples and each of the two steps and sub-steps described in the section on **Task 1 in Chapter III**. Tell the panelists that some assessment items may not match with the GPF since each country has its own standards. That is okay. Make sure they understand that both items with a partial fit or complete fit count toward alignment criteria. Where the grade of the assessment is different from the grade in the GPF being used for policy linking, make sure you explain why this is the case – to ensure the same standard is applied across all countries.

6. Panelists Undertake Alignment Activity (Independent Activity)

MATERIALS: FACILITATION SLIDES (PRESENTATION 6), PANELIST WORKSHOP PACKETS (SPECIFICALLY ANNEX H)

In this activity, you will give the panelists an opportunity to ask questions before they proceed with aligning the assessment items with the GPF statements of knowledge and/or skill(s) using the recording sheet provided.

Figure 22: Tips for Facilitators on Task 1 – Aligning the Assessment(s) with the GPF

While discussion is encouraged during the group work, each panelist should conduct their own individual and independent alignment ratings, or item-statement of knowledge and/or skill(s) ratings, and submit their form to the content facilitators for analysis by the lead facilitators or data analyst. Panelists should only be aligning to statements of knowledge and/or skill(s) that are relevant for the grade level (Grade 2 for SDG 4.1.1(a), Grade 5 for SDG 4.1.1(b) and Grade 8 for SDG 4.1.1(c)), as depicted by each "x" in GPF Table 3.

7. Presentation and Discussion of Alignment Results

MATERIALS: FACILITATION SLIDES (PRESENTATION 7)
(It is recommended that you create additional slides for this presentation, including one slide per item where there was significant disagreement among panelists on which statement(s) of knowledge and/or skill(s) the item aligns with.)

In this presentation, you will cover the results from the alignment activity. You may also address the level of alignment achieved based on the alignment criteria, presented in **Table 4** and **Table 5**. You will also want to review individual items and alignment ratings where there was a considerable amount of disagreement between panelists on which statement(s) of knowledge and/or skill(s) the item aligned. Tips on facilitating this discussion are included in **Table 10** above in the Content Facilitator Training Section.

Figure 23: Tips for Facilitators on Reviewing the Results of Task 1

Although agreement is not required on the alignment task, it is necessary for matching. Content facilitators should have access to the agreed alignment outcomes from the self-assessment process to help guide discussions.

D. TASK 2 - MATCHING

8. Train Panelists on the Matching Task

MATERIALS: FACILITATION SLIDES (PRESENTATION 8), PANELIST WORKSHOP PACKETS

In this presentation, you will build on the alignment conducted during Task 1 (to the statements of knowledge and/or skill[s]) to discuss matching to GPLs and GPDs (also called performance standards) – views on alignment may change through the matching exercise, which is allowed as long as consensus is reached at the end. You will walk the panelists through answering the three questions required under the task (see the section on **Task 2** in Chapter III for the questions) – namely, what knowledge and/or skills are required to answer the item correctly, what makes the item easy/difficult, and what is the lowest GPL that matches with the item. For reading, this will include discussion of the reading passage complexity annex of the GPF, since this will also need to be considered as part of the matching task, to make sure the text is grade appropriate. You will walk the participants through some sample items to ensure they understand the task. There are sample items included in **presentation 8** for both reading and mathematics that you can use for this purpose, or you can select or develop your own.

You may also want panelists to have access to their own curriculum documents to remind them of grade-level expectations. This is particularly the case for reading, where the GPF contains references to, for example, common words for the grade. However, if sharing curriculum documents, make sure all panelists are clear that they are linking the items to the GPF and not to their own curriculum, where the content may be linked to different year groups.

Figure 24: Tips for Facilitators on the Task 2 Matching Presentation

Remind panelists that this activity builds on the understanding of the assessment items and the GPF gained through the alignment activity. The key concept is to match the items with the lowest GPL and GPD that describe the expectations learners must meet to correctly answer the item for the grade under consideration. If the group rated the item as a partial-fit item, they will need to consider the two relevant GPDs and likely select the higher of the two GPLs since learners must meet expectations from both to correctly answer the item. If the group rated the item as 'no fit' panelists should follow the process described in **Figure 10**.

If an item matches with a descriptor from a grade other than the one under consideration (Grade 2 for SDG 4.1.1(a), Grade 5 for SDG 4.1.1(b) and Grade 8 for SDG 4.1.1(c)), this should also be recorded, noting the relevant grade. This information will be used during the benchmarking task. If the item is linked to a lower grade, it is likely that learners at the grade under consideration would be able to answer correctly. If the item is linked to a higher grade, it is unlikely that learners at the grade under consideration would be able to answer correctly.

9. Panelists Undertake Matching Activity (Group Activity)

MATERIALS: FACILITATION SLIDES (PRESENTATION 9), PANELIST WORKSHOP PACKETS

In this activity, you will operationalize the presentation. You will provide an opportunity for the panelists to ask questions on the GPLs and GPDs. You will again clarify the difference between the statements of knowledge and/or skill(s) and GPDs. You will break the panel up into separate panel-level groups for each assessment (grade, subject, and language) being linked through the workshop, and the content facilitators will lead them through matching each item with the lowest GPLs and GPDs. The content facilitators will also work to help them achieve consensus. Where this is proving difficult, you should use the outcome from the alignment activity carried out as part of self-assessment to guide the discussion, since ultimately, this has been determined to be the correct alignment by the country.

Figure 25: Tips for Facilitators on Overseeing the Task 2 Matching Activity

Make sure the panelists go item by item and have discussions on where the items match with the lowest GPDs. It may be helpful for the panelists discuss their matches in small groups and then come together to reach consensus in their panels. Remind them to write the answers to the three questions for the task directly on their assessment instrument/test booklet next to the item.

10. Presentation and Discussion on the Matching Results (Task 2)

MATERIALS: FACILITATION SLIDES (PRESENTATION 10), PANELIST WORKSHOP PACKETS

In this presentation, you will provide the matching results and verify the panelists' understanding of the matching process. You will summarize the consensus answers to the three questions for this activity. Since the matching process is a group activity, you may not need to spend much time reviewing the results. You might just ask whether the panelists focused on the GPDs in making their determinations, if there were any disagreements, and if and how those were resolved. One instance where you would want to spend a lot of time on this activity is if you have two different panels setting benchmarks on a single assessment, presumably at different grade levels. If this is the case, vertical alignment between the benchmarks will be critical, and reviewing GPD matches might help indicate challenges that may arise early on (e.g., if a grade three panel matches an item to a lower grade level than the grade two panel). Additional tips on facilitating this discussion are included in **Table 10**.

Figure 26: Tips for Facilitators on Reviewing the Task 2 Matching Results

The panelists will need to agree on the matches, i.e., reach consensus, prior to moving to the benchmarking process. Note that Tasks 1 and 3 involve individual and independent ratings, but Task 2 involves consensus between the panelists on the matches. Ensure that the results from the matches are recorded by each panelist in their assessment instrument/test booklet.

E. TASK 3 - BENCHMARKING

11. Overview on Global Standards and Benchmarking Approach

MATERIALS: FACILITATION SLIDES (PRESENTATION 11), PANELIST WORKSHOP PACKETS

In this presentation, you will explain the main concepts behind global benchmarking in relation to the GPF using several examples. You will explain the first graphic (slide 84) showing the meets benchmark on the two scales – national assessment and GPF – and how the benchmarks link the scales at the identified score points. You will explain the graphic that shows three national assessments with different benchmarks depending on the difficulty of those assessments (slide 85). You will cover the third graphic in the presentation (slide 86) with the percentages of learners in the GPLs (categories) from the assessment data sets, which is used for comparisons, aggregation, and tracking on SDG 4.1.1 and USAID indicators.

Figure 27: Tips for Facilitators on the Global Benchmarking Presentation

This presentation proceeds step-by-step through the assessment scales and GPF graphic, with one benchmark (two levels and percentages) to three benchmarks (four levels and percentages). Make sure the panelists realize that the placement of the benchmarks depends on the difficulty of the assessment. They also need to know that each assessment has a different difficulty level and therefore has different benchmarks in relation to the common scale.

12. Train Panelists on Angoff Method

MATERIALS: FACILITATION SLIDES (PRESENTATION 12), PANELIST WORKSHOP PACKETS (SPECIFICALLY ANNEX H)

In this presentation, you will explain the standardized process for setting benchmarks using the Yes-No version of the Angoff method (see the section on **Task 3**). You will provide background on the Angoff method and how it is used to set global benchmarks on national and international assessments. You will introduce the idea of two rounds of item ratings. You will say that the panelists need to conduct individual and independent ratings of each item to set their benchmarks, which are then averaged to calculate the benchmarks for the panel. You will show panelists how their individual benchmarks are calculated, by adding the number of 'Yes' decisions for each GPL (noting that if, for example, a 'Yes' decision is recorded for the 'meets' GPL, then it will also be recorded for the 'exceeds' GPL). You will also explain that the overall panel benchmark will be calculated by taking the mean average of all of the individual benchmarks.

If there are polytomous items in the assessment, make sure you include the relevant slides from **presentation 12** to explain that they need to consider whether learners at each GPL will achieve each score-point in turn.

Figure 28: Tips for Facilitators on Presenting the Task 3 Angoff Method

Tell the panelists that the same process occurs for the initial benchmarks (Round 1) and final benchmarks (Round 2). Introduce concepts of learner expectations ("should" according to the GPDs and realistic expectations, and "would," based on reality in test situations) along with the need to set the benchmarks at the lowest GPL that matches the expectations learners must meet to answer the item correctly. A flowchart for the ratings and examples is provided for the panelists in the slides and in **Figure 8**, along with ratings tips.

Provide a clear description of the Just Partially Meets (JP), Just Meets (JM), and Just Exceeds (JE) learners in relation to the GPF and how this relates to their own leaners. Make clear that the pupils in their class may not be representative of those described in the GPF, because of different choices made in the curriculum or specific circumstances in the country, for example.

If there are items that were aligned as 'no fit' the discussion in the matching exercise (see **Figure 10**) should help panelists determine whether learners would answer the item correctly, based on their experience.

13. Panelists Undertake Angoff Method on Practice Items

MATERIALS: FACILITATION SLIDES (PRESENTATION 13), PANELIST WORKSHOP PACKETS (SPECIFICALLY ANNEX I)

In this activity, you will review the presentations on global benchmarking and the Angoff method in the panels. You will go over the examples from the presentation and the flowchart, with the Angoff ratings. You will provide ample time for the panelists to practice their item ratings using pre-selected sample items. There are sample reading and mathematics items included in **presentation 13** that you can use for this purpose, or you can select/develop your own. Note that sample items should not be too similar to the actual assessment items that panelists will rate, as this may bias ratings, but it is helpful if they cover similar subconstructs. You will lead discussions of the panelists' ratings in the panel. You will provide an opportunity for the panelists to ask questions and clarify the process.

Figure 29: Tips for Facilitators on the Task 3 Angoff Practice

Emphasize that a key part of this activity relies on the matching from Task 2, in which the panelists matched their items with the lowest GPLs and GPDs in the GPF. These matches provide information for rating the example items (assuming the same example items were used throughout) and, more importantly, the actual items in the next activity. They should ensure that they are matching with both the statements of knowledge and/or skill(s) (Task 1) and the GPDs (Task 2) as well as considering what makes an item easy or difficult (from Task 2), and whether they are reasonably sure that a minimally proficient learner would answer the item correctly. The panelists need to be clear on the process of rating the items before proceeding to Round 1. You should leave plenty of time for questions during this session.

14. Round 1

MATERIALS: FACILITATION SLIDES (PRESENTATION 14), PANELIST WORKSHOP PACKETS (SPECIFICALLY ANNEX I)

In this activity, you will guide the panelists in applying the Angoff method to rate the assessment items. You will explain the item ratings form (as shown in **Table 7**) that they fill out for Round 1 and Round 2. You will reiterate that the panelists need to rate the items individually and independently, which is different from the matching activity in which they reached consensus. You will tell the panelists that variation between them is expected, but it has to be based on a common understanding of the items and the GPF – linked to the outcomes of the matching task, which all panelists should have access to. You will show the panelists how to calculate their own benchmarks. Then, at the end of the day, the data analyst will check those calculations and average them across panelists to generate benchmarks for the panels (see **Annex T** for details on these calculations). Panelists will complete their Round 1 ratings individually but can ask one-on-one questions of facilitators during the process.

Figure 30: Tips for Facilitators on Overseeing Task 3 – Round 1 Ratings

The panelists need to know that they should take their time with the Round 1 ratings. They should be fully aware that collaboration with the other panelists is not accepted in this activity, but that they will have opportunities to discuss their ratings with other panelists before the final round (Round 2). The panelists should ensure that they are matching with the statements of knowledge and/or skill(s) from the GPF and the GPDs. It is also important that in responding to questions from panelists, facilitators only provide guidance on the methodology but not steer panelists in how to rate a particular item.

15. Presentation and Discussion of Round 1 Results and Impact Data

MATERIALS: FACILITATION SLIDES (PRESENTATION 15), PANELIST WORKSHOP PACKETS ANNEX N, ANNEX O AND ANNEX T

In this presentation, you will explain in detail the analyses of the Round 1 benchmarks (all presented anonymously, using panelist IDs): 1) individual panelists' benchmarks and their distributions, 2) panel-level benchmarks (see **Annex T** for details on how to calculate the benchmarks) and normative information (location statistics) of the panelists' benchmarks (details on how to create this graph are included in **Annex O**), 3) item ratings in relation to actual item difficulty (see **Annex N**), 4) averages of the panelists' benchmarks, and 5) impact data with percentages of learners by

GPL based on the benchmarks set by panelists in Round 1. You will engage the panelists in discussions based on each of these analyses. See **Table 10** for tips on how to run this discussion.

Figure 31: Tips for Facilitators on Sharing Round 1 Results

The analyses in the generic slides will need to be replaced with actual analyses based on panelists' ratings in the workshop. Discuss the differences in the panelists' ratings and the reasons behind those differences. Examine the highest and lowest benchmarks from the panelists. You may also want to review individual items for which there was considerable disagreement. Ask volunteers who scored an item one way to share why and volunteers who scored it another way to share why. The idea is to help panelists better understand the different rating options to better inform their Round 2 ratings. Tips for this discussion are included in **Table 11**. Also, have the panelists compare the actual p-values (difficulty statistics) with their ratings to see whether their ratings are consistent with the data – though remembering to remind them to be careful with the interpretation of p-values of items as they may not necessarily be indicative of 'global' difficulty. And, finally, ask them if the impact data are in line with what they would expect from the assessment population. Explore why results might be different from their expectations. Reinforce the idea that they need to have common understandings but not common ratings, i.e., that variation is normal, and the results are averaged to calculate the panel's benchmarks.

If there are outlier panelists whose views are significantly different from their peers, you should have a separate conversation with them to check that they understand the task. Outliers will be removed as part of the final benchmark setting process.

16. Presentation on Angoff Round 2

MATERIALS: FACILITATION SLIDES (PRESENTATION 16), PANELIST WORKSHOP PACKETS (SPECIFICALLY ANNEX I)

In this presentation, you will briefly review the procedures used in the ratings for Round 1 as guidance for Round 2. You will explain that the panelists should examine the ratings for Round 1, take into consideration the data and discussions, and then revise their ratings for Round 2 (it is okay if the panelists do not change their ratings, but they should go through the process of revising each item). You will tell the panelists that they should use Round 1 as a starting point for making their Round 2 ratings.

Figure 32: Tips for Facilitators on Presenting Angoff Round 2

Any changes in panelist ratings from Round 1 to Round 2 should be based on an increased level of understanding, both for the panelists themselves and for the panels. This should lead the panelists to become self-sufficient and become group participants, with the idea that more understanding should lead to greater accuracy and consistency in the benchmarks.

17. Round 2

MATERIALS: FACILITATION SLIDES (PRESENTATION 17), PANELIST WORKSHOP PACKETS (SPECIFICALLY ANNEX I)

In this activity, you will ask the panelists if they have any questions from Round 1 or from the presentation of the Round 1 results. You will tell the panelists to 1) keep a focus on the item content in relation to the GPLs and GPDs, 2) maintain consideration of item difficulty as a basis for making their judgments, 3) provide adjustments where appropriate to their Round 1 ratings based on their individual and independent judgments, and 4) remember to consider how the learners "would" answer the items rather than how they "should" answer the items, and to ensure they are at least "reasonably sure" of their rating. You will have the panelists submit their rating forms – the same rating forms as in Round 1 – to the content facilitators after making their Round 2 item ratings.

Figure 33: Tips for Facilitators on Overseeing Angoff Round 2 Ratings

It is important to monitor the panelists as they conduct their Round 2 ratings. Some panelists may not adequately consider the discussions and data from Round 1. They should take their time and realize that this is their final opportunity to make the most accurate ratings possible based on their knowledge of the assessments, GPF, data, and discussions.

18. Presentation of Round 2 Results

MATERIALS: FACILITATION SLIDES (PRESENTATION 18), PANELIST WORKSHOP PACKETS

ANNEX T

In this presentation, you will summarize the process that has been followed over the workshop and provide the final benchmarks to the panelists, with comments about the changes between Round 1 and Round 2. You will provide the following analyses: 1) Round 1 and Round 2 averages of the panelists' benchmarks, i.e., the benchmarks for the panel(s), 2) an explanation of changes between the rounds, and 3) impact data on the percentage of learners in the GPLs. You will present the results in both tabular and graphic formats. You will lead a short discussion on the results as the final technical activity of the workshop, though panelists will not be able to change their decisions as a result of any discussion.

During the workshop, analysis is carried out using the results from all panelists. However, there may still be panelists who are outliers following round 2. The process for dealing with them is discussed in **Chapter VI**, but this should take place after the workshop is complete so as not to draw attention to the individuals concerned during the workshop.

Figure 34: Tips for Facilitators on Presenting Final Results

The results are more limited than the presentation after Round 1. The main point is to compare the changes from Round 1 to Round 2, as well as discuss whether the panelists believe that the results are reasonable. Again, the lead facilitators and data analyst will need to replace the table in the slides based on the workshop results.

F. EVALUATION

19. Activity – Workshop Evaluation

MATERIALS: FACILITATION SLIDES (PRESENTATION 19), PANELIST WORKSHOP PACKETS (SPECIFICALLY ANNEX R)

Ideally, the evaluation will have been conducted at the end of every task to enable issues to be identified and addressed. If this is not possible, then a single evaluation can be carried out at the end of the workshop. You will provide instructions to the panelists on completing the workshop evaluation form based on when it is being administered.

Panelist IDs will be collected in case a panelist says on the evaluation form that they are not confident in their ratings, which may bring into question that panelist's ratings. However, you should be sure to emphasize to the panelists that the evaluation feedback will not be shared widely or reflect on their participation in the workshop; so, they are strongly encouraged to share their honest feedback. This information will inform future workshops.

Figure 35: Tips for Facilitators on Presenting the Evaluation Form

The lead facilitators and data analyst will compile the evaluation ratings after the workshop. The ratings are mostly in the format of Likert scales, with some areas for open-ended responses. You will provide the results in the technical documentation after the workshop.

20. Workshop Closing and Logistics

MATERIALS: FACILITATION SLIDES (PRESENTATION 20), PANELIST WORKSHOP PACKETS

In this final workshop session, encourage the government officials, donor education officials (if relevant), and implementing partner representatives (if relevant) to provide their final remarks. Hand out certificates to the panelists and thank them for their participation (see **Annex U** for a certificate template). Complete any final logistics and take a group photo, if appropriate.

Figure 36: Tips for Facilitators on Workshop Closing

The officials should be encouraged to talk about next steps with the benchmarks, i.e., using percentages by category for global reporting. There may need to be additional work on using sampling weights to generalize to the population if the assessment was a sample-based assessment rather than a census.

G. ADDITIONAL TIPS FOR HOSTING REMOTE WORKSHOPS

Tips for hosting remote workshops follow.

Logistics

- Ensure panelists have the printed documents they will need to complete the workshop.
- Ensure panelists are able to join via a laptop (strongly preferred) or smartphone so they can see slides and submit tasks. Allow panelists to submit tasks either as soft copies, photos or scans of forms, or (depending on the task) in the body of the text through email or WhatsApp to ensure panelists are able to complete tasks with limited IT challenges.
- Provide data cards to panelists to ensure they have sufficient data to connect to the sessions, and encourage panelists to assess their service far in advance of the workshop in case they need to explore changing providers (if possible).
- Set up a WhatsApp group in advance of the workshop to facilitate announcements, remind panelists of sessions, and ensure ease of communication between workshop sessions when many panelists do not have regular access to email communications.
- Send out calendar invites for all panelists for the sessions.
- Use a teleconference platform that allows for: 1) presenting slides and sharing one's screen, 2) assigning panelists to break-out groups, 3) recording the sessions (for panelists 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 panelists quickly), 4) muting everyone upon entry in the meeting, 5) typed chats, 6) raising one's hand to indicate a question or comment and 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 panelists who missed significant portions of the workshop due to technological issues).
- Host a series of short pre-workshop calls to check small groups of panelists' abilities to connect and troubleshoot any technology issues.
- Have an administrative assistant (NOT a facilitator) manage the teleconference platform, letting participants in, assigning panelists to small groups, etc., as this task can be quite difficult to manage while leading sessions.
- If using breakout rooms, provide backgrounds for each panelist to use that align to their group to easily identify if they are in the incorrect room

Lead Facilitator(s)

- Engage two (or at least one per grade/subject/language of assessment) lead facilitators to help facilitate the small group break-out sessions, to allow panelists to hear from more than one person, and to allow for one person to be tracking questions that come up in the chat while the other facilitator is presenting.

Content Facilitator Training and Interaction

- Plan for a minimum of an 8-hour remote content facilitator training, split into two sessions. However, if it is possible to increase the length of this training to ensure the content facilitators have time to complete each of the activities themselves, it is recommended.

- Have the lead facilitators lead all plenary sessions unless the content facilitators have previous experience with standard setting.
- In addition to the general content facilitator training, scheduling short preparation sessions with the content facilitators to remind them of key issues just before the sessions where they are leading breakout groups is highly recommended.

Pre-sessions

Remote workshops have an advantage in that they can be extended out over a somewhat longer period of time since project teams need not be concerned with hotel and per diem arrangements (unless panelists are meeting in person with only the lead facilitators attending remotely).

- Plan pre-sessions to allow panelists to become more familiar with the GPF and the assessment before undertaking the learner assessment task with three learners who meet the requirements for each GPL.
- Note, in some cases, it may not be possible for panelists to complete the learner assessment task (e.g., due to safety concerns related to COVID-19). In those cases, ensure panelists have an opportunity to take the assessment themselves during one of the pre-sessions or to administer the assessment to children in their homes or communities (e.g., outside using masks) between the pre-sessions and the regular session.
- To aid with the later tasks, ask panelists to write down the names of learners in their class who are described by meets GPDs as part of their inter-session activity.

Discussions

One major disadvantage of remote workshops is that panelists do not have the opportunity to engage in informal discussions with their neighbors, which often highlight misunderstandings or questions, nor do facilitators have the ability to walk around while panelists complete the tasks and look over panelists' shoulders to identify potential misunderstandings. The tips below are focused on trying to address these shortcomings.

- Record of each session and make recordings available for panelists to review in case of connections issues
- If possible, it would be helpful to identify a way of allowing panelists to have conversations between themselves and then come back together to ask facilitators questions. This might be done by going into breakout groups for 10 minutes after every set of slides to discuss and identify any questions or issues. Sessions may need to be extended to accommodate this possibility.
- If possible, it would also be helpful to identify a way of "looking over panelists' shoulders." This might be done by scheduling individual one-on-one 15–30-minute sessions between a lead facilitator and each panelist after the end of the plenary sessions. During these calls, the facilitators can ask panelists to explain the task and describe how they are aligning/matching/ rating each item. This should help identify and correct misunderstandings. It should also ensure panelists who missed portions of the workshop due to technology issues have time to ask questions and become clear on the task.
- Finally, lead facilitators might stay on the call for each workshop session that includes a task assignment (Task 1 and 3, for both rounds) for an hour or so after the session to allow people to do the task on their own but rejoin the call if they have questions.

CHAPTER VI

CHAPTER VI. SELF-ASSESSMENT OF THE WORKSHOP OUTCOMES

A. PRODUCTION OF THE TECHNICAL DOCUMENTATION (AFTER THE WORKSHOP IS COMPLETED)

At the end of the workshop, the following information should be documented using the template in **Annex V**:

- The panelist demographics
- The benchmark(s) agreed by the workshop following the removal of outlier judges
- The identification of any GPLs that appear to be affected by ceiling or floor effects
- The proportion of learners achieving each of the GPLs for which benchmarks were set
- The precision, accuracy, and consistency of the judgements
- The outcomes of the panelists' evaluation
- Self-assessment of the outcomes against the criteria#

Panelist Demographics

The demographics of the panelists will have been captured using the form in **Annex M**. The project team will need to confirm that all panelists met the requirements for participation and, as a group, were sufficiently representative. The requirements for panelists and the requirements for representation are provided in **Chapter IV**.

Benchmarks

During the workshop, benchmarks were created using all panelists ratings. It is possible, however, that one (or more) panelists may be determined to be an outlier who is disproportionately and inappropriately affecting the final benchmark. Outliers are determined using the Tukey fences model (Tukey, 1977) described in **Annex J**. Once outliers have been removed, the final benchmarks are calculated as the mean of the remaining panelists.

