



POLICY LINKING FOR MEASURING GLOBAL LEARNING OUTCOMES TOOLKIT

Linking Assessments to the Global Proficiency Framework

DECEMBER 2020



This publication was produced for review by the United States Agency for International Development (USAID). It was prepared for the NORC Reading and Access Evaluation Project by Management Systems International (MSI), a Tetra Tech Company.

Contracted under AID-OAA-M-13-00010

Reading and Access Evaluation Project

DISCLAIMER

The author's views expressed in this publication do not necessarily reflect the views of the United States Agency for International Development or the United States Government.

CONTENTS

ACRONYMS.....	VII
GLOSSARY OF TERMS	VIII
ACKNOWLEDGMENTS	X
CHAPTER I. INTRODUCTION TO POLICY LINKING.....	2
A. Rationale for Policy Linking.....	2
B. Audience.....	3
C. Overview of the Global Proficiency Framework.....	3
D. Overview of Policy Linking.....	6
E. Policy Linking Stages	7
F. Uses and Benefits of Policy Linking	9
G. Using the Policy Linking Toolkit.....	9
CHAPTER II. THE POLICY LINKING METHOD.....	12
A. Task 1—Aligning the Assessment to the GPF.....	12
B. Task 2—Matching Assessment Items with GPLs and GPDs.....	16
C. Task 3—The Angoff Method for Setting Benchmarks.....	18
CHAPTER III. PREPARING FOR THE POLICY LINKING WORKSHOP.....	23
A. Select Workshop Facilitators and Analyst.....	23
B. Plan Workshop Logistics	24
C. Select and Invite Workshop Panelists	24
D. Prepare Workshop Materials and Analyses	27
E. Train Content Facilitators	32
CHAPTER IV. IMPLEMENTING THE POLICY LINKING WORKSHOP	35
A. Workshop Day One.....	36
B. Workshop Day Two.....	38
C. Workshop Day Three.....	39
D. Workshop Day Four	40
E. Workshop Day Five.....	42
F. Tips for Hosting Remote Workshops.....	44
CHAPTER V. DOCUMENTING THE WORKSHOP OUTCOMES	48
A. Production of the Technical Documentation (After the Workshop is Completed)	48
CHAPTER VI. REVIEWING AND SUBMITTING WORKSHOP OUTCOMES.....	51
A. Collect Evidence From the Workshop	51
B. Submit Evidence to UIS.....	51
C. Receive a Response from UIS.....	52
BIBLIOGRAPHY	54
ANNEX A—RELATED RESOURCES	58
ANNEX B—GLOBAL MINIMUM PROFICIENCY LEVELS.....	59
ANNEX C—WORKSHOP PREPARATION CHECKLIST	60
ANNEX D—ALIGNMENT RATING FORM FOR TASK I	62
ANNEX E—WORKSHOP FACILITATION SLIDES.....	63
ANNEX F—ITEM RATING FORMS	132
ANNEX G—INTRA- AND INTER-RATER CONSISTENCY AND STANDARD ERROR (SE)	136
ANNEX H—PANELIST DEMOGRAPHIC INFORMATION.....	140

ANNEX I—INVITATION LETTER TEMPLATE FOR OBSERVERS 141

ANNEX J—INVITATION LETTER TEMPLATE FOR WORKSHOP PANELISTS 142

ANNEX K—TEMPLATE FOR PANELIST PRE-WORKSHOP ACTIVITY EXPLANATION 143

ANNEX L—PRE-WORKSHOP STATISTICS 144

ANNEX M—FEEDBACK DATA EXAMPLES AND INSTRUCTIONS 147

ANNEX N—SAMPLE AGENDA FOR AN IN-PERSON WORKSHOP 148

ANNEX O—SAMPLE AGENDA FOR A REMOTE WORKSHOP 150

ANNEX P—WORKSHOP EVALUATION FORM 153

ANNEX Q—CONTENT FACILITATOR SLIDES 158

ANNEX R—BENCHMARK CALCULATIONS FOR THE WORKSHOP 159

ANNEX S—CERTIFICATE OF APPRECIATION TEMPLATE 161

ANNEX T—OUTLINE FOR THE POLICY LINKING TECHNICAL REPORT 163

ANNEX U—4.1.1 CRITERIA FOR POLICY LINKING WORKSHOP VALIDITY 164

ANNEX V—AGREEMENT AND CONSISTENCY COEFFICIENTS 165

ANNEX W—POLICY LINKING PROCESS DOCUMENTATION FORM 167

TABLES

Table 1: Grade 3 Mathematics Example from the GPF	4
Table 2: USAID Foreign Assistance Indicators for Primary-Level Reading and Mathematics	5
Table 3: Policy Linking Stages.....	8
Table 4: Example of Summary Alignment Results for a Grade 3 Assessment by Domain, Construct, and Subconstruct.....	14
Table 5: Mathematics Assessment Alignment Criteria for Grades 1–9.....	16
Table 6: Reading Assessment Alignment Criteria for Grades 1–9	16
Table 7: Item Rating Form for Use with Yes-No Angoff Modification	21
Table 8: Brief Description of the In-Person Workshop Agenda.....	29
Table 9: Discussion Purpose, Do’s, and Don’ts by Task.....	33
Table 10: Summary of Tasks and Activities for the Policy Linking Workshop (Day References are for In-Person Workshops)	35
Table 11: Timeline for Submitting Results to UIS & Receiving Responses	51
Table 12: Workshop Preparation Checklist.....	60
Table 13: Alignment Rating Form Template.....	62
Table 14: Item Rating Form Example for Untimed Assessments.....	132
Table 15: Example Item Rating Form for Assessments with Constructed Response Questions.....	133
Table 16: Example Item Rating Form for Timed Reading Assessment (in Hausa)	134
Table 17: Example Item Rating Form for Conditional Reading Comprehension Questions (in Hausa).....	135
Table 18: Example Data Distribution Table for Oral Reading Passage.....	144
Table 19: Example Data Distribution Table for a Reading Comprehension Subtask.....	145
Table 20: Example Data Distribution Table for Timed Reading Assessment	146
Table 21: Sample Agenda for In-Person Workshop.....	148
Table 22: Example Agenda for Remote Preparation Session 1	150
Table 23: Example Agenda for Remote Preparation Session 2	150
Table 24: Example Agenda for Remote Workshop Session 1	150
Table 25: Example Agenda for remote Workshop Session 2.....	151
Table 26: Example Agenda for Remote Workshop Session 3.....	151
Table 27: Example Agenda for Remote Workshop Session 4.....	151
Table 28: Example Agenda for Remote Workshop Session 5.....	152
Table 29: Example Agenda for Remote Workshop Session 6.....	152
Table 30: Evaluation Form for the Training on the GPF.....	153
Table 31: Evaluation Form for the Assessment Training.....	154
Table 32: Evaluation Form for Task 1—Alignment	154
Table 33: Evaluation Form for Task 2—Matching.....	155
Table 34: Evaluation Form for Task 3—Benchmarking.....	155
Table 35: Evaluation Form for Task 3—Benchmarking Round 2.....	156
Table 36: Criterion 4 for Policy Linking Validity	164
Table 37: Approximate Value of Agreement Coefficient using Absolute Value and Reliability Coefficient	165
Table 38: Approximate Value of Consistency Coefficient using Absolute Value and Reliability Coefficient	166

FIGURES

Figure 1: Setting One versus Three Benchmarks.....	5
Figure 2: Education System Alignment.....	6
Figure 3: Example of Comparable Benchmarks on Various Assessments	7
Figure 4: Policy Linking Process and Benefits.....	9
Figure 5: Alignment Scale and Number of Statements of Knowledge and/or Skill(s) to Which an Item Aligns.....	13
Figure 6: Example Alignment of an Item to the GPF with Complete Fit.....	13
Figure 7: Example Alignment of an Item to the GPF with Partial Fit.....	14
Figure 8: Example of Matching Items to the GPLs and GPDs.....	18
Figure 9: Item Rating Process for Yes-No Angoff Modification.....	19
Figure 10: Grade-Level/Text Complexity of Reading Passages.....	20
Figure 11: Activities to Prepare for the Policy Linking Workshop.....	23
Figure 12: Composition of Panelists.....	25
Figure 13: Assessment Security Considerations	26
Figure 14: Invitation Adaptations for Remote Workshops	27
Figure 15: Translation of the GPF.....	30
Figure 16: Key Differences between Untimed Assessments (Largely CBAs) and Timed Assessments.....	31
Figure 17: Tips for Facilitators on Opening Presentation	36
Figure 18: Tips for Facilitators on Background Presentation.....	37
Figure 19: Tips for Facilitators on Presentation of the GPF.....	37
Figure 20: Tips for Facilitators on the Assessment Presentation.....	37
Figure 21: Tips for Facilitators on the Alignment Presentation.....	38
Figure 22: Tips for Facilitators on Task 1—Aligning the Assessment(s) with the GPF	38
Figure 23: Tips for Facilitators on Reviewing the Results of Task 1	39
Figure 24: Tips for Facilitators on the Task 2 Matching Presentation	39
Figure 25: Tips for Facilitators on Overseeing the Task 2 Matching Activity	40
Figure 26: Tips for Facilitators on Reviewing the Task 2 Matching Results.....	40
Figure 27: Tips for Facilitators on the Global Benchmarking Presentation	40
Figure 28: Tips for Facilitators on Presenting the Task 3 Angoff Method	41
Figure 29: Tips for Facilitators on the Task 3 Angoff Practice.....	41
Figure 30: Tips for Facilitators on Overseeing Task 3—Round 1 Ratings.....	42
Figure 31: Tips for Facilitators on Sharing Round 1 Results.....	42
Figure 32: Tips for Facilitators on Presenting Angoff Round 2.....	43
Figure 33: Tips for Facilitators on Overseeing Angoff Round 2 Ratings	43
Figure 34: Tips for Facilitators on Presenting the Evaluation Form.....	43
Figure 35: Tips for Facilitators on Presenting Final Results.....	44
Figure 36: Tips for Facilitators on Workshop Closing	44
Figure 36: Example Normative Data on Panelist Ratings.....	147
Figure 37: Example Impact Data Table.....	147

ACRONYMS

ACER	Australian Council for Educational Research
AERA	American Educational Research Association
APA	American Psychological Association
CAT	Comparing, Aggregating, and Tracking
CBA	Curriculum-Based Assessments
COR	Contracting Officer’s Representative
CPLV	Criteria for Policy Linking Validity
CR	Constructed Response
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
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
MC	Multiple Choice
M_j	Meets Minimum Proficiency
MSI	Management Systems International
NAEP	National Assessment of Educational Progress
NCME	National Council on Measurement in Education
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
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.

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.

Intra-rater consistency—An index that indicates panelists' overall performance in assessing test item difficulty.

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.

ACKNOWLEDGMENTS

This draft toolkit follows 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, 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 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 this draft toolkit. 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.

Dr. Abdullah Ferdous, Sean Kelly, and Dr. Jeff Davis of Management Systems International (MSI), with 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; and Colin Watson of the U.K. Department for Education, were the primary authors of the toolkit. Carlos Fierros (NORC), along with Nathalie Liautaud and Ryan Aghabozorg (MSI), provided essential management assistance.

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 in Bangladesh, India, and Nigeria, which supported the pilots in those countries and 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.

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, ACER, 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 have 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 NFER evaluation of the method, funded by the Bill & Melinda Gates Foundation, is ongoing and will continue to inform changes to the method.

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)

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 this revised version of the toolkit. The second draft GPF added Grade 1 and Grades 7 through 9.

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 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.

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.

B. AUDIENCE

This toolkit was created for use by country governments and assessment agencies (for multinational assessments) and their partners. Given that a primary focus of the toolkit is helping facilitate country reporting on SDG 4.1.1, all toolkit users, including assessment agencies, should closely coordinate with the relevant country government(s), as the governments are the ones 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 global reporting. 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 (CAT). For instance, all Grade 3 reading assessments can be linked to the grade three reading GPF, which then allows for comparing, aggregating, and tracking outcomes from those grade three 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.

A team of more than 60 reading and math 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

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.

assessment frameworks from 53 countries for mathematics.^{6,7} It was important that the GPF was grounded in the content framework and expert experience in diverse contexts to ensure the standards described within the document are aligned with and do not exceed existing country content standards and curricula.

An example from part of the grade three mathematics GPF is shown in **Table I**. 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 3 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 100.	Count in whole numbers up to 1,000.	Count in whole numbers up to 10,000.
			Compare and order whole numbers	Read and write whole numbers up to 100 in words and in numerals.	Read and write whole numbers up to 1,000 in words and numerals.	Read and write whole numbers up to 10,000 in words and in numerals.
			Skip count forwards or backwards	Compare and order whole numbers up to 100.	Compare and order whole numbers up to 1,000.	Compare and order whole numbers up to 10,000.
		Skip count forwards by twos or tens.		Skip count backwards by tens.	Skip count forwards and backwards by hundreds.	
		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 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 $35 = 3 \text{ tens and } 5 \text{ ones}$, $35 = 30 + 5$, 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 $254 = 2 \text{ hundreds, } 5 \text{ tens, and } 4 \text{ ones}$; $254 = 200 + 50 + 4$; determine the value of a digit in the hundreds place, etc.).

As **Table I** 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.

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).

- **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.

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 allow countries that partner with USAID to report against new 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

Importantly for global reporting, the Meets Global Minimum Proficiency level is directly aligned with SDG 4.1.1 as well as similar indicators for individual donor agencies, such as USAID’s Foreign Assistance (“F”) indicators, as shown in **Table 2** below. Learners with knowledge or skill at the Meets Global Minimum Proficiency level will satisfy SDG 4.1.1 and some of the USAID “F” indicators. For this reason, countries may decide to only set one benchmark at the “meets” level (see **Figure 1** for some criteria countries and assessment organizations may consider to determine 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. 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.

Table 2: USAID Foreign Assistance Indicators for Primary-Level Reading and Mathematics

Indicator Number	Indicator Title
ES.1–1	Percent of learners targeted for USG assistance who attain a minimum grade-level proficiency in reading at the end of Grade 2
ES.1–2	Percent of learners targeted for USG assistance who attain minimum grade-level proficiency in reading at the end of primary school
ES.1–47	Percent of learners with a disability targeted for USG assistance who attain a minimum grade-level proficiency in reading at the end of Grade 2
ES.1–48	Percent of learners targeted for USG assistance with an increase of at least one proficiency level in reading at the end of Grade 2
ES.1–54	Percent of individuals with improved reading skills following participation in USG-assisted programs
Supp–2	Percent of learners targeted for USG assistance with an increase of at least one proficiency level in reading at the end of primary school
Supp–3	Percent of learners targeted for USG assistance who attain minimum grade-level proficiency in math at the end of Grade 2

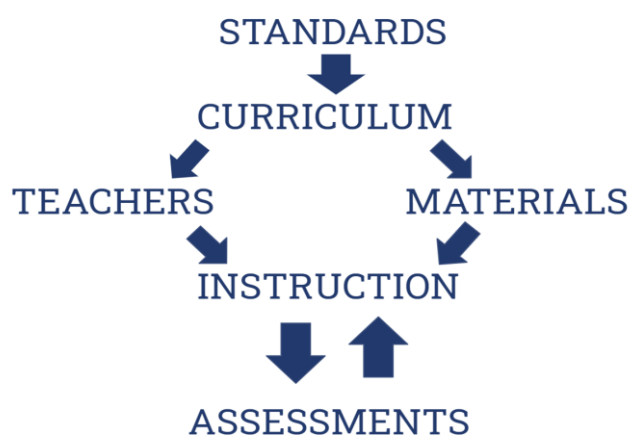
Indicator Number	Indicator Title
Supp-4	Percent of learners with an increase in proficiency in math of at least one level at the end of Grade 2 with USG assistance
Supp-5	Percent of learners targeted for USG assistance attaining minimum grade-level proficiency in math at the end of primary school with USG assistance
Supp-6	Percent of learners with an increase in proficiency in math of at least one level at the end of primary school
Supp-13	Percent of individuals with improved math skills following participation in USG-assisted programs
Supp-14	Percent of individuals with improved digital literacy skills following participation in USG-assisted programs
Supp-15	Education system strengthened: policy reform
Supp-16	Education system strengthened: data systems strengthened

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. However, a country government/assessment agency can elect to use only the Meets GPL, which is sufficient for reporting on SDG 4.1.1.

Additionally, while the GPF was created for use with policy linking and is not intended to be prescriptive in nature, countries can use it as a tool to inform the development or adaptation of national performance standard frameworks for guiding the construction of new or adapted national assessments. Assessments created in this manner are more likely to be aligned with the GPF. The GPF might also be used to inform country content standards and curriculum frameworks, teacher training, and text and materials in countries that are looking to modify their education systems. It is critical that all aspects of an education system are aligned, meaning curricula should reflect the standards, teacher training should be aligned with the curriculum and based on the textbooks, and assessments should test learner knowledge and skills taught in the classroom and described in standards, as shown in **Figure 2**.

Figure 2: Education System Alignment



The GPF offers a lens by which countries can examine alignment between the various components of their education system. During the piloting phase for the GPF between September 2019 and July 2020, several countries used it for these purposes.

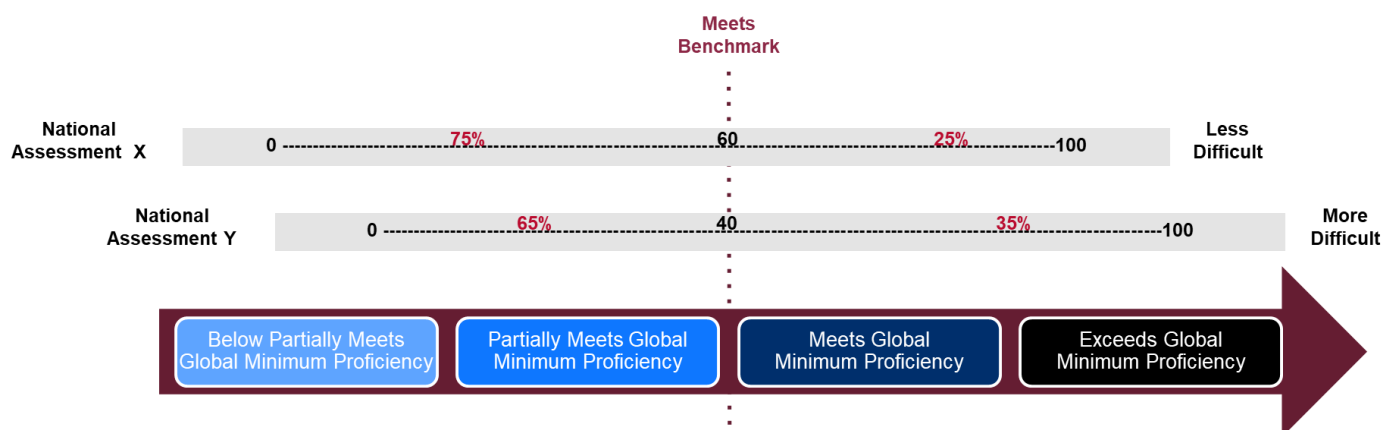
D. OVERVIEW OF POLICY LINKING

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. 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. To establish the numeric thresholds for each proficiency level for different assessments, policy linking requires aligning those assessments to the

GPF, matching assessment items to GPDs, and setting benchmarks. 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 of the assessments.⁸ This means that the benchmarks are set at different places (numbers) on the different assessments (unless the assessments are of equivalent difficulty).

For instance, as **Figure 3** shows, two different assessments using scales of 0 (minimum) to 100 (maximum) points 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. For instance, Country X and Country Y have national assessments with scales of 0 to 100 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. 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.

Figure 3: Example of Comparable Benchmarks on Various Assessments



To set the benchmarks, policy linking uses an internationally recognized, standardized, test-centered, Angoff-based benchmarking procedure. The Angoff procedure requires groups of national SMEs, called panelists, to make judgments on the 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.

E. POLICY LINKING STAGES

There are seven stages to policy linking for measuring global learning outcomes that must be completed to facilitate global reporting, as shown in **Table 3**. 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

⁸ 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.

Stages 4 and 5. **Table 3** provides information on resources available to support the other stages. It is critical that countries receive approval of their assessment(s) from the 4.1.1 Review Panel (Stages 2 and 3) ahead of planning for and implementing the policy linking workshop if they wish to use their outcomes to report on SDG 4.1.1 and/or USAID “F” indicators.

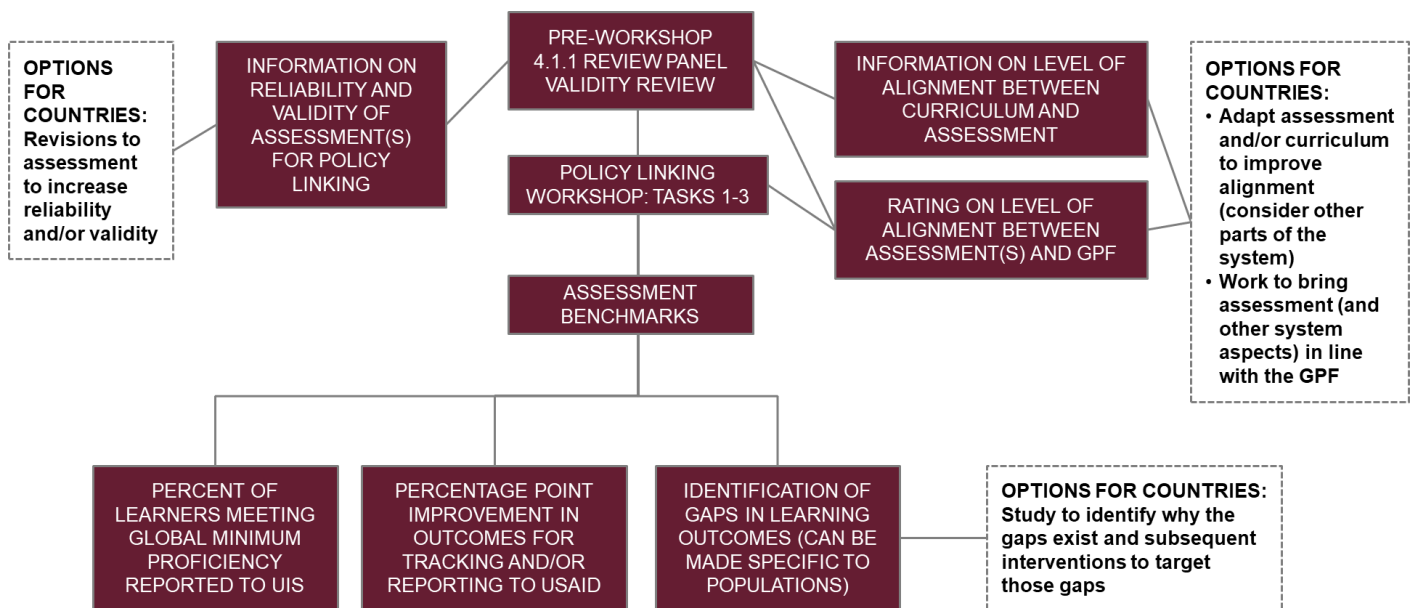
Table 3: Policy Linking Stages

#	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 make the decision whether to move forward with policy linking, either at a national or regional/state level and which assessment(s) they will link to global standards, as well as whether they wish to set three benchmarks for each assessment for the partially meets, meets, and exceeds GPLs (recommended) or only one at the meets level.	Country governments/ assessment agencies may complete this stage themselves or they may request/receive support from their partners—UIS, donors, and/or policy linking contractors. It is critical that country governments own this process either way and that at the end of the process, they are able to run future workshops on their own.	<ul style="list-style-type: none"> SDG 4.1.1 Options SDG 4.1.1 Reporting Decision Tree Policy Linking Overview Policy Linking Overview Slides Policy Linking Memo
2	Collation of evidence of curriculum and assessment validity and alignment	To submit for review by UIS’s 4.1.1 Review Panel to ensure assessments used for global reporting are valid, reliable, and sufficiently aligned to the GPF	Country governments/ assessment agencies with/without support of partners	<ul style="list-style-type: none"> Criteria for Policy Linking Validity (CPLV)
3	Review of evidence by the 4.1.1 Review Panel	To determine whether assessment reliability, validity, and alignment with the GPF meet requirements for proceeding with policy linking for global reporting and that the assessment is of sufficient length to allow for setting three benchmarks or if only one should be set at the meets level	4.1.1 Review Panel	<ul style="list-style-type: none"> Criteria for Policy Linking Validity
4	Preparation for the policy linking workshop (if approval received from UIS following Stage 3 to proceed)	To identify facilitators (if not done), invite panelists, prepare materials, and secure a venue	Country governments/ assessment agencies with/without support of partners	<ul style="list-style-type: none"> Policy Linking Toolkit (Chapter 3) Workshop Preparation Checklist (Annex C)
5	Implementation of policy linking workshop and documentation of outcomes	To set benchmarks and document details regarding reliability and validity of the workshop and country learning outcomes	Country governments/ assessment agencies with/without support of partners	<ul style="list-style-type: none"> Policy Linking Toolkit (Chapters 4, 5, and 6)
6	Review of workshop outcomes by 4.1.1 Review Panel	To determine whether workshop reliability and validity meet with criteria for global reporting	4.1.1 Review Panel	<ul style="list-style-type: none"> Criteria for Policy Linking Validity Policy Linking Toolkit (Chapter 6)
7	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	<ul style="list-style-type: none"> Protocol for Reporting on SDG Global Indicator 4.1.1 Individual donor guidelines

F. 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** in the second and third stages, the country government/assessment agency and its partners will get information from the 4.1.1 Review Panel on indicators of reliability and validity of its assessment(s) as well as the level of alignment between the country’s (“countries” in the case of multinational assessments) curriculum and assessment and between its assessment and the GPF. This information might help inform improvements in country education systems, as described in the GPF section above. 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



G. 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.⁹ It has guidelines for implementing the method.

Chapter II includes details on the policy linking methodology. **Chapter III** 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 IV** provides step-by-step guidance on how to implement a policy linking workshop. **Chapter V** presents key considerations for documenting the outcomes of the policy linking

⁹ 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.

workshop. Finally, **Chapter IV** presents details on the materials country governments/assessment agencies and partners need to submit to the 4.1.1 Review Panel.

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.¹⁰

The annexes provide all the materials and forms needed for applying the policy linking procedures outlined in the toolkit. This includes, among other things, the GPF, 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.

¹⁰ 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. THE POLICY LINKING METHOD

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 review, 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 ratings¹¹

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

A. TASK 1—ALIGNING THE ASSESSMENT TO THE GPF

It is important to distinguish the alignment activity in Task 1 from the alignment work conducted by the government/assessment agency and the 4.1.1 Review Panel in Stages 2 and 3 of the policy linking process. 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 D**.

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.

¹¹ 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.

- **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 Item 1 from a grade three assessment, 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

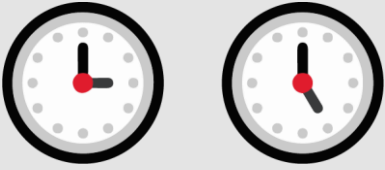
Knowledge or skill (content standard): Count, read, and write in whole numbers

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.

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 best describes the most important of the knowledge and/or skill(s) needed to answer the item correctly). An example of summary results for a grade three assessment with 26 items appears in **Table 4** below.

Table 4: Example of Summary Alignment Results for a Grade 3 Assessment by Domain, Construct, and Subconstruct

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
N2	Fractions	0
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
N2.1	Identify and represent fractions using objects, pictures, and symbols, and identify relative magnitude	0
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 should assess 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 and will report the outcomes of the alignment study according to the following three categories:

- **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.

The criteria for each of the categories are the same as those used by the 4.1.1 Review Panel. The criteria for mathematics are presented in **Table 5** and those for reading are presented in **Table 6**. When summarizing results to the subconstruct level, facilitators and/or data analysts should only consider the subconstructs with knowledge and/or skill(s) expected at the grade level for which alignment is being conducted. As such, when constructing the summary alignment tables, data analysts/facilitators should only list the domains, constructs, subconstructs, and statements of knowledge and/or skill(s) that have an “x” listed under the appropriate grade level column in GPF Table 3. For example, **Table 4** only includes the domains, constructs, and subconstructs relevant for grade three (not all of the subconstructs represented in the GPF).

From the criteria below, it is clear that the example grade three assessment described in **Table 4** would be considered “additionally aligned” since it: 1) contains more than five number items (14 total) and more than five total measurement and geometry items (10 total), and 2) has items covering at least 50 percent of the number, measurement, and geometry subconstructs with knowledge and/or skills expected at grade three (8 out of 12 subconstructs are covered).

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

Level of Alignment	Category	Criteria
Minimally Aligned	Domain (depth):	Number (minimum five items)
	Subconstructs (breadth):	Items covering at least 50 percent of the Number and Operations subconstructs
Additionally Aligned	Domain (depth):	Number (minimum 5 items) and Measurement and Geometry (minimum 5 items)
	Subconstructs (breadth):	Items covering at least 50 percent of the Number, Measurement, and Geometry subconstructs
Strongly Aligned	Domain (depth):	Number (minimum five items) and Measurement and Geometry (minimum five items) and Statistics and Probability and Algebra (minimum five items)
	Subconstructs (breadth):	Items covering at least 50 percent of all subconstructs

Table 6: Reading Assessment Alignment Criteria for Grades 1–9

Level of Alignment	Category	Grade 1–2 Criteria	Grade 3–6 Criteria	Grade 7–9 Criteria
Minimally Aligned	Domain/Construct (depth):	D (minimum five items) C (minimum five items)	R (minimum five items)	R (minimum five items)
	Subconstructs (breadth):	Items covering at least 50 percent of the D and C subconstructs	Items covering at least 50 percent of the R subconstructs	Items covering at least 50 percent of the R subconstructs
Additionally Aligned	Domain/Construct (depth):	N/A	N/A	R: B1 (minimum 5 items) R: B2 (minimum 5 items)
	Subconstructs (breadth):	N/A	N/A	Items covering at least 50 percent of the R subconstructs
Strongly Aligned	Domain/Construct (depth):	R (minimum five items)	R: B1 (minimum five items) R: B2 (minimum five items)	R: B1 (minimum five items) R: B2 (minimum five items) R: B3 (minimum five items)
	Subconstructs (breadth):	Items covering at least 50 percent of the R subconstructs	Items covering at least 50 percent of the R subconstructs	Items covering at least 50 percent of the 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

Following the Policy Linking Workshop, the government/assessment agency, with support from its partners (if relevant), will need to report the results of the alignment exercise to the 4.I.I Review Panel.

B. 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. 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.¹² They 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 8** 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 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 8 provides an example taken from the Workshop Facilitation Slides included in **Annex E**. In this example item, learners are asked to read the following passage:

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. The dog went to sleep. When the dog woke up, Jabu took the dog outside to play again.

Learners are then asked to respond to the question, “Who had a pet dog?” This question matches with the statement of knowledge or skill of retrieving a single piece of explicit information from a grade-level continuous 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 three-level passage, and the item aligns to the Partially Meets Global Minimum Proficiency level at grade three.

¹² 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.

Figure 8: 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 words, “had a pet dog”. Also, Jabu is a common name in this context.

Domain: Reading comprehension

Construct: Retrieve information

Subconstruct: Locate explicitly stated information

Passage grade level: Grade 3

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 continuous 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 3-level continuous 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 multiple pieces of explicit information from a grade 3-level continuous text by direct- or close-word matching when the information required is adjacent to the matched word and 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.

C. TASK 3—THE ANGOFF METHOD FOR SETTING BENCHMARKS

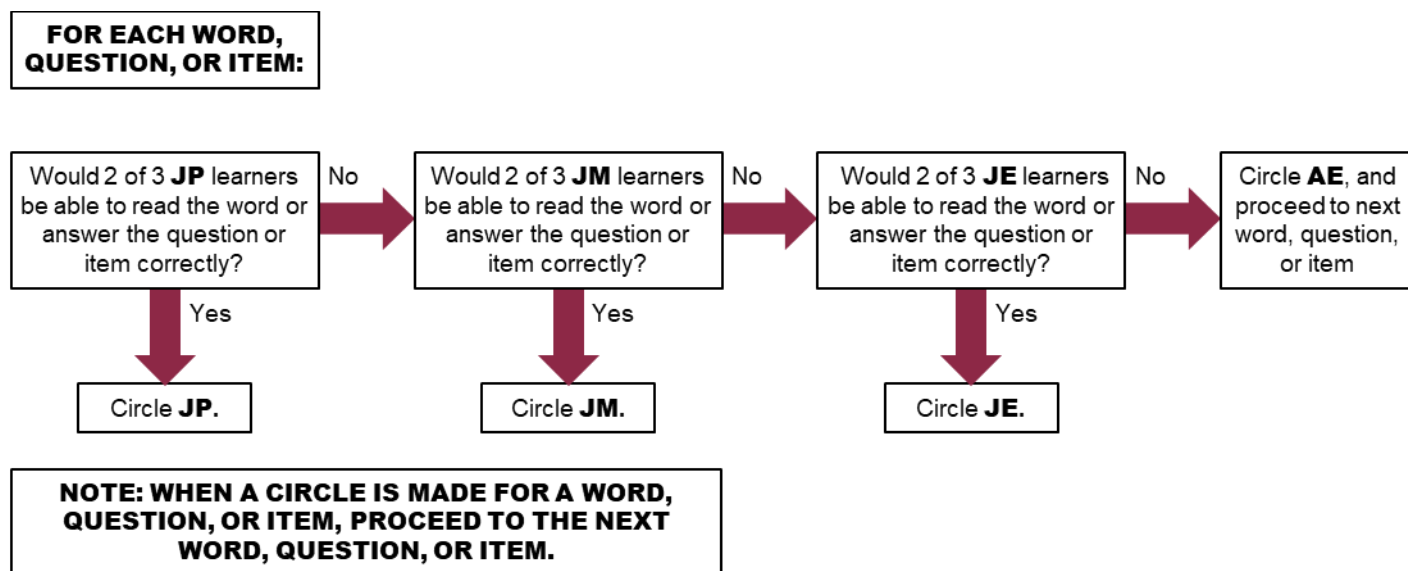
Task 3 is the most important task in the Policy Linking Workshop, as this is where panelists set benchmarks by making their judgements of how learners whose reading or math abilities correspond with the knowledge and/or skill(s) aligned to each item in Task 1 and how the GPDs matched with each item in 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 F**) 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 (or just the meets level if panelists will only be setting one benchmark).¹³ 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. As described in **Chapter III**, unless assessment security protocols prevent doing so, panelists will have an opportunity to assess learners at each of these levels ahead of the workshop, and they can be thinking specifically of those learners and how they performed on the assessment during this step.
- **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).

¹³ 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 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 conceptualize learners at the meets or JM level.

- **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).
- **Step 4:** Based on an understanding of Steps 1–3, follow the procedure shown in the flowchart in **Figure 9** 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 9** 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.¹⁴

Figure 9: 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”)

¹⁴ 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.

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 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 10**.

Figure 10: 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 G** for details on how to calculate these and **Chapter IV** for more details on when/why these are calculated), which have to be reported to the 4.I.1 Review Panel 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 IV**), 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 IV**), 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
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 R**. Details for preparing for the workshop are presented in **Chapter III** below, and facilitator notes for implementing this methodology in an in-person or remote workshop are included in **Chapter IV**.

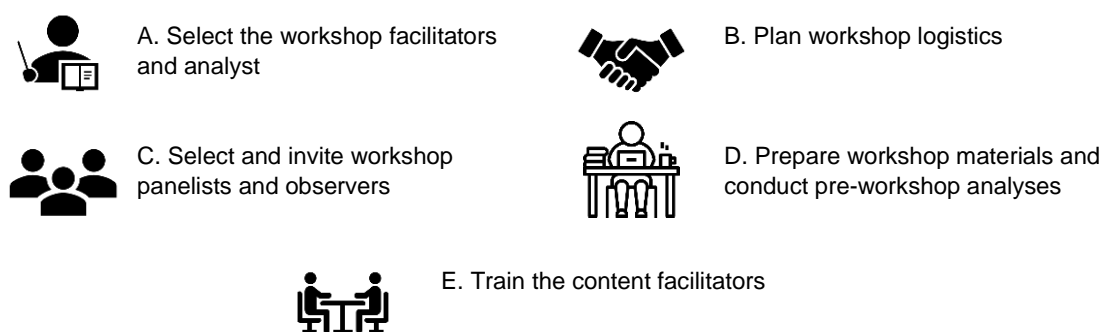
CHAPTER III

CHAPTER III. 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 11**. 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 C**.

Figure 11: 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 for the workshop data plus statistical software, such as Stata or SPSS, 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 C

It is recommended that policy linking workshops be held with the facilitators and panelists gathering in person. However, if that is not possible, it is possible to hold the workshop remotely with either: 1) the panelists and content facilitators gathering in person, in country and the lead facilitators attending remotely (only necessary if the lead facilitators are internationally based) or 2) all panelists and facilitators attending remotely (see tips on hosting a remote workshop in **Chapter IV, Section F**). The project team should work with relevant government and partner stakeholders to select the appropriate gathering 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 in this activity. If it is not possible to gather in person, the project team and government should agree on an appropriate digital 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 (see Activity C, below); etc. More details about each of the relevant steps under this activity are included in **Annex C**.

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 H

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.

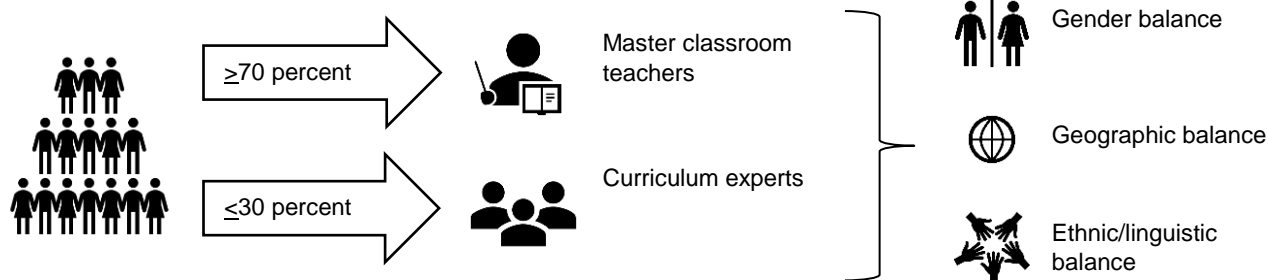
When selecting a panel (or panels) for a policy linking workshop, the number of panelists must be sufficiently large and representative. This is to provide reasonable assurance that the benchmarks 1) will be realistic, attainable, and unbiased and 2) would not vary greatly if the process were repeated 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.¹⁵ As shown in **Figure 12**, the

¹⁵ 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.

panel should be made up of at least 70 percent master classroom teachers and up to 30 percent non-teachers, preferably curriculum experts.

Figure 12: 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 H**. 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 4.I.1 Review Panel sufficient data to address the degree to which the panelists meet the criteria.

Inviting Panelists and Observers and the Pre-workshop Activity

USE ANNEXES I, J, AND K

Panelists should be invited well in advance of the workshop; at least six weeks is recommended. **Annex I** and **Annex J** 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. The panelists' invitation letter should also reference the advance preparation needed to serve as a panelist, the details of which follow and are also included in the form of an invitation addendum that can be sent to panelists in **Annex K**.

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. Prior to the workshop, the panelist will be asked to select nine learners: three who the panelist knows just barely meet the requirements of the GPF's Partially Meets Global Minimum Proficiency level for the grade level of the assessment, three who just barely meet the requirements of the Meets Global Minimum Proficiency level, and three who just barely meet the requirements of the Exceeds Global Minimum Proficiency level. The panelists will record the scores of the learners as well as which assessment items the learners got right and wrong and will bring that information to 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, this is not the preference, as it will not give panelists insight into reasonable benchmarks. See **Figure 13** for more information on assessment security.

- For individually administered timed assessments, such as early grade reading or mathematics assessments (EGRAs or EGMA), 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. **Figure 14** includes details on invitations for remote workshops.

Figure 13: 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.

Figure 14: Invitation Adaptations for Remote Workshops

Invitations will still need to be sent out for remote workshops, but they should include different information, including the following:

- **Information on what platform the workshop will use and how participants will get the link** for each session
- **Information on the preferred hardware for joining** (computers are strongly preferred to allow panelists to see the slides and submit tasks, but smartphones can be used if necessary)
- **Information on how to join a WhatsApp group or another collaboration platform for panelists** (This is a great way to send the group reminders, troubleshoot problems, etc.)
- **Information on which documents need to be printed ahead of the workshop** (See **Chapter IV, Section F** for tips on how to run a remote policy linking workshop).

Remote workshops may also not require panelists to assess learners ahead of time, as this can be done between sessions by creating a gap between the first workshop session(s), which would describe the assessment and how to administer it as well as provide details on the GPF and how to select learners who fall in the partially meets, meets, and exceeds proficiency levels.

D. PREPARE WORKSHOP MATERIALS AND ANALYSES

USE ANNEXES A, D, E, L, M, N, O, P

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.

In order to prepare materials for the policy linking workshop, the facilitators should obtain documentation from the national assessment. The following list of documents and data are required to inform creation/adaptation of the workshop agenda, slides, forms, and templates. Most of these should have been obtained during Stage I; 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).

Materials That Need to be Obtained

- Assessment specifications (optional)
- Assessment instrument
- Assessment data file
- 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. However, the project team will also send either the whole assessment instrument (preferred) or a short sample assessment (back-up option) to the panelists so they can administer the items to learners (as described earlier) either ahead of the workshop (for in-person workshops) or after panelists have been trained on the GPF and how to administer the assessment instrument (for remote workshops; more details on remote workshops are included in **Chapter IV, Section F**).

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 L** for details on how to calculate these statistics using the data from the most recent assessment results.
- **Data distributions**—See **Annex M** 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**—Templates included in **Annex N**, for in-person workshops, and **Annex O**, for remote workshops; these 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 H** but may need to be updated depending on criteria for representativeness of panelists.
- **Relevant grade/subject GPDs**, including the grade below the one being linked and one grade level above. These will be carefully reviewed by the panelists during the workshop, including the grade level for the assessment(s) under consideration and one grade level below and one above the grade level of the assessment(s). (Facilitators will need to cut the GPF back to the relevant grades for the workshop and further 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 for the grade level being linked and one above and one below, as well as example items for the relevant grade levels.
- **Facilitation slides**—Details on how to locate the slide templates are included in **Annex E** for both timed and untimed assessments, but facilitators will need to adapt these; instructions on how to do so are included in the template.
- **Alignment rating forms and item rating forms**—Details on how to create these forms and examples are included in **Annex D** for the alignment form and **Annex F** for the item rating form.
- Workshop evaluation forms—A draft is included in **Annex P**. 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 M**).

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 sample in-person workshop agenda (**Annex N**) provides a day-by-day list of the in-person workshop sessions, time allocations, and facilitation requirements. The structure of the sessions should remain constant for all in-person 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. A brief summary of the five-day in-person workshop agenda is presented in **Table 8**. Note that some of the sessions (including the opening, training, and closing) will be plenary and, thus, led by the lead facilitator, while other sessions (activities, discussions, and feedback) will be panels, preferably led by the content facilitators.

A sample remote workshop agenda (**Annex O**) provides a day-by-day list of the remote workshop sessions, time allocations, and facilitation requirements. The recommendation for remote workshops is that the sessions are shorter (approximately 2–4 hours per session) and 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 assess nine or more learners using the assessment ahead of the workshop as recommended in the “Inviting panelists and the pre-workshop activity” subsection above.