Ceiling and Floor Effects

A ceiling effect is when there are insufficient difficult items on an assessment to set the highest benchmark and is characterized by a benchmark that is close to full marks. A floor effect is when there are insufficient easy items on an assessment to set the lowest benchmark and is characterized by a benchmark that is close to 0 marks. In either case, the setting of the benchmark is inappropriate as there is insufficient information from the assessment on the likely performance of learners at this level. The criteria for alignment in the first self-assessment, ensuring there are sufficient items aligned to each of the GPLs, should avoid this issue.

Shulruf (2016) used a simulation study to consider the reliability of the Angoff method. Interpreting this study, if the benchmark is set within 5 marks of the minimum or maximum score, it is likely to have too much error to be considered reliable. As a result, benchmarks that are within 5 marks of the minimum or maximum score on an assessment shall not be accepted for reporting for SDG 4.1.1.

Outcome Data

The final benchmarks should be used, along with the data distributions (calculated to provide impact data during the workshop using **Annex N**), to determine the proportion of learners achieving each of the GPLs.

Precision, Accuracy and Consistency

The formulas in **Annex J** should be used to calculate the following statistics:

- Inter-rater consistency, with the removal of outlier panelists
- The Standard Error for each benchmark
- The confidence intervals for each benchmark

Although a minimum value is set for inter-rater consistency (values at or above 0.7 are considered acceptable) it is not possible to set similar values for Standard Error or confidence intervals as they are depending on the number of items in the assessment. These should be calculated and reviewed by countries themselves to determine if they are appropriate for the assessment. Evidence from previous pilots shows that for an assessment with around 45 score-points, it is possible to achieve a standard error of less than 1.

Evaluation

The evaluation (**Annex R**) contains a series of statements against which panelists record their level of agreement against a 5-point Likert scale (1 – strongly disagree to 5 – strongly agree) grouped into sections:

- GPF training
- Assessment training
- Alignment task
- Matching task
- Policy linking training
- Round 2 outcomes

There is also an overall evaluation described on a 4-point scale (1 – very uncomfortable to 4 – very comfortable).

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

Self-Assessment Criteria

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

- Criterion 1 – Did all panelists meet the requirements for participation? (YES / NO)
- Criterion 2 – Were the group of panelists sufficiently representative in terms of the characteristics agreed by the country? (YES / NO)
- Criterion 3 – Were all outliers removed before calculating the final benchmarks? (YES / NO)
- Criterion 4 – Were benchmarks only set for GPLS that don't exhibit floor or ceiling effects? (YES / NO)
- Criterion 5 – Is the inter-rater consistency statistic greater than or equal to 0.7? (YES / NO)
- Criterion 6 – Has the Standard Error for each benchmark been calculated and reviewed to be determined as appropriate? (YES / NO)
- Criterion 7 – Has the confidence interval for each benchmark been calculated and reviewed to be determined as appropriate? (YES / NO)
- Criterion 8 – Was the mean average score for each section of the evaluation greater than or equal to 4? (YES / NO)
- Criterion 9 – Was the mean average score for the overall evaluation greater than or equal to 3? (YES / NO)

In order to be considered eligible for reporting against SDG 4.1.1, countries will need to be able to answer YES to each of the criteria questions. If the answer to one or more of the questions is NO, then countries will need to consider running the workshop again or look at other alternatives for reporting.

B. SUBMIT EVIDENCE TO UIS

Table 12 indicates the information countries will need to provide to UIS for SDG 4.1.1 reporting.

Table 12: 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 12** was generated in accordance with the requirements for policy linking:

- Self-assessment – appropriateness of assessment (see **Annex C**)
- Self-assessment – workshop outcomes (see **Annex V**)
- Policy linking workshop report (see **Annex V**)

BIBLIOGRAPHY

BIBLIOGRAPHY

Adams, R., Jackson, J., & Turner, R. (2018). *Learning progressions as an inclusive solution to global education monitoring*. Melbourne, Australia: Australian Council for Educational Research.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

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

Bejar, I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement, 7*(3), 303-310.

Berk, R. (1996). Standard setting: The next generation (where few psychometricians have gone before!). *Applied Measurement in Education, 9*(3), 215-225.

Brennan, R. L. (2004). *BB-CLASS v.1.1 [Computer program]*. Iowa City: Center for Advanced Studies in Measurement and Assessment, University of Iowa.

Brennan, R. L., & Wan, L. (2004). Bootstrap procedures for estimating decision consistency for single administration complex assessments. *CASMA Research Report No. 7*. Iowa City: University of Iowa.

Brown, J.D. (1989). Criterion-referenced test reliability. *University of Hawai'i Working Papers in ESL, 8*(1), 79-113.

Burry-Stock, J.A., Shaw, D.G., Laurie, C., & Chissom, B.S. (1996). Reader agreement indexes for performance assessments. *Educational and Psychological Measurement, 56,* 251-262.

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

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

Engelhard, G. & Stone, G. (1998). Evaluating the quality of ratings obtained from standard-setting judges. *Educational and Psychological Measurement, 58*(2), 179-196.

Ferdous, A. (2019). *Setting performance standards for reading fluency in Lebanon*. Paper presented for the annual meeting of the Comparative and International Education Society. San Francisco, CA.

Ferdous, A. & Buckendahl, C. (2013). Evaluating panelists' standard setting perceptions in a developing nation. *International Journal of Testing, 13(1),* 4-18.

Ferdous, A. & Plake, B. (2005). Understanding the factors that influence decisions of panelists in a standard setting study. *Applied Measurement in Education, 18(3)*, 257-267.

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

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378-382.

Frisbie, D.A. (2003). Checking the alignment of an assessment tool and a set of content standards. Iowa City, IA: University of Iowa.

Giraud, G., Impara, C., & Plake, B. (2005). Teachers' conceptions of the target examinee in Angoff standard setting. *Applied Measurement in Education, 18*(3), 223–232.

Halpin, G. & Halpin, G. (1983). *Reliability and validity of ten different standard setting procedures*. Paper presented at the American Psychological Association, Anaheim, CA.

Hambleton, R. (2001). The next generation of the ITC Test Translation and Adaptation Guidelines. *European Journal of Psychological Assessment, 17*(3), 164-172.

Hambleton, R. (2008). Psychometric models, test designs and item types for the next generation of educational and psychological tests. In D. Bartram and R. Hambleton (Eds.) *Computer-Based Testing and the Internet: Issues and Advances* (pp. 77-89). New York, NY: John Wiley & Sons Ltd.

Hambleton, R. & Bourque, M. (1991). The LEVELS of mathematics achievement: Initial performance standards for the 1990 NAEP mathematics assessment: Vol. III. Technical report. Washington, D.C.: National Assessment Governing Board.

Hambleton, R. & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.) *Educational Measurement* (4th ed.). Westport: American Council on Education & Praeger Publishers.

Hambleton, R. & Plake, B. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8*(1), 41-55.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement, 27*, 345-359.

Hurtz, G. & Hertz, N. (1999). How many raters should be used for establishing cutoff scores with the Angoff method? A generalizability theory study. *Educational and Psychological Measurement, 59*(6), 885–897.

Jaeger, R. (1989). Certification of learner competence. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Jaeger, R. (1995). Setting performance standard through two-stage judgmental policy capturing. *Applied Measurement in Education, 8*(1), 15-40.

Kahl, S., Crockett, T., DePascale, C., & Rindfleisch, S. (1995). *Setting standards for performance levels using learner-based constructed-response method*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Kane, M. (1998). Choosing between examinee-centered and test-centered standard-setting methods. *Educational Assessment, 5*(3), 129-145.

Lewis, D., Mitzel, H., Green, D., & Patz, R. (1999). *The bookmark standard setting procedure*. Monterey, CA: McGraw-Hill.

Livingston, S. & Zieky, M. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.

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

Lorge, I. & Kruglov, L. (1953). The improvement of estimates of test difficulty. *Educational and Psychological Measurement, 13*(1), 34-46.

Management Systems International (2019). *Policy linking method: Linking assessments to a global standard*. U.S. Agency for International Development (USAID), Washington, D.C.

Mehrens, W. & Popham, W. (1992). How to evaluate the legal defensibility of high-stakes tests. *Applied Measurement in Education, 5*(3), 265-283.

Norcini, J., Shea, J., & Grasso, L. (1991). The effect of numbers of experts and common items on cutting score equivalents based on expert judgement. *Applied Psychological Measurement, 15*(3), 241-246.

Pitoniak, M. (2003). Standard setting methods for complex licensure examinations. *Doctoral Dissertations 1896 – February 2014*. University of Massachusetts, Amherst.

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

Plake, B. & Hambleton, R. (2000). A standard setting method designed for complex performance assessments: Categorical assignments of learner work. *Educational Assessment, 6*(3), 197-215.

Plake, B., Hambleton, R., & Jaeger, R. (1997). A new standard-setting method for performance assessments: The dominant profile judgment method and some field-test results. *Educational and Psychological Measurement, 57*(3), 400-411.

Rasch, G. (1960) Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institut.

Schaeffer, G. & Collins, J. (1984). *Setting performance standards for high-stakes tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Shulruf, B., Wilkinson, T., Weller, J. et al. (2016) Insights into the Angoff method: results from a simulation study. *BMC Med Educ* 16, 134.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement, 13*, 265-276.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability for mastery tests. *Journal of Educational Measurement, 25*, 47-55.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, Massachusetts.

UNESCO. (2018a). *Global content framework of reference for reading: Global consultation*. Paper presented at the fifth meeting of the Global Alliance to Monitor Learning (GAML), Hamburg, Germany.

UNESCO. (2018b). *Global content framework of reference for mathematics: Global consultation*. Paper presented at fifth meeting of the Global Alliance to Monitor Learning (GAML), Hamburg, Germany.

ANNEXES

ANNEX A – RELATED RESOURCES

- Global Proficiency Framework for Mathematics: Grades 1 to 9[24]
- Global Proficiency Framework for Reading: Grades 1 to 9[25]
- Workshop Facilitation Slides: Policy Linking for Measuring Global Learning Outcomes with the Timed Assessment(s)
- Workshop Facilitation Slides: Policy Linking for Measuring Global Learning Outcomes with the Untimed Assessment(s)
- Content Facilitator Slides
- Workshop Preparation Checklist
- Alignment Rating Form for Task 1
- Item Rating Forms
- Panelist Demographic Information Form
- Self-Assessment Template Summary Forms
- Templates
    - Invitation Letter for Observers
    - Invitation Letter for Workshop Panelists
    - Certificate of Appreciation

[24] https://gaml.uis.unesco.org/wp-content/uploads/sites/2/2021/03/Global-Proficiency-Framework-Math.pdf

[25] https://gaml.uis.unesco.org/wp-content/uploads/sites/2/2021/03/Global-Proficiency-Framework-Reading.pdf

ANNEX B – GLOBAL MINIMUM PROFICIENCY LEVELS

Below Partially Meets Minimum Proficiency: Learners lack the most basic knowledge and skills. As a result, they generally cannot complete the most basic tasks.

Partially Meets Minimum Proficiency: Learners have partial knowledge and skills. As a result, they can partially complete basic tasks.

Meets Minimum Proficiency: Learners have sufficient knowledge and skills. As a result, they can successfully complete basic tasks.

Exceeds Minimum Proficiency: Learners have superior knowledge and skills. As a result, they can successfully complete complex tasks.

A fuller description of the MPLs can be found in the MPLs Unpacked[26] document.

[26] https://tcg.uis.unesco.org/wp-content/uploads/sites/4/2022/11/WG_GAML_4_MPLs-Unpacked_ACER.pdf

ANNEX C – SELF-ASSESSMENT TEMPLATE SUMMARY (APPROPRIATENESS OF ASSESSMENT)

Assessment Instrument	[Insert name of instrument]
Jurisdiction	[Insert jurisdiction where assessment instrument is administered]
Grade	[Insert grade assessed by instrument]
SDG 4.1.1 level	a / b / c [delete as appropriate]
Subject	Mathematics / Reading [delete as appropriate]
GPLs being set	Partially meets / Meets / Exceeds [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]
Does the assessment instrument meet the quantitative requirements of the specification?	Yes / No [delete as appropriate]
Are the assessment and curriculum frameworks aligned?	Yes / No [delete as appropriate]

Criterion 1 rating: Insufficient / Minimal / Good / Excellent [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]
Is there evidence that the items have been reviewed to ensure appropriateness for relevant subgroups of the population?	Yes / No [delete as appropriate]

Criterion 2 rating: Insufficient / Minimal / Excellent [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]
Was the minimum detectable effect size calculated and thought through ahead of finalizing sample size calculations?	Yes / No / Not Applicable [delete as appropriate]

Criterion 3 rating: Insufficient / Minimal / Excellent [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 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 / Minimal [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]
Is the level of agreement between human-scorers and pre-agreed scores, or double-marked scores, over 80%?	Yes / No / Not Applicable [delete as appropriate]
Are other measures of reliability on the assessment, e.g., classification constancy, classification accuracy or inter-rater reliability, at levels that are consistent with international best practice?	[Insert details of other reliability measures] Yes / No / Not Applicable [delete as appropriate]

Criterion 5 rating: Insufficient / Minimal / Good / Excellent [delete as appropriate]

Overall Self-Assessment Rating

Criteria	Insufficient	Minimal	Good	Excellent
Criterion 1 – Alignment	☐	☐	☐	☐
Criterion 2 – Item review	☐	☐		☐
Criterion 3 – Sample	☐	☐		☐
Criterion 4 – Administration	☐	☐		
Criterion 5 – Reliability	☐	☐	☐	☐
Overall Self-Assessment Rating	☐	☐	☐	☐

ANNEX D – WORKSHOP PREPARATION CHECKLIST

Table 13: Workshop Preparation Checklist

Activity	Responsible	Deadline	✓	Comments
1. Select and contract facilitators and data analyst				
a. Identify and contract lead facilitators				
b. Identify and contract content facilitators				
c. Identify and contract coordinator, if needed for logistical preparation				
2. Prepare workshop logistics				
a. Determine whether workshop will be in person, mixed (panelists and content facilitators in person and lead facilitators remote), or remote				
b. Identify and secure physical space or remote conferencing service				
c. Determine what food/refreshments will be provided to participants and procure				
d. Arrange or procure materials, such as notebooks, pens, flipcharts, folders, name tags/tents, banners				
e. Identify per diems, travel budget, phone card/data allowances (for remote workshops), hotel costs, etc., and agree on amounts for panelists and observers with government/ assessment agency and donor officials (if applicable)				
f. Make hotel arrangements, if needed				
g. Make facilitator travel arrangements, if needed				
h. Make panelist/observer travel arrangements, if needed				
i. Inspect venue to plan for workshop and locations of breakout rooms				
j. Identify method for receiving funds in country (if necessary); this might involve a wire or cash transfer				
k. Make cash/wire transfer, if needed				
l. Transfer funds to participants; for in-person workshops, this is often done during the workshop				
3. Select and invite participants				
a. Finalize teacher panelist list				
b. Finalize curriculum specialist panelist list				
c. Finalize observer list				
Draft pre-workshop assessment activity instructions, if the workshop will be in person				
e. Prepare a practice assessment if assessment security is a concern (See **Figure 13** for more information)				
f. Prepare and distribute invitations, with pre-workshop assessment instructions, to teacher panelists				
g. Prepare and distribute invitations, with pre-workshop assessment instructions, to specialist panelists				

Activity	Responsible	Deadline	✓	Comments
h. Prepare and distribute invitations for observers				
4. Prepare Materials				
a. Finalize and print the agenda (and distribute, if the workshop will be remote)				
b. Finalize and print the acronym list (and distribute, if the workshop will be remote)				
c. Finalize and print the glossary (and distribute, if the workshop will be remote)				
d. Assign panelist IDs (and distribute, if the workshop will be remote)				
e. Translate reading GPF into local language, if necessary				
f. Tailor the GPF to the relevant grades/ subjects, print, (and distribute, if the workshop will be remote)				
g. Develop practice passages/questions for the slides				
h. Finalize ratings forms (alignment and item rating forms), print, (and distribute if the workshop will be remote)				
i. Print/distribute assessment instruments, following security protocols				
j. Finalize and print workshop evaluation forms				
k. Analyze data to produce data distributions, item difficulty data, etc.				
l. Finalize facilitation slides and print				
m. Finalize daily attendance forms and print				
5. Train Content Facilitators				
a. Finalize slides for content facilitator training				
b. Make logistical arrangements for content facilitator training				
c. Train content facilitators				

Coordinator: _____

Lead Facilitator: _____

ANNEX E – WORKSHOP ACTIVITY PLANNER

Table 14: Workshop Activity Planner

WEEK-BY-WEEK TIMELINE FOR PREPARATION ACTIVITIES FOR POLICY LINKING WORKSHOPS				
Number	**Activity**	**Role/Responsibility**	**Workshop format for which activity is relevant**	**Deadline (number of weeks before workshop)**
Four Weeks before the Workshop				
1	Initiate contact with country	UIS/Donor Organization (DO)	All	At least 4 weeks
2	Decide on which assessment, grade level, and language to focus	Country with UIS/DO and Delivery Partner (DP) support	All	4 weeks
3	Decide what format the workshop will take (in person, all remote or hybrid with participants gathering in one or multiple places) and the timing of the workshop	Country with UIS/DO/DP support	All	4 weeks
4	Identify local Content Facilitators	Country	All	4 weeks
5	Identify interpreters (if relevant)	Country	All	4 weeks
6	Identify logistician (if needed)	Country	All	4 weeks
7	Identify panelists (both teachers and content specialists), including collecting their contact information; ensure panel is representative	Country	All	4 weeks
8	Tailor the GPF to the relevant grades/subjects	DP	All	4 weeks
3-4 Weeks before the Workshop (the earlier the better)				
9	Draft agenda	DP	All	3-4 weeks
10	Provide feedback on draft agenda	Country	All	3-4 weeks
11	Finalize agenda	DP	All	3-4 weeks
12	Invite panelists	Country, UIS/DO, or DP - depending on country's preference	All	3-4 weeks (depending on country norms)
3 Weeks before the Workshop				
13	Identify and invite any workshop observers - from other donors, Ministries, etc.	Country with UIS/DO/DP support	All	3 weeks
14	Identify other potential costs for the workshop, including phone/internet cards, transportation, lodging, per diems, meals, water, and materials during the workshop	Country	All	3 weeks
15	Submit budget to UIS/DO	Country	All	3 weeks
16	UIS/DO and DP complete NDAs	UIS/DO and DP	All	3 weeks
17	Send assessment instruments to UIS/DO and DP	Country	All	3 weeks
18	Provide Ministry logo for certificates and banner (the latter only for in person and hybrid workshops) and determine who from the Ministry will sign	Country	All	3 weeks

19	Translate GPF into local language, if necessary and back-translate to check quality	Country	All	3 weeks
Number	**Activity**	**Role/Responsibility**	**Workshop format for which activity is relevant**	**Deadline (number of weeks before workshop)**
20	Draft workshop slides, including example items, and rating forms to send to UIS/DO for review	DP	All	3 weeks
21	Identify and secure physical space for workshop	Country	Hybrid	3 weeks
2 Weeks before the Workshop				
22	Finalize contracts with local Content Facilitators, interpreters, and logistician (the latter two, if applicable)	UIS/DO	All	2 weeks
23	Review workshop slides, including example items, and rating forms and send feedback to DP	UIS/DO	All	2 weeks
24	Finalize MOU with country based on approved budget	UIS/DO	All	2 weeks
25	Draft certificates and banner	DP	All	2 weeks
26	Finalize item rating forms and slides based on UIS/DO feedback	DP	All	2 weeks
27	Send data to UIS/DO and DO (if possible)	Country	All	2 weeks
28	Confirm panelist participation	Country	All	2 weeks
29	Reserve hotel rooms for panelists, if needed	Country	In Person and Hybrid	2 weeks
30	Decide on remote conferencing service for workshop	Country	Remote	2 weeks
31	Transfer funds and/or phone/internet cards to participants	Country	Remote	2 weeks
32	Finalize slides for content facilitator training	DP	All	2 weeks
33	Make logistical arrangements for content facilitator training	DP	All	2 weeks
1 Week before the Workshop				
34	Prepare funds to disperse to participants for per diems, travel, etc.	Country	In Person and Hybrid	1 week
35	Determine what food/refreshments will be provided to participants and procure	Country	In Person and Hybrid	1 week
36	Arrange or procure materials, such as notebooks, pens, flipcharts, folders, name tags/tents	Country	In Person and Hybrid	1 week
37	Inspect venue to plan for workshop, locations of breakout rooms, and to test remote access (if applicable, e.g., if not a government facility)	Country	In Person and Hybrid	1 week
38	Finalize the agenda (with any last-minute changes)	DP	All	1 week
39	Finalize the agenda, acronym list, glossary, assessment, GPF, rating forms, evaluation forms, slides with notes fields, certificates, banners, and any other documents	DP	All	1 week
40	Print the agenda, acronym list, glossary, assessment, GPF, rating forms, evaluation forms, slides with notes fields, certificates, banners, and any other documents	Country	All	1 week
41	Distribute the agenda, acronym list, glossary, assessment, GPF, rating forms, evaluation forms, slides with notes fields, certificates, banners, and any other documents	Country	Remote	1 week
42	Assign panelist IDs	DP	All	1 week
43	Distribute panelist IDs	Country	Remote	1 week
44	Analyze data to produce data distributions, item difficulty data, etc.	DP	All	1 week

45	Finalize facilitation slides and print	DP	In Person and Hybrid	1 week
Number	**Activity**	**Role/Responsibility**	**Workshop format for which activity is relevant**	**Deadline (number of weeks before workshop)**
46	Finalize daily attendance forms	DP	All	1 week
47	Print daily attendance forms	Country	In Person and Hybrid	1 week
48	Train Content Facilitators	DP	All	1 week
A Few Days before the Workshop				
49	Remote platform testing with panelists or venue to make sure are participants can access the platform and don't need technical support	All	All	A few days
Workshop begins				

ANNEX F – BUDGET ESTIMATION TEMPLATE

COST ESTIMATION TEMPLATE FOR POLICY LINKING WORKSHOP					
Proposed Date/Time :					
Sl. No.	**Activities**	**No. of Individuals**	**No. of Days**	**Unit cost**	**Estimated Budget**
Staff and Consultant Time Costs					
1	Content Facilitators				
2	Logistician, if needed				
3	Interpreters, if needed				
Staff and Consultant Participation Costs					
4	DSA for above participants/facilitators/interpreters and logistician				
5	Honorarium Participants				
6	Honorarium Facilitators				
7	Transportation for participants/facilitators/interpreters				
8	Hotel rooms (if it is in-person)				
9	Workshop Kit (includes folder, note pad, pen, pencil, eraser, sharper, etc.)				
10	Lunch, if relevant				
11	Tea/Coffee, snacks, and water				
Pre-Workshop Logistics Costs					
12	Printing and delivering (sending) invitations/nominations, if necessary				
13	Translating GPF into the local language				
14	Translating any other materials in local language				
Workshop Venue and Materials Costs					
15	Venue rental, if needed				
16	Printing of slides, agenda, GPF, assessment, rating forms, acronyms list, glossary, attendance sheets, and certificates				
17	Printing of banner				
18	Hire of Video Conferencing Devices				
19	Communications and Internet (lump sum)				
20	Administrative support (3%)				
Total Budget					**-**

ANNEX G – WORKSHOP FACILITATION SLIDES

Facilitators will need to update/adapt all slides marked with a yellow plus sign for use in their specific context.

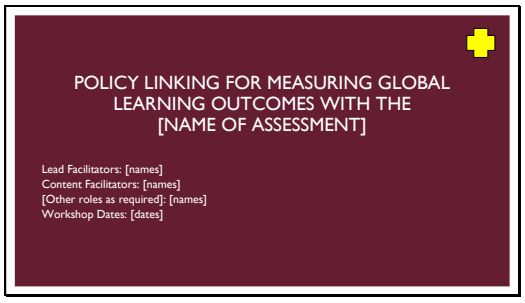
Instructions on how to do so are included in **BOLD** in the notes section of each slide.

Facilitator notes are also included in the notes section.

Brackets, like these [ ] have been used to designate areas that need updating/adapting on the actual slides.