Table 8: Brief Description of the In-Person Workshop Agenda

Day	Descriptions
Day 1	This day was requested by country governments and other stakeholders during piloting. The focus is on introducing and carefully reviewing the GPF and assessment instrument(s) ahead of diving into activities where these documents will be used. The lead facilitators open the workshop with introductions. 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 summary of the method. Then, the majority of the day is spent reviewing the GPF and the assessment instrument. Facilitators may even have the panelists administer the assessment to one another for practice (especially if not all panelists were able to assess learners ahead of the workshop).
Day 2	The lead facilitators review what the group covered in the previous day, answer any questions, and then make the Task 1 presentation on the GPF and alignment exercise. The content facilitators lead the Task 1 activity on aligning the assessments with the GPF, which is an individual and independent activity.
Day 3	The lead facilitators present the alignment results. They make the Task 2 presentation on the assessments and the GPLs/GPDs. The content facilitators lead the Task 2 activity on matching the assessments with the GPDs/GPLs, which is a group activity.
Day 4	The lead facilitators present the matching results (This is only necessary if the workshop seeks to set benchmarks for more than one grade level using the same assessment). They make the first Task 3 presentation on global benchmarking. They make the second Task 3 presentation on the Angoff method. The content facilitators lead the first Task 3 activity with Angoff practice. They lead the second Task 3 activity with Angoff Round 1.
Day 5	The lead facilitators present the Round 1 results. The content facilitators lead the third Task 3 activity with Angoff Round 2. They lead the fourth Task 3 activity with the workshop evaluation. The lead facilitators present the Round 2 results. 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 H** to ensure that panelists are representative of the population being assessed. This information must also be reported to the 4.I.I Review Panel along with details on the population being assessed and the teachers of that population. For instance, the 4.I.I 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 H** 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—those for which benchmarks are being set—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 GPL thresholds. 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 15** for details), but facilitators should not make any other changes to the content or language of the GPF.

Figure 15: 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

The facilitators will present the slides during Days 1 to 5 of the workshop (for in-person workshops) or through a series of eight workshop sessions (for remote workshops). The slides are included in **Annex E** 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 there are two sets of slides depending on the type of assessment, e.g., timed assessments such as the EGRA/EGMA or untimed assessments (CBAs usually fall into the category of untimed assessments, as do untimed, individual or group-administered regional and international assessments). Details on the differences between implementing a policy linking workshop for an untimed versus a timed assessment are included in **Figure 16**.

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 can interpret as needed.

Figure 16: Key Differences between Untimed Assessments (Largely CBAs) and Timed Assessments

Given test security considerations, common with untimed assessments, facilitators may not be able to send a full CBA or other group-administered assessment to panelists in advance of the workshop. Facilitators may send a sample assessment in lieu of the full assessment and allot an appropriate amount of time to review the assessment during the workshop in this case.

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 an item (following the typical steps for Task 3 described in **Chapter II**).

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 D)**—These will be used for the panelists' ratings of the alignment between the assessments and the GPF.
- **Item rating forms (Annex F)**—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 (such as EGRA/EGMA) 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 F** 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, some pilots have found it useful to have panelists complete a shorter daily evaluation form to check in on knowledge acquisition, areas that may need further clarity, facilitation techniques that are working/not working, etc. **Annex P** includes the minimum evaluation questions that should be asked of panelists at the end of the workshop. It is designed to capture their views on the policy linking process. 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. A summary of the results (e.g., average Likert scale ratings per question) will be included in the workshop report and presented to the 4.1.1 Review Panel and the government/assessment agency and its partners as an indicator of the strengths and weaknesses of the activities and as an indicator of the validity of the ratings by the panelists.

If the project team opts not to include a daily evaluation (which could be adapted from the form in Annex P—Workshop Evaluation Form by adding in additional day and activity-specific questions), 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 M**. 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 Days 4 and 5 of the workshop. The same process will need to be completed following Round 2 ratings. The data analyst will need to conduct that analysis during the actual workshop day on Day 5, either during lunch, a certificate award ceremony, or another appropriate time.

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
- Assessment instrument (should only be included if assessment security protocols allow for it; see **Figure 13** 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 Q**. 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. It should also include a discussion of lead and content facilitator roles and responsibilities 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 9** (the same rules apply to answering panelist questions and facilitating practice ratings).

Table 9: 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.	<ul style="list-style-type: none"> • 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. 	<ul style="list-style-type: none"> • 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.	<ul style="list-style-type: none"> • 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. 	<ul style="list-style-type: none"> • 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.	<ul style="list-style-type: none"> • 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. 	<ul style="list-style-type: none"> • 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.	<ul style="list-style-type: none"> • Make sure everyone has the opportunity to speak and ask questions. 	<ul style="list-style-type: none"> • Make unsubstantiated claims about how the government/regional or international assessment agency will use the benchmarks.

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.

CHAPTER IV

CHAPTER IV. IMPLEMENTING THE POLICY LINKING WORKSHOP

While **Chapter II** 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. As described in **Chapters II and III**, facilitators will lead presentations and activities over a period of five days for in-person workshops and eight sessions for remote workshops. During that time, they will introduce the workshop methodology, the GPF, and the assessment 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)¹⁶

Table 10 has the workshop tasks, with the presentations and activities by day. Day references are for in-person workshops; tips for remote workshops are included in **Section F** at the end of this chapter. Timing suggestions for all activities are included in the sample agendas in **Annex N** (for in-person workshops) and **Annex O** (for remote workshops). There are a total of 20 presentations and activities that are conducted in a step-by-step process, culminating in the production of the final global benchmarks and the documentation of workshop outcomes, i.e., calculating the indicators and writing the technical report. 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 (more details are included in the “**workshop evaluation form**” subsection). Regardless of what is decided for the daily check-ins/evaluations, panelists must complete a written evaluation at the end of the workshop for reporting purposes.

Table 10: Summary of Tasks and Activities for the Policy Linking Workshop (Day References are for In-Person Workshops)

Task	Day	Presentation or Activity
Opening	Day 1	1. Opening, introductions, logistics, and agenda
		2. Presentation on the background, objective, and tasks
		3. Presentation on the GPF
		4. Presentation on the assessment, discussion of pre-workshop activity, and optional opportunity for panelists to take the assessment if they were unable to complete the pre-workshop exercise with learners or to further clarify the assessment
Task 1	Day 2	5. Presentation on the alignment exercise
	Day 3	6. Activity on aligning the assessments with the GPF
Task 2	Day 3	7. Presentation and discussion on the alignment results
		8. Presentation on matching assessment items with the GPLs/GPDs
Task 3	Day 4	9. Activity on matching assessment items with the GPLs/GPDs
		10. Presentation and discussion on the matching results
		11. Presentation on global benchmarking
		12. Presentation on the Angoff method
		13. Activity on Angoff practice
		14. Activity on Angoff Round 1

¹⁶ 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	Day	Presentation or Activity
	Day 5	15. Presentation and discussion of the Round 1 results
		16. Presentation on Angoff Round 2
		17. Activity on Angoff Round 2
		18. Workshop evaluation
		19. Presentation on Round 2 results
Closing		20. Closing and logistics
Documentation	After the workshop	Production of the technical documentation

Information on each of the above presentations and activities (1–20) is provided below, along with tips for the facilitators. There are references to the facilitation slides for the opening and presentations. There are two sets of slides:

- Group- or individually-administered untimed assessments with multiple choice and constructed response items, such as CBAs (164 slides)
- Individually-administered assessments with timed subtasks, such as EGRA/EGMAs (166 slides)

The slides, with notes, are provided as attachments to the toolkit (see **Annex E**) and contain additional facilitator details and tips.

A. WORKSHOP DAY ONE

1. Opening, Introductions, Logistics, and Agenda

MATERIALS: FACILITATION SLIDES, PANELIST WORKSHOP PACKETS (SEE CHAPTER 3, SECTION D)
SLIDES: 1–11 (TIMED AND UNTIMED ASSESSMENTS)
ANNEXES: N AND O

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 pertaining to the venue, plenary and breakout rooms, lodging, meals, per diem, and transportation. Finally, you will provide an overview of the workshop agenda to the participants.

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. Presentation on the Background, Objective, and Tasks

MATERIALS: FACILITATION SLIDES, PANELIST WORKSHOP PACKETS
SLIDES: 12–27 (TIMED AND UNTIMED ASSESSMENTS)

In this presentation, you will provide background information to the panelists 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.

3. Presentation on the GPF

**MATERIALS: FACILITATION SLIDES, RELEVANT GRADE/SUBJECT GPDs FROM THE GPF SLIDES: 28–40 (TIMED ASSESSMENTS) AND 28–41 (UNTIMED ASSESSMENTS)
ANNEXES: A AND B**

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. 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.

Figure 19: Tips for Facilitators on Presentation of the GPF

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.

4. Presentation on the Assessment Instrument

**MATERIALS: FACILITATION SLIDES, ASSESSMENT INSTRUMENT
SLIDES: 41-44 (TIMED ASSESSMENTS) AND 42-45 (UNTIMED ASSESSMENTS)
(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. During this process, you will ask panelists to report on how the learners they assessed prior to the workshop performed on the assessment (e.g., how did learners who meets the partially meets, meets, and exceeds descriptors perform?) and each of the items (e.g., what were some of their common stumbling blocks?). If there is time and it makes sense based on whether all panelists were able to assess learners ahead of time, 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

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.

B. WORKSHOP DAY TWO

5. Presentation on the Alignment Exercise (Task 1)

MATERIALS: FACILITATION SLIDES, PANELIST WORKSHOP PACKETS

SLIDES: 45–65 (TIMED ASSESSMENTS) AND 46–65 (UNTIMED ASSESSMENTS)

ANNEXES: A AND D

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** 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 will 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 items included in the timed assessment slides (slides 57–59) and sample math items included in the untimed assessment slides (slides 57–59) 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 will share the alignment threshold criteria listed on page 14 (**Table 5** and **Table 6**).

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 II**. 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.

6. Activity on Aligning the Assessment(s) with the GPF (Task 1)

MATERIALS: FACILITATION SLIDES, PANELIST WORKSHOP PACKETS

SLIDES: 66–70 (TIMED AND UNTIMED ASSESSMENTS)

ANNEXES: A AND D

In this activity, you will give the panelists an opportunity to ask questions, after which, if you have more than one panel, you may split the group into panel-level groups and have the content facilitators re-explain the task before panelists proceed with aligning the assessment items with the GPF statements of knowledge and/or skill(s). You will explain to the panelists that alignment is conducted between the items and the GPF statements of knowledge and/or skill(s) and at the end, there must be sufficient breadth and depth of alignment for policy linking to work well.

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, as depicted by each “x” in GPF Table 3.

C. WORKSHOP DAY THREE

7. Presentation and Discussion of Alignment Results from Day Two (Task 1)

MATERIALS: FACILITATION SLIDES, PANELIST WORKSHOP PACKETS

SLIDES: 71–79 (TIMED AND UNTIMED ASSESSMENTS)

(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 will address the level of alignment achieved based on the threshold criteria, presented in **Table 5** and **Table 6** above. 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 9** above in the Content Facilitator Training Section.

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

Reiterate that most (at least 50 percent) of the domains, constructs, and subconstructs for the relevant domains (as detailed in **Table 5** and **Table 6**) need to be covered by items (called breadth), and there need to be at least five items per relevant domain (called depth). Review the summary table. Discuss the implications of items that do not align with any statements of knowledge and/or skill(s) in the GPF, namely that the assumption will be that global minimum proficiency learners will get these items wrong on the assessment, since this issue will become apparent in Task 2 on matching.

8. Presentation on Matching Assessment Items with the GPLs/GPDs (Task 2)

MATERIALS: FACILITATION SLIDES, PANELIST WORKSHOP PACKETS

SLIDES: 80–88 (TIMED ASSESSMENTS) AND 80–89 (UNTIMED ASSESSMENTS)

ANNEX: A

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). You will walk the panelists through answering the three questions required under the task (see the section on **Task 2** 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. You will walk the participants through some sample items to ensure they understand the task. There are sample reading items included in the timed assessment slides (slides 83–86) and sample math items included in the untimed assessment slides (slides 83–86) that you can use for this purpose, or you can select or develop your own.

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. 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.

9. Activity on Matching the Assessment Items with the GPLs/GPDs (Task 2)

MATERIALS: FACILITATION SLIDES, PANELIST WORKSHOP PACKETS

SLIDES: 89–95 (TIMED ASSESSMENTS) AND 90–96 (UNTIMED ASSESSMENTS)

ANNEX: A

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.

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 is important that the panelists discuss their matches in small groups and then 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.

D. WORKSHOP DAY FOUR

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

MATERIALS: FACILITATION SLIDES, PANELIST WORKSHOP PACKETS

SLIDES: 96–103 (TIMED ASSESSMENTS) AND 97–104 (UNTIMED ASSESSMENTS)

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 9**.

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.

11. Presentation on Global Benchmarking (Task 3)

MATERIALS: FACILITATION SLIDES, PANELIST WORKSHOP PACKETS

SLIDES: 104–111 (TIMED ASSESSMENTS) AND 105–112 (UNTIMED ASSESSMENTS)

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 106) 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 107). You will cover the third graphic in the presentation (slide 108) 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. Presentation on the Angoff Method (Task 3)

MATERIALS: FACILITATION SLIDES, PANELIST WORKSHOP PACKETS

SLIDES: 112–127 (TIMED ASSESSMENTS) AND 113–127 (UNTIMED ASSESSMENTS)

ANNEXES: A, F, AND R

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 the benchmarks are calculated, both for the panelists and for the panels.

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.

13. Activity on Angoff method practice (Task 3)

MATERIALS: FACILITATION SLIDES, PANELIST WORKSHOP PACKETS

SLIDES: 128–132 (TIMED ASSESSMENTS) AND 128–133 (UNTIMED ASSESSMENTS)

ANNEXES: A, F, AND R

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 items included in the timed assessments slides (slides 130–133) and sample math items included in the untimed assessments slides (slides 132–135) 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. Activity on the Angoff Method Round 1 (Task 3)

MATERIALS: FACILITATION SLIDES, PANELIST WORKSHOP PACKETS

SLIDES: 133–141 (TIMED ASSESSMENTS); 134–141 (UNTIMED ASSESSMENTS)

ANNEXES: A, F, AND R

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. 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 R** 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.

E. WORKSHOP DAY FIVE

15. Presentation and Discussion of Round 1 Results and Item Difficulty and Impact Data (Task 3)

MATERIALS: FACILITATION SLIDES, PANELIST WORKSHOP PACKETS

SLIDES: 142–155 (TIMED AND UNTIMED ASSESSMENTS)

ANNEXES: L, M, AND R

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 R** 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 M**), 3) item ratings in relation to actual item difficulty (see page 25 and **Annex L**), 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 9** 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 9**. Also, have the panelists compare the actual p-values (difficulty statistics) with their ratings to see whether their ratings are consistent with the data. 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.

16. Presentation on the Angoff Method Round 2 (Review) (Task 3)

MATERIALS: FACILITATION SLIDES, PANELIST WORKSHOP PACKETS

SLIDES: 156–159 (TIMED AND UNTIMED ASSESSMENTS)

ANNEXES: A, F, AND R

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. Activity on the Angoff Method Round 2 (Task 3)

MATERIALS: FACILITATION SLIDES, PANELIST WORKSHOP PACKETS

SLIDES: 160–161 (TIMED AND UNTIMED ASSESSMENTS)

ANNEXES: A, F, AND R

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. Activity—Workshop Evaluation

MATERIALS: FACILITATION SLIDES, PANELIST WORKSHOP PACKETS

SLIDES: 162–163 (TIMED AND UNTIMED ASSESSMENTS)

ANNEX: P

In this presentation, you will provide instructions to the panelists on completing the workshop evaluation form. You will tell the panelists to take their time, while noting that the evaluation takes place while the lead facilitators and the data analyst are compiling the ratings from Round 2 (unless the analyst has another opportunity to do this, e.g., during lunch or a break; if that is the case, presentations 18 and 19 can be swapped). You will explain to the panelists that they should complete their evaluation forms to share their opinions about the following aspects of the workshop: 1) orientation and training, 2) Round 1 ratings, 3) Round 2 ratings, 4) benchmarks, and 5) the overall workshop. 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 34: 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.

19. Presentation and Discussion on the Angoff Method Round 2 Results

MATERIALS: FACILITATION SLIDES, PANELIST WORKSHOP PACKETS

SLIDES: 164–166 (TIMED AND UNTIMED ASSESSMENTS)

ANNEX: R

In this presentation, you will 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. Additional tips on how to lead this discussion are included in **Table 5**.

Figure 35: 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.

20. Workshop Closing and Logistics

MATERIALS: FACILITATION SLIDES, PANELIST WORKSHOP PACKETS

SLIDES: 167–171 (TIMED AND UNTIMED ASSESSMENTS)

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 S** 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.

F. TIPS FOR HOSTING REMOTE WORKSHOPS

Tips for hosting remote workshops follow. These tips are based on the first policy linking pilot workshop held remotely, with the People's Action for Learning (PAL) Network's International Common Assessment of Numeracy (ICAN) and panelists from Kenya and Nigeria, during the COVID-19 pandemic in August–September 2020.

Logistics

- Ensure panelists have the printed documents they will need to complete the workshop (see the subsection on **Panelist Packets in Chapter III** for details).
- 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.

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.

- 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 V

CHAPTER V. DOCUMENTING THE WORKSHOP OUTCOMES

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

MATERIALS: N/A

SLIDES: N/A

The lead facilitators and data analyst will need to produce the workshop technical documentation, which is critical for defending the benchmarks set by the panelists. An often-cited source of this type of documentation is the technical report on setting benchmarks for the National Assessment of Educational Progress (NAEP).¹⁷ **Annex T** provides an example of a benchmarking technical report outline adapted from NAEP that countries can use to report to global bodies.

The documentation includes the process, benchmarks (see **Annex R** for details on how to calculate these), panelist ratings and impact data (see **Annex M**), statistics, such as intra-rater and inter-rater consistency indices and the SE (see **Annex G** for details on how to calculate these), and evaluation feedback results. The intra-rater consistency index evaluates the panelists' overall consistency in estimating item difficulty. The inter-rater consistency index evaluates the panelists' overall agreement or consensus across all possible pairs of panelists. The SE provides details on the panelists' consistency in estimating the benchmarks.¹⁸

Intra-rater consistency is calculated for each panelist across all items on the assessment. The value ranges between 0 and 1. A lower value indicates high consistency and a higher value indicates low consistency. **Annex G** provides the formal equations and steps for calculating it.

Inter-rater consistency is calculated at the item level and for the entire assessment. The value ranges between 0 and 1 with values of 0.80 or greater desirable, as they indicate substantial agreement between the panelists. **Annex G** provides the formal equations and steps for calculating it.

The SE is calculated at the benchmark level. High SE values indicate a lack of consistency in panelists' estimated benchmarks, and low values indicate a high level of consistency in panelists' estimated benchmarks; acceptable values depend on the length of the assessment, among other factors. **Annex G** provides the formal equations and steps for calculating it.

Results of the panelists' workshop evaluations provide evidence of how well the policy linking method was implemented and to what extent panelists understood, applied, and had confidence in their benchmarks (see **Annex P** for the evaluation form and **Annex U** for details on how this information should be summarized and presented to the 4.1.1 Review Panel). Other potential sources of validity evidence are provided in the literature.¹⁹

Statistical processes to measure the accuracy and consistency of the benchmarking decisions that classify learners as meeting global minimum proficiency are also required. Several research studies have estimated the consistency and accuracy of learner classifications due to the benchmarks set on an assessment.²⁰ A method for calculating accuracy and consistency of the classifications is provided in **Annex V**.

¹⁷ See Hambleton & Bourque (1991) for an often-cited example of a benchmarking technical report.

¹⁸ See Cohen, 1960; Fleiss, 1971; Burry-Stock, Shaw, Lurie, & Chissom, 1996; Chang, 1999; and Ferdous & Plake, 2007 for calculating these indices and interpreting the results.

¹⁹ See Pitoniak, 2003; Hambleton & Pitoniak, 2006 for sources of validity evidence and methods for evaluating it.

²⁰ See Cohen, 1960; Subkoviak, 1976, 1988; Hanson & Brennan, 1990; Livingston & Lewis, 1995; Brennan, 2004; and Brennan & Wan, 2004 for methods on calculating classification accuracy and consistency. Subkoviak's method in **Annex P** is computationally straightforward.

Technical documentation (see **Annex W**) should be provided to the donor agency (if relevant) and the government (which will submit a report to the 4.I.I Review Panel) for reporting on the SDG and/or USAID indicators.

Finally, if the workshop is a pilot, the Policy Linking Global Working Group highly encourages countries and stakeholders to fill out the process documentation form included in **Annex W** to help inform updates to the toolkit and/or GPF.

CHAPTER VI

CHAPTER VI. REVIEWING AND SUBMITTING WORKSHOP OUTCOMES

After completing the policy linking workshop, a host government that wants to use the results for reporting against SDG Indicator 4.1.1. or USAID’s “F” indicators will need to submit the results to the 4.1.1 Review Panel for review and determination of workshop validity for reporting (Stage 6 of the Policy Linking for Global Reporting process). The process entails 1) collecting the evidence from the policy linking workshop, 2) submitting the evidence to UIS for review, and 3) waiting to receive a response back from UIS on whether the workshop results will be accepted for reporting. Note that the information needed to complete each of these steps is laid out in much more detail in the Criteria for Policy Linking Validity (CPLV) document.

A. COLLECT EVIDENCE FROM THE WORKSHOP

MATERIALS: N/A

SLIDES: N/A

RESOURCES: CPLV

Host governments sponsoring policy linking are invited to submit evidence from the workshop to UIS for review by its 4.1.1 Review Panel. Information submission is required if a host government wants to use the results from the policy linking workshop to report against SDG Indicator 4.1.1 and/or USAID’s “F” Indicators. The CPLV contains the information needed for submission, the source materials for that information, and the validity criteria.

B. SUBMIT EVIDENCE TO UIS

MATERIALS: N/A

SLIDES: N/A

RESOURCES: CPLV

UIS has quarterly submission deadlines: March 31, June 30, September 30, or December 31. If a government wants to report its results to UIS for the current year, then the government should complete the policy linking workshop and submit its evidence according to the timeline indicated in **Table 11**.

Table 11: Timeline for Submitting Results to UIS & Receiving Responses

Stage 3 Document Submission	Decision from the CPLV and UNESCO (Stage 4)	Policy Linking Workshop (Stage 5)	Stage 6 Document Submission	Decision from CPLV and UNESCO (Stage 7)
January	March 31	April–June	By June 30	September 30
February	March 31	April–June	By June 30	September 30
March	March 31	April–June	By June 30	September 30
April	June 30	July–Sept.	By Sept. 30	December 31
May	June 30	July–Sept.	By Sept. 30	December 31
June	June 30	July–Sept.	By Sept. 30	December 31
July	September 30	Oct.–Dec.	By Dec. 31	March 31
August	September 30	Oct.–Dec.	By Dec. 31	March 31
September	September 30	Oct.–Dec.	By Dec. 31	March 31
October	December 31	Jan. –March	By March 31	June 30
November	December 31	Jan. –March	By March 31	June 30
December	December 31	Jan. –March	By March 31	June 30

Similarly, USAID has annual deadlines for its congressional reporting, along with reporting requirements in terms of quality. Project teams should check with their Contracting Officer’s Representative (COR) at USAID to determine the appropriate timeline for results submission.

C. RECEIVE A RESPONSE FROM UIS

MATERIALS: N/A

SLIDES: N/A

RESOURCES: ANNEX T AND CPLV

The 4.1.1 Review Panel will review the workshop outcomes (see **Annex T** for the policy linking workshop validity criteria the review panel will use in evaluating the outcomes) and make one of three recommendations to UIS:

1. Policy linking carried out appropriately and reported outcomes are validated; as with Stage 2, the 4.1.1 CPLV will also provide a grade for the adequacy of the policy linking workshop. Grades follow:
 - a. **Excellent**—All six criteria are met.
 - b. **Good**—Four of the six criteria are met, two of which must be criteria b and c (inter-rater reliability and SE).
2. More evidence required to confirm whether policy linking was carried out appropriately before outcomes can be validated.
3. Policy linking not carried out appropriately and/or outcomes cannot be validated (in this case, the workshop would need to be re-run).

The Review Panel will produce a report to explain the rationale for its recommendation, including stipulating any additional documentation that must be submitted before it can recommend validated outcomes. UIS will share the outcomes with the government and confirm next or final steps.

Once the policy linking outcomes have been validated by the 4.1.1 CPLV and accepted by UIS, the government can submit the data for reporting against SDG 4.1.1 and/or USAID’s “F” Indicators (Stage 7). Data will be reported with associated grades (based on the results of the 4.1.1 Review Panel recommendations and UIS decisions in Stages 3 and 6), assigned as follows:

- **Excellent**—Country/assessment agency received an “excellent” rating on both the suitability of the assessment used for policy linking and the adequacy of the policy linking workshop.
- **Good**—Country/assessment agency either received “good” ratings for both the suitability of the assessment and the adequacy of the policy linking workshop or a “good” rating for one and an “excellent” rating for the other.
- **Sufficient**—Country/assessment agency received a “sufficient” rating for the suitability of the assessment and a “good” or “excellent” rating for the adequacy of the policy linking workshop.

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.

Plake, B., Melican, G., & Mills, C. (1991). Factors influencing intra-judge consistency during standard setting. *Educational Measurement: Issues and Practice*, 10(2), 15-16, 22-26.

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.

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.

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
- Global Proficiency Framework for Reading: Grades 1 to 9
- 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
- Workshop Evaluation Form
- Policy Linking Process Documentation Form
- Templates
 - Invitation Letter for Observers
 - Invitation Letter for Workshop Panelists
 - Certificate of Appreciation

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.

ANNEX C—WORKSHOP PREPARATION CHECKLIST

Table 12: 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 D—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 13: Alignment Rating Form Template

Question	Domain	Construct reference	Subconstruct reference	Knowledge or skill	Fit	These columns are only required where there is partial fit. You can use these to record any other domains, constructs, and subconstructs that relate to the item.				
						Domain	Construct reference	Subconstruct reference	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 E—WORKSHOP FACILITATION SLIDES

There are two sets of workshop facilitation slides:

- 1) The untimed assessment slides (which can also be used for most CBAs and other group assessments); you will also find math example items in these slides that can be used for either timed or untimed assessment workshops.
- 2) The timed assessment slides (which can be used for EGRA and EGMA, among other timed assessments); you will also find reading example items in these slides that can be used for either type of assessment workshop.

UNTIMED ASSESSMENT SLIDES

Slide 1

FACILITATOR NOTES FOR ADAPTING THE 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 and can be referenced in **Chapter IV** of the Toolkit.

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

Slide 2

POLICY LINKING FOR MEASURING GLOBAL LEARNING OUTCOMES WITH THE [UNTIMED ASSESSMENT(S) (UA(s))]

Lead Facilitators: [names]
Content Facilitators: [names]
Workshop Dates: [dates]

Slide 3

WELCOME AND INTRODUCTIONS

Workshop Participants

- Ministry of Education (MOE) officials—[name, location, position]
- Government assessment officials—[name, location, position]
- Panelists (groups)—[name, location, position]
- Resource persons/observers—[name, location, position]


Slide 4

WELCOME AND INTRODUCTIONS

Project Team


- [Donor, if applicable] education officials [name, position]
- [Implementing partner (IP), if applicable] representatives [name, position]
- Workshop coordinator(s) [name, position]
- Lead facilitator(s) [name, position]
- Content (group) facilitators [name, position]
- Administrative staff [name, position]

Slide 5

WORKSHOP OVERVIEW 

- 5 days: 9:00 a.m.–5:00 p.m. [Adjust times as needed].
- Morning/afternoon tea breaks; lunch break.
- The workshop will include **presentations by facilitators and activities** for panelists to complete in groups.
- **We will go over three main tasks over the course of 5 days.**

Slide 6

WORKSHOP OBJECTIVES 


By the end of this workshop, we aim to:

- Understand how well the [UA(s)] align with global minimum proficiency in [subjects] for [grades] as defined in the Global Proficiency Framework (GPF)
- Set benchmarks a learner would need to achieve on the [UA(s)] to demonstrate that they have met global minimum proficiency levels for [grades]
- Allow reporting of [UA(s)] to [SDG 4.1.1, USAID "F" Indicators, and/or other indicators]

Slide 7


DAY 1

Slide 8

FIVE-DAY OVERVIEW 

Day 1—[Date]	Day 4—[Date]
Opening, introductions, logistics, and agenda	Task 2 Presentation: Matching results
Background, objective, and tasks	Task 3 Presentation: Global benchmarking & Angoff method
Overview Presentation: Policy linking and the GPF	Task 3 Activity: Practice Angoff ratings
Overview Presentation: [UA(s)]	Task 3 Activity: Conduct Angoff Round 1
Day 2—[Date]	Day 5—[Date]
Task 1 Presentation: GPF and alignment	Task 3 Presentation: Round 1 results
Task 1 Activity: Align UA(s) and the GPF	Task 3 Presentation: Angoff method (review)
Day 3—[Date]	Task 3 Activity: Conduct Angoff Round 2
Task 1 Presentation: Alignment results	Task 3 Activity: Evaluate workshop
Task 2 Presentation: Matching [UA(s)] and Global Proficiency Descriptors/Levels (GPDs/GPLs)	Task 3 Presentation: Round 2 results
Task 2 Activity: Match [UA(s)] and GPDs/GPLs	Closing and logistics

Slide 9

PARTICIPANT PACKET 

1. Agenda
2. Panelist ID
3. Glossary of Terms
4. Acronym list
5. [Relevant grade/subject] GPDs from the GPF
6. Assessment instrument(s) [UA(s)]
7. Slides (printed in notes format)
8. Alignment rating form
9. Item rating form

Slide 10

KEY OBJECTIVES OF DAY ONE

- Understand Global Proficiency Framework
- Understand the purpose of policy linking
- Briefly review [UA(s)]

Slide 11

DAY ONE AGENDA

Time	Task, Presentations, and Activities	Facilitation
08:30-09:00	Registration	Administrators
09:00-10:00	Opening, introductions, logistics, and agenda	MOE, [donor], IP, and PLI
10:00-11:00	Presentation: Background, objective, and PL overview	Lead facilitators
11:00-11:15	Tea break	--
11:15-13:00	Presentation: Overview of the GPF and review of GPDs	Lead facilitators
13:00-14:00	Lunch break	--
14:00-14:30	Remaining questions on the GPF	Panelists
14:30-15:15	Presentation: Overview of the [UA(s)]	All facilitators
15:15-15:30	Tea break	--
15:30-16:30	Presentation: Overview of the [UA(s)]	All facilitators
16:30-17:00	Day 1 closing and preview of Day 2	All facilitators

Slide 12

PRESENTATION

WHAT IS POLICY LINKING?

Slide 13

WHAT IS POLICY LINKING?

- A low-cost, practical method that relies on panelist's judgment to link assessments (like the [UA]) to the **Global Proficiency Framework (GPF)** for reporting on Sustainable Development Goal 4.1.1 and other donor indicators

Slide 14

BACKGROUND ON POLICY LINKING: SDG 4.1.1

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."

- Reporting requires setting benchmarks for global minimum proficiency on all national and cross-national assessments.
- Policy linking was proposed as a method for linking assessments to the GPF and SDG 4.1.1 that includes a benchmarking task.

Slide 15

BACKGROUND ON POLICY LINKING: USAID STANDARD INDICATORS	
ES.1-1	Percent of learners targeted for USG assistance who attain a minimum grade-level proficiency in reading at the end of grade 2
ES.1-2	Percent of learners targeted for USG assistance who attain minimum grade-level proficiency in reading at the end of primary school
ES.1-47	Percent of learners with a disability targeted for USG assistance who attain a minimum grade-level proficiency in reading at the end of grade 2
ES.1-48	Percent of learners targeted for USG assistance with an increase of at least one proficiency level in reading at the end of grade 2
ES.1-54	Percent of individuals with improved reading skills following participation in USG-assisted programs

Slide 16

BACKGROUND ON POLICY LINKING: USAID SUPPLEMENTAL INDICATORS	
Supp-2	Percent of learners targeted for USG assistance with an increase of at least one proficiency level in reading at the end of primary school
Supp-3	Percent of learners targeted for USG assistance who attain minimum grade-level proficiency in math at the end of grade 2
Supp-4	Percent of learners with an increase in proficiency in math of at least one level at the end of grade 2 with USG assistance
Supp-5	Percent of learners targeted for USG assistance attaining minimum grade-level proficiency in math at the end of primary school with USG assistance
Supp-6	Percent of learners with an increase in proficiency in math of at least one level at the end of primary school
Supp-13	Percent of individuals with improved math skills following participation in USG-assisted programs

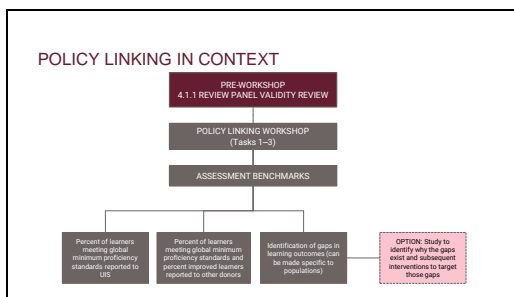
Slide 17

- 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

Slide 18

- BACKGROUND ON POLICY LINKING: TIMELINE**
- **April/May 2019:** Global Proficiency Framework (GPF) drafted
 - **September 2019:** Draft Policy Linking for Measuring Global Learning Outcomes Toolkit (PLT) written
 - **October 2019–September 2020:** Five pilot workshops conducted
 - **June–October 2020:** GPF and PLT updated based on pilots
 - **October 2020 and afterwards:** Additional workshops held to revise the GPF and PLT; training sessions for stakeholders planned

Slide 19



Slide 20

THREE KEY TASKS FOR POLICY LINKING WORKSHOP

Background

Today (Day 1) Begin with review of [UA] and GPF

↓

Alignment

Task 1 (Day 2). Check content alignment between [UA(s)] and the GPF
Task 2 (Day 3). Match [UA] items with the GPF

↓

Benchmarking

Task 3 (Days 4 and 5). Set [3] global benchmarks for each [UA] through two rounds of ratings

Slide 21

ALIGNMENT IN POLICY LINKING

Which rectangle is $\frac{1}{2}$ shaded?

Ⓐ

Ⓑ

Ⓒ

Ⓓ

Domain: ?

Construct: ?

Subconstruct: ?

Knowledge and/or skill(s): ?

Alignment: ?

Source: Mullis, I. V. S., Martin, M. O., Foy, P., & Hagem, M. (2016). TIMSS 2015 International Results in Mathematics. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: <http://timssandpirls.bc.edu/timss2015/Content/assessment.html>

Slide 22

ALIGNMENT IN POLICY LINKING

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
		N1.2 Represent whole numbers in standard form	Understanding of place value (ones, tens, hundreds, and thousands); understanding of standard form
		N1.3 Order operations using whole numbers	Ordering of numbers; understanding of place value (ones, tens, hundreds, and thousands); understanding of standard form
		N1.4 Represent word problems involving addition and subtraction	Understanding of addition and subtraction; understanding of place value (ones, tens, hundreds, and thousands); understanding of standard form
	N2: Fractions	N2.1 Represent word problems involving addition and subtraction	Understanding of addition and subtraction; understanding of place value (ones, tens, hundreds, and thousands); understanding of standard form
		N2.2 Identify and represent fractions using number models and number lines	Understanding of fractions; understanding of place value (ones, tens, hundreds, and thousands); understanding of standard form
		N2.3 Compare and order fractions	Understanding of fractions; understanding of place value (ones, tens, hundreds, and thousands); understanding of standard form
		N2.4 Represent word problems involving addition and subtraction	Understanding of addition and subtraction; understanding of place value (ones, tens, hundreds, and thousands); understanding of standard form
	N3: Decimals	N3.1 Represent decimals in equivalent forms (tenths and hundredths)	Understanding of decimals; understanding of place value (ones, tens, hundreds, and thousands); understanding of standard form
		N3.2 Compare and order decimals, fractions, and percents, including when they are positive and negative	Understanding of decimals; understanding of place value (ones, tens, hundreds, and thousands); understanding of standard form
		N3.3 Order operations using decimals	Understanding of decimals; understanding of place value (ones, tens, hundreds, and thousands); understanding of standard form
		N3.4 Represent word problems involving addition and subtraction	Understanding of addition and subtraction; understanding of place value (ones, tens, hundreds, and thousands); understanding of standard form

Slide 23

SETTING BENCHMARKS IN POLICY LINKING

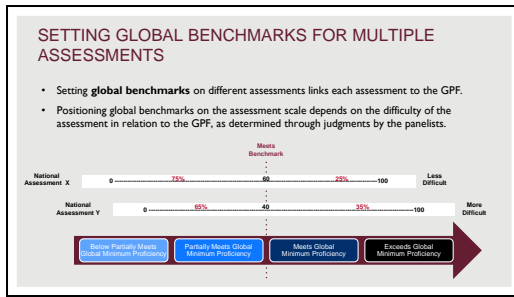
N: NUMBER AND OPERATIONS		Question	Item	Would a minimally proficient learner answer this item correctly?
N1: WHOLE NUMBERS				
N1.1: Identify and count in whole numbers, and identify their relative magnitude				
	N1.1.1_P	Count in whole numbers up to 100.	1 Circle the number 99	No
	N1.1.2_P	Read and write whole numbers up to 100 in words and in numerals	2 Sort the following numbers from smallest to largest: 99, 21, 64, 72, 46	Yes
	N1.1.3_P	Compare and order whole numbers up to 100.	3 Count the number of objects (note these are 14)	
			4 $12 + 7 =$	
			5 $42 + 19 =$	
	N1.1.4_P	Skip count forwards by twos or tens.	6 Identify which pencil is the longest	
			7 Identify which container will hold the most maize	

Slide 24

SETTING BENCHMARKS IN POLICY LINKING

- Once you have made your ratings for each item, you will then add up your yeses to get your panelist-level benchmark for the assessment.
- We will then average all of the panelist benchmarks to get the overall panel benchmark.

Slide 25



Slide 26

BENEFITS OF POLICY LINKING

- Enable **three types of analyses (CAT)** with the global benchmarks:
 - Compare** assessment results across contexts/languages within the country and with outcomes from other countries
 - Aggregate** assessment results across different assessments in the country and with those of other countries
 - Track** assessment results over time to monitor progress
- To allow for country ownership of outcomes—benchmarks set by countries for countries.
- To determine if learners have developed the knowledge and skills we should expect for their grade.

Slide 27

TEA BREAK

Slide 28

PRESENTATION

WHAT IS THE GLOBAL PROFICIENCY FRAMEWORK (GPF)?

Slide 29

THE GLOBAL PROFICIENCY FRAMEWORK

- Created by global reading and math experts and revised based on pilots
- Sets out global minimum proficiency (how much learners should be able to know and do) in reading and math for grades 1–9
- Evidence-based and:
 - Relies on developmental progressions
 - Relies on data from curriculum and assessments frameworks from across approximately 50 countries
- Not prescriptive

Slide 30

GLOBAL PROFICIENCY LEVELS (GPLs)

As part of their work on reporting against Sustainable Development Goal 4.1.1, UNESCO-UIS and its partners set four Global Proficiency Levels (GPLs) for the GPF:

- Below partially meets global minimum proficiency
- Partially meets global minimum proficiency
- Meets global minimum proficiency
- Exceeds global minimum proficiency

Slide 31

GLOBAL PROFICIENCY LEVELS (GPLs)

As part of their work on reporting against Sustainable Development Goal 4.1.1, UNESCO-UIS and its partners set four Global Proficiency Levels (GPLs):

- Below partially meets global minimum proficiency
- Partially meets global minimum proficiency
- **Meets global minimum proficiency** ← GPL used for SDG 4.1.1 reporting
- Exceeds global minimum proficiency

Slide 32

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 33

GPF DOMAINS

There are [X] domains in the GPF for [reading/mathematics]:

- [Domain]
- [Domain]
- [Domain]
- [Domain]
- [Domain]

Slide 34

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 Some operations using whole numbers
		N1.4 Some real-world problems involving whole numbers
	N2 Fractions	N2.1 Identify and represent fractions using words, pictures, and symbols, and identify relative magnitude
		N2.2 Some operations using fractions
		N2.3 Some real-world problems involving fractions
		N2.4 Identify and represent decimals using words, pictures, and symbols, and identify relative magnitude
	N3 Decimals	N3.1 Identify and represent decimals using words, pictures, and symbols, and identify relative magnitude
		N3.2 Some operations using decimals
		N3.3 Some real-world problems involving decimals
		N3.4 Some real-world problems involving decimals
N4 Integers	N4.1 Identify and represent integers using words, pictures, or symbols, and identify relative magnitude	
	N4.2 Some operations using integers	
	N4.3 Some real-world problems involving integers	
	N4.4 Some real-world problems involving integers	
N5 Exponents and roots	N5.1 Identify and represent exponents using exponents and roots, and identify the relative magnitude	
	N5.2 Some operations involving exponents and roots	
	N5.3 Some real-world problems involving exponents and roots	
	N5.4 Some operations involving exponents, fractions, decimals, percentages, and exponents	

Slide 35

GPF CONSTRUCTS AND SUBCONSTRUCTS

Domain	Construct	Subconstruct
M Measurement	M1 Length, weight, capacity, area, mass, and temperature	M1.1 Use non-standard and standard units to measure, compare, and order.
	M2 Time	M2.1 Tell time.
	M3 Estimation	M3.1 Use different capacity units to create amounts.
G Geometry	G1 Perimeter of shapes and figures	G1.1 Recognize and describe shapes and figures.
	G2 Area and perimeter	G2.1 Measure and estimate the length and area of shapes.
S Statistics and probability	S1 Data management	S1.1 Collect and organize data presented in displays.
	S2 Chance and probability	S2.1 Describe the likelihood of events in different ways.
A Algebra	A1 Patterns	A1.1 Recognize, describe, extend, and generate patterns.
	A2 Equations	A2.1 Solve problems involving relations (ratio, proportion, and percentages).
	A3 Relations and functions	A3.1 Represent an understanding of equations.

Slide 36

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 37

GPF KNOWLEDGE AND/OR SKILLS

Domain	Construct	Subconstruct	Knowledge or Skill
N Number and operations	N1 Whole numbers	N1.1 Identify and count a whole number and describe whole numbers.	Count, read and write whole numbers.
		N1.2 Represent whole numbers in equivalent ways.	Represent whole numbers in different ways (base ten blocks, number lines, number names, etc.).
		N1.3 Perform operations using whole numbers.	Add, subtract, multiply, and divide whole numbers.
		N1.4 Understand and explain the relationship between addition and subtraction.	Use addition to solve subtraction problems and vice versa.
N2 Fractions	N2.1 Identify and represent fractions using objects, pictures, and symbols.	N2.1.1 Identify and represent fractions.	Identify and represent fractions using objects, pictures, and symbols.
		N2.1.2 Compare and order fractions.	Compare and order fractions.
		N2.1.3 Add and subtract fractions.	Add and subtract fractions.
		N2.1.4 Multiply and divide fractions.	Multiply and divide fractions.
N3 Decimals	N3.1 Represent decimals in equivalent ways.	N3.1.1 Identify and represent decimals.	Identify and represent decimals using objects, pictures, and symbols.
		N3.1.2 Compare and order decimals.	Compare and order decimals.
		N3.1.3 Add and subtract decimals.	Add and subtract decimals.
		N3.1.4 Multiply and divide decimals.	Multiply and divide decimals.

Slide 38

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 39

GLOBAL PROFICIENCY DESCRIPTORS (MATH)

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

Grade	Partially meets	Meets	Exceeds
Grade 2 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 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 40


GLOBAL PROFICIENCY DESCRIPTORS 

[Insert reading/math GPF table here—may take more than one slide, perhaps one per domain or one per construct]

Slide 41


LUNCH

Slide 42

PRESENTATION 

REVIEW OF THE [UA(S)]

Slide 43

ASSESSMENT ACTIVITY 

- How did the pre-workshop assessment activity go?
- Were you able to assess:
 - 3 learners you classified as partially meeting global minimum proficiency
 - 3 learners who meet global minimum proficiency
 - 3 learners who exceed global minimum proficiency?
- How did the learners do on the assessment?
 - Which items did they do well on, which were more difficult?
 - What were some of the typical mistakes they made?