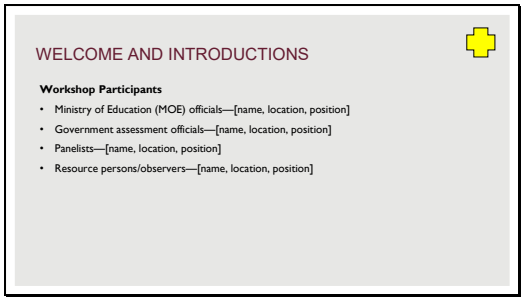
PRESENTATION 1 – WELCOME AND INTRODUCTIONS

Slide 1

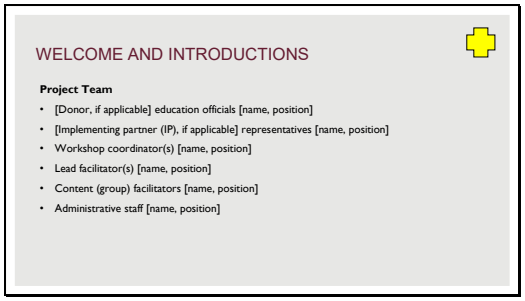


Slide 2

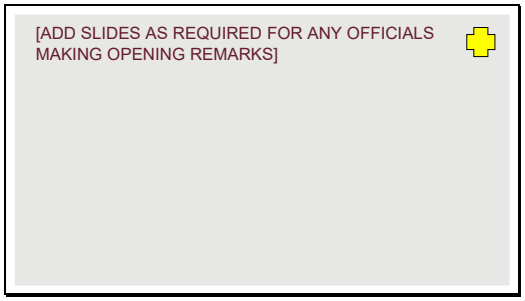
Slide 3



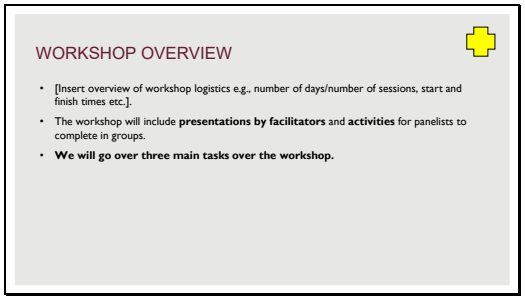
Slide 4



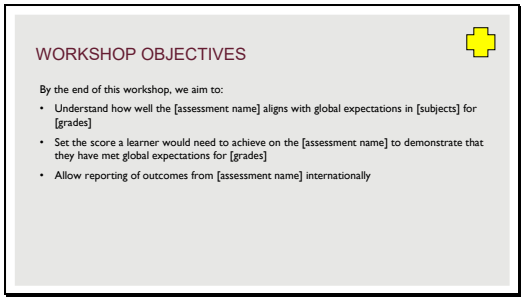
Slide 5



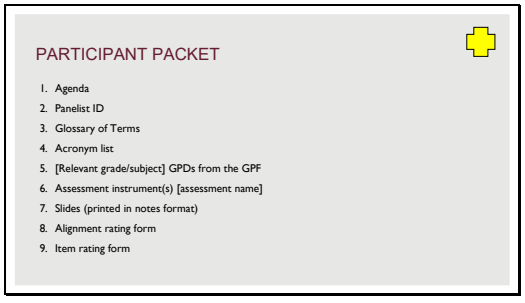
Slide 6



Slide 7



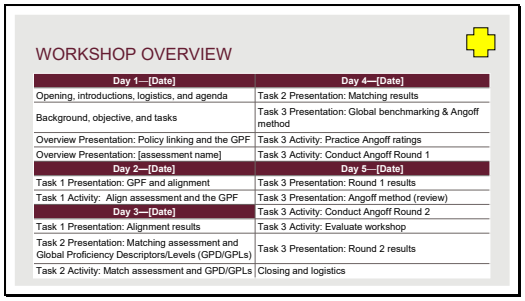
Slide 8



PRESENTATION 2 – OVERVIEW OF AGENDA, OBJECTIVES AND METHOD

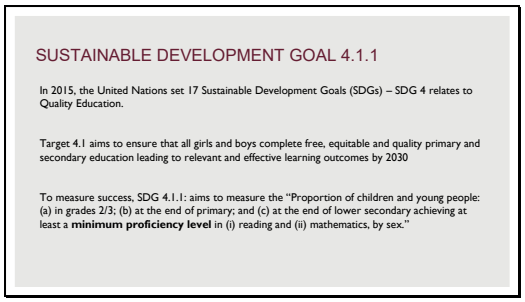
Slide 9

Slide 10

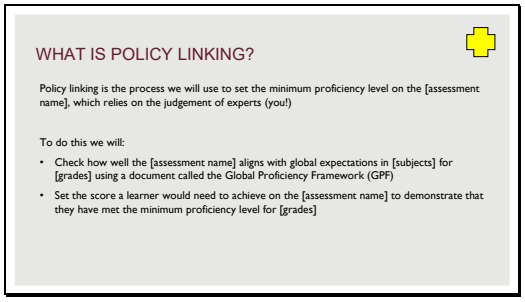


Slide 11

Slide 12



Slide 13



Slide 14

KEY TERMINOLOGY

- **Global Proficiency Framework (GPF)** – a document that describes globally-agreed expectations at the end of each grade in reading and mathematics
- **Minimum Proficiency Levels (MPL)** – the minimum standards expected of learners at the end of grade2/3, end of primary and end of lower secondary
- **Alignment** – whether the content of an assessment is similar to the globally-agreed definitions of what constitutes reading and mathematics
- **Benchmark** – the score a leaner needs to achieve on an assessment to demonstrate that they are achieving a particular standard e.g., the MPL
- **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.

Slide 15

POLICY LINKING TIMELINE

- September 2017: A UNESCO Institute for Statistics (UIS) stakeholder workshop proposed policy linking as a method for setting global benchmarks on each assessment based on a common proficiency scale
- August 2018: Joint U.S. Agency for International Development (USAID)—UIS stakeholder workshop discussed policy linking for reporting minimum proficiency through SDG 4.1.1 and USAID indicators
- April/May 2019: Global Proficiency Framework (GPF) drafted
- September 2019: Draft Policy Linking for Measuring Global Learning Outcomes Toolkit (PLT) written

Slide 16

BACKGROUND ON POLICY LINKING: TIMELINE

- October 2019 – September 2020: Five pilot workshops conducted
- June – October 2020: GPF and PLT updated based on pilots
- October 2020 – August 2022: Additional workshops held to continue to pilot the GPF and PLT
- August 2022 – March 2023: PLT updated

Slide 17

THE KEY TASKS FOR POLICY LINKING WORKSHOP

Familiarization → Get to know the GPF and [assessment name]

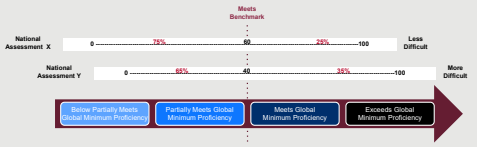
Alignment and Matching → Check the alignment between the [assessment name] and the GPF (TASK 1) and match the assessment items to the descriptors in the GPF (TASK 2)

Benchmarking → Set the benchmarks on the [assessment name] through two rounds of rating (TASK 3)

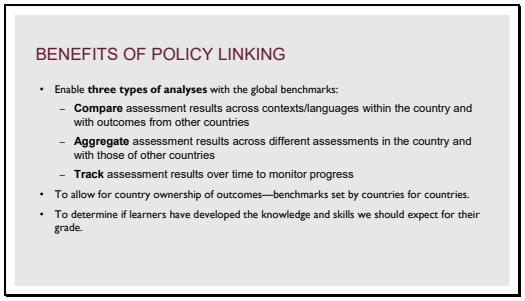
Slide 18

SETTING GLOBAL BENCHMARKS FOR MULTIPLE ASSESSMENTS

- Setting **global benchmarks** on different assessments links each assessment to the GPF.
- Positioning global benchmarks on the assessment scale depends on the difficulty of the assessment in relation to the GPF, as determined through judgments by the panelists.

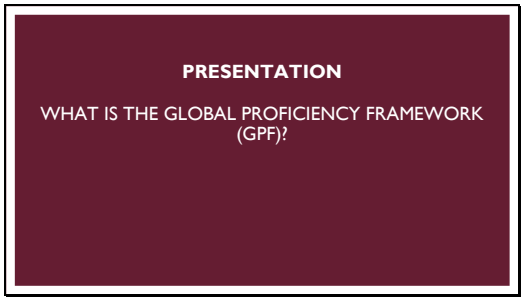


Slide 19

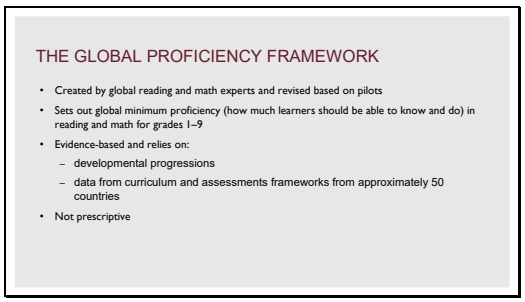


PRESENTATION 3 – FAMILIARIZATION WITH GPF

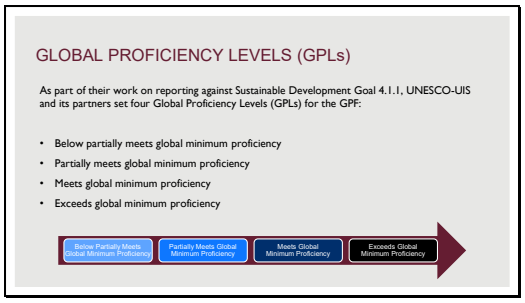
Slide 20



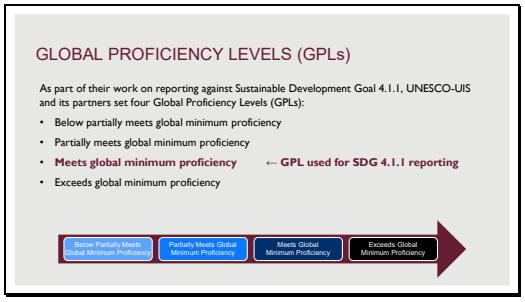
Slide 21



Slide 22



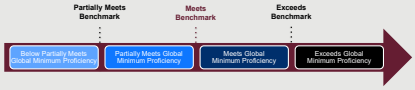
Slide 23



Slide 24

GPF OVERVIEW

- The Global Proficiency Framework (GPF) sets out the agreed domains, constructs, subconstructs, and knowledge and/or skills (sometimes called content standards) for each grade level
- For each knowledge and/or skill, there are Global Proficiency Descriptors (GPDs) (sometimes called performance standards) that detail expectations for the top 3 GPLs (partially meets, meets, and exceeds).



Slide 25

GPF DOMAINS

There are 5 domains in the GPF for mathematics:

- Number and Operations
- Measurement
- Geometry
- Statistics and Probability
- Algebra

Slide 26

GPF DOMAINS

There are 3 domains in the GPF for reading:

- Comprehension of spoken or signed language
- Decoding
- Reading Comprehension

Slide 27

GPF CONSTRUCTS AND SUBCONSTRUCTS

Domain		Construct		Subconstruct
N	Number and operations	N1	Whole numbers	N1.1 Identify and count in whole numbers, and identify their relative magnitude
				N1.2 Represent whole numbers in equivalent ways
				N1.3 Solve operations using whole numbers
				N1.4 Solve real-world problems involving whole numbers
		N2	Fractions	N2.1 Identify and represent fractions using objects, pictures, and a symbols, and identify relative magnitude
				N2.2 Solve operations using fractions
				N2.3 Solve real-world problems involving fractions
		N3	Decimals	N3.1 Identify and represent decimals using objects, pictures, and symbols, and identify relative magnitude
				N3.2 Represent decimals in equivalent ways (including fractions and percentages)
				N3.3 Solve operations using decimals
				N3.4 Solve real-world problems involving decimals
		N4	Integers	N4.1 Identify and represent integers using objects, pictures, or symbols, and identify relative magnitude
				N4.2 Solve operations using integers
				N4.3 Solve real-world problems involving integers
		N5	Exponents and roots	N5.1 Identify and represent quantities using exponents and roots, and identify the relative magnitude
				N5.2 Solve operations involving exponents and roots
		N6	Operations across number	N6.1 Solve operations involving integers, fractions, decimals, percentages, and exponents

Slide 28

GPF CONSTRUCTS AND SUBCONSTRUCTS

Domain		Construct		Subconstruct
M	Measurement	M1	Length, weight, capacity, volume, area, and perimeter	M1.1 Use non-standard and standard units to measure, compare, and order
				M1.2 Solve problems involving measurement
		M2	Time	M2.1 Tell time
				M2.2 Solve problems involving time
		M3	Currency	M3.1 Use different currency units to create amounts
G	Geometry	G1	Properties of shapes and figures	G1.1 Recognize and describe shapes and figures
		G2	Spatial visualizations	G2.1 Compose and decompose shapes and figures
		G3	Position and direction	G3.1 Describe the position and direction of objects in space
S	Statistics and probability	S1	Data management	S1.1 Retrieve and interpret data presented in displays
				S1.2 Calculate and interpret central tendency
		S2	Chance and probability	S2.1 Describe the likelihood of events in different ways
				S2.2 Identify permutations and combinations
A	Algebra	A1	Patterns	A1.1 Recognize, describe, extend, and generate patterns
		A2	Expressions	A2.1 Evaluate, model, and compute with expressions
		A3	Relations and functions	A3.1 Solve problems involving variation (ratio, proportion, and percentage)
				A3.2 Demonstrate an understanding of equivalency
				A3.3 Solve equations and inequalities
				A3.4 Interpret and evaluate functions

Slide 29

GPF CONSTRUCTS AND SUBCONSTRUCTS

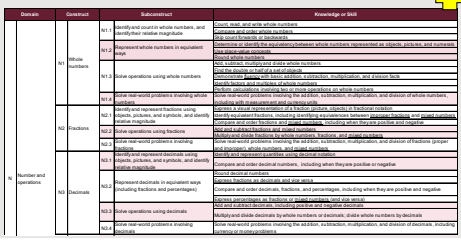
Domain		Construct		Subconstruct	
C	Comprehension of spoken or signed language	C1	Retrieve information at word level	C1.1	Comprehend spoken and signed language at the word or phrase level
				C1.2	Recognize the meaning of common grade-level words in a short, grade-level continuous text read to or signed for the learner
		C2	Retrieve information at sentence or text level	C2.1	Retrieve explicit information in a short grade-level continuous text read to or signed for the learner
		C3	Interpret information at sentence or text level	C3.1	Interpret information in a short grade-level continuous text read to or signed for the learner
D	Decoding	D1	Precision	D1.1	Identify symbol-sound/fingerspelling and/or symbol-morpheme correspondences
				D1.2	Decode isolated words
		D2	Fluency	D2.1	Say or sign a grade-level continuous text at pace and with accuracy
R	Reading comprehension	R1	Retrieve information	R1.1	Recognize the meaning of common grade-level words
				R1.2	Retrieve explicit information in a grade-level text by direct- or close-word matching
				R1.3	Retrieve explicit information in a grade-level text by synonymous word matching
		R2	Interpret information	R2.1	Identify the meaning of unknown words and expressions in a grade-level text
				R2.2	Make inferences in a grade-level text
				R2.3	Identify the main and secondary ideas in a grade-level text
		R3	Reflect on information	R3.1	Identify the purpose and audience of a text
				R3.2	Evaluate a text with justification
				R3.3	Evaluate the status of claims made in a text
				R3.4	Evaluate the effectiveness of a text

Slide 30

GPF KNOWLEDGE, SKILLS, AND STANDARDS

- **Statements of knowledge and/or skills (content standards)**: WHAT content learners are expected to know and be able to do as described in the GPF.
  - Example: Grade 3 learners **should be able to** demonstrate fluency with basic addition, subtraction, multiplication, and division facts.
- **Global Proficiency Descriptors (performance standards)**: HOW MUCH content do learners need to know and be able to demonstrate in relation to knowledge or skills.
  - Example: Grade 3 learners who **"meet global minimum proficiency"** should be able to demonstrate fluency with addition and subtraction within 20 and add and subtract within 100 (i.e., where the sum or minuend does not surpass 100), with and without regrouping, and represent these operations with objects, pictures, or symbols.

Slide 31



Domain		Construct		Subconstruct		Knowledge or Skill
N	Number and operations	N1	Whole numbers	N1.1	Identify and count in whole numbers, and identify their relative magnitude	Count, read, and write whole numbers
						Compare and order whole numbers
						Skip count forwards or backwards
				N1.2	Represent whole numbers in equivalent ways	Determine or identify the equivalency between whole numbers represented as objects, pictures, and numerals
						Use place-value concepts
						Round whole numbers
				N1.3	Solve operations using whole numbers	Add, subtract, multiply and divide whole numbers
						Use the double or half of a set of objects
						Decontextualize basics with basic addition, subtraction, multiplication, and division facts
						Estimate and check answers
						Perform calculations involving two or more operations on whole numbers
				N1.4	Solve real-world problems involving whole numbers	Solve real-world problems involving the addition, subtraction, multiplication, and division of whole numbers, including with measurement and currency units
		N2	Fractions	N2.1	Identify and represent fractions using objects, pictures, and symbols, and identify relative magnitude	Express a visual representation of a fraction (picture, objects) in fraction notation
						Identify equivalent fractions, including identifying equivalences between improper fractions and mixed numbers
						Compare and order fractions and mixed numbers, including when they are positive and negative
				N2.2	Solve operations using fractions	Add and subtract fractions and mixed numbers
						Multiply and divide fractions by whole numbers, fractions, and mixed numbers
				N2.3	Solve real-world problems involving fractions	Solve real-world problems involving the addition, subtraction, multiplication, and division of fractions (proper and improper), whole numbers, and mixed numbers
		N3	Decimals	N3.1	Identify and represent decimals using objects, pictures, and symbols, and identify relative magnitude	Identify and represent quantities using decimal notation
						Compare and order decimal numbers, including when they are positive or negative
						Round decimal numbers
				N3.2	Represent decimals in equivalent ways (including fractions and percentages)	Express fractions as decimals and vice versa
						Compare and order decimals, fractions, and percentages, including when they are positive and negative
						Express percentages as fractions or mixed numbers (and vice versa)
				N3.3	Solve operations using decimals	Add and subtract decimals, including positive and negative decimals
						Multiply and divide decimals by whole numbers or decimals; divide whole numbers by decimals
				N3.4	Solve real-world problems involving decimals	Solve real-world problems involving the addition, subtraction, multiplication, and division of decimals, including currency or money problems

Slide 32

GLOBAL PERFORMANCE DESCRIPTORS (GPDs)

- For each subconstruct and knowledge or skill, there are descriptions of performance at the partially meets, meets, and exceeds GPLs.
- For example, in grade [X] in the [name] domain, for the [name] construct, and [name] subconstruct of the GPF has the following:

Subconstruct	Partially Meets	Meets	Exceeds
Represent whole numbers in equivalent ways	Identify and represent the equivalence between whole quantities up to 30 represented as objects, pictures, and numerals *(e.g., when given a picture of 30 flowers, identify the picture that has the number of butterflies that would be needed for each flower to have a butterfly; given a picture of 19 shapes, draw 19 more shapes).*	Use place-value concepts for tens and ones *(e.g., compose or decompose a two-digit whole number using a number sentence such as 35 = 3 tens and 5 ones, 35 =30 + 5, or using number bonds; determine the value of a digit in the tens and ones place).*	Use place-value concepts for hundreds, tens, and ones *(e.g., compose or decompose a three-digit whole number using a number sentence such as 254 = 2 hundreds, 5 tens, and 4 ones; 254 = 200 + 50 + 4; determine the value of a digit in the hundreds place, etc.).*

Slide 33

GLOBAL PROFICIENCY DESCRIPTORS

- In general, there is a connection between the descriptors across grades:
- Exceeds at grade 2 → Meets at grade 3 Meets at grade 2 → Partially meets at grade 3

Grade 2	Partially meets	Meets	Exceeds
Time Tell Time	Sequence and describe events in time using informal comparisons	Tell time using an analog clock to the nearest hour.	Tell time using an analog clock to the nearest half hour.

Grade 3	Partially meets	Meets	Exceeds
Time Tell Time	Tell time using an analog clock to the nearest hour.	Tell time using an analog clock to the nearest half hour.	Tell time using an analog clock to the nearest minute.

Slide 34

GPF KNOWLEDGE, SKILLS, AND STANDARDS

- **Statements of knowledge and/or skills (content standards)**: WHAT content learners are expected to know and be able to do as described in the GPF.
  - Example: Grade 3 learners **should be able to** retrieve explicit information in a grade-level text by direct- or close-word matching.
- **Global Proficiency Descriptors (performance standards)**: HOW MUCH content do learners need to know and be able to demonstrate in relation to knowledge or skills.
  - Example: Grade 3 learners who **"meet global minimum proficiency"** should be able to retrieve a single piece of explicit information from a grade 3-level text by direct- or close-word matching when the information required is adjacent to the matched word and there is limited competing information. This will generally be in response to a "who," "what," "when," or "where" question.

Slide 35

GPF KNOWLEDGE AND SKILLS (READING)

Domain	Construct	Subconstruct	Knowledge or Skill
Reading comprehension	R1 Retrieve information	R1.1 Recognize the meaning of common grade-level words	Recognize the meaning of common grade-level words.
		R1.2 Retrieve explicit information in a grade-level text by direct- or close-word matching	Retrieve a single piece of explicit information from a grade-level text by direct- or close-word matching
			Retrieve a single piece of explicit information from a grade-level continuous text by direct- or close-word matching
			Retrieve a single piece of explicit information from a grade-level non-continuous text (tables, diagrams, graphs) by direct- or close-word matching
		R1.3 Retrieve explicit information in a grade-level text by synonymous matching	Retrieve a single piece of explicit information from a grade-level text by synonymous word matching
			Retrieve a single piece of explicit information from a grade-level continuous text by synonymous word matching
			Retrieve a single piece of explicit information from a grade-level non-continuous text (e.g., simple diagrams and tables) by synonymous word matching
	R2 Interpret information	R2.1 Identify the meaning of unknown words and expressions in a grade-level text	Identify the meaning of unknown words (including familiar words used in unfamiliar ways) and idiomatic and figurative expressions in a grade-level text
			Identify the meaning of unknown words and expressions in a grade-level text
		R2.2 Make inferences in a grade-level text	Make simple inferences in a grade-level text by relating pieces of explicit and/or implicit information in the text
			Make inferences in a grade-level continuous text by relating pieces of explicit and/or implicit information in the text
			Make inferences in a grade-level non-continuous text (e.g., tables, diagrams, graphs) by relating pieces of explicit and/or implicit information
			Identify the sequence of events/actions/steps in a grade-level text
			Identify, compare, or contrast points of view in a grade-level text
			Identify, compare, or contrast evidence in a grade-level text to support or explain an idea, action, or statement
			Draw a basic conclusion from a grade-level text by synthesizing information in the text (grades 6 to 9)
			Apply information from a grade-level text to a new example or situation

Slide 36

GLOBAL PERFORMANCE DESCRIPTORS (GPDs)

- For each subconstruct and knowledge or skill, there are descriptions of performance at the partially meets, meets, and exceeds GPLs.
- For example, in grade [X] in the [name] domain, for the [name] construct, and [name] subconstruct of the GPF has the following:

Subconstruct	Partially Meets	Meets	Exceeds
Recognize the meaning of common grade-level words in a short, grade-level continuous text read to or signed for the learner	When listening to a short grade 2-level continuous text, identify the meaning of very common words (See example items in Appendix A).	When listening to a short grade 2-level continuous text, identify the meaning of common words (See example items in Appendix A).	When listening to a short grade 2-level continuous text, identify the meaning of less common words (See example items in Appendix A).

Slide 37

GLOBAL PROFICIENCY DESCRIPTORS

[Insert reading/math GPF table here for the relevant grade (Grade 2 for SDG 4.1.19(a), Grade 5 for SDG 4.1.1(b) and Grade 8 for SDG 4.1.1(c))—may take more than one slide, perhaps one per domain or one per construct]

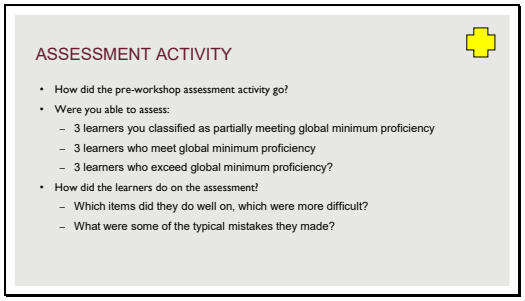
PRESENTATION 4 – FAMILIARIZATION WITH ASSESSMENT INSTRUMENT

Slide 38

PRESENTATION

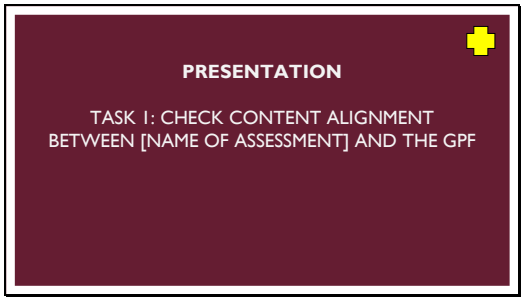
REVIEW OF THE [ASSESSMENT NAME]

Slide 39

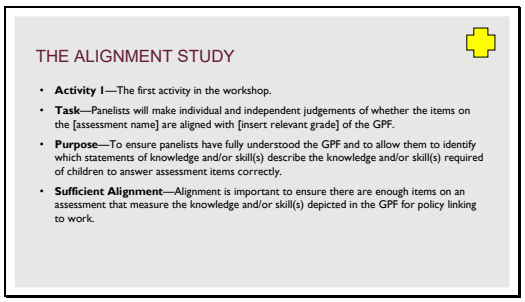


PRESENTATION 5 – TRAINING ON ALIGNMENT EXERCISE

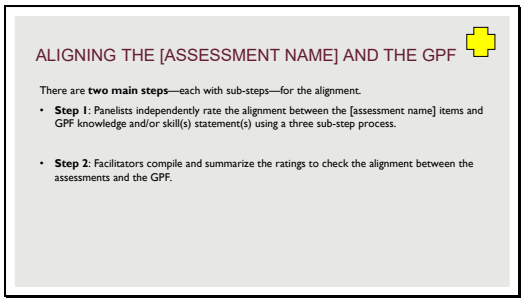
Slide 40



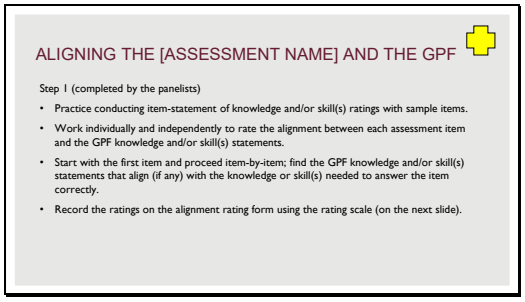
Slide 41



Slide 42



Slide 43



Slide 44

ALIGNING THE [ASSESSMENT NAME] AND THE GPF

Rate each item using a scale of **Complete Fit, Partial Fit**, and **No Fit** as follows:

- **Complete Fit (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 completely use the knowledge and/or skill(s) described in the statement.
- **Partial Fit (P)** signifies that part of the content required to answer the item correctly is contained in the statement of knowledge and/or skills, i.e., if the learner answers the item correctly, it is because they partially use knowledge and/or skill(s) described in the statement.
- **No Fit (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 do not use knowledge and/or skill(s) described in the GPF.