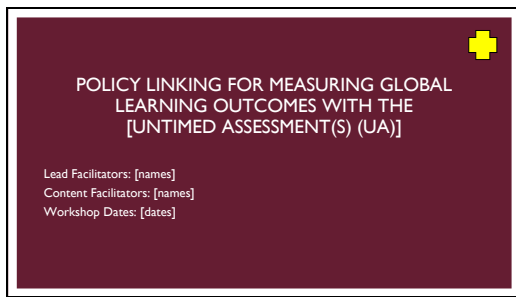
Slide 44

DAILY CHECK-IN

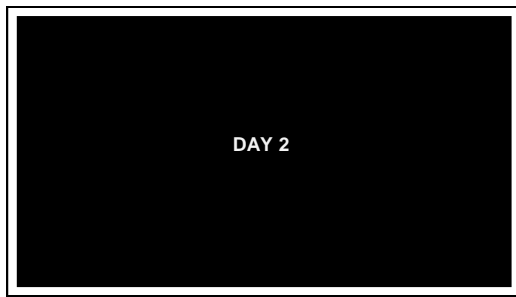
Slide 45



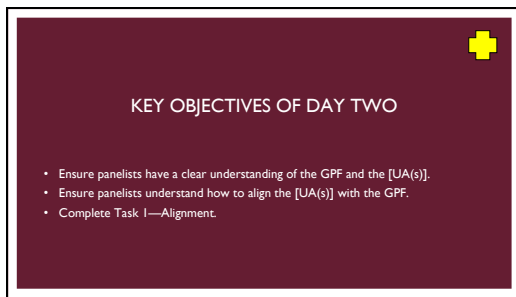
Slide 46



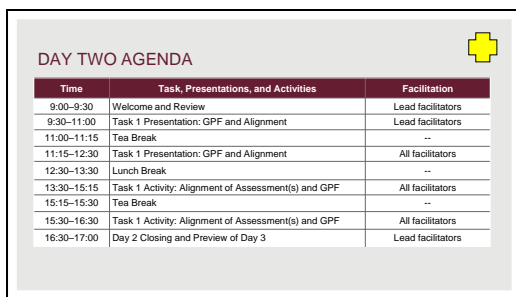
Slide 47




Slide 48



Slide 49




Slide 50

REVIEW OF DAY 1 


- Purpose of the workshop
- Policy linking
- Global Proficiency Framework
- [UA]

Slide 51

PRESENTATION 


TASK 1: CHECK CONTENT ALIGNMENT
BETWEEN [UA(S)] AND THE GPF

Slide 52

THE ALIGNMENT STUDY 

- **Activity 1**—The first activity in the workshop.
- **Task**—Panelists will make individual and independent judgements of whether the items on the [UA] are aligned with the GPF.
- **Purpose**—To ensure panelists have fully understood the GPF and to allow them to identify which statements of knowledge and/or skill(s) describe the knowledge and/or skill(s) required of children to answer assessment items correctly.
- **Sufficient Alignment**—Alignment is important to ensure there are enough items on an assessment that measure the knowledge and/or skill(s) depicted in the GPF for policy linking to work.
 - 4.1.1 Review Panel determined there was sufficient alignment


Slide 53

ALIGNING THE [UA(S)] AND THE GPF 

There are **two main steps**—each with sub-steps—for the alignment.

- **Step 1:** Panelists independently rate the alignment between the [UA] items and GPF knowledge and/or skill(s) statement(s) using a three sub-step process.
- **Step 2:** Facilitators compile and summarize the ratings to check the alignment between the assessments and the GPF.

Slide 54

ALIGNING THE [UA(S)] AND THE GPF 

Step 1 (completed by the panelists)

- Practice conducting item-statement of knowledge and/or skill(s) ratings with sample items.
- Work individually and independently to rate the alignment between each UA item and the GPF knowledge and/or skill(s) statements.
- Start with the first item and proceed item-by-item; find the GPF knowledge and/or skill(s) statements that align (if any) with the knowledge or skill(s) needed to answer the item correctly.
- Record the ratings on the alignment rating form using the rating scale (on the next slide).

Slide 55

ALIGNING THE [UA(S)] 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 56

ALIGNING THE [UA(S)] 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 57

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 58

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 59

EXAMPLE: NO FIT 

3. What is $2\frac{3}{4} - 1\frac{1}{3}$?

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 60



Slide 61

ALIGNMENT RATING FORM

These columns are only required where there is partial fit. You can use these to record the number of items that are not aligned.

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 62

BREADTH AND DEPTH ALIGNMENT LEVELS

- **Minimal alignment** if there are at least five number items covering at least 50 percent of the number subconstructs relevant at the grade level
- **Additional alignment** if there are at least five number and five measurement/geometry items covering at least 50 percent of the number, measurement, and geometry subconstructs relevant at the grade level
- **Strong alignment** if there are at least five number, five measurement/geometry, and five statistics and probability/algebra items covering at least 50 percent of all subconstructs relevant at the grade level

Slide 63

BREADTH AND DEPTH ALIGNMENT LEVELS

Level of Alignment	Category	Grade 1-2 Criteria	Grade 3-4 Criteria	Grade 7-9 Criteria
Minimally Aligned	Domain/Construct (depth):	D (minimum five items) C (minimum five items)	R (minimum five items)	R (minimum five items)
	Subconstructs (breadth):	Items covering at least 50 percent of the D and C subconstructs	Items covering at least 50 percent of the R subconstructs	Items covering at least 50 percent of the R subconstructs
Additionally Aligned	Domain/Construct (depth):	N/A	N/A	R: B1 (minimum 5 items) R: B2 (minimum 5 items)
	Subconstructs (breadth):	N/A	N/A	Items covering at least 50 percent of the R subconstructs
Strongly Aligned	Domain/Construct (depth):	R (minimum five items)	R: B1 (minimum five items) R: B2 (minimum five items) R: B3 (minimum five items)	R: B1 (minimum five items) R: B2 (minimum five items) R: B3 (minimum five items)
	Subconstructs (breadth):	Items covering at least 50 percent of the R subconstructs	Items covering at least 50 percent of the R subconstructs	Items covering at least 50 percent of the R subconstructs

Key:
D—Decoding
C—Comprehension of spoken or signed language
R—Reading comprehension
B1—Inferential information
B2—Explicit information
B3—Detail or information

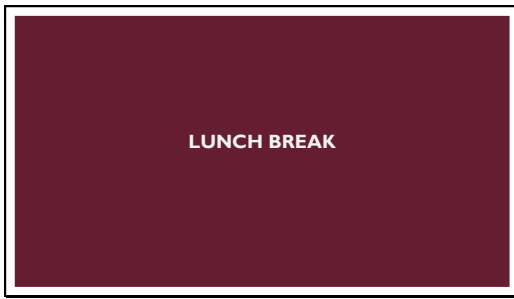
Slide 64

ALIGNING THE [UA(S)] AND THE GPF

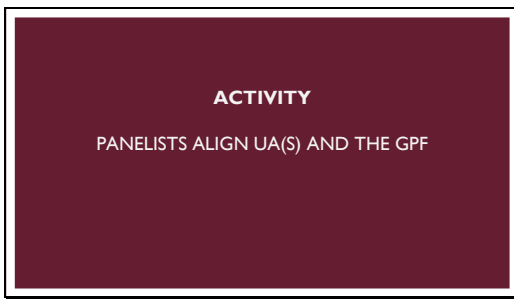
Step 2 (completed by the facilitators)

- **Compile, analyze, and summarize** the alignment ratings
- **Calculate totals, averages, and medians**
- **Answer these questions** on the alignment between the [UA(s)] and the GPF:
 - Are there at least 5 items from the [UA] that cover the domains relevant for the grade?
 - Are 50 percent of the subconstructs within the relevant domains covered?

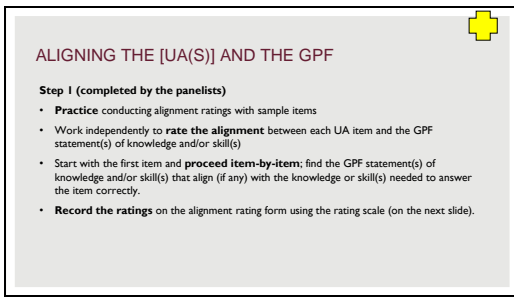
Slide 65



Slide 66



Slide 67



ALIGNING THE [UA(S)] AND THE GPF

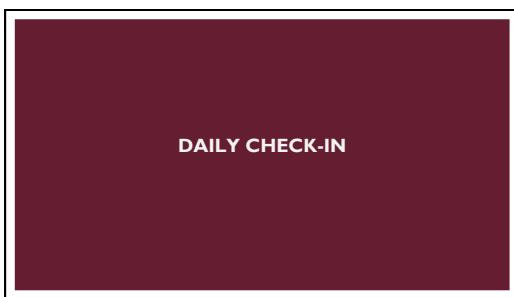
Step 1 (completed by the panelists)

- **Practice** conducting alignment ratings with sample items
- Work independently to **rate the alignment** between each UA item and the GPF statement(s) of knowledge and/or skill(s)
- Start with the first item and **proceed item-by-item**: find the GPF statement(s) of knowledge and/or skill(s) that align (if any) with the knowledge or skill(s) needed to answer the item correctly.
- **Record the ratings** on the alignment rating form using the rating scale (on the next slide).

Slide 68



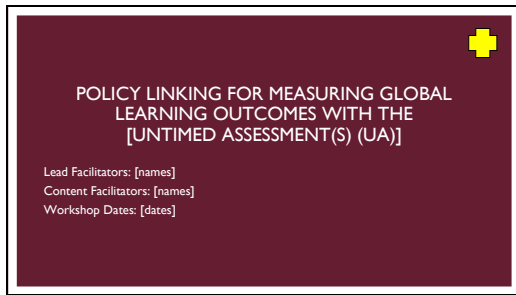
Slide 69



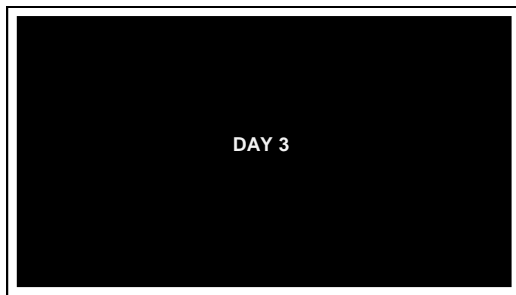
Slide 70



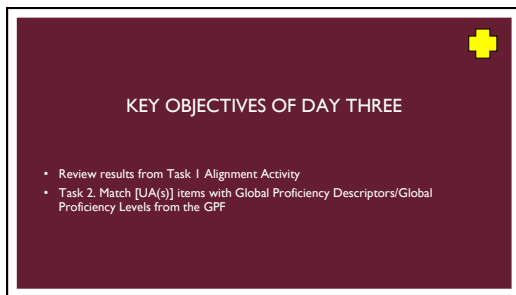
Slide 71



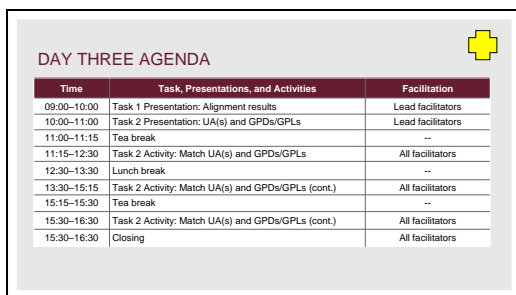
Slide 72




Slide 73



Slide 74



Slide 75



REVIEW OF DAY 2


- Purpose of Task 1
- Statements of knowledge and/or skill(s) (content standards) versus global proficiency descriptors (performance standards)
- Global Proficiency Framework
- [UA]

Slide 76

PRESENTATION

REVIEW PANELIST ALIGNMENT RESULTS FROM TASK 1


Slide 77



ALIGNMENT RATINGS

Domain	Items
N	Number and operations
M	Measurement
G	Geometry
S	Statistics and probability
A	Algebra
Total	26
Construct	Items
N1	Whole numbers
N2	Fractions
M1	Length, weight, capacity, volume, area, and perimeter
M2	Time
M3	Currency
G1	Properties of shapes and figures
G2	Spatial visualizations
G3	Position and direction
S1	Data management
A1	Patterns
A3	Relations and functions
Total	26


Slide 78



ALIGNMENT RATINGS

Subconstruct	Items
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
N1.4	Identify and represent fractions using objects, pictures, and symbols, and identify relative magnitude
M1.1	Use non-standard and standard units to measure, compare, and order
M2.1	Tell time
M2.2	Solve problems involving time
M3.1	Use different currency units to create amounts
G1.1	Recognize and describe shapes and figures
G2.1	Compose and decompose shapes and figures
G3.1	Describe the position and directions of objects in space
S1.1	Retrieve and interpret data presented in displays
A1.1	Recognize, describe, extend, and generate patterns
A3.2	Demonstrate an understanding of equivalency
Total	26


Slide 79




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

 In this picture, which child is farthest from the tree?



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

Slide 80

PRESENTATION

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

Slide 81

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 82

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 NA	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 83

MATCHING ITEMS WITH GPLS AND GPDS

- Build** on your understanding of the [UA] 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 84

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 85

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).

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 100 in words and in numerals.
Meets: Read and write whole numbers up to 1,000 in words and in numerals.
Exceeds: Read and write whole numbers up to 10,000 in words and in numerals.

Slide 86

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?
 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 and half-hours.

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 87

MATCHING ITEMS WITH GPLS AND GPDS

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



What makes the item easy/difficult?
 Difficult—understanding that the shaded portion must be $\frac{1}{3}$ 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 objects or pictures in fractional notation

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

Slide 88

MATCHING ITEMS WITH GPLS AND GPDS

Job had 16 peaches.
 He gave away 4 peaches.
 Then Job divided the remaining peaches equally between 2 baskets.
 How many peaches did Job 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?
 Grade 4 Meets—Solve simple real-world problems involving the multiplication of two whole numbers to 5, and associated division facts

Slide 89

TEA BREAK

Slide 90

ACTIVITY 

PANELIST GROUPS MATCH
[UA(S)] ITEMS WITH LEVELS AND
DESCRIPTORS IN THE GPF

Slide 91

MATCHING ITEMS TO GPLS/GPDS

1. **Work in panel-level groups;** start with the first item on the assessment and proceed item by item.
2. **Review the knowledge or skill in the GPF** (from Task 1) that matches with each item.
3. **Come to consensus on the statement of knowledge and/or skill(s) required and the lowest GPL and GPD** (performance standard) necessary to answer the word, question, or item correctly.
4. **Also identify what makes the item easy or difficult.**
5. **Write the GPL and GPD and what makes the item easy or difficult** on the test booklet next to the item, question, or word number on the GPF that matches with the item.

Slide 92

LUNCH BREAK

Slide 93

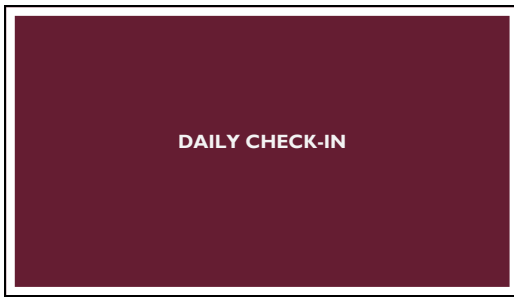
MATCHING ITEMS TO GPLS/GPDS

1. **Work in panel-level groups;** start with the first item on the assessment and proceed item by item.
2. **Review the knowledge or skill in the GPF** (from Task 1) that matches with each item.
3. **Come to consensus on the statement of knowledge and/or skill(s) required and the lowest GPL and GPD** (performance standard) necessary to answer the word, question, or item correctly.
4. **Also identify what makes the item easy or difficult.**
5. **Write the GPL and GPD and what makes the item easy or difficult** on the test booklet next to the item, question, or word number on the GPF that matches with the item.

Slide 94

TEA BREAK

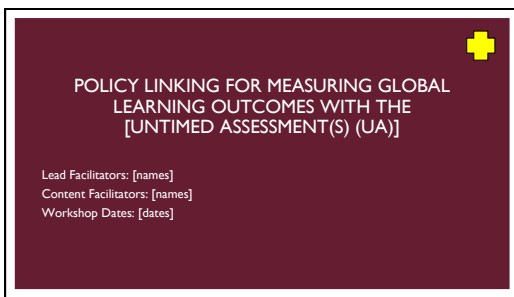
Slide 95



Slide 96



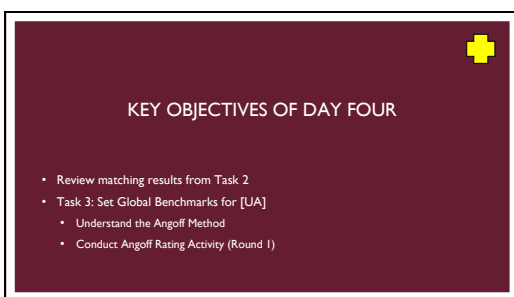
Slide 97




Slide 98



Slide 99




Slide 100



DAY FOUR AGENDA

Time	Task, Presentations, and Activities	Facilitation
09:00-10:00	Task 2 Presentation: Matching results	Lead facilitators
10:00-11:00	Task 3 Presentation: Global benchmarking	Lead facilitators
11:00-11:15	Tea break	---
11:15-12:30	Task 3 Presentation: Angoff method	Lead facilitators
12:30-13:30	Lunch break	---
13:30-15:00	Task 3 Activity: Practice Angoff method	All facilitators
15:00-16:15	Tea break	---
16:15-17:00	Task 3 Activity: Conduct Angoff Round 1	All facilitators

Slide 101



REVIEW OF DAY 3

- Purpose of Task 2
- Knowledge and skills (content standards) versus performance standards
- Global Proficiency Framework
- [UA]

Slide 102

PRESENTATION


REVIEW PANELIST GROUP MATCHING RESULTS FROM TASK 2

Slide 103

DISCUSSION OF GROUP MATCHING TASK 2

1. **Did you focus on this key aspect?**
Descriptions of levels of global minimally proficient learners (GPLs and GPDs) that match with the items
2. **Was it difficult to achieve consensus on some items? If so, which items and why?**
3. **Did you all agree with the group decisions? Why or why not?**

Slide 104



GROUP MATCHING RESULTS FROM TASK 2

Let's go through your group's matching results on items on the [UA(s)]

Summarize the answers to these questions for each item (based on group consensus):

1. What **knowledge and skills** are required to answer the items correctly?
2. Is the item **easy or difficult**?
3. What is the **lowest GPL and GPD** that is most appropriate for the item?

Slide 105

PRESENTATION

TASK 3. SET GLOBAL BENCHMARKS ON THE [JA(S)]

Slide 106

SETTING GLOBAL BENCHMARKS

- Use a standardized benchmarking procedure (the **Modified Angoff method**) for setting global benchmarks that will link the [JA(s)] to the GPF.
- Focus on setting the **Meets Benchmark** to separate the [JA] scores into two levels.
- For instance, imagine a Meets Benchmark of 50 points on a scale of 0 to 100 points.
- Determine the **score ranges for two levels**:
 - Below Partially Meets/Partially Meets** = 0 to 49 points
 - Meets/Exceeds** = 50 to 100 points.

Slide 107

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 108

CALCULATING GLOBAL MINIMUM PROFICIENCY PERCENTAGES

- Applying the global benchmarks to the data (and generalizing from a sample) for each assessment gives the **percentages of learners** meeting global minimum proficiency.
- Reporting on these percentages is **required** for the SDG and USAID indicators.

Slide 109

COMPARING, AGGREGATING, AND TRACKING RESULTS

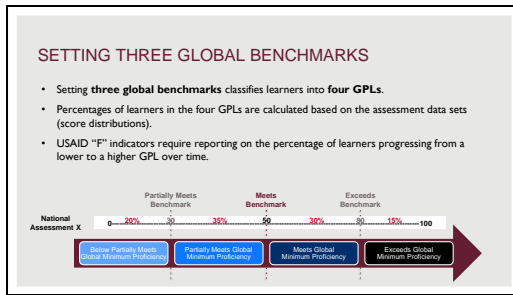
- Results from different countries can be **compared** by examining the percentages of learners meeting (and not meeting) global minimum proficiency.
- Results will be **aggregated both within and across** countries for global reporting.
- Results will be **tracked** over time (by country) to examine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

Slide 110

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%

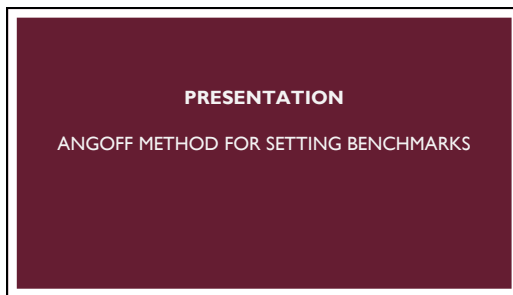
Slide 111



Slide 112



Slide 113



Slide 114

IMPLEMENTING THE ANGOFF METHOD FOR SETTING BENCHMARKS

- The **Modified Angoff method** is used to set the benchmarks:
 - Most popular 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

Slide 115

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 116

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 117

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 118

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 119

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.

Slide 120

ROUND 1: RATING PROCEDURE 

Step 2: Carefully read the first item on the assessment and consider the **knowledge and/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.

How is eight hundred and seventy written in standard form?

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

Knowledge and Skills Required: Count, read, and write whole numbers up to 1,000.

Item Stem: It is clearly stated.

Item Distractors: Options A and C are strong.

Possible Errors: Learners may confuse seventy with seven or seventeen.

Slide 121

ROUND 1: RATING PROCEDURE

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

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 whole numbers
GPLs and GPDs (performance standards):
Grade Level: Grade 3
Partially Meets: Read and write whole numbers up to 100 in words and in numerals.
Meets: Read and write whole numbers up to 1,000 in words and in numerals.
Exceeds: Read and write whole numbers up to 10,000 in words and in numerals.