Slide 45

ALIGNING THE [ASSESSMENT NAME] AND THE GPF

Follow these **additional instructions** for the alignment ratings:

- If an item has a rating of **Complete Fit (C)** with a particular statement of knowledge and/or skill(s), the panelists should not match it with other statements of knowledge and/or skill(s), meaning it is aligned to only one statement in the GPF;
- If an item has a rating of **Partial Fit (P)** with a particular statement of knowledge and/or skill(s), the panelists should generally match it to one or two additional statements of knowledge and/or skill(s); and
- If an item has a rating of **No Fit (N)** with any statements of knowledge and/or skill(s), the panelists should not match it to any statements of knowledge and/or skill(s).

Slide 46

EXAMPLE: COMPLETE FIT

1. How is eight hundred and seventy written in standard form?

A. 807
B. 870
C. 817
D. 871

Domain: Number and Operations
Construct: Whole Numbers
Subconstruct: Identify, count in, and identify the relative magnitude of whole numbers
Knowledge or skill(s) statement: Count, read, and write whole numbers

To answer this item correctly, the learner needs to be able to identify and count whole numbers. Therefore, the item can be rated as "complete fit" with the statement of knowledge and/or skill(s) since it only requires the knowledge or skills from that single statement.

Slide 47

EXAMPLE: PARTIAL FIT

2. What is the largest sum?

A. 22 + 37
B. 21 + 39
C. 23 + 38
D. 24 + 36

Domain: Number and Operations
Construct: Whole Numbers
Subconstruct: Solve operations using whole numbers
Knowledge or skill statement: Add, subtract, multiply, and divide whole numbers

To answer this item correctly, the learner needs to be able to compare and order whole numbers as well as add and subtract whole numbers. Therefore, the item can be rated as "partial fit" with the statements of knowledge and/or skill(s) since it requires knowledge or skills from both statements.

Domain: Number and Operations
Construct: Whole Numbers
Subconstruct: Identify, count in, and identify the relative magnitude of whole numbers
Knowledge or skill statement: Compare and order whole numbers

Slide 48

EXAMPLE: NO FIT

3. What is 2/3–1/3 ?

A. 1/0
B. 1/3
C. 2/3
D. 3/0

To answer this item correctly, the learner needs to be able to add and subtract fractions. This knowledge or skill is not expected until the upper primary grades. Therefore, the item can be rated as "no fit" since it requires knowledge or skill that is not expected at (or before) the grade level.

Slide 49

EXAMPLE: COMPLETE FIT

Grade and Subject: Grade 3 Reading

Oral Reading Fluency: Read this passage aloud, quickly but carefully, in a minute.

Jabu had a pet dog. He took the dog outside to play. The dog ran away and got lost. Jabu was sad. After a while, the dog came back. Jabu took the dog inside. He gave the dog some food. Then the dog went to sleep. When the dog woke up, Jabu took the dog outside to play again. (59 words)

Domain: Decoding

Construct: Fluency

Subconstruct: Say or sign a grade-level continuous text at pace and with accuracy

Knowledge or skill(s) statement: Say or sign fluently a grade-level continuous text

The learner needs to read the passage aloud, quickly but carefully, in a minute. Therefore, the item can be rated as "complete fit" with the statement of knowledge and/or skill(s) since it only requires the knowledge or skills from that single statement.

Slide 50

EXAMPLE: COMPLETE FIT (CONSIDERING GRADE LEVEL OF PASSAGE)

Feature	Scope	Elaboration	Contextualization
Length	Very short	A few sentences: approximately 20-30 words in English	Fewer words in agglutinative or highly synthetic languages
Familiarity	Very familiar	Everyday experiences, events and objects that are likely to be familiar to the students	Context dependent
Predictability	Medium	Context or setting is familiar and somewhat predictable, but includes details that cannot be predicted to ensure that students are required to make meaning from the text.	
Challenge	As little as possible	Little or no implied information, minimal competing information and possibly also supportive illustrations	
Text structure	Very simple	Familiar structure with a clear main idea, only one or two characters, few details	
Vocabulary	Very common	Simple words that are likely to have been encountered often and typically describe concrete concepts; may include a highly-supported uncommon word	Depends on the transparency of the orthography and the language background of the students
Sentence structure	Simple and common	Simple sentences or a simple compound sentence that is commonly encountered.	Language dependent

Slide 51

EXAMPLE: PARTIAL FIT

Grade and Subject: Grade 3 Reading

Reading Comprehension: What did Jabu do when the dog woke up? (Note this question is only asked if the learner reads this far in the story within 1 minute).

Answer: He took the dog outside to play again.

Domain: Decoding

Construct: Fluency

Subconstruct: Say or sign a grade-level continuous text at pace and with accuracy

Knowledge or skill statement: Say or sign fluently a grade-level continuous text

To answer this item correctly, the learner needs to be able to decode the passage quickly (in less than a minute) and to retrieve a single piece of explicit information. Therefore, the item can be rated as "partial fit" since it requires knowledge of two different statements of knowledge and/or skill(s).

Domain: Reading Comprehension

Construct: Retrieve Information

Subconstruct: Retrieve explicit information in a grade-level text by direct- or close-word matching

Knowledge or skill: Retrieve a single piece of explicit information from a grade-level continuous text by direct- or close-word

Slide 52

EXAMPLE: NO FIT

Grade and Subject: Grade 2 Reading

Which of these sentences is punctuated correctly as a question?

A. Where is the dog!

B. Where is the dog.

C. Where is the dog?

D. Where is the dog:

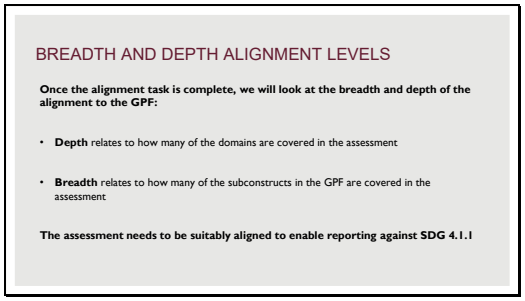
This item is a "no fit" item, as it requires the learner to explicitly demonstrate their attainment in relation to punctuation, which is not referenced in the GPF

Slide 53

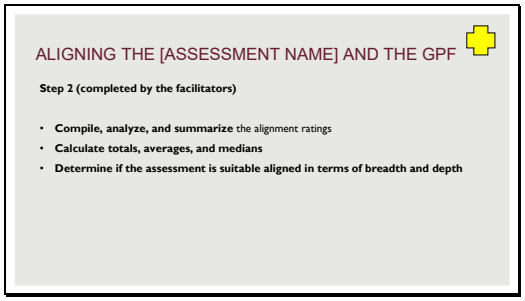
ALIGNMENT RATING FORM

							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.				
Question	Domain	Construct reference	Subconstruct reference	Knowledge or skill	Fit	Domain	Construct reference	Subconstruct reference	Knowledge or skill	Fit
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										

Slide 54

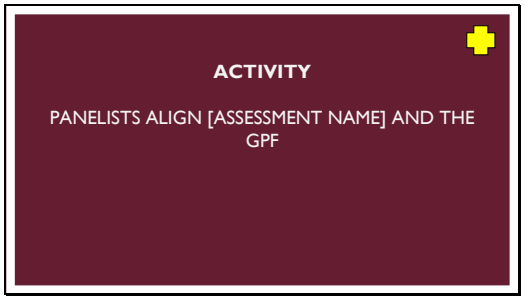


Slide 55

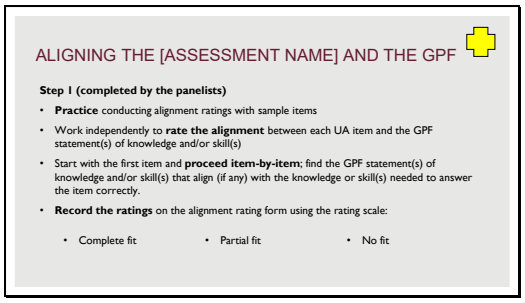


PRESENTATION 6 – UNDERTAKE ALIGNMENT EXERCISE

Slide 56



Slide 57



PRESENTATION 7 – PRESENTATION AND DISCUSSION OF ALIGNMENT RESULTS

Slide 58

PRESENTATION

REVIEW PANELIST ALIGNMENT RESULTS
FROM TASK 1

Slide 59

ALIGNMENT RATINGS

Domain		Items
N	Number and operations	14
M	Measurement	7
G	Geometry	3
S	Statistics and probability	2
A	Algebra	0
Total		**26**
Construct		**Items**
N1	Whole numbers	14
M1	Length, weight, capacity, volume, area, and perimeter	3
M2	Time	4
M3	Currency	0
G1	Properties of shapes and figures	2
G2	Spatial visualizations	0
G3	Position and direction	1
S1	Data management	2
A1	Patterns	0
A3	Relations and functions	0
Total		**26**

Slide 60

ALIGNMENT RATINGS

Subconstruct		Items
N1.1	Identify and count in whole numbers, and identify their relative magnitude	4
N1.2	Represent whole numbers in equivalent ways	0
N1.3	Solve operations using whole numbers	8
N1.4	Solve real-world problems involving whole numbers	2
M1.1	Use non-standard and standard units to measure, compare, and order	3
M2.1	Tell time	2
M2.2	Solve problems involving time	2
M3.1	Use different currency units to create amounts	0
G1.1	Recognize and describe shapes and figures	2
G2.1	Compose and decompose shapes and figures	0
G3.1	Describe the position and directions of objects in space	1
S1.1	Retrieve and interpret data presented in displays	2
A1.1	Recognize, describe, extend, and generate patterns	0
A3.2	Demonstrate an understanding of equivalency	0
Total		**26**

Slide 61

ALIGNMENT RATINGS

Domain		Items
C	Comprehension of spoken or signed language	14
D	Decoding	7
R	Reading comprehension	3
Total		**24**
Construct		**Items**
C1	Retrieve information at word level	14
C2	Retrieve information at sentence or text level	0
C3	Interpret information at sentence or text level	3
D1	Precision	4
D2	Fluency	0
R1	Retrieve information	2
R2	Interpret information	0
R3	Reflect on information	1
Total		**24**

Slide 62

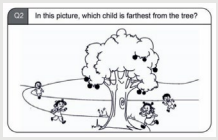
ALIGNMENT RATINGS

Subconstruct		Items
C1.1	Comprehend spoken and signed language at the word or phrase level	4
C1.2	Recognize the meaning of common grade-level words in a short, grade-level continuous text read to or signed for the learner	0
C2.1	Retrieve explicit information in a short grade-level continuous text read to or signed for the learner	8
C3.1	Interpret information in a short grade-level continuous text read to or signed for the learner	2
D1.1	Identify symbol-sound/fingerspelling and/or symbol-morpheme correspondences	0
D1.2	Decode isolated words	3
D2.1	Say or sign a grade-level continuous text at pace and with accuracy	2
R1.1	Recognize the meaning of common grade-level words	2
R1.2	Retrieve explicit information in a grade-level text by direct- or close-word matching	0
R1.3	Retrieve explicit information in a grade-level text by synonymous word matching	2
R2.1	Identify the meaning of unknown words and expressions in a grade-level text	0
R2.2	Make inferences in a grade-level text	1
R2.3	Identify the main and secondary ideas in a grade-level text	2
R3.1	Identify the purpose and audience of a text	0
R3.2	Evaluate a text with justification	
R3.3	Evaluate the status of claims made in a text	0
Total		**26**

Slide 63

ALIGNMENT RATINGS

Some aligned this item to:
- **Domain:** Measurement
- **Construct:** Length, weight, capacity, volume, area and perimeter
- **Subconstruct:** Use non-standard and standard units to measure, compare, and order
- **Knowledge and/or Skills:** Use non-standard units to estimate, measure, and compare length, weight, volume, and capacity
- **Fit:** Complete



Others aligned it to:
- **Domain:** Geometry
- **Construct:** Position and direction
- **Subconstruct:** Describe the position and direction of objects in space
- **Knowledge and/or Skills:** Use positional terms to describe the location of an object
- **Fit:** Complete

PRESENTATION 8 – TRAIN PANELISTS ON THE MATCHING TASK

Slide 64

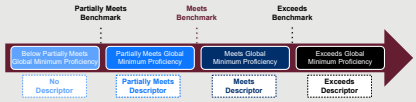
PRESENTATION

TASK 2. MATCH [ASSESSMENT NAME] ITEMS WITH PROFICIENCY LEVELS AND DESCRIPTORS IN THE GPF

Slide 65

GLOBAL PROFICIENCY FRAMEWORK (REVIEW)

- The GPF has **GPLs** (Levels) and **GPDs** (Descriptors) for grades 1 through 9 in reading and mathematics:



- The GPDs describe **minimum proficiency** for the GPLs, i.e., the minimum knowledge or skill(s) necessary for classification into each GPL (by grade and subject).
- The GPDs are organized hierarchically **by domains, constructs, subconstructs**, and **knowledge and skills**, with descriptors for each of the knowledge and skills.

Slide 66

GLOBAL PROFICIENCY FRAMEWORK (EXAMPLE)

Partially Meets Global Minimum Proficiency	Meets Global Minimum Proficiency	Exceeds Global Minimum Proficiency
M2: TIME		
M2.1: Tell time		
M2.1.1_P Identify, sequence, and describe activities/events that take place at different parts of the day *(e.g., morning and afternoon).*	M2.1.1_M Tell time using an analog clock to the nearest hour.	M2.1.1_E Tell time using an analog clock to the nearest half hour.
M2.1.2_P N/A	M2.1.2_M Recognize the number of days in a week and months in a year.	M2.1.2_E Recognize the number of hours in a day, minutes in an hour, and seconds in a minute.

Slide 67

GLOBAL PROFICIENCY FRAMEWORK (EXAMPLE)

Partially Meets Global Minimum Proficiency	Meets Global Minimum Proficiency	Exceeds Global Minimum Proficiency
R. READING COMPREHENSION		
R1: RETRIEVE INFORMATION		
R1.1: Recognize the meaning of common grade-level words		
R1.1.1_P Recognize the meaning of very common grade 2-level words. *(e.g., match a given word to an illustration or synonym or provide a brief spoken/signed definition).*	R1.1.1_M Recognize the meaning of common grade 2-level words. *(e.g., match a given word to an illustration or synonym or provide a brief spoken/signed definition).*	R1.1.1_E Recognize the meaning of less common grade 2-level words. *(e.g., match a given word to an illustration or synonym or provide a brief spoken/signed definition).*

Slide 68

MATCHING ITEMS WITH GPLS AND GPDS

- **Build on** your understanding of the [assessment name] items and the GPF gained through the alignment activity in Task 1.
- **Group Activity**: You should work to achieve consensus.
- **Focus on one key aspect**: Descriptors (GPDs) of global minimum proficiency that match with the items.

Slide 69

MATCHING ITEMS WITH GPLS AND GPDS

Answer these questions for each item (based on consensus in the groups):

- What **knowledge and/or skill(s)** is/are required to answer the items correctly?
- What makes the item **easy or difficult**?
- What is the **lowest GPL** and GPD that are most appropriate for the item?

Slide 70

MATCHING "NO FIT" ITEMS

For "no fit" items we will follow these steps:

- Imagine a group of learners who are best described by the GPDs in the 'partially meets' level.
- Using your experience of teaching such learners, determine whether this is an appropriate item for those learners and if they would be likely to answer the item correctly.
- If the item is determined to be appropriate for learners at the partially meets level, this can be recorded.
- If not, then the process should be repeated for the 'meets' level and then the 'exceeds' level, if required.
- If the item is determined to be too difficult for the grade, then it should be recorded as above the exceeds level.

Slide 71

MATCHING ITEMS WITH GPLS AND GPDS

The item is from a grade 3 assessment and is therefore being linked to grade 2 for SDG 4.1.1(a) reporting. It matches with the Meets GPL and GPD (performance standard) for grade 2.

How is eighty-seven written in standard form?

A. 80
B. 87
C. 807
D. 870

Domain: Number and Operations
Construct: Whole Numbers
Subconstruct: Identify and count in whole numbers, and identify their relative magnitude
Knowledge or skill (content standard): Count, read, and write whole numbers

What makes it easy or difficult: the other answer choices are strong distractors, especially C.

GPL and GPD (performance standard):

Partially Meets: Read and write whole numbers up to 30 in words and in numerals.

Meets: Read and write whole numbers up to 100 in words and in numerals.

Exceeds: N/A

Slide 72

MATCHING ITEMS WITH GPLS AND GPDS

Jeb had 16 peaches.
He gave away 4 peaches.
Then Jeb divided the remaining peaches equally between 2 baskets.
How many peaches did Jeb put in each basket?

A. 6
B. 8
C. 10
D. 12

Domain: Number and Operations
Construct: Whole Numbers
Subconstruct: Solve real-world problems involving whole numbers
Knowledge/skills: Solve real-world problems involving the addition, subtraction, multiplication, and division of whole numbers

What makes the item easy/difficult?

Difficult—Since this is a real-world problem, learners have to identify which operations need to be completed, and there are two operations/steps.

Lowest GPD to answer correctly?

Above exceeds for grade 2 as matches with grade 4 Meets—Solve simple real-world problems involving the multiplication of two whole numbers to 5, and associated division facts

Slide 73

MATCHING ITEMS WITH GPLS AND GPDS

What is the difference in time shown between these two clocks?

What makes the item easy/difficult?
Difficult—it is a two-step problem, and the numbers are not shown on the clocks

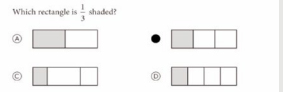
Lowest GPD to answer correctly?
Although 'Tell time using an analog clock to the nearest hour' is "meets" at grade 2, grade 2 students are not expected to solve problems, involving elapsed time, which is "meets" at grade 3.

Domain: Measurement
Construct: Time
Subconstruct: Tell time AND solve problems involving time
Knowledge or skill (content standard): Tell time using an analog clock AND identify or solve problems involving equivalences between different units of time

Slide 74

MATCHING ITEMS WITH GPLS AND GPDS

Which rectangle is $\frac{1}{3}$ shaded?



Source: Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 International Results in Mathematics*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2015/international-results/

Domain: Number and Operations
Construct: Fractions
Subconstruct: Identify and represent fractions using objects, pictures, and symbols, and identify relative magnitude
Knowledge/skills: Express a visual representation of a fraction (picture, objects) in fractional notation

What makes the item easy/difficult? easy/difficult?

Difficult—understanding that the shaded portion must be 1/3 shaded rather than just that 1 out of 3 pieces must be shaded. C is a strong distractor.

Lowest GPD to answer correctly?

Although this item matches with grade 3 Partially Meets—Identify everyday unit fractions represented as objects or pictures in fractional notation, it is 'no fit' for grade 2

Slide 75

MATCHING ITEMS WITH GPLS AND GPDS

This item uses the text from slide 49 in presentation 5. The text is grade 3 appropriate, so this needs to be considered when determining the match.

Item: Who has a pet dog?

What makes it easy or difficult: It is easy because this question comes from the first sentence of the passage and uses direct-word matching.

Domain: Reading Comprehension
Construct: Retrieve Information
Subconstruct: Retrieve explicit information in a grade-level text by direct- or close-word matching
Knowledge or skill: Retrieve a single piece of explicit information from a grade-level continuous text by direct- or close-word matching

GPL and GPD (performance standard):

Partially Meets: Retrieve a single piece of prominent, explicit information from a grade 3-level text by direct- or close-word matching when the information required is adjacent to the matched word and there is no competing information.

Meets: ". . ." and there is limited competing information" . . ."

Exceeds: Retrieve multiple pieces of explicit information
. . . when the information required is adjacent to the matched word and there is limited competing information . . .

Slide 76

MATCHING ITEMS WITH GPLS AND GPDS

The item matches with this grade 3 statement of knowledge and/or skill(s) and the Partially Meets GPL and GPD (performance standard).

Item: Decoding passage—Word "Jabu"

What makes it easy or difficult: It is easy because this is a simple, short word following standard orthographical rules; it might be difficult if it is not a common name.

Domain: Decoding
Construct: Fluency
Subconstruct: Say or sign a grade-level continuous text at pace and with accuracy
Knowledge or skill: Say or sign fluently a grade-level continuous text

GPL and GPD (performance standard):

Partially Meets: Say or sign accurately a grade 3-level continuous text, at a pace that is slow by country standards for fluency for the language in which the assessment is administered (e.g., often word-by-word).

Meets: Say or sign accurately a grade 3-level continuous text, at a pace that meets minimal country standards for fluency for the language in which the assessment is administered.

Exceeds: Say or sign accurately a grade 3-level continuous text, at a pace that exceeds minimal country standards for fluency for the language in which the assessment is administered.

Slide 77

MATCHING ITEMS WITH GPLS AND GPDS

The item matches with this grade 3 statement of knowledge and/or skill(s) and the Exceeds GPL and GPD (performance standard).

Item: Why did the dog come back?

What makes it easy or difficult: It is difficult because there is space between the clues, and there could be other reasons the dog came back.

Domain: Reading Comprehension
Construct: Interpret Information
Subconstruct: Make inferences in a grade-level text
Knowledge or skill: Make simple inferences in a grade-level text by relating pieces of explicit and/or implicit information in the text

GPL and GPD (performance standard):

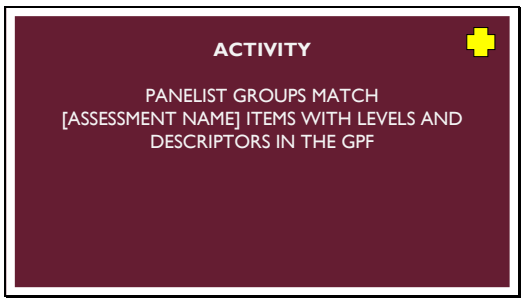
Partially Meets: Make simple inferences in a grade 3-level text by relating two pieces of explicit information in consecutive sentences when there is no competing information. This will generally be in response to a 'why' or 'how' question. (See example items in Appendix C).

Meets: ". . ." when there is limited competing information" . . ."

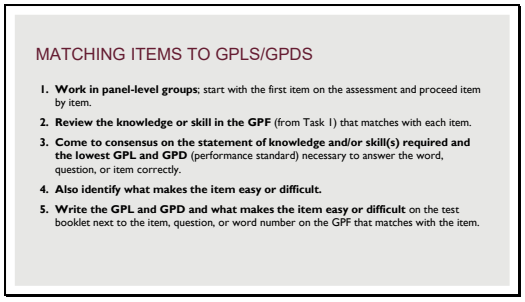
Exceeds: ". . ." in one or more paragraphs when there is more distance between the pieces of information that need to be related and/or a lot of competing information ". . ."