Slide 122

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 123

ROUND 1: RATING PROCEDURE

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

FOR EACH WORD, QUESTION, OR ITEM:

```

    graph LR
      Q1{Would 2 of 3 JP learners be able to read the word or answer the question or item correctly?}
      Q2{Would 2 of 3 JM learners be able to read the word or answer the question or item correctly?}
      Q3{Would 2 of 3 JE learners be able to read the word or answer the question or item correctly?}
      Q4{Circle AE, and proceed to next word, question, or item.}
      
      Q1 -- Yes --> JP[Circle JP.]
      Q1 -- No --> Q2
      Q2 -- Yes --> JM[Circle JM.]
      Q2 -- No --> Q3
      Q3 -- Yes --> JE[Circle JE.]
      Q3 -- No --> Q4
  
```

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

Slide 124

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			
	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

Slide 125

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 126

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 127

LUNCH BREAK

Slide 128

TASK 3 ACTIVITY
PRACTICE ANGOFF METHOD

Slide 129

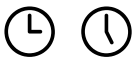
RATING PRACTICE ITEM 1 

<p>How is eighty-seven written in standard form?</p> <p>A. 80 B. 87 C. 807 D. R70</p> <p>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</p>	<p>What makes it easy or difficult: the other answer choices are strong distractors, especially C.</p> <p>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 . . .</p>
--	--

Slide 130

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.


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 131

RATING PRACTICE ITEM 3

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



What makes the item easy/difficult?
Difficult—understanding that the shaded portion must be $\frac{1}{3}$ 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

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

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

Slide 132

RATING PRACTICE ITEM 4

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. 4
B. 8
C. 10
D. 12

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?
Grade 4 Meets—Solve simple real-world problems involving the multiplication of two whole numbers to 5, and associated division facts.

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

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 . . .

Slide 133


TEA BREAK

Slide 134

TASK 3 ACTIVITY

CONDUCT ANGOFF BENCHMARKING ROUND I

Slide 135



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.

Slide 136

ANGOFF PROCEDURE: FOUR STEPS

Step 2: Carefully read the first item on the assessment and, building from Task 1, consider the **knowledge and/or skill(s)** 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.

Slide 137

ANGOFF PROCEDURE: FOUR STEPS

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

Slide 138

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 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 139

ROUND 1: RATING INSTRUCTIONS

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

FOR EACH WORD, QUESTION, OR ITEM:

```

graph TD
    Start[FOR EACH WORD, QUESTION, OR ITEM] --> Q1{Would 2 of 3 JP learners be able to read the word or answer the question or item correctly?}
    Q1 -- No --> Q2{Would 2 of 3 JM learners be able to read the word or answer the question or item correctly?}
    Q1 -- Yes --> JP[Circle JP.]
    Q2 -- No --> Q3{Would 2 of 3 JE learners be able to read the word or answer the question or item correctly?}
    Q2 -- Yes --> JM[Circle JM.]
    Q3 -- No --> AE[Circle AE, and proceed to next word, question, or item.]
    Q3 -- Yes --> JE[Circle JE.]
  
```

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

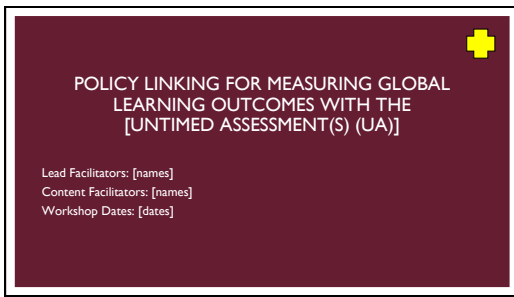
Slide 140



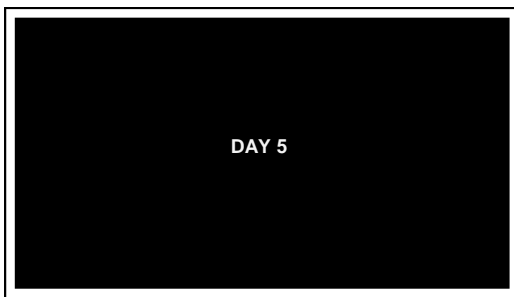
Slide 141



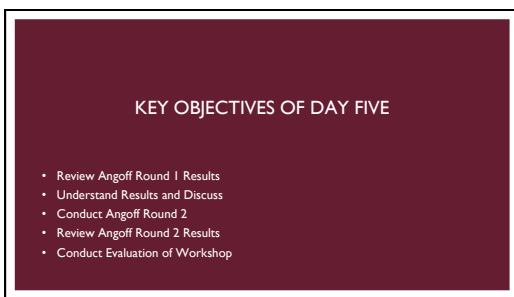
Slide 142




Slide 143



Slide 144




Slide 145



DAY FIVE AGENDA

Time	Task, Presentations, and Activities	Facilitation
09:00-10:00	Task 3 Presentation: Round 1 results	Lead facilitators
10:00-11:00	Task 3 Presentation: Discuss Round 1 results	Lead facilitators
11:00-11:15	Tea break	--
11:15-11:45	Task 3 Presentation: Angoff Round 2	All facilitators
11:45-12:30	Task 3 Activity: Angoff Round 2	All facilitators
12:30-13:30	Lunch break	--
13:30-14:30	Task 3 Activity: Angoff Round 2	All facilitators
14:30-15:00	Task 3 Activity: Workshop evaluation	All facilitators
15:00-16:15	Tea break	--
16:15-16:00	Task 3 Presentation: Round 2 results	Lead facilitators
16:00-17:00	Closing and logistics	MOE, USAID, IP, and PLI

Slide 146

- 
- ### REVIEW OF DAY 4
- Review matching results from Task 2
 - Task 3: Setting Global Benchmarks
 - Understand the Angoff Method
 - Conduct Angoff Rating Activity (Round 1)

Slide 147

PRESENTATION

REVIEW ANGOFF ROUND 1 ACTIVITY RESULTS FROM TASK 3

Slide 148

- ### ROUND 1 ITEM RATINGS AND BENCHMARKS
- We will review round 1 results in a few different ways:
- Individual panelists' **initial benchmarks** and their distributions
 - Differences in individual item ratings
 - **Location statistics** on panelists' item ratings
 - Item ratings in relation to **item difficulty values** (p-values)
 - **Impact data** showing percentage of learners falling into each GPL based on initial benchmarks

Slide 149

- ### ROUND 1 ITEM RATINGS AND BENCHMARKS
- We will review round 1 results in a few different ways:
- **Averages** of the panelists' benchmarks, i.e., the panel's initial benchmarks
 - **Differences in ratings** on specific items
 - **Impact data** with percentages of scores by GPL given the panel's benchmarks

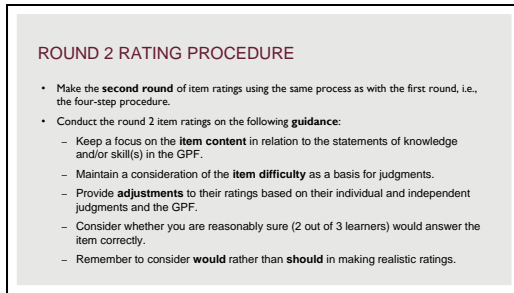
Slide 155



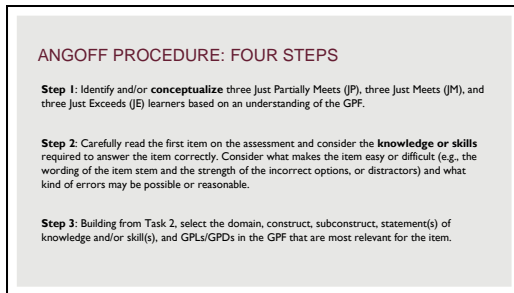
Slide 156



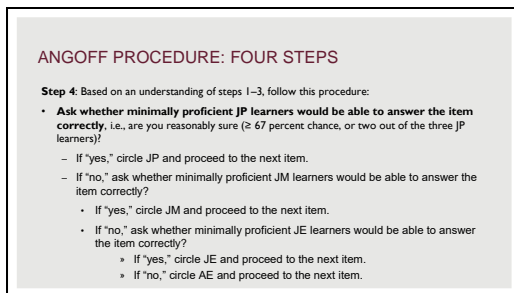
Slide 157



Slide 158



Slide 159



Slide 160

TASK 3 ACTIVITY
CONDUCT ANGOFF BENCHMARKING ROUND 2

Slide 161

LUNCH BREAK

Slide 162

WORKSHOP EVALUATION

Slide 163

WORKSHOP EVALUATION INSTRUCTIONS

- You will now complete an **evaluation form** to share your opinions about the following aspects of the workshop:
 - **Orientation and training** (guidance on setting benchmarks, practice with the method, interpretation of feedback information, adequacy of training time)
 - **Round 1 ratings** (confidence, comfort, and time allocation)
 - **Round 2 ratings** (confidence, comfort, and time allocation)
 - **Benchmarks** (calculations, feedback, and discussion)
 - **Workshop Overall** (organization, facilitation, and time allocation)

Slide 164

PRESENTATION
REVIEW ANGOFF ROUND 2 RESULTS

Slide 165

FINAL RESULTS AND SHIFT BETWEEN ROUNDS +

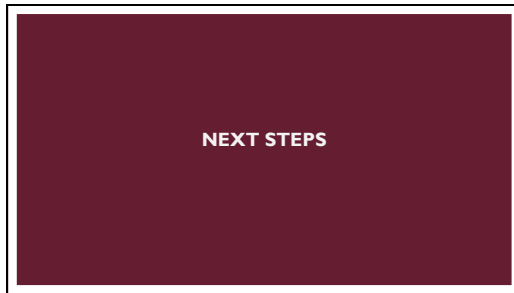
Impact data:

Minimum Proficiency Levels	ROUND 1			ROUND 2		
	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%			

Slide 166



Slide 167



Slide 168

- ### USE OF WORKSHOP RESULTS
- Enable **three types of analyses (CAT)** with the global benchmarks:
 - **Compare** assessment results across contexts/languages within the country and with outcomes from other countries.
 - **Aggregate** assessment results across different assessments in the country and with those of other countries.
 - **Track** assessment results over time to monitor progress.
 - Understand which learners most need support in the country.
 - Could inform a study into why gaps in learning exist and how best to address those.
 - How will you use the results?

Slide 169

DISCUSSION

- What do you think of the results?
- What, if anything, did you learn from this process?
- Has this informed your thinking about what learners in grade [X] should be able to accomplish? In what way(s)?

Slide 170



Slide 171



TIMED ASSESSMENT SLIDES

Slide 1




FACILITATOR NOTES FOR ADAPTING THE 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 and can be referenced in **Chapter IV** of the Toolkit.

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


Slide 2



POLICY LINKING FOR MEASURING GLOBAL LEARNING OUTCOMES WITH THE [TIMED ASSESSMENT(S) (TA(s))]

Lead Facilitators: [names]
 Content Facilitators: [names]
 Workshop Dates: [dates]


Slide 3

WELCOME AND INTRODUCTIONS 

Workshop Participants

- Ministry of Education (MOE) officials [name, location, position]
- Government assessment officials [name, location, position]
- Panelists (groups) [name, location, position]
- Resource persons/observers [name, location, position]


Slide 4

WELCOME AND INTRODUCTIONS 

Project Team


- [Donor, if applicable] education officials [name, position]
- [Implementing partner (IP), if applicable] representatives [name, position]
- Workshop coordinator(s) [name, position]
- Lead facilitator(s) [name, position]
- Content (group) facilitators [name, position]
- Administrative staff [name, position]

Slide 5

WORKSHOP OVERVIEW 

- 5 days: 9:00 a.m.–5:00 p.m. [Adjust times as needed].
- Morning/afternoon tea breaks; lunch break.
- The workshop will include **presentations by facilitators** and **activities** for panelists to complete in groups.
- **We will go over three main tasks over the course of 5 days.**

Slide 6

WORKSHOP OBJECTIVES 


By the end of this workshop, we aim to:

- Understand how well the [TA(s)] align with global minimum proficiency in [subjects] for [grades] as defined in the Global Proficiency Framework
- Set benchmarks a learner would need to achieve on the [TA(s)] to demonstrate that they have met global minimum proficiency levels for [grades]
- Allow reporting of [TA(s)] to [SDG 4.1.1, USAID "F" Indicators, and/or other indicators]

Slide 7

DAY 1


Slide 8



FIVE-DAY OVERVIEW

Day 1--(Date)	Day 4--(Date)
Opening, introductions, logistics, and agenda	Task 2 Presentation: Matching results
Background, objective, and tasks	Task 3 Presentation: Global benchmarking & Angoff method
Overview Presentation: Policy linking and the GPF	Task 3 Activity: Practice Angoff ratings
Overview Presentation: [TA(s)]	Task 3 Activity: Conduct Angoff Round 1
Day 2--(Date)	Day 5--(Date)
Task 1 Presentation: GPF and alignment	Task 3 Presentation: Round 1 results
Task 1 Activity: Align TA(s) and the GPF	Task 3 Presentation: Angoff method (review)
Day 3--(Date)	Task 3 Activity: Conduct Angoff Round 2
Task 1 Presentation: Alignment results	Task 3 Activity: Evaluate workshop
Task 2 Presentation: TA(s) and Global Proficiency Descriptors/Global Proficiency Levels (GPDs/GPLs)	Task 3 Presentation: Round 2 results
Task 2 Activity: Match [TA(s)] and GPDs/GPLs	Closing and logistics


Slide 9



PARTICIPANT PACKET

1. Agenda
2. Panelist ID
3. Glossary of Terms
4. Acronym list
5. [Relevant grade/subject] GPDs from the GPF
6. Assessment instrument(s) [TA(s)]
7. Slides (printed in notes format)
8. Alignment rating form
9. Item rating form


Slide 10



KEY OBJECTIVES OF DAY ONE

- Understand Global Proficiency Framework
- Understand the purpose of policy linking
- Briefly review [TA(s)]

Slide 11



DAY ONE AGENDA


Time	Task, Presentations, and Activities	Facilitation
08:30–09:00	Registration	Administrators
09:00–10:00	Opening, introductions, logistics, and agenda	MOE, [donor], IP, and PLJ
10:00–11:00	Presentation: Background, objective, and PL overview	Lead facilitators
11:00–11:15	Tea break	---
11:15–13:00	Presentation: Overview of the GPF and review of GPDs	Lead facilitators
13:00–14:00	Lunch break	---
14:00–14:30	Remaining questions on the GPF	Panelists
14:30–15:15	Presentation: Overview of the [TA(s)]	All facilitators
15:15–15:30	Tea break	---
15:30–16:30	Presentation: Overview of the [TA(s)]	All facilitators
16:30–17:00	Day 1 closing and preview of Day 2	All facilitators

Slide 12

PRESENTATION

WHAT IS POLICY LINKING?

Slide 13

WHAT IS POLICY LINKING? 

- A low-cost, practical method that relies on panelist’s judgment to link assessments (like the [TA(s)]) to the **Global Proficiency Framework (GPF)** for reporting on Sustainable Development Goal 4.1.1 and other donor indicators


Slide 14

BACKGROUND ON POLICY LINKING: SDG 4.1.1

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.”


- Reporting requires setting benchmarks for global minimum proficiency on all national and cross-national assessments.
- Policy linking was proposed as a method for linking assessments to the GPF and SDG 4.1.1 that includes a benchmarking task.

Slide 15

BACKGROUND ON POLICY LINKING: USAID STANDARD INDICATORS 

ES.1-1	Percent of learners targeted for USG assistance who attain a minimum grade-level proficiency in reading at the end of grade 2
ES.1-2	Percent of learners targeted for USG assistance who attain minimum grade-level proficiency in reading at the end of primary school
ES.1-47	Percent of learners with a disability targeted for USG assistance who attain a minimum grade-level proficiency in reading at the end of grade 2
ES.1-48	Percent of learners targeted for USG assistance with an increase of at least one proficiency level in reading at the end of grade 2
ES.1-54	Percent of individuals with improved reading skills following participation in USG-assisted programs

Slide 16

BACKGROUND ON POLICY LINKING: USAID SUPPLEMENTAL INDICATORS 

Supp-2	Percent of learners targeted for USG assistance with an increase of at least one proficiency level in reading at the end of primary school
Supp-3	Percent of learners targeted for USG assistance who attain minimum grade-level proficiency in math at the end of grade 2
Supp-4	Percent of learners with an increase in proficiency in math of at least one level at the end of grade 2 with USG assistance
Supp-5	Percent of learners targeted for USG assistance attaining minimum grade-level proficiency in math at the end of primary school with USG assistance
Supp-6	Percent of learners with an increase in proficiency in math of at least one level at the end of primary school
Supp-13	Percent of individuals with improved math skills following participation in USG-assisted programs

Slide 17

POLICY LINKING TIMELINE

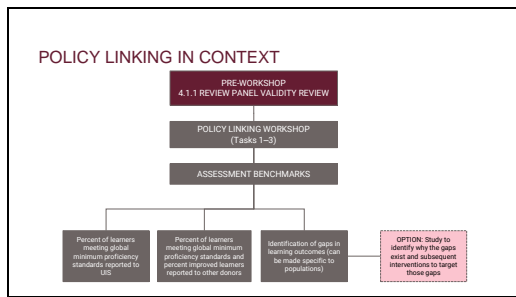
- **September 2017:** A UNESCO Institute for Statistics (UIS) stakeholder workshop proposed policy linking as a method for setting global benchmarks on each assessment based on a common proficiency scale
- **August 2018:** Joint U.S. Agency for International Development (USAID)–UIS stakeholder workshop discussed policy linking for reporting minimum proficiency through SDG 4.1.1 and USAID indicators

Slide 18

BACKGROUND ON POLICY LINKING: TIMELINE

- **April/May 2019:** Global Proficiency Framework (GPF) drafted
- **September 2019:** Draft Policy Linking for Measuring Global Learning Outcomes Toolkit (PLT) written
- **October 2019–September 2020:** Five pilot workshops conducted
- **June–October 2020:** GPF and PLT updated based on pilots
- **October 2020 and afterwards:** Additional workshops held to revise the GPF and PLT; training sessions for stakeholders planned

Slide 19



Slide 20

THREE KEY TASKS FOR POLICY LINKING WORKSHOP

Background Today (Day 1) Begin with review of [TA(s)] and GPF

Alignment

- Task 1 (Day 2). Check content alignment between [TA(s)] and the GPF
- Task 2 (Day 3). Match [TA(s)] items with the GPF

Benchmarking Task 3 (Days 4 and 5). Set [3] global benchmarks for each [TA(s)] through two rounds of ratings

Slide 21

ALIGNMENT IN POLICY LINKING

Grade and Subject: Grade 3 Reading

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

Passage:
 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: ?
Construct: ?
Subconstruct: ?
Knowledge and/or skill(s): ?
Alignment: ?

Slide 22

ALIGNMENT IN POLICY LINKING

Domain	Construct	Subconstruct	Knowledge or Skill
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	Integrate the meaning of grade-level words in a short grade-level passage and read or sign for the learner
	C2: Interpret information at sentence or text level	C2.1	Retrieve explicit information in a short grade-level passage and read or sign for the learner
		C2.2	Interpret information in a short grade-level passage and read or sign for the learner
D: Decoding	D1: Precision	D1.1	Identify grade-level phonemes
		D1.2	Identify grade-level words
	D2: Fluency	D2.1	Play or sign a grade-level continuous text with ease and accuracy
		D2.2	Play or sign fluently a grade-level continuous text

Slide 23

SETTING BENCHMARKS IN POLICY LINKING

R: READING COMPREHENSION

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 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. (See example items in Appendix C.)

R1.3: Retrieve explicit information in a grade-level text by synonymous word matching

Retrieve a single piece of explicit information from a grade 3-level text by synonymous word matching when there is no competing information. This will generally be in response to a "who," "what," "when," or "where" question. (See example items in Appendix C.)

R2: INTERPRET INFORMATION

R2.2: Make inferences in a grade-level text

Make simple inferences in a grade 3-level text by relating two pieces of explicit information in contrasting applications when there is limited competing information. This will generally be in response to a "why" or "how" question. (See example items in Appendix C.)

Item #	Item	Would a "Meets Global Minimum Proficiency" Learner answer this item correctly?
1	Who had a pet dog?	Yes
2	Why do you think the dog came back?	No
3	What is the topic of this story?	

Slide 24

SETTING BENCHMARKS IN POLICY LINKING

- Once you have made your ratings for each item, you will then add up your yeses to get your panelist-level benchmark for the assessment.
- We will then average all of the panelist benchmarks to get the overall panel benchmark.

Slide 25

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 26

BENEFITS OF POLICY LINKING

- Enable **three types of analyses (CAT)** with the global benchmarks:
 - Compare** assessment results across contexts/languages within the country and with outcomes from other countries
 - Aggregate** assessment results across different assessments in the country and with those of other countries
 - Track** assessment results over time to monitor progress
- To allow for country ownership of outcomes - benchmarks set by countries for countries.
- To determine if learners have developed the knowledge and skills we should expect for their grade.

Slide 27

TEA BREAK

Slide 28

PRESENTATION

WHAT IS THE GLOBAL PROFICIENCY FRAMEWORK (GPF)?