PRESENTATION 9 – UNDERTAKE MATCHING ACTIVITY

Slide 78

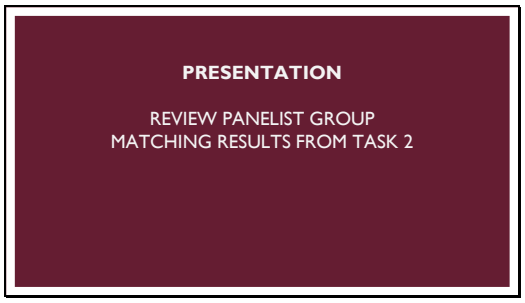


Slide 79

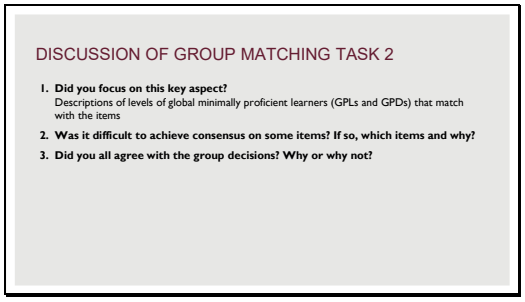


PRESENTATION 10 – PRESENTATION AND DISCUSSION OF MATCHING RESULTS

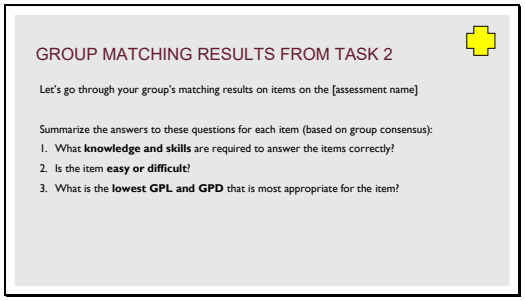
Slide 80



Slide 81

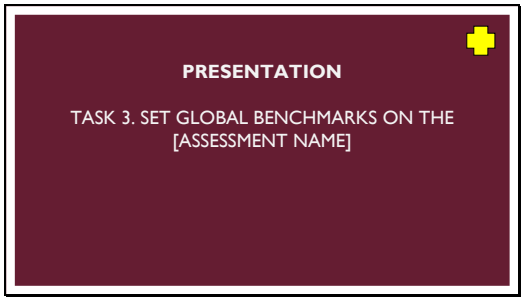


Slide 82

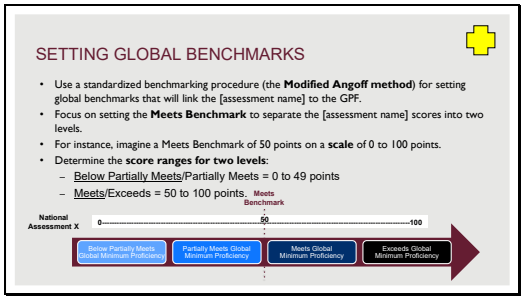


PRESENTATION 11 – OVERVIEW OF GLOBAL STANDARDS AND BENCHMARKING APPROACH

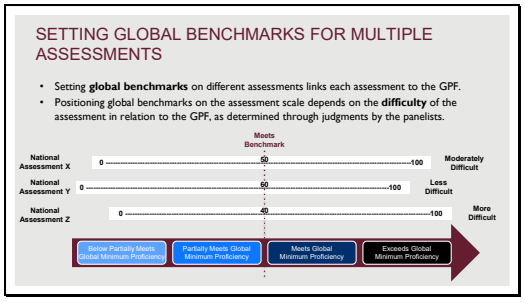
Slide 83



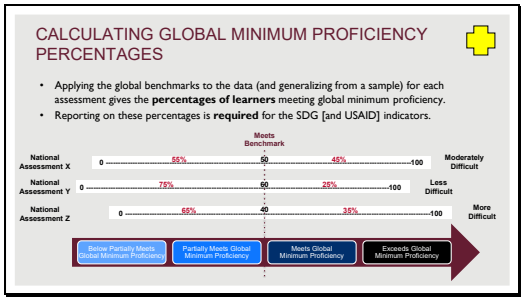
Slide 84



Slide 85



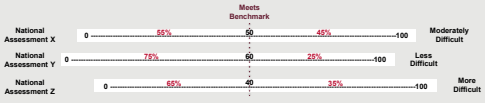
Slide 86



Slide 87

COMPARING, AGGREGATING, AND TRACKING RESULTS

- Results from different countries can be **compared** by examining the percentages of learners meeting (and not meeting) global minimum proficiency.
- Results will be **aggregated both within and across** countries for global reporting.
- Results will be **tracked** over time (by country) to examine changes in the percentages of learners meeting global minimum proficiency.



Slide 88

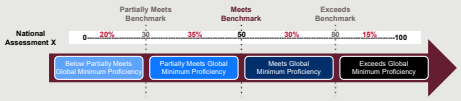
COMPARING, AGGREGATING, AND TRACKING RESULTS

Country and Assessment	Global Minimum Proficiency Levels			
	Below Partially Meets/ Partially Meets		Meets/Exceeds	
	Score Range	Percentage	Score Range	Percentage
National Assessment X	0-49	55%	50-100	45%
National Assessment Y	0-59	75%	60-100	25%
National Assessment Z	0-39	65%	40-100	35%

Slide 89

SETTING THREE GLOBAL BENCHMARKS

- Setting **three global benchmarks** classifies learners into **four GPLs**.
- Percentages of learners in the four GPLs are calculated based on the assessment data sets (score distributions).
- [USAID "FI" indicators require reporting on the percentage of learners progressing from a lower to a higher GPL over time.]



PRESENTATION 12 – TRAIN PANELISTS ON ANGOFF METHOD

Slide 90

PRESENTATION

ANGOFF METHOD FOR SETTING BENCHMARKS

Slide 91

IMPLEMENTING THE ANGOFF METHOD FOR SETTING BENCHMARKS

- The **Modified Angoff method** is used to set the benchmarks:
  - Widely-used benchmarking method
  - Relies on judgements by expert panelists
  - Item-centered method, i.e., panelists rate each item, estimating whether minimally proficient learners at each GPL would answer the item correctly
- Critical to focus on the **definitions of minimum proficiency from the GPDs** in the GPF at [(Grade 2 for SDG 4.1.1(a), Grade 5 for SDG 4.1.1(b) and Grade 8 for SDG 4.1.1(c))]

Slide 92

IMPLEMENTING THE ANGOFF METHOD FOR SETTING BENCHMARKS

- Ratings for Task 3 should be **individual and independent**.
- **Consensus on ratings is not needed**, though consistency is desired.
- **Benchmarks represent the panel's estimates of scores** that a minimally proficient learner at each level would obtain on the assessment.
- Angoff uses **two rounds** of item ratings, with discussions and feedback between rounds.
- **Global benchmarks** are calculated based on the total ratings by each panelist and the averages across all the panelists.

Slide 93

IMPLEMENTING THE ANGOFF METHOD FOR SETTING BENCHMARKS

Two Rounds

- **Round 1**: Make **beginning ratings** for each item on the assessment.
  - After Round 1, total the ratings to calculate each panelist's **initial global benchmarks**, and then average them to calculate the panel's initial benchmarks.
- **Round 2**: Make **revised ratings** for each item on the assessment.
  - After Round 2, total the ratings to calculate each panelist's **final global benchmarks**, and then average them to calculate the panel's final benchmarks.

Slide 94

IMPLEMENTING THE ANGOFF METHOD FOR SETTING BENCHMARKS

Between Below Partially Meets and Partially Meets Global Minimum Proficiency (**Partially Meets Benchmark**)	At or Slightly Above Partially Meets Global Minimum Proficiency (Just Partially Meets or **JP**)
Between Partially Meets and Meets Global Minimum Proficiency (**Meets Benchmark**)	At or Slightly Above Meets Global Minimum Proficiency (Just Meets or **JM**)
Between Meets and Exceeds Global Minimum Proficiency (**Exceeds Benchmark**)	At or Slightly Above Exceeds Global Minimum Proficiency (Just Exceeds or **JE**)

Slide 95

IMPLEMENTING THE ANGOFF METHOD FOR SETTING BENCHMARKS

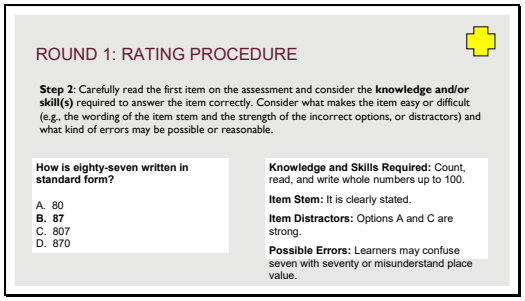
- Item ratings are based on **four expectations**, i.e., chances of whether a minimally proficient learner (based on the GPDs in the GPF) would answer each item correctly:
  - Probably not ("no")
  - Somewhat possible ("no")
  - **Reasonably sure or ≥ 67 percent chance ("yes")**
  - Absolutely positive ("yes")
- Item ratings are not based on "should" but on "would" for **realistic expectations**:
  - **Should** refers to performance based only based on the statements of knowledge and/or skill(s) from the GPF.
  - **Would** is influenced by assessment constraints, e.g., difficulty of an item for a particular learner, testing conditions, learner anxiety, and random errors.

Slide 96

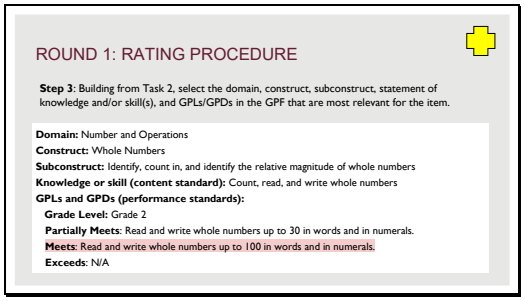
ROUND 1: RATING PROCEDURE

Step 1: Identify and/or **conceptualize** three Just Partially Meets (JP), three Just Meets (JM), and three Just Exceeds (JE) learners based on an understanding of the GPF at [Grade 2 for SDG 4.1.1(a), Grade 5 for SDG 4.1.1(b) and Grade 8 for SDG 4.1.1(c)].

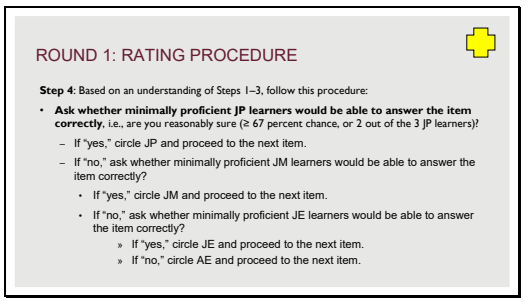
Slide 97



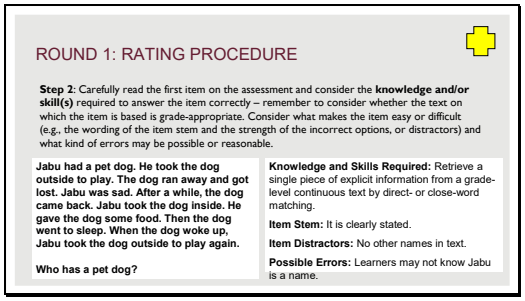
Slide 98



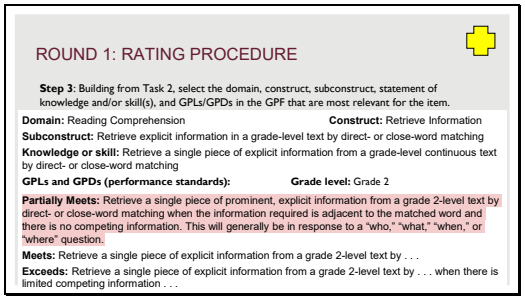
Slide 99



Slide 100



Slide 101



Slide 102

ROUND 1: RATING PROCEDURE

Step 4: Based on an understanding of Steps 1–3, follow this procedure:

- **Ask whether minimally proficient JP learners would be able to answer the item correctly**, i.e., are you reasonably sure (≥ 67 percent chance, or 2 out of the 3 JP learners)?
  - If "yes," circle JP and proceed to the next item.
  - If "no," ask whether minimally proficient JM learners would be able to answer the item correctly?
    - If "yes," circle JM and proceed to the next item.
    - If "no," ask whether minimally proficient JE learners would be able to answer the item correctly?
      - » If "yes," circle JE and proceed to the next item.
      - » If "no," circle AE and proceed to the next item.

Slide 103

ROUND 1: RATING PROCEDURE

Step 2: Estimate number of items JP, JM, and JE learners would be able to complete within the time limit (e.g., words in the oral reading passage the learners would attempt to read in a minute).

Word No.	Reading passage (Word)	Round 1: No. of words learners would attempt to read in a minute			Round 1 individual and independent ratings				Round 2: No. of words learners would attempt to read in a minute			Round 2 individual and independent ratings			
		JP	JM	JE	JP	JM	JE	AE	JP	JM	JE	JP	JM	JE	AE
1	Kande	1	1	1	JP	JM	JE	AE	1	1	1	JP	JM	JE	AE
2	da	2	2	2	JP	JM	JE	AE	2	2	2	JP	JM	JE	AE
3	abokiyarta	3	3	3	JP	JM	JE	AE	3	3	3	JP	JM	JE	AE
4	Delu	4	4	4	JP	JM	JE	AE	4	4	4	JP	JM	JE	AE
5	sukan	5	5	5	JP	JM	JE	AE	5	5	5	JP	JM	JE	AE

Slide 104

ROUND 1: RATING PROCEDURE

Step 3: Carefully read the first word or question on the [TA] and consider the **knowledge and/or skill(s)** required to read or answer the word or question correctly. Consider what makes the word or question easy or difficult (e.g., the type of knowledge and skills required, the wording of the question) and what kind of errors may be possible or reasonable.

Item: Decoding passage—Word "Jabu"

What makes it easy or difficult: It is easy because this is a simple, short word following standard orthographical rules; it might be difficult if it is not a common name.

Domain: Decoding
Construct: Fluency
Subconstruct: Say or sign a grade-level continuous text at pace and with accuracy
Knowledge or skill: Say or sign fluently a grade-level continuous text

Slide 105

ROUND 1: RATING PROCEDURE

Step 4: Building from Task 2, select the domain, construct, subconstruct, statement of knowledge and/or skill(s), and GPLs/GPDs in the GPF that are most relevant for the item.

Domain: Decoding
Construct: Fluency
Subconstruct: Say or sign a grade-level continuous text at pace and with accuracy
Knowledge or skill: Say or sign fluently a grade-level continuous text
GPLs and GPDs (performance standards):

Grade Level: Grade 2

Partially Meets: Say or sign accurately a grade 2-level continuous text, at a pace that is slow by country standards for fluency for the language in which the assessment is administered (e.g., word-by-word).

Meets: Say or sign accurately a grade 2-level continuous text, at a pace that meets minimal country standards for fluency for the language in which the assessment is administered.

Exceeds: Say or sign accurately a grade 2-level continuous text, at a pace that exceeds minimal country standards for fluency for the language in which the assessment is administered.

Slide 106

ROUND 1: RATING PROCEDURE

Step 5: Based on an understanding of Steps 1–3, follow this procedure:

- **Ask whether minimally proficient JP learners would be able to answer the item correctly**, i.e., are you reasonably sure (≥ 67 percent chance, or 2 out of the 3 JP learners)?
  - If "yes," circle JP and proceed to the next item.
  - If "no," ask whether minimally proficient JM learners would be able to answer the item correctly?
    - If "yes," circle JM and proceed to the next item.
    - If "no," ask whether minimally proficient JE learners would be able to answer the item correctly?
      - » If "yes," circle JE and proceed to the next item.
      - » If "no," circle AE and proceed to the next item.

Slide 107

ROUND 1: RATING PROCEDURE

There are slightly different steps when considering items with more than one score point:

- **Step 1:** Identify and/or conceptualize three Just Partially Meets (JP), three Just Meets (JM), and three Just Exceeds (JE) learners based on an understanding of the GPF.
- **Step 2:** Carefully read the item and consider the knowledge and/or skill(s) required to answer the item correctly [– remember to consider whether the text on which the item is based is grade-appropriate.] Think carefully about what the learner needs to demonstrate to achieve each score point.
- **Step 3:** Building from Task 2, select the domain, construct, subconstruct, statement of knowledge and/or skill(s), and GPLs/GPDs in the GPF that are most relevant for the item.

Slide 108

ROUND 1: RATING PROCEDURE

- **Step 4:** Ask whether minimally proficient JP learners would score the first score point?
  - If "Yes," circle JP for the first score point and move onto the second score point
  - If "No," ask whether minimally proficient JM learners would score the first score point?
    - If "Yes," circle JM for the first score point and move onto the second score point
    - If "No," ask whether minimally proficient JE learners would score the first score point?
      - » If "Yes," circle JE and move onto the second score point
      - » If "No," circle AE for all score-points
- **Step 5:** Repeat step 4 for each score point. At the end, you should have identified how many score-points you think a JP, JM and JE would achieve on the item.

Slide 109

ROUND 1: RATING PROCEDURE

There are slightly different steps when considering items that aligned as "No fit":

- **Step 1:** Identify and/or conceptualize three Just Partially Meets (JP), three Just Meets (JM), and three Just Exceeds (JE) learners based on an understanding of the GPF.
- **Step 2:** Carefully read the item and consider the knowledge and/or skill(s) required to answer the item correctly and how the JP, JM and JE might perform on the item

Slide 110

ROUND 1: RATING PROCEDURE

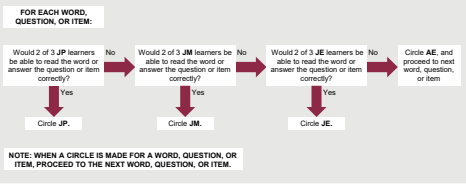
There are slightly different steps when considering items that aligned as "No fit":

- **Step 3:** Ask whether minimally proficient JP learners would answer the item correctly?
  - If "Yes," circle JP and move onto the next item
  - If "No," ask whether minimally proficient JM learners would answer the item correctly?
    - If "Yes," circle JM and move onto the next item
    - If "No," ask whether minimally proficient JE learners would answer the item correctly?
      - » If "Yes," circle JE and move onto the next item
      - » If "No," circle AE and move onto the next item

Slide 111

ROUND 1: RATING PROCEDURE

Flowchart for item ratings with words, questions, or items:

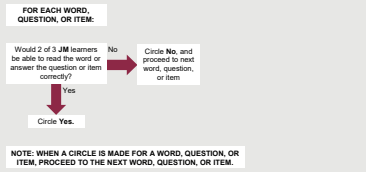


NOTE: WHEN A CIRCLE IS MADE FOR A WORD, QUESTION, OR ITEM, PROCEED TO THE NEXT WORD, QUESTION, OR ITEM.

Slide 112

ROUND 1: RATING PROCEDURE

Flowchart for item ratings with words, questions, or items:



NOTE: WHEN A CIRCLE IS MADE FOR A WORD, QUESTION, OR ITEM, PROCEED TO THE NEXT WORD, QUESTION, OR ITEM.

Slide 113

ROUND 1: ITEM RATING FORM

Directions: For each item, circle either Just Partially Meets (**JP**), Just Meets (**JM**), or Just Exceeds (**JE**) Global Minimum Proficiency, depending on whether the minimally proficient learners at each level would answer the item correctly ("yes"). Circle Above Exceeds Global Minimum Proficiency (**AE**) for items that even a JE learner would not be able to answer correctly.

ITEM NO.	ROUND 1				ROUND 2			
1	JP	JM	JE	AE	JP	JM	JE	AE
2	JP	JM	JE	AE	JP	JM	JE	AE
3	JP	JM	JE	AE	JP	JM	JE	AE
4	JP	JM	JE	AE	JP	JM	JE	AE
5	JP	JM	JE	AE	JP	JM	JE	AE
6	JP	JM	JE	AE	JP	JM	JE	AE
7	JP	JM	JE	AE	JP	JM	JE	AE
8	JP	JM	JE	AE	JP	JM	JE	AE
9	JP	JM	JE	AE	JP	JM	JE	AE
10	JP	JM	JE	AE	JP	JM	JE	AE

Slide 114

ROUND 1: ITEM RATING FORM

Directions: For each item, circle either Just Partially Meets (**JP**), Just Meets (**JM**), or Just Exceeds (**JE**) Global Minimum Proficiency, depending on whether the minimally proficient learners at each level would answer the item correctly ("yes"). Circle Above Exceeds Global Minimum Proficiency (**AE**) for items that even a JE learner would not be able to answer correctly.

Word No.	Reading passage (Word)	Round 1: No. of words learners would attempt to read in a minute			Round 1 individual and independent ratings				Round 2: No. of words learners would attempt to read in a minute			Round 2 individual and independent ratings			
		JP	JM	JE	JP	JM	JE	AE	JP	JM	JE	JP	JM	JE	AE
1	Kande	1	1	1	JP	JM	JE	AE	1	1	1	JP	JM	JE	AE
2	da	2	2	2	JP	JM	JE	AE	2	2	2	JP	JM	JE	AE
3	abokiyarta	3	3	3	JP	JM	JE	AE	3	3	3	JP	JM	JE	AE
4	Delu	4	4	4	JP	JM	JE	AE	4	4	4	JP	JM	JE	AE
5	sukan	5	5	5	JP	JM	JE	AE	5	5	5	JP	JM	JE	AE

Slide 115

ROUND 1: HELPFUL TIPS FOR CONDUCTING ITEM RATING

- Base the first round of item ratings on the **following guidance**:
  - Conduct ratings based on **individual and independent** judgments of the items and the GPF.
  - Focus on the **item content** in relation to the statements of knowledge and/or skill(s) in the GPF.
  - Take into consideration the **difficulty of the item**, including possible and reasonable errors by the learners.
  - Consider **would** rather than **should** in making realistic ratings.

Slide 116

ROUND 1: CALCULATING THE GLOBAL BENCHMARKS

- Calculate totals for the initial benchmarks **for each panelist**:
  - **Partially Meets** = Total of "yeses" in the JP column of the rating form
  - **Meets** = Total of "yeses" in the JP and JM columns of the rating form
  - **Exceeds** = Total of "yeses" in the JP, JM, and JE columns of the rating form
- Calculate averages for the initial global benchmarks **for the panel**:
  - **Partially Meets** = Average of the partially meets benchmarks across all panelists
  - **Meets** = Average of the meets benchmarks across all panelists
  - **Exceeds** = Average of the exceeds benchmarks across all panelists

Slide 117

ROUND 1: CALCULATING THE GLOBAL BENCHMARKS

ITEM NO.	ROUND 1				ROUND 2			
1	(JP)	JM	JE	AE	JP	JM	JE	AE
2	JP	(JM)	JE	AE	JP	JM	JE	AE
3	(JP)	JM	JE	AE	JP	JM	JE	AE
4	JP	JM	(JE)	AE	JP	JM	JE	AE
5	JP	(JM)	JE	AE	JP	JM	JE	AE
6	(JP)	JM	JE	AE	JP	JM	JE	AE
7	JP	JM	(JE)	AE	JP	JM	JE	AE
8	JP	(JM)	JE	AE	JP	JM	JE	AE
9	(JP)	JM	JE	AE	JP	JM	JE	AE
10	JP	JM	JE	(AE)	JP	JM	JE	AE

Partially Meets = Total of "yeses" in the JP column of the rating form = 4

Meets = Total of "yeses" in the JP and JM columns of the rating form = 4 + 3 = 7

Exceeds = Total of "yeses" in the JP, JM, and JE columns of the rating form = 4 + 3 + 2 = 9

PRESENTATION 13 – UNDERTAKE ANGOFF METHOD WITH PRACTICE ITEMS

Slide 118

TASK 3 ACTIVITY

PRACTICE ANGOFF METHOD

Slide 119

RATING PRACTICE ITEM 1

How is eighty-seven written in standard form?

A. 80
B. 87
C. 807
D. 870

Domain: Number and Operations

Construct: Whole Numbers

Subconstruct: Identify and count in whole numbers, and identify their relative magnitude

Knowledge or skill (content standard): Count, read, and write whole numbers

What makes it easy or difficult: the other answer choices are strong distractors, especially C.

GPL and GPD (performance standard):

Lowest GPD to answer correctly—Partially Meets: Read and write whole numbers up to 100 in words and in numerals.

Would 2 out of 3 JP learners answer the item correctly? . .
If yes, then circle JP
If no, then ask about JM . . .

Slide 120

RATING PRACTICE ITEM 2

What is the difference in time shown between these two clocks?

What makes the item easy/difficult?
Difficult—it is a two-step problem, and the numbers are not shown on the clocks

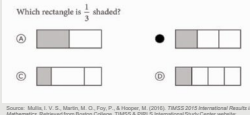
Lowest GPD to answer correctly?
Grade 3 Meets—Tell time using an analog clock to the nearest hour AND Solve problems, including real-world problems, involving elapsed time in hours. So above exceeds for grade 2

Would 2 out of 3 JP learners answer the item correctly? . .

Domain: Measurement
Construct: Time
Subconstruct: Tell time AND solve problems involving time
Knowledge or skill (content standard): Tell time using an analog clock AND identify or solve problems involving equivalences between different units of time

Slide 121

RATING PRACTICE ITEM 3



Source: Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 International Results in Mathematics*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2015/international-results/

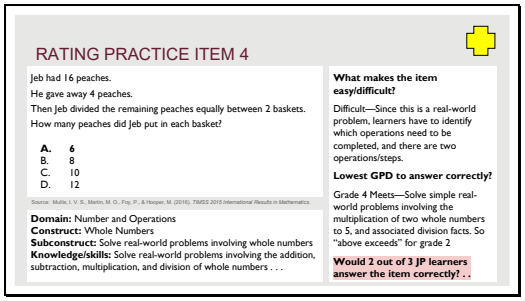
Domain: Number and Operations
Construct: Fractions
Subconstruct: Identify and represent fractions using objects, pictures, and symbols, and identify relative magnitude
Knowledge/skills: Express a visual representation of a fraction (picture, objects) in fractional notation

What makes the item easy/difficult? easy/difficult?
Difficult—understanding that the shaded portion must be 1/3 shaded rather than just that 1 out of 3 pieces must be shaded. C is a strong distractor.

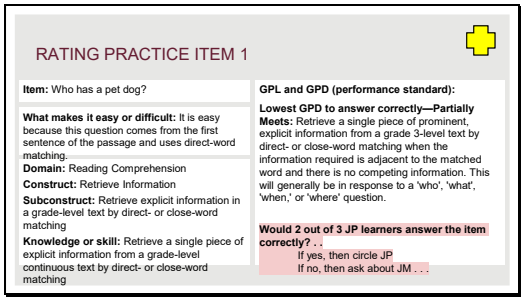
Lowest GPD to answer correctly?
Grade 3 Partially Meets—Identify everyday unit fractions represented as pictures in fractional notation. "No fit" for grade 2

Would 2 out of 3 JP learners answer the item correctly? . .

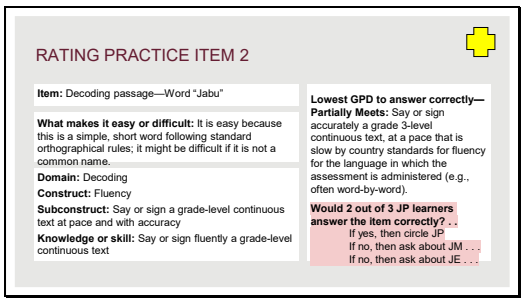
Slide 122



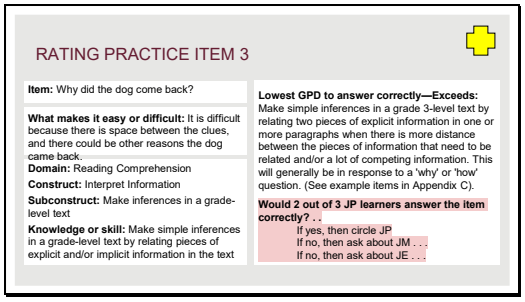
Slide 123



Slide 124



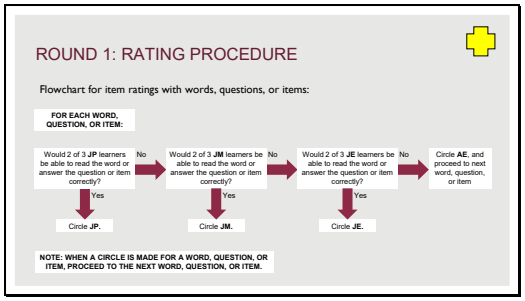
Slide 125



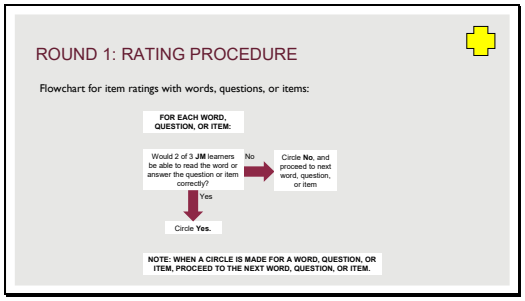
PRESENTATION 14 – ROUND 1

Slide 126

Slide 127

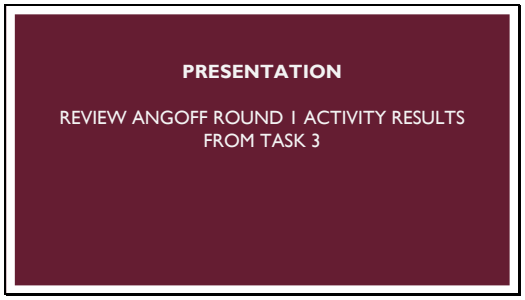


Slide 128

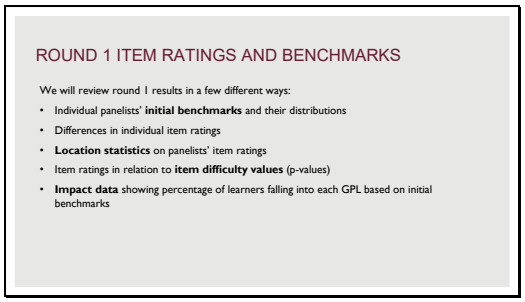


PRESENTATION 15 – PRESENTATION AND DISCUSSION OF ROUND 1 RESULTS

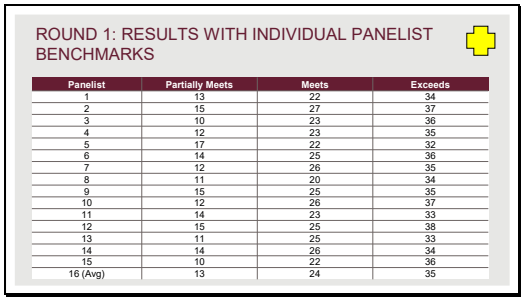
Slide 129



Slide 130



Slide 131

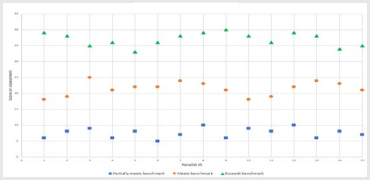


Panelist	Partially Meets	Meets	Exceeds
1	13	22	34
2	15	27	37
3	10	23	36
4	12	23	35
5	17	22	32
6	14	25	36
7	12	26	35
8	11	20	34
9	15	25	35
10	12	26	37
11	14	23	33
12	15	25	38
13	11	25	33
14	14	26	34
15	10	22	36
16 (Avg)	13	24	35

Slide 132

ROUND 1: RESULTS USING LOCATION STATISTICS

Location statistics for benchmarks:



Slide 133

ROUND 1: RESULTS BY ITEM

GRADE 2 RATING DISCUSSION

Where did we disagree?

Question: 12 + 7

Responses: JP: 1 JM: 6 JE: 5 AE: 0

"Meets": Add and subtract within 20

"Exceeds": Add and subtract within 30

Slide 134

ROUND 1: COMPARING RESULTS WITH ITEM DIFFICULTY

Item difficulty:

Item Number	P-Value	Item Number	P-Value
1	0.72	21	0.40
2	0.38	22	0.38
3	0.52	23	0.35
4	0.58	24	0.36
5	0.75	25	0.57
6	0.55	26	0.54
7	0.69	27	0.69
8	0.59	28	0.17
9	0.70	29	0.56
10	0.31	30	0.44
11	0.47	31	0.71
12	0.36	32	0.41
13	0.47	33	0.58
14	0.71	34	0.35
15	0.40	35	0.39
16	0.42	36	0.44
17	0.34	37	0.29
18	0.71	38	0.34
19	0.49	39	0.53
20	0.43	40	0.26

Slide 135

ROUND 1: RESULTS USING IMPACT DATA

Impact data:

Minimum Proficiency Levels	Round 1 Benchmark	Score Range	Percentage of Learners
Below Partially Meets	N/A	0–12	44.5%
Partially Meets	13	13-23	34.7%
Meets	24	24-34	17.6%
Exceeds	35	35-40	3.2%
Total			100.0%

PRESENTATION 16 – PRESENTATION ON ANGOFF ROUND 2

Slide 136

PRESENTATION

TASK 3: ANGOFF BENCHMARKING ROUND 2

Slide 137

ROUND 2 RATING PROCEDURE

- Make the **second round** of item ratings using the same process as with the first round, i.e., the four-step procedure.
- Conduct the round 2 item ratings on the following **guidance**:
  - Keep a focus on the **item content** in relation to the statements of knowledge and/or skill(s) in the GPF.
  - Maintain a consideration of the **item difficulty** as a basis for judgments.
  - Provide **adjustments** to their ratings based on their individual and independent judgments and the GPF.
  - Consider whether you are reasonably sure (2 out of 3 learners) would answer the item correctly.
  - Remember to consider **would** rather than **should** in making realistic ratings.

Slide 138

ANGOFF PROCEDURE: FOUR STEPS

Step 1: Identify and/or **conceptualize** three Just Partially Meets (JP), three Just Meets (JM), and three Just Exceeds (JE) learners based on an understanding of the GPF.

Step 2: Carefully read the first item on the assessment and consider the **knowledge or skills** required to answer the item correctly [– remember to consider whether the text on which the item is based is grade-appropriate]. Consider what makes the item easy or difficult (e.g., the wording of the item stem and the strength of the incorrect options, or distractors) and what kind of errors may be possible or reasonable.

Step 3: Building from Task 2, select the domain, construct, subconstruct, statement(s) of knowledge and/or skill(s), and GPLs/GPDs in the GPF that are most relevant for the item.

Slide 139

ANGOFF PROCEDURE: FOUR STEPS

Step 4: Based on an understanding of steps 1–3, follow this procedure:

- **Ask whether minimally proficient JP learners would be able to answer the item correctly**, i.e., are you reasonably sure (≥ 67 percent chance, or two out of the three JP learners)?
  - If "yes," circle JP and proceed to the next item.
  - If "no," ask whether minimally proficient JM learners would be able to answer the item correctly?
    - If "yes," circle JM and proceed to the next item.
    - If "no," ask whether minimally proficient JE learners would be able to answer the item correctly?
      - » If "yes," circle JE and proceed to the next item.
      - » If "no," circle AE and proceed to the next item.

Slide 140

ANGOFF PROCEDURE: FIVE STEPS

Step 1: Identify and/or **conceptualize** three Just Partially Meets (JP), three Just Meets (JM), and three Just Exceeds (JE) learners based on an understanding of the GPF.

Step 2: Estimate number of items JP, JM, and JE learners would be able to complete within the time limit (e.g., words in the oral reading passage the learners would attempt to read in a minute).

Step 3: Carefully read the first item on the assessment and consider the **knowledge or skills** required to answer the item correctly. Consider what makes the item easy or difficult (e.g., the wording of the item stem and the strength of the incorrect options, or distractors) and what kind of errors may be possible or reasonable.

Step 4: Building from Task 2, select the domain, construct, subconstruct, statement(s) of knowledge and/or skill(s), and GPLs/GPDs in the GPF that are most relevant for the item.

Slide 141

ANGOFF PROCEDURE: FIVE STEPS

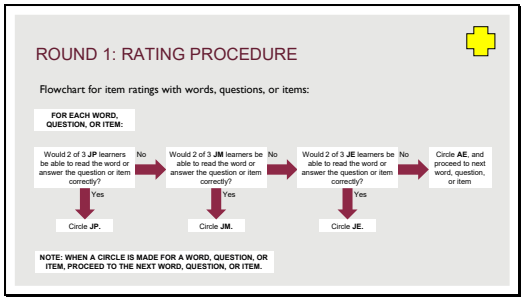
Step 5: Based on an understanding of steps 1–3, follow this procedure:

- **Ask whether minimally proficient JP learners would be able to answer the item correctly**, i.e., are you reasonably sure (≥ 67 percent chance, or two out of the three JP learners)?
  - If "yes," circle JP and proceed to the next item.
  - If "no," ask whether minimally proficient JM learners would be able to answer the item correctly?
    - If "yes," circle JM and proceed to the next item.
    - If "no," ask whether minimally proficient JE learners would be able to answer the item correctly?
      - » If "yes," circle JE and proceed to the next item.
      - » If "no," circle AE and proceed to the next item.

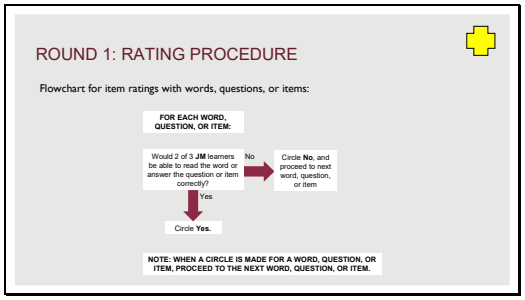
PRESENTATION 17 – ROUND 2

Slide 142

Slide 143



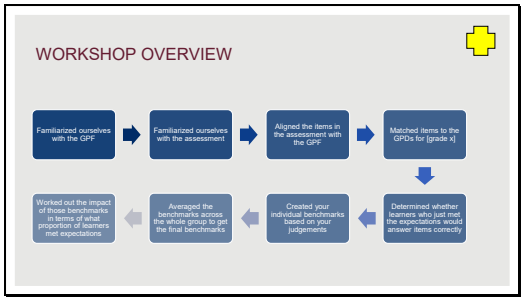
Slide 144



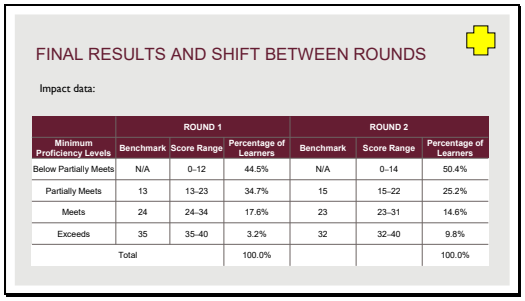
PRESENTATION 18 – PRESENTATION OF ROUND 2 RESULTS

Slide 145

Slide 146



Slide 147

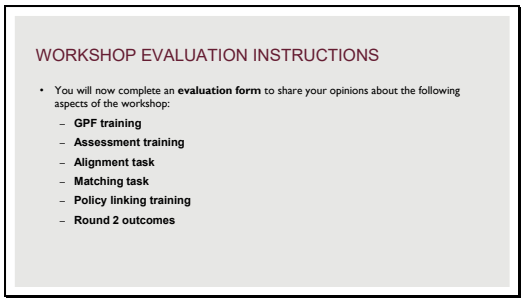


	ROUND 1			ROUND 2		
Minimum Proficiency Levels	Benchmark	Score Range	Percentage of Learners	Benchmark	Score Range	Percentage of Learners
Below Partially Meets	N/A	0–12	44.5%	N/A	0–14	50.4%
Partially Meets	13	13–23	34.7%	15	15–22	25.2%
Meets	24	24–34	17.6%	23	23–31	14.6%
Exceeds	35	35–40	3.2%	32	32–40	9.8%
Total			100.0%			100.0%

PRESENTATION 19 – WORKSHOP EVALUATION

Slide 148

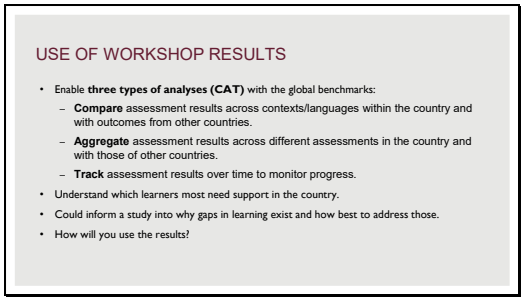
Slide 149



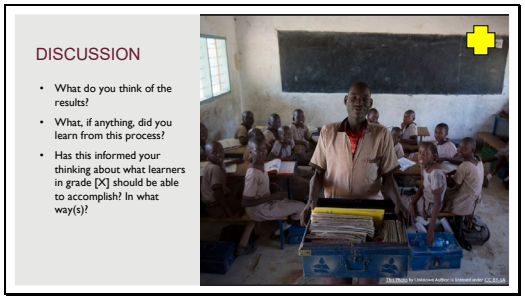
PRESENTATION 20 – CLOSING REMARKS AND PRESENTATION OF CERTIFICATES

Slide 150

Slide 151



Slide 152



Slide 153

Slide 154

ANNEX H – ALIGNMENT RATING FORM FOR TASK 1

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 15: Alignment Rating Form Template

						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.				
Question	**Domain**	**Construct reference**	**Subconstruct reference**	**Knowledge or skill**	**Fit**	**Domain**	**Construct reference**	**Subconstruct reference**	**Knowledge or skill**	**Fit**
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										

ANNEX I – ITEM RATING FORMS

Several example item rating forms are included below. Sample Form 1, including **Table 14**, is a form that can be used for setting three benchmarks on a 45-item **untimed assessment**. Additional items can be added, as needed. To adapt this form to set just one benchmark, facilitators need only remove the JP and JE columns and rename the AE column AM (Above Meets). Other sample forms are included below.

SAMPLE FORM 1. ASSESSMENT WITH 20 OBJECTIVE ITEMS (MULTIPLE CHOICE):

3 JP learners: _____

3 JM learners: _____

3 JE learners: _____

Name of the Panelist: _____

Panelist Code: _____

Directions: For each item, circle either a Just Partially Meeting Minimum Proficiency (JP), Just Meeting Minimum Proficiency (JM), Just Exceeding Minimum Proficiency (JE), or Above Exceeding Minimum Proficiency (AE).

Table 16: Item Rating Form Example for Untimed Assessments

Item no.	Round 1 individual and independent predictions				Round 2 individual and independent predictions			
	JP	JM	JE	AE	JP	JM	JE	AE
1	JP	JM	JE	AE	JP	JM	JE	AE
2	JP	JM	JE	AE	JP	JM	JE	AE
3	JP	JM	JE	AE	JP	JM	JE	AE
4	JP	JM	JE	AE	JP	JM	JE	AE
5	JP	JM	JE	AE	JP	JM	JE	AE
6	JP	JM	JE	AE	JP	JM	JE	AE
7	JP	JM	JE	AE	JP	JM	JE	AE
8	JP	JM	JE	AE	JP	JM	JE	AE
9	JP	JM	JE	AE	JP	JM	JE	AE
10	JP	JM	JE	AE	JP	JM	JE	AE
11	JP	JM	JE	AE	JP	JM	JE	AE
12	JP	JM	JE	AE	JP	JM	JE	AE
13	JP	JM	JE	AE	JP	JM	JE	AE
14	JP	JM	JE	AE	JP	JM	JE	AE
15	JP	JM	JE	AE	JP	JM	JE	AE
16	JP	JM	JE	AE	JP	JM	JE	AE
17	JP	JM	JE	AE	JP	JM	JE	AE
18	JP	JM	JE	AE	JP	JM	JE	AE
19	JP	JM	JE	AE	JP	JM	JE	AE
20	JP	JM	JE	AE	JP	JM	JE	AE
21	JP	JM	JE	AE	JP	JM	JE	AE
22	JP	JM	JE	AE	JP	JM	JE	AE
23	JP	JM	JE	AE	JP	JM	JE	AE
24	JP	JM	JE	AE	JP	JM	JE	AE
25	JP	JM	JE	AE	JP	JM	JE	AE
26	JP	JM	JE	AE	JP	JM	JE	AE
27	JP	JM	JE	AE	JP	JM	JE	AE
28	JP	JM	JE	AE	JP	JM	JE	AE
29	JP	JM	JE	AE	JP	JM	JE	AE

Item no.	Round 1 individual and independent predictions				Round 2 individual and independent predictions			
	JP	JM	JE	AE	JP	JM	JE	AE
30	JP	JM	JE	AE	JP	JM	JE	AE
31	JP	JM	JE	AE	JP	JM	JE	AE
32	JP	JM	JE	AE	JP	JM	JE	AE
33	JP	JM	JE	AE	JP	JM	JE	AE
34	JP	JM	JE	AE	JP	JM	JE	AE
35	JP	JM	JE	AE	JP	JM	JE	AE
36	JP	JM	JE	AE	JP	JM	JE	AE
37	JP	JM	JE	AE	JP	JM	JE	AE
38	JP	JM	JE	AE	JP	JM	JE	AE
39	JP	JM	JE	AE	JP	JM	JE	AE
40	JP	JM	JE	AE	JP	JM	JE	AE
41	JP	JM	JE	AE	JP	JM	JE	AE
42	JP	JM	JE	AE	JP	JM	JE	AE
43	JP	JM	JE	AE	JP	JM	JE	AE
44	JP	JM	JE	AE	JP	JM	JE	AE
45	JP	JM	JE	AE	JP	JM	JE	AE

Sample Form 2 should be used with **constructed response/open-ended items** on **untimed assessments**. Facilitators will need to make adjustments based on the number of points possible for each question and the total number of questions (the example below only includes space to rate five questions). There should be a row included for every possible point value per question and every question. Adjustments are also necessary if workshops will only include setting one benchmark (as described above).

SAMPLE FORM 2. ASSESSMENT WITH FIVE OPEN-ENDED ITEMS
(Item 1 has a score of 2 points, items 2 and 3 have a score of 4 points, item 4 has a score of 3 points, and item 5 has a score of 5 points).

3 JP learners: _____

3 JM learners: _____

3 JE learners: _____

Name of the Panelist: _____

Panelist Code: _____

Directions: For each item, circle either a Just Partially Meeting Minimum Proficiency (JP), Just Meeting Minimum Proficiency (JM), Just Exceeding Minimum Proficiency (JE), or Above Exceeding Minimum Proficiency (AE).

Table 17: Example Item Rating Form for Assessments with Constructed Response Questions

Item no.	Score point	Round 1 individual and independent predictions				Round 2 individual and independent predictions			
		JP	JM	JE	AE	JP	JM	JE	AE
1	1–1	JP	JM	JE	AE	JP	JM	JE	AE
1	1–2	JP	JM	JE	AE	JP	JM	JE	AE
2	2–1	JP	JM	JE	AE	JP	JM	JE	AE
2	2–2	JP	JM	JE	AE	JP	JM	JE	AE
2	2–3	JP	JM	JE	AE	JP	JM	JE	AE
2	2–4	JP	JM	JE	AE	JP	JM	JE	AE
3	3–1	JP	JM	JE	AE	JP	JM	JE	AE
3	3–2	JP	JM	JE	AE	JP	JM	JE	AE

Item no.	Score point	Round 1 individual and independent predictions				Round 2 individual and independent predictions			
		JP	JM	JE	AE	JP	JM	JE	AE
3	3–3	JP	JM	JE	AE	JP	JM	JE	AE
3	3–4	JP	JM	JE	AE	JP	JM	JE	AE
4	4–1	JP	JM	JE	AE	JP	JM	JE	AE
4	4–2	JP	JM	JE	AE	JP	JM	JE	AE
4	4–3	JP	JM	JE	AE	JP	JM	JE	AE
5	5–1	JP	JM	JE	AE	JP	JM	JE	AE
5	5–2	JP	JM	JE	AE	JP	JM	JE	AE
5	5–3	JP	JM	JE	AE	JP	JM	JE	AE
5	5–4	JP	JM	JE	AE	JP	JM	JE	AE
5	5–5	JP	JM	JE	AE	JP	JM	JE	AE

Sample Form 3 provides an example of a form that can be used for **timed assessments**. The example comes from a policy linking workshop focused on setting benchmarks for EGRA. There are additional columns necessary for timed assessments, as panelists need to first determine how many items/words a learner will attempt in the time allotted and then determine whether learners will answer each of the items/read each of those words correctly or not (only up to the number the panelist determines learners in that performance level will attempt). For example, if a panelist says that a JP learner will attempt 10 words, in the second step of the rating process for timed assessments, they will only rate whether the learner would correctly answer those first ten words (e.g., up to the word Wata, in the example below). Similar to the forms above, this form needs to be adjusted based on the total number of items as well as the number of benchmarks that will be set in the workshop. Another difference with this form is that rather than just including the item number, in this case, it includes the actual item (in this case "word" in a reading passage). The items could also be added to the above forms for clarify. This is usually only necessary when item numbers are not clearly marked on the assessment.

SAMPLE FORM 3. ORAL READING FLUENCY SUBTASK WITH 35 WORDS AND 5 READING COMPREHENSION ITEMS

3 JP learners: _____

3 JM learners: _____

3 JE learners: _____

Name of the Panelist: _____

Panelist Code: _____

Directions: For each item, circle either Just Partially Meeting Minimum Proficiency (JP), Just Meeting Minimum Proficiency (JM), Just Exceeding Minimum Proficiency (JE), or Above Exceeding Minimum Proficiency (AE).

Table 18: Example Item Rating Form for Timed Reading Assessment (in Hausa)

Word No.	Reading passage (Word)	Round 1: No. of words learners would attempt to read in a minute			Round 1 individual and independent ratings				Round 2: No. of words learners would attempt to read in a minute			Round 2 individual and independent ratings			
		JP	JM	JE	JP	JM	JE	AE	JP	JM	JE	JP	JM	JE	AE
1	Kande	1	1	1	JP	JM	JE	AE	1	1	1	JP	JM	JE	AE
2	da	2	2	2	JP	JM	JE	AE	2	2	2	JP	JM	JE	AE
3	abokiyarta	3	3	3	JP	JM	JE	AE	3	3	3	JP	JM	JE	AE
4	Delu	4	4	4	JP	JM	JE	AE	4	4	4	JP	JM	JE	AE
5	sukan	5	5	5	JP	JM	JE	AE	5	5	5	JP	JM	JE	AE
6	tafi	6	6	6	JP	JM	JE	AE	6	6	6	JP	JM	JE	AE
7	Makaranta	7	7	7	JP	JM	JE	AE	7	7	7	JP	JM	JE	AE

Word No.	Reading passage (Word)	Round 1: No. of words learners would attempt to read in a minute			Round 1 individual and independent ratings				Round 2: No. of words learners would attempt to read in a minute			Round 2 individual and independent ratings			
		JP	JM	JE	JP	JM	JE	AE	JP	JM	JE	JP	JM	JE	AE
8	tare	8	8	8	JP	JM	JE	AE	8	8	8	JP	JM	JE	AE
9	kullum.	9	9	9	JP	JM	JE	AE	9	9	9	JP	JM	JE	AE
10	Wata	10	10	10	JP	JM	JE	AE	10	10	10	JP	JM	JE	AE
11	rana	11	11	11	JP	JM	JE	AE	11	11	11	JP	JM	JE	AE
12	Kande	12	12	12	JP	JM	JE	AE	12	12	12	JP	JM	JE	AE
13	ta	13	13	13	JP	JM	JE	AE	13	13	13	JP	JM	JE	AE
14	zo	14	14	14	JP	JM	JE	AE	14	14	14	JP	JM	JE	AE
15	da	15	15	15	JP	JM	JE	AE	15	15	15	JP	JM	JE	AE
16	aiki	16	16	16	JP	JM	JE	AE	16	16	16	JP	JM	JE	AE
17	daga	17	17	17	JP	JM	JE	AE	17	17	17	JP	JM	JE	AE
18	makaranta.	18	18	18	JP	JM	JE	AE	18	18	18	JP	JM	JE	AE
19	Delu	19	19	19	JP	JM	JE	AE	19	19	19	JP	JM	JE	AE
20	ta	20	20	20	JP	JM	JE	AE	20	20	20	JP	JM	JE	AE
21	taimaka	21	21	21	JP	JM	JE	AE	21	21	21	JP	JM	JE	AE
22	mata.	22	22	22	JP	JM	JE	AE	22	22	22	JP	JM	JE	AE
23	Kande	23	23	23	JP	JM	JE	AE	23	23	23	JP	JM	JE	AE
24	ta	24	24	24	JP	JM	JE	AE	24	24	24	JP	JM	JE	AE
25	samu	25	25	25	JP	JM	JE	AE	25	25	25	JP	JM	JE	AE
26	yabo	26	26	26	JP	JM	JE	AE	26	26	26	JP	JM	JE	AE
27	a	27	27	27	JP	JM	JE	AE	27	27	27	JP	JM	JE	AE
28	ajinsu.	28	28	28	JP	JM	JE	AE	28	28	28	JP	JM	JE	AE
29	Kande	29	29	29	JP	JM	JE	AE	29	29	29	JP	JM	JE	AE
30	da	30	30	30	JP	JM	JE	AE	30	30	30	JP	JM	JE	AE
31	Delu	31	31	31	JP	JM	JE	AE	31	31	31	JP	JM	JE	AE
32	Sun	32	32	32	JP	JM	JE	AE	32	32	32	JP	JM	JE	AE
33	ji	33	33	33	JP	JM	JE	AE	33	33	33	JP	JM	JE	AE
34	dadi	34	34	34	JP	JM	JE	AE	34	34	34	JP	JM	JE	AE
35	sosai.	35	35	35	JP	JM	JE	AE	35	35	35	JP	JM	JE	AE
Total															

The second part of Sample Form 3 can also be used with timed assessments, such as EGRA/EGMA, or other assessments with conditional questions. This example comes from the reading comprehension subtask of the EGRA. The EGRA reading comprehension subtask requires that enumerators only read the number of reading comprehension questions to learners that align with the number of words the learner attempted, as shown in the "condition" column of the below form. As such, it is important that when rating a subtask, such as the reading comprehension subtask from EGRA, that panelists consider the number they estimated learners in a specific performance level would have attempted. Thus, expanding on the above example, this would mean that if a panelist estimates that JP learners would read 10 words in the passage, then those JP learners would only be asked the first question from the table below (per the criteria listed in the "condition" column). So, they should only rate the first question as yes/no for JP learners. This form will need to be adapted based on the number of items, the conditions for those items, the items themselves, and the number of benchmarks.