Slide 29

THE GLOBAL PROFICIENCY FRAMEWORK

- Created by global reading and math experts and revised based on pilots
- Sets out global minimum proficiency (how much learners should be able to know and do) in reading and math for grades 1-9
- Evidence-based and:
 - Relies on developmental progressions
 - Relies on data from curriculum and assessments frameworks from across approximately 50 countries
- Not prescriptive

Slide 30

GLOBAL PROFICIENCY LEVELS (GPLs)

As part of their work on reporting against Sustainable Development Goal 4.1.1, UNESCO-UIS and its partners set four Global Proficiency Levels (GPLs) for the GPF:

- Below partially meets global minimum proficiency
- Partially meets global minimum proficiency
- Meets global minimum proficiency
- Exceeds global minimum proficiency

Slide 31

GLOBAL PROFICIENCY LEVELS (GPLs)

As part of their work on reporting against Sustainable Development Goal 4.1.1, UNESCO-UIS and its partners set four Global Proficiency Levels (GPLs):

- Does not meet global minimum proficiency
- Partially meets global minimum proficiency
- **Meets global minimum proficiency** ← GPL used for SDG 4.1.1 reporting
- Exceeds global minimum proficiency

Slide 32

GPF OVERVIEW


- The Global Proficiency Framework (GPF) sets out the agreed domains, constructs, subconstructs, 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 33

GPF DOMAINS

There are [X] domains in the GPF for [reading/mathematics]:

- [Domain]
- [Domain]
- [Domain]
- [Domain]



Slide 34

GPF CONSTRUCTS AND SUBCONSTRUCTS

Domain	Construct	Subconstruct	Information
C	Comprehension of spoken or signed language	C1	Comprehend spoken and signed language at the word or phrase level
		C2	Recognize the meaning of common grade-level words in a short grade-level continuous text read to or signed for the learner
		C3	Interpret information in a short grade-level continuous text read to or signed for the learner
D	Decoding	D1	Identify graphemes representing English phonemes and orthographic patterns
		D2	Decode isolated words
R	Reading comprehension	R1	Draw upon a purpose , comprehension skill or process and with accuracy
		R1.1	Recognize the meaning of common grade-level words
		R1.2	Recognize the meaning of academic grade-level words
		R1.3	Recognize the meaning of academic grade-level words read to or signed for the learner
		R1.4	Recognize the meaning of academic grade-level words in a short grade-level continuous text read to or signed for the learner
		R1.5	Recognize the meaning of academic grade-level words in a short grade-level continuous text read to or signed for the learner
R2	Interpret information	R2.1	Identify the main and secondary ideas in a grade-level text
		R2.2	Identify the main and secondary ideas in a grade-level text
		R2.3	Identify the explicit and implicit ideas in a grade-level text
		R2.4	Draw on a text with justification
		R2.5	Draw on the ideas of common words in a text
R3	Reflect on information	R3.1	Draw on the ideas of common words in a text
		R3.2	Draw on the effectiveness of a text

Slide 35

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** identify the main idea in a grade-level text when it is not explicitly stated
- **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 identify the general topic of a grade 3-level continuous text when it is prominent but not explicitly stated.

Slide 36

GPF KNOWLEDGE AND SKILLS (READING)

Domain	Construct	Subconstruct	Knowledge or Skill
R	Retrieve information	R1	Recognize the meaning of common grade-level words
		R1.1	Recognize the meaning of common grade-level words
		R1.2	Recognize the meaning of academic grade-level words
		R1.3	Recognize the meaning of academic grade-level words read to or signed for the learner
		R1.4	Recognize the meaning of academic grade-level words in a short grade-level continuous text read to or signed for the learner
	Interpret information	R2	Identify the main and secondary ideas in a grade-level text
		R2.1	Identify the main and secondary ideas in a grade-level text
		R2.2	Identify the explicit and implicit ideas in a grade-level text
		R2.3	Draw on a text with justification
		R2.4	Draw on the ideas of common words in a text


Slide 37

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 38

GLOBAL PROFICIENCY DESCRIPTORS (MATH) 

- 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

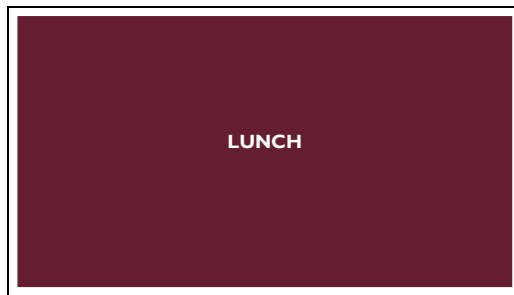
Grade 2	Partially meets	Meets	Exceeds
Time Tell Time	Sequence and describe events in time using informal comparisons	Tell time using an analog clock to the nearest hour.	Tell time using an analog clock to the nearest half hour.
Grade 3	Partially meets	Meets	Exceeds
Time Tell Time	Tell time using an analog 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 39

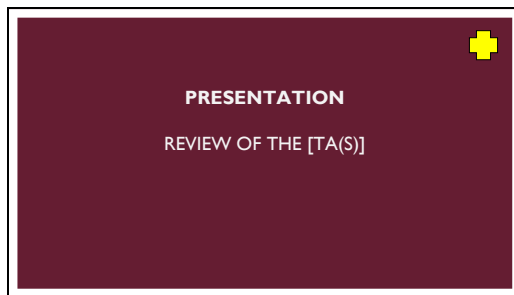
GLOBAL PROFICIENCY DESCRIPTORS 

[Insert reading/math GPF table here—may take more than one slide, perhaps one per domain or one per construct]


Slide 40



Slide 41



Slide 42

ASSESSMENT ACTIVITY 

- How did the pre-workshop assessment activity go?
- Were you able to assess:
 - 3 learners you classified as partially meeting global minimum proficiency
 - 3 learners who meet global minimum proficiency
 - 3 learners who exceed global minimum proficiency?
- How did the learners do on the assessment?
 - Which items did they do well on, which were more difficult?
 - What were some of the typical mistakes they made?

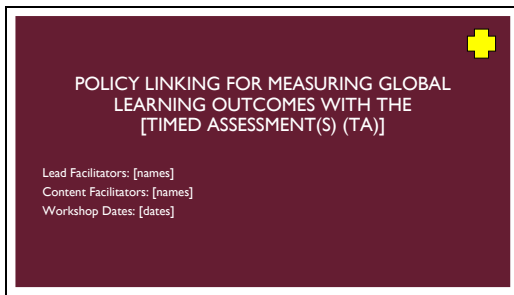
Slide 43



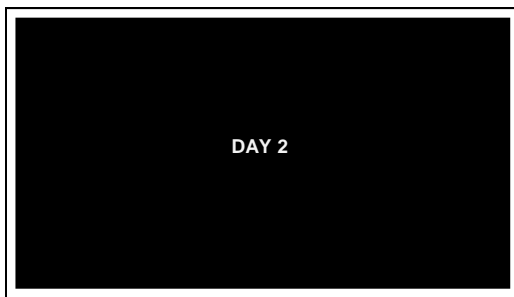
Slide 44



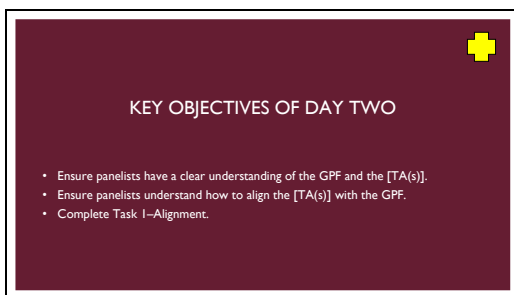
Slide 45




Slide 46



Slide 47




Slide 48



DAY TWO AGENDA

Time	Task, Presentations, and Activities	Facilitation
9:00–9:30	Welcome and Review	Lead facilitators
9:30–11:00	Task 1 Presentation: GPF and Alignment	Lead facilitators
11:00–11:15	Tea Break	—
11:15–12:30	Task 1 Presentation: GPF and Alignment	All facilitators
12:30–13:30	Lunch Break	—
13:30–15:15	Task 1 Activity: Alignment of Assessment(s) and GPF	All facilitators
15:15–15:30	Tea Break	—
15:30–16:30	Task 1 Activity: Alignment of Assessment(s) and GPF	All facilitators
16:30–17:00	Day 2 Closing and Preview of Day 3	Lead facilitators


Slide 49



REVIEW OF DAY 1

- Purpose of the workshop
- Policy linking
- Global Proficiency Framework
- [TA(s)]


Slide 50



PRESENTATION

TASK 1: CHECK CONTENT ALIGNMENT BETWEEN [TA(S)] AND THE GPF


Slide 51



THE ALIGNMENT STUDY

- **Activity 1**—The first activity in the workshop.
- **Task**—Panelists will make individual and independent judgements of whether the items on the [TA(s)] are aligned with the GPF.
- **Purpose**—To ensure panelists have fully understood the GPF and to allow them to identify which statements of knowledge and/or skill(s) describe the knowledge and/or skill(s) required of children to answer assessment items correctly.
- **Sufficient Alignment**—Alignment is important to ensure there are enough items on an assessment that measure the knowledge and/or skill(s) depicted in the GPF for policy linking to work.
 - 4.1.1 Review Panel determined there was sufficient alignment

Slide 52



ALIGNING THE [TA(S)] AND THE GPF

There are **two main steps**—each with sub-steps—for the alignment.

- **Step 1:** Panelists independently rate the alignment between the [TA] items and GPF knowledge and/or skill(s) statement(s) using a three sub-step process.
- **Step 2:** Facilitators compile and summarize the ratings to check the alignment between the assessments and the GPF.

Slide 53

ALIGNING THE [TA(S)] AND THE GPF

Step 1 (completed by the panelists)

- Practice conducting item-statement of knowledge and/or skill(s) ratings with sample items.
- Work individually and independently to rate the alignment between each TA item and the GPF knowledge and/or skill(s) statements.
- Start with the first item and proceed item-by-item; find the GPF knowledge and/or skill(s) statements that align (if any) with the knowledge or skill(s) needed to answer the item correctly.
- Record the ratings on the alignment rating form using the rating scale (on the next slide).

Slide 54

ALIGNING THE [TA(S)] 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 55

ALIGNING THE [TA(S)] 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).
- 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 56

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 57

EXAMPLE: COMPLETE FIT (CONSIDERING GRADE LEVEL OF PASSAGE)

Feature	Scope	Elaboration	Contextualization
Length	short	Six or more sentences; approximately 60-80 words in English	Fewer words in agglutinative or highly synthetic languages; fewer sentences if long sentences are commonly used
Familiarity	Familiar	Common everyday experiences, events and objects.	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	Minimal	Limited competing information; simple implied information	
Text structure	Very simple	Familiar, straightforward structure; a clear main idea with some supporting details; logical progression	
Vocabulary	Very common	A range of words with familiar meanings that typically describe concrete concepts and some common abstract 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	A variety of simple sentence structures that are commonly encountered.	Language dependent

Slide 58

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.

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: 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

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 59

EXAMPLE: NO FIT 


Grade and Subject: Grade 3 Reading
Reading Comprehension: Put the following items in the order they happened in the story: Jabu was sad. The dog ran away and got lost. The dog went to sleep. Jabu gave the dog some food.
Answer: The dog ran away and got lost. Jabu was sad. Jabu gave the dog some food. The dog went to sleep.

This item is a "no fit" item, as it requires the learner to sequence events from a text, which is not expected until a higher grade. There are no statements of knowledge and/or skill for sequencing events at the grade 3-level.

Slide 60

TEA BREAK


Slide 61

ALIGNMENT RATING FORM 

These columns are only required where there is partial fit. You can use these to record any grade-level assessment or performance on the basis of the fit.

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 62

BREADTH ALIGNMENT LEVELS--(MATHEMATICS) 

- Minimal alignment** if there are at least five number items covering at least 50 percent of the number of subconstructs relevant at the grade level
- Additional alignment** if there are at least five number and five measurement/geometry items covering at least 50 percent of the number, measurement, and geometry subconstructs relevant at the grade level
- Strong alignment** if there are at least five number, five measurement/geometry, and five statistics and probability/algebra items covering at least 50 percent of all subconstructs relevant at the grade level

Slide 63

BREADTH ALIGNMENT LEVELS – (READING)

Level of Alignment	Category	Grade 1–2 Criteria	Grade 3–6 Criteria	Grade 7–9 Criteria
Minimally Aligned	Domain/Construct (depth):	D (minimum five items) C (minimum five items)	R (minimum five items)	R (minimum five items)
	Subconstructs (breadth):	Items covering at least 50 percent of the D and C subconstructs	Items covering at least 50 percent of the R subconstructs	Items covering at least 50 percent of the R subconstructs
Additionally Aligned	Domain/Construct (depth):	N/A	N/A	R: B1 (minimum 5 items) R: B2 (minimum 5 items)
	Subconstructs (breadth):	N/A	N/A	Items covering at least 50 percent of the R subconstructs
Strongly Aligned	Domain/Construct (depth):	R (minimum five items)	R: B1 (minimum five items) R: B2 (minimum five items)	R: B1 (minimum five items) R: B2 (minimum five items) R: B3 (minimum five items)
	Subconstructs (breadth):	Items covering at least 50 percent of the R subconstructs	Items covering at least 50 percent of the R subconstructs	Items covering at least 50 percent of the R subconstructs

Key:
D—Decoding
C—Comprehension of spoken or signed language
R—Reading comprehension
B1—Retrieval information
B2—Inferential information
B3—Reflective information

Slide 64

ALIGNING THE [TA(S)] AND THE GPF

Step 2 (completed by the facilitators)

- **Compile, analyze, and summarize** the alignment ratings
- **Calculate totals, averages, and medians**
- **Answer these questions** on the alignment between the [TA(s)] and the GPF:
 - Are there at least 5 items from the [TA(s)] that cover the domains relevant for the grade?
 - Are 50 percent of the subconstructs within the relevant domains covered?

Slide 65

LUNCH BREAK

Slide 66

ACTIVITY

PANELISTS ALIGN TA(S) AND THE GPF

Slide 67

ALIGNING THE [TA(S)] AND THE GPF

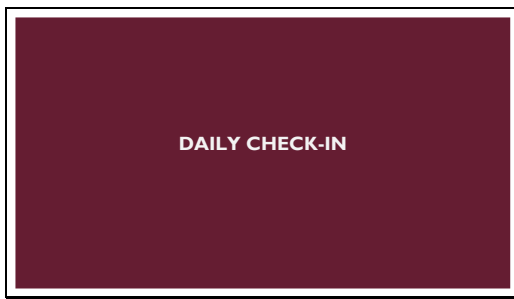
Step 1 (completed by the panelists)

- **Practice** conducting alignment ratings with sample items.
- Work independently to **rate the alignment** between each TA item and the GPF statement(s) of knowledge and/or skill(s).
- Start with the first item and **proceed item-by-item**: find the GPF statement(s) of knowledge and/or skill(s) that align (if any) with the knowledge or skill(s) needed to answer the item correctly.
- **Record the ratings** on the alignment rating form using the rating scale (on the next slide).

Slide 68



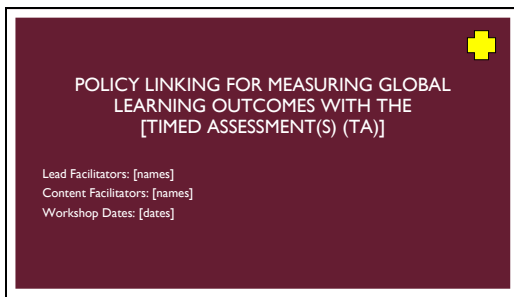
Slide 69



Slide 70



Slide 71



Slide 72



Slide 73

KEY OBJECTIVES OF DAY THREE

- Review results from Task 1 Alignment Activity
- Task 2. Match [TA(s)] items with Global Proficiency Descriptors/Global Proficiency Levels from the GPF

Slide 74

DAY THREE AGENDA

Time	Task, Presentations, and Activities	Facilitation
09:00–10:00	Task 1 Presentation: Alignment results	Lead facilitators
10:00–11:00	Task 2 Presentation: TA(s) and GPDs/GPLs	Lead facilitators
11:00–11:15	Tea break	--
11:15–12:30	Task 2 Activity: Match TA(s) and GPDs/GPLs	All facilitators
12:30–13:30	Lunch break	--
13:30–15:15	Task 2 Activity: Match TA(s) and GPDs/GPLs (cont.)	All facilitators
15:15–15:30	Tea break	--
15:30–16:30	Task 2 Activity: Match TA(s) and GPDs/GPLs (cont.)	All facilitators
16:30–16:30	Closing	All facilitators

Slide 75

REVIEW OF DAY 2

- Purpose of Task 1
- Statements of knowledge and/or skill(s) (content standards) versus global proficiency descriptors (performance standards)
- Global Proficiency Framework
- [TA(s)]

Slide 76

PRESENTATION


REVIEW PANELIST ALIGNMENT RESULTS FROM TASK 1

Slide 77

ALIGNMENT RATINGS

Domain	Items
C	14
D	7
R	3
Total	24
Construct	Items
C1	14
C2	0
C3	0
D1	4
D2	0
R1	2
R2	0
R3	1
Total	24

Slide 83

MATCHING ITEMS WITH GPLS AND GPDS 

- **Build** on your understanding of the [TA] 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 84

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** that is most appropriate for the item?


Slide 85

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: Who has a pet dog?	GPL and GPD (performance standard):
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.	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. This will generally be in response to a "who," "what," "when," or "where" question.
Domain: Reading Comprehension	Meets: "... and there is limited competing information" ...
Construct: Retrieve Information	Exceeds: Retrieve multiple pieces of explicit information ... when the information required is adjacent to the matched word and there is limited competing 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.	


Slide 86

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"	GPL and GPD (performance standard):
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.	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).
Domain: Decoding	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.
Construct: Fluency	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.
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 87

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?	GPL and GPD (performance standard):
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.	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 G).
Domain: Reading Comprehension	Meets: "... when there is limited competing information" ...
Construct: Interpret 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" ...
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	

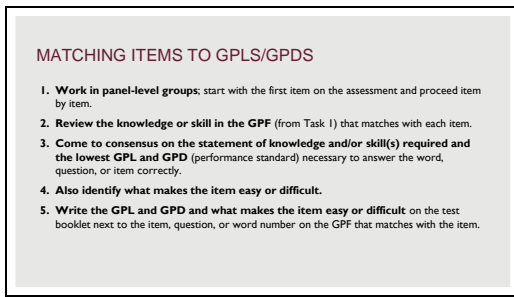
Slide 88



Slide 89

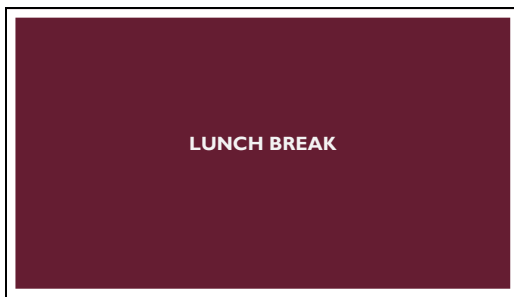


Slide 90

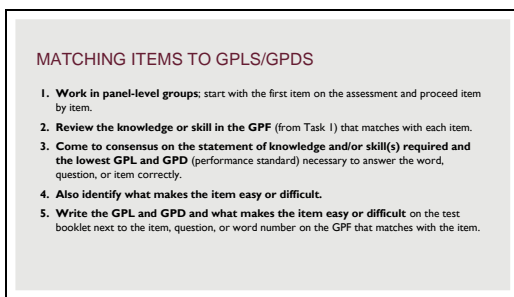


- MATCHING ITEMS TO GPLS/GPDS**
1. **Work in panel-level groups;** start with the first item on the assessment and proceed item by item.
 2. **Review the knowledge or skill in the GPF** (from Task 1) that matches with each item.
 3. **Come to consensus on the statement of knowledge and/or skill(s) required and the lowest GPL and GPD** (performance standard) necessary to answer the word, question, or item correctly.
 4. **Also identify what makes the item easy or difficult.**
 5. **Write the GPL and GPD and what makes the item easy or difficult** on the test booklet next to the item, question, or word number on the GPF that matches with the item.

Slide 91



Slide 92

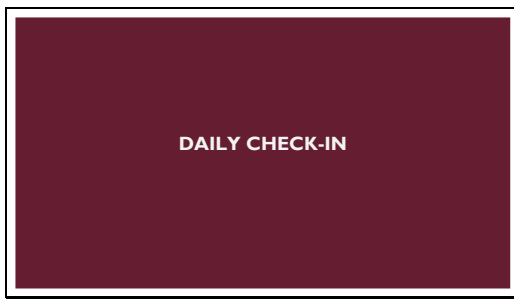


- MATCHING ITEMS TO GPLS/GPDS**
1. **Work in panel-level groups;** start with the first item on the assessment and proceed item by item.
 2. **Review the knowledge or skill in the GPF** (from Task 1) that matches with each item.
 3. **Come to consensus on the statement of knowledge and/or skill(s) required and the lowest GPL and GPD** (performance standard) necessary to answer the word, question, or item correctly.
 4. **Also identify what makes the item easy or difficult.**
 5. **Write the GPL and GPD and what makes the item easy or difficult** on the test booklet next to the item, question, or word number on the GPF that matches with the item.

Slide 93



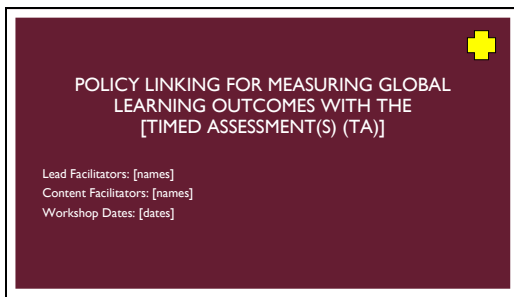
Slide 94



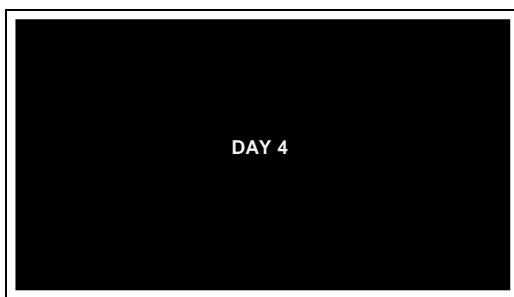
Slide 95



Slide 96



Slide 97




Slide 98



KEY OBJECTIVES OF DAY FOUR

- Review matching results from Task 2
- Task 3: Set Global Benchmarks for [TA(s)]
 - Understand the Angoff Method
- Conduct Angoff Rating Activity (Round 1)


Slide 99



DAY FOUR AGENDA

Time	Task, Presentations, and Activities	Facilitation
09:00–10:00	Task 2 Presentation: Matching results	Lead facilitators
10:00–11:00	Task 3 Presentation: Global benchmarking	Lead facilitators
11:00–11:15	Tea break	--
11:15–12:30	Task 3 Presentation: Angoff method	Lead facilitators
12:30–13:30	Lunch break	--
13:30–15:00	Task 3 Activity: Practice Angoff method	All facilitators
15:00–15:15	Tea break	--
15:15–17:00	Task 3 Activity: Conduct Angoff Round 1	All facilitators

Slide 100



REVIEW OF DAY 3

- Purpose of Task 2
- Knowledge and skills (content standards) versus performance standards
- Global Proficiency Framework
- [TA(s)]

Slide 101

PRESENTATION

REVIEW PANELIST GROUP MATCHING RESULTS FROM TASK 2

Slide 102

DISCUSSION OF GROUP MATCHING TASK 2

1. **Did you focus on this key aspect?**
Descriptions of levels of global minimally proficient learners (GPLs and GPDs) that match with the items
2. **Was it difficult to achieve consensus on some items? If so, which items and why?**
3. **Did you all agree with the group decisions? Why or why not?**

Slide 103

GROUP MATCHING RESULTS FROM TASK 2

Let's go through your group's matching results on items on the [JA(s)]

Summarize the answers to these questions for each item (based on group consensus):

1. What **knowledge and skills** are required to answer the items correctly?
2. Is the item **easy or difficult**?
3. What is the **lowest GPL and GPD** that is most appropriate for the item?