Table 19: Example Item Rating Form for Conditional Reading Comprehension Questions (in Hausa)

Item no.	Condition	Questions	Round 1 individual and independent ratings				Round 2 individual and independent ratings			
			JP	JM	JE	AE	JP	JM	JE	AE
1	≤ 9 words attempted	Su waye abokan juna? *{Kande da Delu}*	JP	JM	JE	AE	JP	JM	JE	AE

2	≤ 18 words attempted	Ina suke tafiya kullum? *{Makaranta}*	JP	JM	JE	AE	JP	JM	JE	AE
3	≤ 22 words attempted	Me Kande ta zo da shi daga makaranta? *{Aiki}*	JP	JM	JE	AE	JP	JM	JE	AE
4	≤ 28 words attempted	Wa ya taimaka wa Kande? *{Delu}*	JP	JM	JE	AE	JP	JM	JE	AE
5	≤ 35 words attempted	Me ya faru a ajin su Kande? *{Kande ta Samu yabo/ yabo}*	JP	JM	JE	AE	JP	JM	JE	AE
Total										

ANNEX J – PRECISION, ACCURACY AND CONSISTENCY STATISTICS

OUTLIER PANELISTS

To identify panelist whose ratings are clear outliers relative to the other member of the panel, the Interquartile or Tukey's fences model will be used.

All cut scores bellow $Q_1 - K(Q_3 - Q_1)$ or above $Q_1 + K(Q_3 - Q_1)$ will be considered to be outliers.

Where,

$Q_1 =$ lower quartile

$Q_3 =$ upper quartile

$K = 1.5$

$K = 1.5$ was prosed as the multiplier by Tukey (1977) and has been predominantly used since.

INTER-RATER CONSISTENCY

Inter-rater consistency is calculated using Ferdous & Plake's (2005) generalized formula for multiple benchmarks. The procedure is based on the absolute difference between two panelists' responses for all possible pairs of panelists. This index can be calculated both at the item level (i.e., for panelists' ratings of items) and for the entire test. The inter-rater consistency for an item *i* is defined as the proportion of the total observed consistencies to the total number of possible consistencies. Total observed consistency is defined by the sum of the absolute differences of all possible pair of panelists' responses.

Inter-rater consistency for item *i* is,

$$I_i = 1 - \frac{TOI_i}{TI} \tag{4}$$

$$TOI_i = \sum_{a,b=1\ a\neq b}^{\frac{Z!}{2(z-2)!}} |R_{ai} - R_{bi}| \tag{5}$$

$$TI = d * \frac{z!}{2(z-2)!} \tag{6}$$

Where,

I_i = Inter-rater consistency for item *i*. High number (0.80 and above) indicates high consistency and low number indicates low consistency

TOI_i = Total observed inter-rater inconsistency for item *i*

TI = Total possible inter-rater inconsistency for each item

Z = Number of panelists in the standard setting study

R_{ai} = Panelist a's response to item *i*; k= 1, 2, 3, 4 (1= partially meets, 4=above exceeds) or 1, 2 (1= meets, 2 = above meets for one benchmark)

R_{bi} = Panelist b's response to item *i*; k = 1, 2, 3, 4 (1= partially meets, 4=above exceeds) or 1, 2 (1 = below meets, 2 = meets for one benchmark)

d = Maximum absolute possible difference between two judges' ratings.

If there are four achievement level categories, one judge may give a rating of 1 (partially meets) to the item and the other judge may give a rating of 4 (above exceeds minimum proficiency); so, the possible maximum absolute difference is 3. If there are two achievement level categories, one judge may give a rating of 1 (meets) to the item and the other judge may give a rating of 2 (above meets); so, the possible maximum absolute difference is 1.

Overall consistency for *n* number of items on the test across all the panelists is:

$$I = n^{-1} \sum_{i=1}^{n} \quad I_i \tag{7}$$

How to Calculate Inter-Rater Consistency

Calculate inter-rater consistency for one item and the entire assessment.

Step 1: Calculate the total possible inter-rater inconsistency.

i. Calculate the factorial of the number of panelists.
ii. Calculate the factorial of two multiplied by the number of panelists minus two.
iii. Divide the results from sub-step 1 by the result from sub-step 2.
iv. Multiply the maximum absolute possible difference between two judges' ratings by the result from sub-step 3. This result is the total possible inter-rater inconsistency.

Step 2: Calculate the inter-rater consistency for one item.

i. Take the absolute value of the difference in ratings between each panelist.
ii. Add together all of the absolute values. The result is the total observed inter-rater inconsistency for the item.
iii. Divide the total observed inter-rater inconsistency for the item by the total possible inter-rater inconsistency. The result is the inter-rater consistency for the item.
iv. Repeat sub-steps 1 through 3 for each item of the assessment.

Step 3: Calculate the inter-rater consistency for the assessment.

i. Add together the inter-rater inconsistency of each item.
ii. Divide the sum by the number of items on the assessment. The result is the inter-rater consistency.

STANDARD ERROR (SE)

The standard error (SE) is calculated for each benchmark separately using the following formulas:

$$SE(Partially\ Meets\ Benchmark) = \frac{SD_{(1)}}{\sqrt{z-1}} \tag{8}$$

$$SE(Meets\ Benchmark) = \frac{SD_{(2)}}{\sqrt{z-1}} \tag{9}$$

$$SE(Exceeds\ Minimum\ Proficiency\ Benchmark) = \frac{SD_{(3)}}{\sqrt{z-1}} \tag{10}$$

Where,

$SD_{(1)}$ = Standard deviation of partially meets benchmark for all z panelists

$SD_{(2)}$ = Standard deviation of meets benchmark for all z panelists

$SD_{(3)}$ = Standard deviation of exceeds minimum proficiency benchmark for all z panelists

z = Total number of panelists

How to Calculate Standard Error of Benchmarks

Calculate the SE for one benchmark.

1) Take the benchmarks of all the panelists and calculate the standard deviation of the panelists' benchmarks.
2) Subtract 1 from the total number of panelists.
3) Calculate the square root of the result from step 2.
4) Divide the result from step 1 by the results from step 3. The result is the SE for that benchmark.
5) Repeat steps 1 through 4 as necessary for each benchmark.

CONFIDENCE INTERVALS

The 95% confidence interval for each benchmark is calculated using the following formula:

$$Confidence\ Interval = B_i \pm 1.96 \times SE_i$$

Where,

$B_i =$ The relevant benchmark for the MPL

$SE_i =$ The relevant Standard Error for the MPL

ANNEX K – INVITATION LETTER TEMPLATE FOR OBSERVERS

This annex includes a letter template for observers from the government/assessment agency and other stakeholder organizations. All details that need to be filled in are included in brackets. The letter should be modified as needed to fit the context.

[Date]

[Name]
[Role]
[Agency]
[Address/location]

Invitation to a Policy Linking for Measuring Global Learning Outcomes Workshop

Dear [Name],

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 "Policy Linking for Measuring Global Learning Outcomes" (called Policy Linking throughout). 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. [USAID is using similar indicators for its global reporting].

Through Policy Linking, countries/assessment agencies link their national assessments to a common global reporting scale using benchmarks. Setting the benchmarks requires judgments on learner performance by panels of curriculum experts and teachers. The benchmarks will allow determinations of the percentage of learners achieving minimum proficiency in reading and mathematics.

[Country/Assessment Agency] is planning to host [a/an in-person/remote] Policy Linking Workshop from **[start date] to [end date]**. **Registration will be at [time] on [date]**. The workshop will focus on linking [Assessment Name(s)] with SDG 4.1.1 for [Grades X and Y]. There will be [X number] panels, [one for Grade Assessment Language X and one for Grade Assessment Language Y – may include more than two as well]. Panelists will be guided through a systematic process that involves reviewing assessment materials and setting benchmarks for [Grade Assessment Language(s)].

Up to [number] administrators from [Agency] are invited to participate as observers. Participation in the workshop will provide an opportunity for the selected administrators to: 1) build on the outputs from the National Reading Framework Workshop, 2) learn more about the global policy linking method for reporting on SDG 4.1.1, and 3) provide background and experience so policy linking can be scaled up [in/with Country/Assessment] to assessments for other grade levels, subject areas, and languages.

Activity Name	Arrival Date	Departure Date	Venue
[Name of workshop]	[Date] Registration at [Time]	[Date] Last session ends by [Time]	[Venue] for workshop and [Hotel] for accommodations for out-of-town participants

[Logistical details, e.g., who will cover transportation costs, accommodation, per diems, lunches]

If you have questions or require further clarification, please contact [Name] via phone [number]. Please kindly confirm your participation by [Date]. Your participation in this workshop is crucial and we look forward to collaborating with you.

Sincerely,

[Name and Title]

ANNEX L – INVITATION LETTER TEMPLATE FOR WORKSHOP PANELISTS

This annex includes a letter template for panelists, both curriculum experts and teachers. 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 a Policy Linking Workshop

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 "Policy Linking for Measuring Global Learning Outcomes" (called Policy Linking throughout). 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 Policy Linking, countries/assessment agencies will link their national assessments to a common global reporting scale using benchmarks. Setting the benchmarks requires judgments by panels of teachers.

[Country/Assessment Agency] is planning to host [a/an in-person/remote] Policy Linking Workshop from **[start date] to [end date]. Registration will be at [time] on [date]**. The workshop will focus on linking [Assessment Name(s)] with SDG 4.1.1 for [Grades X and Y]. There will be [X number] panels, [one for Grade Assessment Language X and one for Grade Assessment Language Y – may include more than two as well]. Panelists will include master teachers and curriculum experts, and they will be guided through a systematic process that involves reviewing assessment materials and setting benchmarks for [Grade Assessment Language(s)].

[Government Ministry/Assessment Agency] needs a total of [Number of Panelists] 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 panelists, who will gain an increased understanding of international standards for learner performance.

Activity Name	Arrival Date	Departure Date	Venue
[Name of workshop]	[Date] Registration at [Time]	[Date] Last session ends by [Time]	[Venue] for workshop and [Hotel] for accommodations for out-of-town participants

[Logistical details, e.g., who will cover transportation costs, accommodation, per diems, lunches]

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 M – PANELIST DEMOGRAPHIC INFORMATION

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: 1) Reading
2) Mathematics

Grade level: _____

Language: _____

Name: _____

Occupation: _____

Region where you teach/work: _____

District where you teach/work: _____

Email: _____

Mobile Number: _____

Gender: 1) Female
2) Male

Ethnicity (if relevant): _____

Education Level: _____

Years of Experience/Expertise: _____

Years Teaching/Working with Relevant Grade and Subject Level: _____

Professional Organization/Affiliation (e.g., school, ministry, etc.): _____

Prior Training(s) in Reading/Mathematics (answer only for the subject for which you are serving as a panelist:

1) No
2) Yes

Experience teaching learners with disabilities:

1) No
2) Yes

Experience working with conflict- and crisis-affected population:

1) No
2) Yes

Native Language: _____

Language(s) Used for Classroom Instruction (for teachers only): _____

ANNEX N – PRE-WORKSHOP STATISTICS

The data analyst and/or lead facilitator should calculate the following statistics before the policy linking workshop. The method used to calculate these statistics will vary and is dependent on the model of assessment used. For assessments where classical test theory statistics are valid (for example, where the whole cohort or a representative sample of learners takes all items in the assessment), then the classical test theory (CTT) approach should be followed. Where a complex sampling design is used with item response theory (IRT) analysis, then the IRT approach should be followed.

ITEM DIFFICULTY

Item difficulty informs facilitators and panelists on how difficult an item is based on how learners performed on the item in the most recent iteration of the assessment. The data analyst should calculate the empirical item difficulty level using the following steps:

Classical Test Theory

1. Calculate empirical item difficulty level for each item by calculating the proportion of learners who get the item right. This is the information you will present panelists between benchmark rating Round 1 and Round 2.

Item Response Theory

1. Depending on the IRT model used, three key parameters are typically reported for an item: difficulty discrimination and guessing. If a 1-parameter IRT model is used, or if it can be assumed that items were rotated across the testing cohort randomly and with approximately equal exposure rate, then item difficulty can be calculated using the same approach as CTT.
2. Where a 2- or 3-parameter model is used and there are concerns that there are differences in the ability of learners who took each item, the IRT item difficulty parameter should be calculated with a 0.50 response probability. When sharing this information with panelist, care should be taken to explain the IRT scale as item difficulty calculated from IRT is less intuitive than for CTT.

DATA DISTRIBUTIONS

The data analyst can prepare information on the data distributions from the most recent iteration of the assessment being linked to the GPF and SDG 4.1.1 ahead of the workshop, though the data is not needed until Day 4, between Round 1 and 2 ratings. Preparing ahead of time saves a step during the usually constrained timeline during the workshop.

Classical Test Theory

To prepare the distributions, the data analyst will analyze the number and percentage of learners who took the assessment that received an overall score of zero, the same for learners that received an overall score of one, and so on through the highest score possible on the assessment. They will use that information to prepare a table like the one presented in **Table 20**, that contains the appropriate formulas for use in excel.

Table 20: Template Data Distribution Table (CTT)

	A	B	C	D
1	**Score**	**Frequency**	**Percent**	**Cumulative Percent**
2	0	Insert frequency from most recent iteration of assessment	=B2/B(n+1)*100	=C2
3	1	Insert frequency from most recent iteration of assessment	=B3/B(n+1)*100	=D2+C3
4	2	Insert frequency from most recent iteration of assessment	=B4/B(n+1)*100	=D3+C4
…	…	…	…	…

n	Maximum score	Insert frequency from most recent iteration of assessment	=Bn/B(n+1)*100	=D(n-1)+Cn (this should equal 100)
n+1	**Total**	=SUM(B2:Bn)		

Item Response Theory

Number-correct scoring approach will be used translate the benchmarks from the Angoff process to the IRT scale location that will be then used to calculate the percentage of students reaching each of the attainment benchmarks. To prepare the distributions, the data analyst will need to complete the following steps:

1. Order the items being used in the policy linking workshop by difficulty on the underlying IRT scale adjusted for the response probability of 0.67.
2. Calculate the proportion of students meeting and exceeding that ability estimate using the approach that was used in the most recent iteration of the assessment.

They will use that information to prepare a table like the one presented in **Table 21**.

Table 21: Template Data Distribution Table (IRT)

	A	B	C
1	**Item ID**	**Item difficulty**	**Estimated percentage achieving this difficulty of higher**
2	Easiest item	Insert IRT difficulty from most recent iteration of assessment	Insert percentage from most recent iteration of assessment
3	Next easiest item	Insert IRT difficulty from most recent iteration of assessment	Insert percentage from most recent iteration of assessment
4	Next easiest item	Insert IRT difficulty from most recent iteration of assessment	Insert percentage from most recent iteration of assessment
…	…	…	…
n	Hardest item	Insert IRT difficulty from most recent iteration of assessment	Insert percentage from most recent iteration of assessment

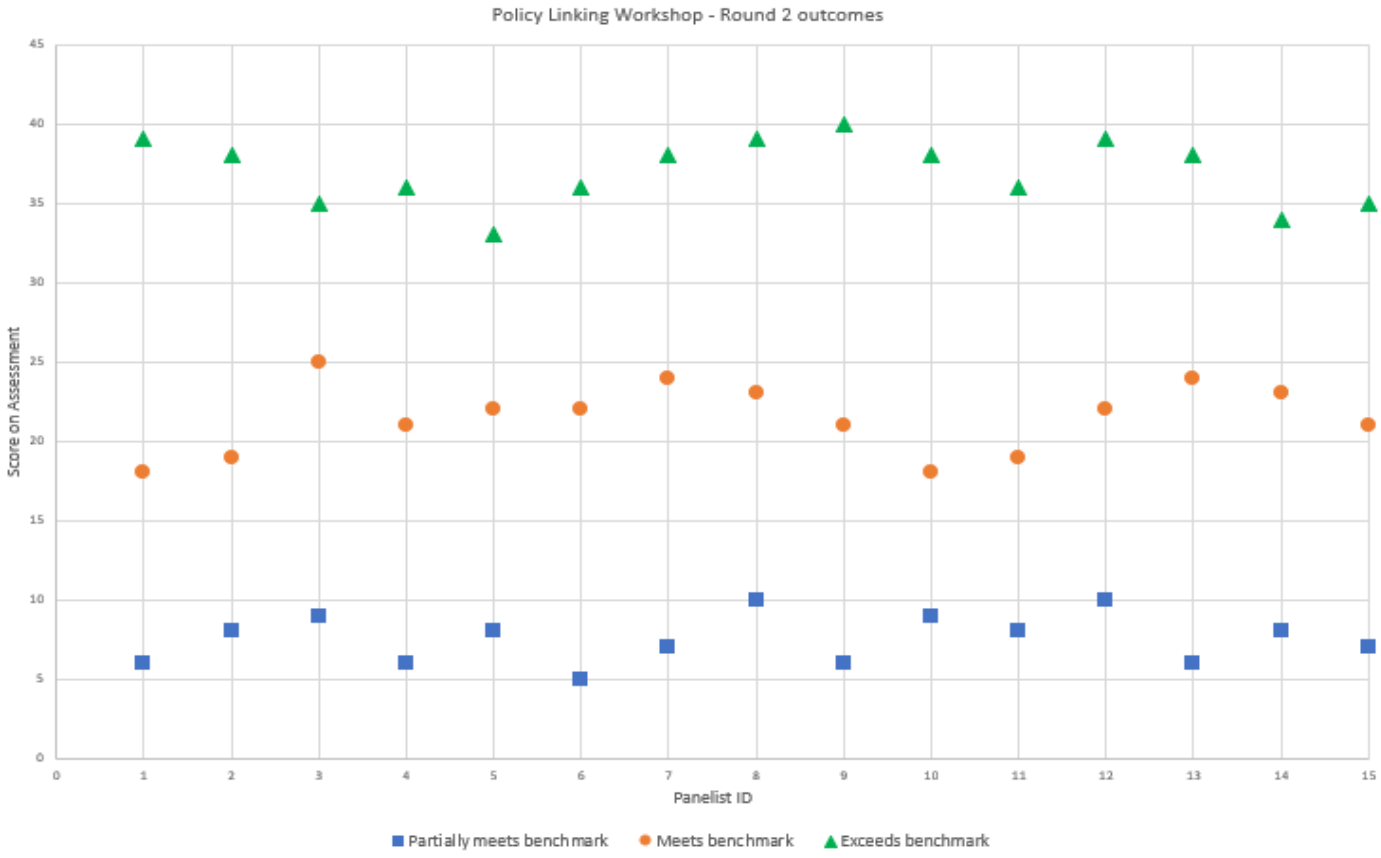
In the workshop, the impact data for each MPL will be determined by using the benchmark from the Angoff process to identify which if the items in **Table 21**, which are in item difficulty order, should be used to determine the proportion of learners meeting the MPL. For example, if the benchmark for 'meets' minimum proficiency is 13, then the 13th item in the ordered item booklet will be used to determine the proportion of learners meeting minimum proficiency.

ANNEX O – FEEDBACK DATA EXAMPLES AND INSTRUCTIONS

NORMATIVE INFORMATION (SOMETIMES CALLED LOCATION STATISTICS)

After each round of ratings, the data analyst should create a graph like the one in **Figure 37** that shows each of the panelists' unique panelist numbers (known only to them) and their benchmark for each of the GPLs. The graph can be created by using the Scatterplot chart type in Excel with data on the panelist-level benchmarks by GPL.

Figure 37: Example Normative Data on Panelist Ratings



IMPACT INFORMATION

To generate the impact information, the data analyst should take the panel-level benchmarks set by the panelists for each GPL and, using the data distributions, identify the percentage of learners who would fall into each GPL based on the most recent iteration of the assessment.

Table 22: Template Impact Data Table

MPL	Benchmark	Score Range	Percentage of Learners
Below partially meets	N/A	0 – (x-1)	Insert percentage calculated from the data distribution
Partially meets	x	x – (y-1)	Insert percentage calculated from the data distribution
Meets	y	y – (z-1)	Insert percentage calculated from the data distribution
Exceeds	z	z – maximum score	Insert percentage calculated from the data distribution
Total			100.0

ANNEX P – AGENDA TIMINGS FOR WORKSHOP

Table 23 shows the sessions and timings for each task in the workshop. Some timings are flexible since the length of the session will depend on things like the number of items being used and the speed at which consensus is reached, where required. The project team will need to use these to create their agenda. Any specific requirements for the session are shown in the 'Notes' column. Sample agendas for in-person and remote workshops are shown in **Annex Q** and can be adapted by countries if required.

Table 23: Agenda for Workshop

Task	Session	Presentation	Facilitator	Time required	Notes
Opening	Welcome and introductions	1	Lead facilitator	30 minutes	Should be completed in one session
	Address by government(s) representatives, assessment agency (if relevant), and donor organization (if relevant)	1		To be determined by country	
	Overview of agenda, objectives and high-level summary of method	2	Lead facilitator	30 minutes	
Familiarization	Familiarization with GPF	3	All facilitators	120 minutes	May take place in advance of workshop
	Familiarization with Assessment Instrument	4	Content Facilitator	120 minutes	May take place in advance of workshop
	Evaluation	19	Lead facilitator	10 minutes	
Alignment (Task 1)	Train panelists on the alignment task (including practice items)	5	Lead facilitator	60 minutes	
	Panelists undertake alignment activity (independent activity)	6	Content facilitator	120 minutes	Time will vary depending on number of items. Activity should take place at the end of a session to allow flexibility and time for facilitators to collate results before next activity
	Presentation and discussion on the alignment results	7	Lead facilitator	45 minutes	
	Evaluation	19	Lead facilitator	10 minutes	
Matching (Task 2)	Train panelists on matching task	8	Lead facilitator	60 minutes	
	Panelists undertake matching task (group activity)	9	Content facilitator	120 minutes	Time will vary depending on number of items. Activity should take place at the end of a session to allow flexibility, particularly as achieving consensus can take longer
	Presentation and discussion on the matching results	10	Lead facilitator	30 minutes	
	Evaluation	19	Lead facilitator	10 minutes	

Task	Session	Presentation	Facilitator	Time required	Notes
Benchmarking (Task 3)	Overview of global standards and benchmarking approach	11	Lead facilitator	30 minutes	
	Train panelists on Angoff method	12	Lead facilitator	60 minutes	
	Panelists undertake Angoff method with practice items	13	Content facilitator	30 minutes	
	Round 1	14	Content facilitator	120 minutes	Time will vary depending on number of items. Activity should take place at the end of a session to allow flexibility and to provide time for facilitators to collate results before next activity
	Presentation and discussion of Round 1 results and impact data	15	Lead facilitator	60 minutes	
	Presentation on Angoff Round 2	16	Lead facilitator	15 minutes	
	Round 2	17	Content facilitators	120 minutes	Time will vary depending on number of items. Activity should take place at the end of a session to allow flexibility and to provide time for facilitators to collate results before next activity
	Presentation of round 2 results	18	Lead facilitators	30 minutes	
	Evaluation	19	Lead facilitator	10 minutes	
Close	Closing remarks and presentation of certificates	20	To be determined by country	To be determined by country	

ANNEX Q – SAMPLE AGENDAS FOR A IN-PERSON AND REMOTE WORKSHOPS

Table 24: Sample Agenda for In-Person Workshop

Time	Activity	Presentation	Facilitation
Time	**Day 1**	**Presentation**	**Facilitation**
09:00–09:30	Registration		Project team
09:30–10:45	Opening, introductions, agenda, and logistics	1	Government/ assessment agency, donors, and implementing partners (if relevant) as well as lead facilitators
10:45–11:00	Tea break		--
11:00–11:45	Background, objective, and overview of method	2	Lead facilitator
11:45–12:30	Overview of the GPF and review of the GPDs	3	All facilitator
12:30–13:30	Lunch break		--
13:30–14:45	Review the GPF	3	All facilitator
14:45–15:30	Overview of the assessment(s)	4	Content facilitator
15:30–15:45	Tea break		--
15:45–16:45	Review of assessment items	4	Content facilitator
16:45–17:00	Day 1 closing and preview of Day 2		Lead facilitator
Time	**Day 2**	**Presentation**	**Facilitation**
09:00–09:30	Introduction of Day 2 and solving issues of Day 1		Lead facilitator
09:30–10:30	Taking the assessment		Content facilitator
10:30–10:45	Tea break		--
10:45–12:30	Review GPF and identify any remaining issues		Lead facilitator
12:30–13:30	Lunch break		--
13:30–14:15	Discussion to clarify outstanding issues from the morning session		All facilitator
14:15–15:00	Task 1: Training on alignment exercise	5	Lead facilitator
15:00–15:15	Tea break		--
15:15–16:00	Task 1: Small group discussions on first 5 items	6	Content facilitator
16:00–16:45	Task 1: Plenary discussion on questions that came up in groups	6	Content facilitator
16:45–17:00	Day 2 closing and preview of Day 3		Lead facilitator
Time	**Day 3**	**Presentation**	**Facilitation**
09:00–09:15	Introduction of Day 3 and solving issues of Day 2		Lead facilitator
09:15–10:30	Task 1: Alignment exercise	6	Content facilitator
10:30–10:45	Tea break		--
10:45–12:30	Task 1: Alignment exercise (continued)	6	Content facilitator
12:30–13:30	Lunch break		--
13:30–14:15	Task 1: Presentation and discussion of alignment results	7	Lead facilitator
14:15–15:00	Task 2: Training on matching exercise	8	Content facilitator
15:00–15:15	Tea break		--
15:15–16:45	Task 2: Start matching exercise	9	Content facilitator
16:45–17:00	Day 3 closing and preview of Day 4		

Time	Activity	Presentation	Facilitation
Time	**Day 4**	**Presentation**	**Facilitation**
09:00–09:15	Introduction of Day 4 and solving issues of Day 3		Lead facilitator
09:15–10:30	Task 2: Small groups complete matching exercise together	9	Content facilitator
10:30–10:45	Tea break		--
10:45–12:30	Task 2: Plenary discussion on matching results	10	Content facilitator
12:30–13:30	Lunch break		--
13:30–14:00	Overview of global standards and benchmarking approach	11	Lead facilitator
14:00–14:45	Task 3: Training on Angoff method	12	Lead facilitator
14:45–15:15	Task 3: Angoff practice in small groups	13	Lead facilitator
15:15–15:30	Tea break		--
15:30–16:15	Task 3: Plenary discussion on questions that arose in small groups		Lead facilitator
16:15–16:45	Task 3: Round 1 Angoff	14	Content facilitator
16:45–17:00	Day 4 closing and preview of Day 5		Lead facilitator
17:00–18:00	Consultation hour in which panelists can consult facilitators		All facilitators
Time	**Day 5**	**Presentation**	**Facilitation**
09:00–09:15	Introduction of Day 5 and solving issues of Day 4		Lead facilitator
09:15–10:30	Task 3: Continue Round 1 ratings	14	Content facilitator
10:30–10:45	Tea break		--
10:45–12:30	Task 3: Complete Round 1 ratings	14	Content facilitator
12:30–13:30	Lunch break		--
13:30–15:15	Task 3: Presentation and discussion of Round 1 results	15	Lead facilitator
15:15–15:30	Tea break		--
15:30–16:30	Task 3: Review Round 1 ratings in small groups to discuss all items where there was disagreement		Content facilitator
16:30–17:00	Task 3: Share and discuss item difficulty and impact data	15	Lead facilitator
17:00–17:15	Day 5 closing and preview of Day 6		Lead facilitator
17:15–18:15	Consultation hour in which panelists can consult facilitators		All facilitators
Time	**Day 5**	**Presentation**	**Facilitation**
09:00–09:15	Introduction of Day 6 and solving issues of Day 5		Lead facilitator
09:15–09:30	Task 3: Presentation on Round 2 Angoff	16	Lead facilitator
09:30–10:30	Task 3: Round 2 Angoff	17	Content facilitator
10:30–10:45	Tea break		--
10:45–12:30	Task 3: Complete Round 2 ratings	17	Content facilitator
12:30–13:30	Lunch break		--
13:30–14:30	Workshop evaluation	19	Lead facilitator
14:30–15:15	Task 3: Presentation of round 2 results	18	Lead facilitator
15:15–15:30	Tea break		--
15:30–16:30	Discuss outcomes and final panelist questions		Lead facilitator
16:30–17:00	Closing and logistics	20	Lead facilitator

Table 25: Example Agenda for Remote Preparation Session 1

(Recommend holding two weeks before the workshop)

Timing	Activity	Presentation	Facilitator
0–15 minutes	Welcome and introductions	1	Lead facilitator
15–40 minutes	Overview of policy linking	2	Lead facilitator
40–55 minutes	Purpose of preparation session		Process facilitator
55–60 minutes	**Comfort break**		
60–80 minutes	Overview of the GPF	3	Lead or content facilitator
80–100 minutes	[Grade and Subject] GPF Review	3	Lead or content facilitator
100–110 minutes	Explanation of inter-session activities		Lead facilitator
110–120 minutes	Closing remarks		Lead facilitator

Panelist inter-session activities:

- Review [Grade and Subject] GPF and identify any elements that are unclear (submit one week prior to workshop)

Table 26: Example Agenda for Remote Preparation Session 2

(Recommend holding two days after the first preparatory session)

Timing	Activity	Presentation	Facilitator
0–15 minutes	Welcome and purpose of the preparation session		Lead facilitator
15–30 minutes	Overview of the [assessment name]	4	Content or lead facilitator
30–55 minutes	Review each item on the [assessment]	4	Content or lead facilitator
55–60 minutes	**Comfort break**		
60–100 minutes	Continue reviewing items and discuss [assessment] administration	4	Content or lead facilitator
100–110 minutes	Explanation of inter-session activities		Lead facilitator
110–120 minutes	Closing remarks		Lead facilitator

Panelist inter-session activities:

- Administer the [assessment] to three learners (from the appropriate grade/age group for each GPL)

Table 27: Example Agenda for Remote Workshop Session 1

Timing	Activity	Presentation	Facilitator
0–10 minutes	Welcome and purpose of session 1		Lead facilitator
10–55 minutes	Review GPF activity and provide clarification		Content or lead facilitator
55–60 minutes	**Comfort break**		
60–105 minutes	Discussion of [assessment] administration activity		Content or lead facilitator
105–120 minutes	Evaluation approach and completion of evaluation 1	19	Lead facilitator

Table 28: Example Agenda for remote Workshop Session 2

Timing	Activity	Presentation	Facilitator
0–10 minutes	Welcome and purpose of session 2		Lead facilitator
10–20 minutes	Address any concerns raised in evaluation 1		Content or lead facilitator
20–55 minutes	Introduction to alignment task (Task 1)	5	Lead facilitator
55–60 minutes	**Comfort break**		
60–90 minutes	Small group discussions on first 5 items[27]	6	Content facilitators[2]
90–110 minutes	Plenary discussion on questions that came up in the groups	6	Lead facilitator
110–120 minutes	Explanation of inter-session activities and close		Lead facilitator

Panelist inter-session activities:

- Complete Task 1 - alignment review on all remaining items (submit four hours after session)
- Complete evaluation 2 (submit with alignment review)

Table 29: Example Agenda for Remote Workshop Session 3

Timing	Activity	Presentation	Facilitator
0–10 minutes	Welcome and purpose of session 3		Lead facilitator
10–40 minutes	Review inter-session activities and provide clarification	7	Content facilitator
40–55 minutes	Introduction to Task 2 – Matching to GPLs and GPDs	8	Lead facilitator
55–120 minutes	Practice with Task 2		Lead facilitator
120–130 minutes	**Comfort break**		
130–230 minutes	Small groups complete Task 2 together (groups organized by grade/subject/language)[28]	9 & 10	Content facilitator
230–240 minutes	Explanation of inter-session activities and close		Lead facilitator

Panelist inter-session activities:

- Complete evaluation 3 (submit one hour after close of session)

Table 30: Example Agenda for Remote Workshop Session 4

Timing	Activity	Presentation	Facilitator
0–10 minutes	Welcome and purpose of session 4		Lead facilitator
10–40 minutes	Present Angoff methodology and Task 4 and provide clarification	11 & 12	Lead facilitator
40–75 minutes	Small group Angoff ratings using practice items	13	Content or lead facilitator
75–80 minutes	**Comfort break**		
80–100 minutes	Plenary discussion of questions that arose in small groups		Lead facilitator
100–110 minutes	Start Round 1 ratings (raise questions that come up)	14	Independent work
110–120 minutes	Explanation of inter-session activities and close		Lead facilitator

Panelist inter-session activities:

- One-on-one meetings between each panelist and a lead facilitator (during these meetings, facilitators answer panelist questions and will ask panelists how they are rating each item and why and check to make sure the reasoning follows the flow of the steps required for this task)

[27]Each small group will have a content facilitator; we recommend the lead facilitator(s) stay out of the small groups so the small groups can identify what questions they have and bring them back to the plenary.

[28] Ibid.

- Complete Round 1 ratings on all remaining items (submit four hours after close of session or one hour after one-on-one meeting with lead facilitators, whichever comes later)
- Complete evaluation 4 (submit with Round 1 ratings)

Table 31: Example Agenda for Remote Workshop Session 5

Timing	Activity	Presentation	Facilitator
0–10 minutes	Welcome and purpose of session 5		Lead facilitator
10–45 minutes	Review and discuss Round 1 ratings in plenary	15	Content facilitator
45–50 minutes	**Comfort break**		
50–110 minutes	Review Round 1 ratings in small groups (organized by grade/subject/language), going through each item where there was disagreement	15	Content facilitator
110–150 minutes	Share and discuss item difficulty and impact data	15	Lead facilitator
150–180 minutes	Explanation of inter-session activities (reminder of methodology) and close	16	Lead facilitator

Panelist inter-session activities:

- One-on-one meetings between each panelist and a lead facilitator (during these meetings, facilitators answer panelist questions and will ask panelists how they are rating each item and why and check to make sure the reasoning follows the flow of the steps required for this task)
- Complete Round 2 ratings (submit four hours after close of session or one hour after one-on-one meeting with lead facilitators, whichever comes later)
- Complete evaluation 5

Table 32: Example Agenda for Remote Workshop Session 6

Timing	Activity	Presentation	Facilitator
0–10 minutes	Welcome and purpose of session 6		Lead facilitator
10–30 minutes	Review Round 2 ratings and share final outcomes	18	Content facilitator
30–90 minutes	Discuss outcomes and final panelist questions	18	Lead facilitator
90–100 minutes	Complete evaluation 6	19	Independent work
100–120 minutes	Thank you and close	20	Lead facilitator

ANNEX R – WORKSHOP EVALUATION FORM

This form can either be cut up so that each of the sections is administered after the day/session in the workshop in which the topic is presented or administered in its entirety on the last day/session of the workshop. Administering the workshop over the course of the workshop will help facilitators identify gaps in understanding and adapt their presentations as needed, but this may also be overly burdensome on panelists. Facilitators should make a decision in consultation with key stakeholders based on the context of the workshop. The introductory language for each section should be adapted based on when the questions are being presented. You will also need to fill in all of the brackets. Finally, some questions may need to be moved to another session for remote workshops where activities don't always occur on the same day as training. No matter how the form is presented, it is important to include the panelist ID on the entire form (if it is administered in one setting) or at least for the Round 1 and Round 2 ratings (if it is administered over the course of the workshop).

PART 1: TRAINING ON THE GLOBAL PROFICIENCY FRAMEWORK

Today, you have been trained on the Global Proficiency Descriptors (GPDs). Please read the following statements carefully and place a mark in that category indicating your level of agreement.

Table 33: Evaluation Form for the Training on the GPF

GPD training	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I understand the purpose of the GPF					
I understand the relationship between domains, constructs, subconstructs, knowledge and skills, and GPDs					
The GPDs were clear and easy to understand					
The discussion of the GPDs helped me understand what is expected of learners in [insert subject] at the end of [insert grade]					
The practical exercise using the GPDs was useful to improve my understanding					
There was an equal opportunity for everyone to contribute their ideas and opinions					
There was an equal opportunity for everyone to ask questions					
The amount of time spent on the GPD training was sufficient					

Please describe in your own terms what the purpose of the GPF is and what the GPDs tell you.

Please list any questions or areas of confusion you have about the GPF.

Please list any tips/requests for facilitators that would make the training work better for you.

PART II: TRAINING ON THE ASSESSMENT(S)

Today, you have been trained on the assessment(s) that we will use for policy linking. Please read the following statements carefully and place a tick in each category to indicate the degree to which you agree with each statement.

Table 34: Evaluation Form for the Assessment Training

Assessment training	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I understand the purpose of the assessment					
I understand the constructs assessed in the assessment					
I understand how the assessment is administered					
Administering the assessment helped me to understand how minimally proficient learners would perform on the assessment (this is only applicable if the panelists were able to assess learners ahead of the workshop)					
I feel I have a good sense of how minimally proficient learners would perform on the assessment					
The amount of time spent on the assessment training was sufficient					

Please list any questions you have about the assessment(s).

Please list any tips/requests for facilitators that would make the training work better for you.

PART III: TRAINING ON ALIGNMENT METHODOLOGY

Today you have been trained on the alignment methodology. Please read the following statements carefully, and place a tick in each category to indicate the degree to which you agree with each statement.

Table 35: Evaluation Form for Task 1 – Alignment

Alignment training	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I understand the purpose of alignment					
I understand the alignment methodology					
I understand the difference between no fit, partial fit, and complete fit					
I feel confident with my alignment ratings					
The amount of time spent on the assessment training was sufficient					

Please list any questions or areas of confusion you have about the alignment methodology/process.

Please list any tips/requests for facilitators that would make the training work better for you.

PART IV: TRAINING ON MATCHING METHODOLOGY

Today you have been trained on the matching methodology. Please read the following statements carefully, and place a tick in each category to indicate the degree to which you agree with each statement.

Table 36: Evaluation Form for Task 2 – Matching

Alignment training	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I understand the purpose of matching					
I understand the matching methodology					
I understand how the alignment activity links to the matching activity					
I agree with the group consensus on the GPLs and GPDs to which we aligned each item (expand below if not)					
The amount of time spent on the matching training was sufficient					

Please describe any group decisions on matching with which you don't agree and why.

Please list any questions or areas of confusion you have about the matching methodology/process.

Please list any tips/requests for facilitators that would make the training work better for you.

PART V: TRAINING ON THE BENCHMARK-SETTING (ANGOFF) METHODOLOGY

Today, you have been trained on the benchmark-setting methodology. Please read the following statements carefully and place a tick in each category to indicate the degree to which you agree with each statement.

Table 37: Evaluation Form for Task 3 – Benchmarking

Policy linking training	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I understand the process I need to follow to complete the benchmarking exercise					
I understand how the benchmarking methodology links to the steps on alignment and matching					

Policy linking training	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I understand the difficulty level of the assessment items					
The discussion of the procedure was sufficient to allow me to feel confident in the methodology					
I understand how my ratings will result in a final benchmark					
There was an equal opportunity for everyone to contribute their ideas and opinions					
There was an equal opportunity for everyone to ask questions					
The amount of time spent on the policy linking method training was sufficient					
I feel confident in my Round 1 ratings					

Please describe the benchmarking methodology in your own terms.

Please list any questions or areas of confusion you have about the benchmarking methodology/process.

Please list any tips/requests for facilitators that would make the training work better for you.

PART VI: BENCHMARK ROUND 2 EVALUATION

During Round 2, you were given actual performance information and data about the impact of using the Round 1 results. Then, you were asked to give revised performance predictions. Please select the best answer below.

Table 38: Evaluation Form for Task 3 – Benchmarking Round 2

Round 2	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I understand the data on others' ratings					
I understand the item difficulty data and how it relates to this process					
I understand the impact data and how it relates to this process					
I am confident about the performance predictions I made during Round 2					
My performance predictions were influenced by the information showing the ratings of other panelists					
My performance predictions were influenced by the item difficulty data showing the actual performance of learners on the assessment					
My performance predictions were influenced by the impact information showing the outcomes for the sample of learners					

Round 2	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I was given sufficient time to complete the Round 2 performance predictions					

Do you have any additional comments on Round 2?

Part V: Overall Evaluation

How comfortable are you with your final performance predictions?

Very uncomfortable	Somewhat uncomfortable	Fairly comfortable	Very comfortable

If you marked either of the uncomfortable options, please explain why.

Overall, how would you rate the success of the policy linking workshop?

a. Totally Successful

b. Successful

c. Unsuccessful

d. Totally Unsuccessful

How would you rate the organization of the workshop?

a. Totally Successful

b. Successful

c. Unsuccessful

d. Totally Unsuccessful

Please provide any comments you feel would be helpful to us in planning future policy linking workshops.

Thank you for your participation in the workshop.

ANNEX S – CONTENT FACILITATOR SLIDES

It is critical that all facilitators be trained on the policy linking methodology. Generally speaking, however, the lead facilitators will have been trained in advance of the policy linking process, so it is likely that only the content facilitators will need to be trained. The lead facilitators should derive the content facilitator training slides from the workshop slide decks included in **Annex G**. We recommend at least eight hours of training for the content facilitators ahead of the workshop, though this may vary depending on their experience with standard setting in general, the assessment, and the modified Angoff method. Slides should be reduced to allow time to get through all of the major technical content, with a focus on the following presentations:

- Presentation 3
- Presentation 5
- Presentation 7
- Presentation 8
- Presentation 10
- Presentation 11
- Presentation 12
- Presentation 15
- Presentation 18.

It is especially critical that the content facilitators have an in-depth understanding of the GPF and the assessment, as understanding and relaying that content and putting it in the local context is their main responsibility. It is helpful if the content facilitators also have an understanding of when different topics/vocabulary/etc. are taught in schools in the local context, what terminology is used in the classroom, etc.

The training should also cover the dos and don'ts of running the workshop provided in **Table 11**.

If there is sufficient time between the training and the workshop, it may be helpful to undertake a rehearsal of the relevant sections of the workshop with the content facilitators, with lead facilitators acting as panelists, to ensure understanding.

ANNEX T – BENCHMARK CALCULATIONS FOR THE WORKSHOP

BENCHMARK CALCULATION FOR THE ANGOFF METHOD

The benchmarks for partially meets, meets, and exceeds minimum proficiency are computed using a set of six equations. Equations one through three are used to calculate benchmarks for each panelist and equations four through six are used to calculate benchmarks recommended by the panel. For these equations, *i* indicates the items or words, *j* indicates panelists, *l* indicates the number of item or words attempted by JP, *m* indicates the number of items or words attempted by JM, and *n* indicates the number of items or words attempted by JE. When only setting one benchmark, as opposed to three, the calculation is much easier. In that case, you need only add up the total yeses for meets by panelists and then average those totals across panelists.

Equation 1 shows the Partially Meets Minimum Proficiency benchmark for one panelist after Round 1.

$$PM_j = \sum_{i=1}^{l} \quad JP_{ij} \tag{1}$$

Equation 2 shows the Meets Minimum Proficiency benchmark for one panelist after Round 1.

$$M_j = PM_j + \sum_{i=l+1}^{m} \quad JM_{ij} \tag{2}$$

Equation 3 shows the Exceeds Minimum Proficiency benchmark for one panelist after Round 1.

$$E_j = M_j + \sum_{i=m+1}^{n} \quad JE_{ij} \tag{3}$$

Equation 4 is the Partially Meets Minimum Proficiency benchmark for all panelists after Round 1.

$$P = \frac{1}{z}\sum_{j=1}^{z} \quad \sum_{i=1}^{l} \quad PM_{ij}$$

Equation 5 is the Meets Minimum Proficiency benchmark for all panelists after Round 1.

$$M = \frac{1}{z}\sum_{j=1}^{z} \quad (PM_j + \sum_{i=l+1}^{m} \quad M_{ij}) \tag{5}$$

Equation 6 is the Exceeds Minimum Proficiency benchmark for all panelists after Round 1.

$$E = \frac{1}{z}\sum_{j=1}^{z} \quad (M_j + \sum_{i=m+1}^{n} \quad E_{ij}) \tag{6}$$

How to Calculate Benchmarks

Step 1: Calculate the Partially Meets Minimum Proficiency score (PM_j) for one panelist after Round 1.

i. Determine how many items or words the panelist decided two of three just meets minimum proficiency learners can attempt to answer or read in a minute (only applicable for timed task).
ii. Considering only those items or words two of the three just partially meets minimum proficiency (JP) learners can answer or read correctly according to the panelist, add together all the items or words from that subset that the panelist rated as just partially meets minimum proficiency.
iii. PMj for that one panelist is the sum from sub-step 2
iv. Repeat sub-steps 1 and 2 for each panelist to calculate PMj for each one

Step 2: Calculate the Meets Minimum Proficiency score (Mj) for one panelist after Round 1.

i. Determine how many items or words the panelist decided two of the three just meets minimum proficiency learner can attempt to answer or read in a minute (only applicable for timed task).
ii. Considering only those items or words two of three just meets minimum proficiency learner can answer or read correctly according to the panelist, add together the all the items from that subset that the panelist rated as just partially meets and just meets minimum proficiency.
iii. Mj for that one panelist is the sum from sub-step 2.
iv. Repeat sub-steps 1 and 2 for each panelist to calculate Mj for each one.

Step 3: Calculate the Exceeds Minimum Proficiency score (Ej) for one panelist after Round 1.

i. Determine how many items or words the panelist decided two of the three just exceeds minimum proficiency learner can attempt to answer or read in a minute (only applicable for timed task).
ii. Considering only those items or words two of three just exceeds minimum proficiency learner can answer or read correctly according to the panelist, add together all the items from that subset that the panelist rated as just partially meets, just meets, and just exceeds minimum proficiency.
iii. Ej for that one panelist is the sum from sub-step 2.
iv. Repeat sub-steps 1 and 2 for each panelist to calculate Ej for each one.

Step 4: Calculate the Partially Meets Minimum Proficiency cut score (P) for all panelists after Round 1.

i. Add up all the PM_j cut scores from the panelists.
ii. Divide the sum of PM_j cut scores and divide by the total number of panelists.
iii. This result is a simple average equivalent to P.

Step 5: Calculate the Meets Minimum Proficiency cut score (M) for all panelists after Round 1.

i. Add up all the Mj cut scores from the panelists.
ii. Divide the sum of Mj cut scores and divide by the total number of panelists.
iii. This result is a simple average equivalent to M.

Step 6: Calculate the Exceeds Minimum Proficiency cut score (E) for all panelists after Round 1.

i. Add up all the Ej cut scores from the panelists.
ii. Divide the sum of Ej cut scores and divide by the total number of panelists.
iii. This result is a simple average equivalent to E.

ANNEX U – CERTIFICATE OF APPRECIATION TEMPLATE

[Insert Government and Development parners logo]

CERTIFICATE OF PARTICIPATION

This is to certify that

[Insert Name]

has successfully participated in the **[Insert country name]** POLICY LINKING WORKSHOP FOR PRODUCING AND REPORTING ON SDG 4.1 for grade [Insert grade], in [Insert date, Month, Year] organized by [Insert name of government organizing agency] in partnership with [Insert name of development partner]

at [Insert venue].

Name

Position
DP's name

Name

Position
Government, organizer

DATE

ANNEX V – SELF-ASSESSMENT TEMPLATE SUMMARY (WORKSHOP OUTCOMES)

Assessment Instrument	[Insert name of instrument]
Jurisdiction	[Insert jurisdiction where assessment instrument is administered]
Grade	[Insert grade assessed by instrument]
SDG 4.1.1 level	a / b / c [delete as appropriate]
Subject	Mathematics / Reading [delete as appropriate]
MPLs being set	Partially meets / Meets / Exceeds [delete as appropriate]
Date of Policy Linking Workshop	[Insert date on which workshop was undertaken]
Assessors	[Insert names and organizations of assessors]

Criterion 1 – Did all panelists meet the requirements for participation?	Yes / No [delete as appropriate]
Criterion 2 – Were the group of panelists sufficiently representative in terms of the characteristics agreed by the country?	Yes / No [delete as appropriate]
Criterion 3 – Were all outliers removed before calculating the final benchmarks	Yes / No [delete as appropriate]
Criterion 4 – Were benchmarks only set for MPLS that don't exhibit floor or ceiling effects?	Yes / No [delete as appropriate]
Criterion 5 – Is the inter-rater consistency statistic greater than or equal to 0.7?	Yes / No [delete as appropriate]
Criterion 6 – Has the Standard Error for each benchmark been calculated and reviewed to be determined as appropriate?	Yes / No [delete as appropriate]
Criterion 7 – Has the confidence interval for each benchmark been calculated and reviewed to be determined as appropriate?	Yes / No [delete as appropriate]
Criterion 8 – Was the minimum score for each section of the evaluation greater than or equal to 4?	Yes / No [delete as appropriate]
Criterion 9 – Was the mean average score for the overall evaluation greater than or equal to 3?	Yes / No [delete as appropriate]

Overall Self-Assessment Rating

Did the Policy Linking Workshop meet all 10 Self-Assessment Criteria?	Yes / No [delete as appropriate]

Policy Linking Workshop Report

In addition to the self-assessment summary, the project team may wish to produce a report on the outcomes of the Policy Linking Workshop. 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. 4Self-Assessment Results
    a. Criterion 1: Alignment
    b. Criterion 2: Item Review
    c. Criterion 3: Sample
    d. Criterion 4: Administration
    e. Criterion 5: Reliability
4. Policy Linking Methodology
    a. Selection and Description of Panelists
    b. Procedure
        i. Preparation for the Policy Linking Workshop