Slide 104

PRESENTATION

TASK 3. SET GLOBAL BENCHMARKS ON THE [TA(S)]

Slide 105

SETTING GLOBAL BENCHMARKS

- Use a standardized benchmarking procedure (the **Modified Angoff method**) for setting global benchmarks that will link the [TA(s)] to the GPF.
- Focus on setting the **Meets Benchmark** to separate the [TA] scores into two levels.
- For instance, imagine a Meets Benchmark of 50 points on a scale of 0 to 100 points.
- Determine the **score ranges for two levels**:
 - **Below/Partially Meets/Partially Meets** = 0 to 49 points
 - **Meets/Exceeds** = 50 to 100 points.

Slide 106

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 107

CALCULATING GLOBAL MINIMUM PROFICIENCY PERCENTAGES

- Applying the global benchmarks to the data (and generalizing from a sample) for each assessment gives the **percentages of learners** meeting global minimum proficiency.
- Reporting on these percentages is **required** for the SDG and USAID indicators.

Slide 108

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 109

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 110

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 "F" indicators require reporting on the percentage of learners progressing from a lower to a higher GPL over time.

Slide 111

TEA BREAK

Slide 112

PRESENTATION

ANGOFF METHOD FOR SETTING BENCHMARKS

Slide 113

IMPLEMENTING THE ANGOFF METHOD FOR SETTING BENCHMARKS

- The **Modified Angoff method** is used to set the benchmarks:
 - Most popular 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

Slide 114

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 115

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 116

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 117

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 118


ROUND 1: RATING PROCEDURE



Just Partially Meets (JP)



Just Meets (JM)



Just Exceeds (JE)

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 GFP.

Slide 119

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	AE	JP	JM	JE	AE	JP	JM	JE	AE	JP	JM	JE	AE
1	Kanda	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	Debu	4	4	4		JP	JM	JE	AE	4	4	4		JP	JM	JE	AE
5	Isukan	5	5	5		JP	JM	JE	AE	5	5	5		JP	JM	JE	AE

Slide 120

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 121

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 GFP 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 3

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., 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.

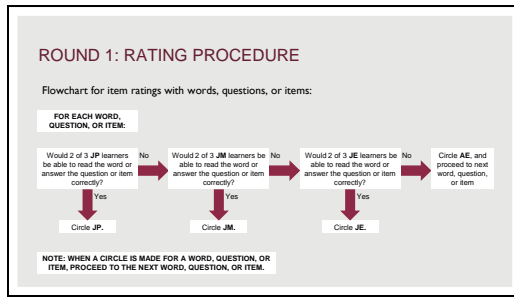
Slide 122

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 (only consider “attempted” items)
 - 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 123



Slide 124

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	AE	JP	JM	JE	AE	JP	JM	JE	AE	JP	JM	JE	AE
1	Konde	1	1	1		JP	JM	JE	AE	1	1	1		JP	JM	JE	AE
2	ds	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	Datu	4	4	4		JP	JM	JE	AE	4	4	4		JP	JM	JE	AE
5	Isukan	5	5	5		JP	JM	JE	AE	5	5	5		JP	JM	JE	AE

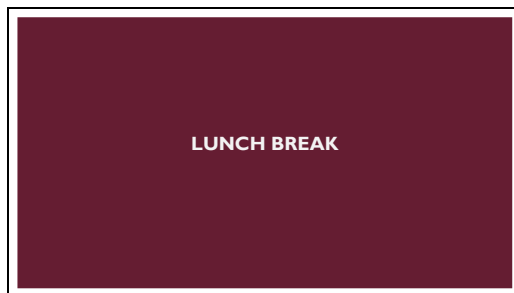
Slide 125

- ### 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 126

- ### 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 127




Slide 128

TASK 3 ACTIVITY

PRACTICE ANGOFF METHOD


Slide 129



RATING PRACTICE ITEM 1

<p>Item: Who has a pet dog?</p> <p>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.</p> <p>Domain: Reading Comprehension</p> <p>Construct: Retrieve Information</p> <p>Subconstruct: Retrieve explicit information in a grade-level text by direct- or close-word matching</p> <p>Knowledge or skill: Retrieve a single piece of explicit information from a grade-level continuous text by direct- or close-word matching</p>	<p>GPL and GPD (performance standard):</p> <p>Lowest GPD to answer correctly—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. This will generally be in response to a 'who', 'what', 'when', or 'where' question.</p> <p>Would 2 out of 3 JP learners answer the item correctly? . . .</p> <p style="background-color: #f8d7da; padding: 2px;">If yes, then circle JP</p> <p style="background-color: #f8d7da; padding: 2px;">If no, then ask about JM . . .</p>
---	---


Slide 130



RATING PRACTICE ITEM 2

<p>Item: Decoding passage—Word "Jabu"</p> <p>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.</p> <p>Domain: Decoding</p> <p>Construct: Fluency</p> <p>Subconstruct: Say or sign a grade-level continuous text at pace and with accuracy</p> <p>Knowledge or skill: Say or sign fluently a grade-level continuous text</p>	<p>Lowest GPD to answer correctly—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).</p> <p>Would 2 out of 3 JP learners answer the item correctly? . . .</p> <p style="background-color: #f8d7da; padding: 2px;">If yes, then circle JP</p> <p style="background-color: #f8d7da; padding: 2px;">If no, then ask about JM . . .</p> <p style="background-color: #f8d7da; padding: 2px;">If no, then ask about JE . . .</p>
---	---

Slide 131



RATING PRACTICE ITEM 3

<p>Item: Why did the dog come back?</p> <p>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.</p> <p>Domain: Reading Comprehension</p> <p>Construct: Interpret Information</p> <p>Subconstruct: Make inferences in a grade-level text</p> <p>Knowledge or skill: Make simple inferences in a grade-level text by relating pieces of explicit and/or implicit information in the text</p>	<p>Lowest GPD to answer correctly—Exceeds: Make simple inferences in a grade 3-level text by relating two pieces of explicit information 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. This will generally be in response to a 'why' or 'how' question. (See example items in Appendix C).</p> <p>Would 2 out of 3 JP learners answer the item correctly? . . .</p> <p style="background-color: #f8d7da; padding: 2px;">If yes, then circle JP</p> <p style="background-color: #f8d7da; padding: 2px;">If no, then ask about JM . . .</p> <p style="background-color: #f8d7da; padding: 2px;">If no, then ask about JE . . .</p>
--	---

Slide 132


TEA BREAK

Slide 133

TASK 3 ACTIVITY


CONDUCT ANGOFF BENCHMARKING ROUND 1

Slide 134

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.

Slide 135

ANGOFF PROCEDURE: FIVE STEPS 

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).

Slide 136

ANGOFF PROCEDURE: FIVE STEPS

Step 3: Carefully read the first item on the assessment and, building from Task 1, consider the **knowledge and/or skill(s)** 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.

Slide 137

ANGOFF PROCEDURE: FIVE STEPS

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

Slide 138

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 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 139

ROUND 1: RATING INSTRUCTIONS

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

FOR EACH WORD, QUESTION, OR ITEM:

```

graph LR
    A[Would 2 of 3 JP learners be able to read the word or answer the question or item correctly?] -- No --> B[Would 2 of 3 JM learners be able to read the word or answer the question or item correctly?]
    A -- Yes --> C[Circle JP.]
    B -- No --> D[Would 2 of 3 JE learners be able to read the word or answer the question or item correctly?]
    B -- Yes --> E[Circle JM.]
    D -- No --> F[Circle AE, and proceed to next word, question, or item]
    D -- Yes --> G[Circle JE.]
  
```

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


Slide 140

DAILY CHECK-IN

Slide 141

CLOSING

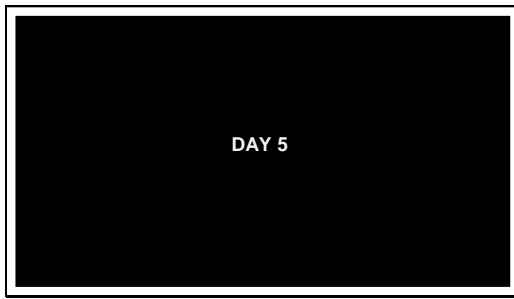
Slide 142



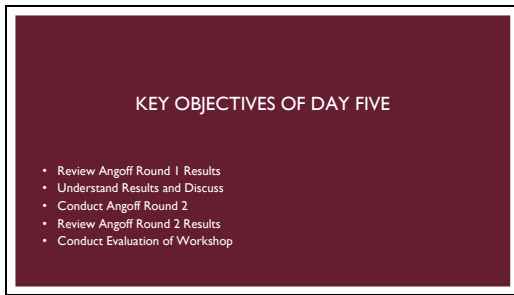
POLICY LINKING FOR MEASURING GLOBAL LEARNING OUTCOMES WITH THE [TIMED ASSESSMENT(S) (TA)]

Lead Facilitators: [names]
 Content Facilitators: [names]
 Workshop Dates: [dates]

Slide 143



Slide 144

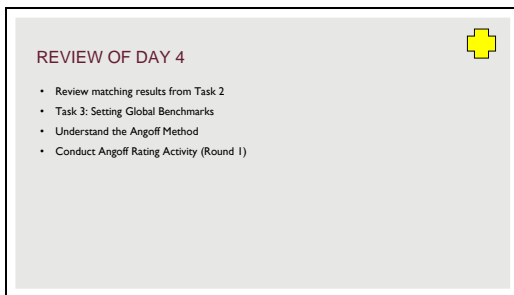


Slide 145

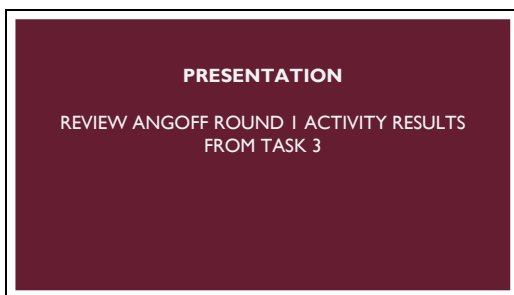
 A slide titled "DAY FIVE AGENDA" containing a table with three columns: Time, Task, Presentations, and Activities, and Facilitation.

Time	Task, Presentations, and Activities	Facilitation
09:00–10:00	Task 3 Presentation: Round 1 results	Lead facilitators
10:00–11:00	Task 3 Presentation: Discuss Round 1 results	Lead facilitators
11:00–11:15	Tea break	--
11:15–11:45	Task 3 Presentation: Angoff Round 2	All facilitators
11:45–12:30	Task 3 Activity: Angoff Round 2	All facilitators
12:30–13:30	Lunch break	--
13:30–14:30	Task 3 Activity: Angoff Round 2	All facilitators
14:30–15:00	Task 3 Activity: Workshop evaluation	All facilitators
15:00–15:15	Tea break	--
15:15–16:00	Task 3 Presentation: Round 2 results	Lead facilitators
16:00–17:00	Closing and logistics	MOE, USAID, IP, and PLI

Slide 146



Slide 147



Slide 148

ROUND 1 ITEM RATINGS AND BENCHMARKS

We will review round 1 results in a few different ways:

- Individual panelists' **initial benchmarks** and their distributions
- Differences in individual item ratings
- Location statistics** on panelists' item ratings
- Item ratings in relation to **item difficulty values** (p-values)
- Impact data** showing percentage of learners falling into each GPL based on initial benchmarks

Slide 149

ROUND 1 ITEM RATINGS AND BENCHMARKS

We will review round 1 results in a few different ways:

- Averages** of the panelists' benchmarks, i.e., the panel's initial benchmarks
- Differences in ratings** on specific items
- Impact data** with percentages of scores by GPL given the panel's benchmarks

Slide 150

ROUND 1: RESULTS USING INDIVIDUAL PANELIST BENCHMARKS

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 151

ROUND 1: RESULTS USING LOCATION STATISTICS

Location statistics for benchmarks:

Slide 152

ROUND 1: RESULTS BY ITEM

GRADE 3 RATING DISCUSSION

WHERE DID WE DISAGREE?

Solve the following questions.

Q21	Q22	Q23	Q24
$\begin{array}{r} 56 \\ +17 \\ \hline \end{array}$	$\begin{array}{r} 78 \\ -29 \\ \hline \end{array}$	$\begin{array}{r} 42 \\ \times 6 \\ \hline \end{array}$	$\begin{array}{r} 7 \overline{)93} \\ \underline{7} \\ 23 \\ \underline{21} \\ 23 \\ \underline{21} \\ 2 \\ \hline \end{array}$

ITEM 23:

- 10 J14
- 3 JE
- 2 AE

ITEM 24:

- 5 J14
- 2 JE
- 8 AE

"Meets" - Multiply and divide within 100 (i.e., up to 10 x 10 and 100 ÷ 10, no remainder)
 "Exceeds" - Multiply and divide within 144 (i.e., up to 12 x 12 and 144 ÷ 12, without a remainder)

Slide 153

ROUND 1: COMPARING RESULTS WITH ITEM DIFFICULTY

Item difficulty:

Item Number	P-Value	Item Number	P-Value
1	0.77	21	0.40
2	0.38	22	0.38
3	0.52	23	0.36
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	0.69	28	0.57
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	0.46	35	0.39
16	0.42	36	0.44
17	0.34	37	0.29
18	0.71	38	0.34
19	0.48	39	0.53
20	0.43	40	0.26

Slide 154

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%

Slide 155

TEA BREAK

Slide 156

PRESENTATION

TASK 3: ANGOFF BENCHMARKING ROUND 2

Slide 157

- 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 158

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 159

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.

Slide 160

TASK 3 ACTIVITY

CONDUCT ANGOFF BENCHMARKING ROUND 2

Slide 161

LUNCH BREAK

Slide 162

WORKSHOP EVALUATION

Slide 163

WORKSHOP EVALUATION INSTRUCTIONS


- You will now complete an **evaluation form** to share your opinions about the following aspects of the workshop:
 - Orientation and training** (guidance on setting benchmarks, practice with the method, interpretation of feedback information, adequacy of training time)
 - Round 1 ratings** (confidence, comfort, and time allocation)
 - Round 2 ratings** (confidence, comfort, and time allocation)
 - Benchmarks** (calculations, feedback, and discussion)
 - Workshop Overall** (organization, facilitation, and time allocation)

Slide 164

PRESENTATION

REVIEW ANGOFF ROUND 2 RESULTS

Slide 165

FINAL RESULTS AND SHIFT BETWEEN ROUNDS 

Impact data:

Minimum Proficiency Levels	ROUND 1			ROUND 2		
	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%

Slide 166

TEA BREAK

Slide 167

NEXT STEPS

Slide 168


USE OF WORKSHOP RESULTS

- Enable **three types of analyses (CAT)** with the global benchmarks:
 - **Compare** assessment results across contexts/languages within the country and with outcomes from other countries.
 - **Aggregate** assessment results across different assessments in the country and with those of other countries.
 - **Track** assessment results over time to monitor progress.
- Understand which learners most need support in the country.
- Could inform a study into **why gaps** in learning exist and how best to address those.
- How will you use the results?

Slide 169

DISCUSSION

- What do you think of the results?
- What, if anything, did you learn from this process?
- Has this informed your thinking about what learners in grade [X] should be able to accomplish? In what way(s)?



Slide 170

CLOSING

Slide 171

THANK YOU

ANNEX F—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 20-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 14: 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

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 15: 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
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 16: 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
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	dafi	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 17: 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? <i>{Kande da Delu}</i>	JP	JM	JE	AE	JP	JM	JE	AE
2	≤ 18 words attempted	Ina suke tafiya kullum? <i>{Makaranta}</i>	JP	JM	JE	AE	JP	JM	JE	AE
3	≤ 22 words attempted	Me Kande ta zo da shi daga makaranta? <i>{Aiki}</i>	JP	JM	JE	AE	JP	JM	JE	AE
4	≤ 28 words attempted	Wa ya taimaka wa Kande? <i>{Delu}</i>	JP	JM	JE	AE	JP	JM	JE	AE
5	≤ 35 words attempted	Me ya faru a ajin su Kande? <i>{Kande ta Samu yabo/ yabo}</i>	JP	JM	JE	AE	JP	JM	JE	AE
Total										

ANNEX G—INTRA- AND INTER-RATER CONSISTENCY AND STANDARD ERROR (SE)

INTRA-RATER CONSISTENCY

Chang's (1999) intra-rater consistency index was created for the traditional Angoff method (panelists estimate probability of giving correct response by minimally proficient learners to the item, not a yes-no decision). It is calculated as:

$$d_j = 1 - \frac{1}{n} \sum_i^n |P_{ij} - P_{ie}| \quad (1)$$

Where,

d_j = Intra-rater consistency for panelist j across all items on the test; a higher number indicates high consistency and a lower number means low consistency

P_{ij} = Panelist j item performance estimate (i.e., probability of correct response to the item i by minimally proficient learners)

P_{ie} = Empirical p-value (item difficulty level) for item i

n = Number of items

For a yes-no variation of Angoff method for multiple benchmarks, we have extended Chang's formula for four performance levels. The intra-rater consistency for each judge j is,

$$d_j = 1 - \frac{1}{n} \sum_i^n |P_{ijk} - P_{ie}| \quad (2)$$

Where,

d_j = Intra-rater consistency for panelist j across all items on the test; the lower number indicates high consistency and higher number means low consistency

P_{ijk} = Panelist j item performance estimate (i.e., panelist gave a yes rating to the k^{th} category for item i); $k=1$ (partially meets), $k=2$ (meets), $k=3$ (exceeds minimum proficiency), and $k=4$ (above exceeds minimum proficiency) (**note:** when setting one benchmark, use $k=1$ (meets) and $k=2$ (above meets))

P_{ij1} = If panelist j gave a yes rating to partially meets category ($k=1$) for item i then it is calculated as conditional item difficulty level for learners who obtain 0–25 percent of observed scores on the subtask or the entire test (**note:** for workshops focused on setting just one benchmark, this variable should be for if panelist j gave a yes rating to the meets category ($k=1$); if so, it is the conditional item difficulty level for learners who obtained 0–50 percent of observed scores on the subtask or the entire test)

P_{ij2} = If panelist j gave a yes rating to meets category ($k=2$) for item i then it is calculated as conditional item difficulty level for learners who obtain 26–50 percent scores on the subtask or the entire test (**note:** for workshops focused on setting just one benchmark, this variable should be for if panelist j gave a yes rating to the above meets category ($k=2$); if so, it is the conditional item difficulty level for learners who obtained 51–100 percent of observed scores on the subtask or the entire test)

P_{ij3} = If panelist j gave a yes rating to exceeds category ($k=3$) for item i then it is calculated as conditional item difficulty level for learners who obtain 51–75 percent scores on the subtask or the entire test (**note:** this variable is not relevant for workshops focused on setting just one benchmark)

P_{ijk} = If panelist j gave a yes rating to above exceeds category ($k=4$) for item i then it is calculated as conditional item difficulty level for learners who obtain 76–100 percent scores on the subtask or the entire test (**note**: this variable is not relevant for workshops focused on setting just one benchmark)

P_{ie} = Empirical item difficulty level for item i

n = Number of items

Overall, intra-rater consistency for the entire panel is calculated by taking average of d_j for m number of panelists.

$$d = \frac{1}{m} \sum_j^m d_j \quad (3)$$

How to Calculate Intra-Rater Consistency

Step 1: Before the policy linking workshop, calculate empirical item difficulty level (P_{ie}) and conditional item difficulty levels (P_{ijk}) for learners with 0–25 percent, 26–50 percent, 51–75 percent, and 76–100 percent of observed scores on a given subtask (individually administered) or on an entire test (group administered).

- i. Calculate empirical item difficulty level for each item by taking proportion of learners who get the item right.
- ii. Calculate raw score for each learner by taking the sum of correct responses to the items.
- iii. Divide maximum observed score by four when setting three benchmarks for four performance levels (or by two when setting one benchmark for two performance levels, e.g., one benchmark for Meets Global Minimum Proficiency) to calculate score ranges for four (or two) categories (0–25 percent for partially meets, 26–50 percent for meets, 51–75 percent for exceeds, and 76–100 percent for above exceeds for four performance levels or 0–50 percent for meets and 51–100 percent for above meets for two performance levels).
- iv. Sort observed scores in ascending order and split learner item response data file into four groups (or two, as described above) by including learners with 0–25 percent scores for partially meets, 26–50 percent for meets, 51–75 percent for exceeds, and 76–100 percent for above exceeds (or 0–50 percent for meets and 51–100 percent for above meets).
- v. For each partially meets, meets, exceeds, and above exceeds group when setting three benchmarks (or for just meets and above meets when setting one benchmark), calculate conditional item difficulty level (P_{ijk}) for each item by calculating the proportion of learners who get the item right.

Step 2: During the policy linking workshop, calculate absolute value $|P_{ijk} - P_{ie}|$ and its sum across the items d_j for each panelist.

- i. For each item, calculate absolute value by taking conditional item difficulty level for panelist's item performance rating (partially meets, meets, exceeds, and above exceeds, or just the two levels if only setting one benchmark) minus the empirical item difficulty level.
- ii. Calculate sum of the absolute values across the items on the subtask or the test.
- iii. Divide the sum by number of items on the subtask or the test to calculate average absolute difference of the panelist.
- iv. Subtract average absolute difference from one to calculate intra-rater consistency of the panelist.

Step 3: Calculate intra-rater consistency for the entire panel (including all the panelists).

- i. Calculate sum of the intra-rater consistencies across the panelists.
- ii. Divide the sum by total number of panelists to calculate an average intra-rater consistency for the panel.

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,

$$I_i = 1 - \frac{TOI_i}{TI} \quad (4)$$

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

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

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:

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

How to Calculate Inter-Rater Consistency

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

Step I: 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:

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

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

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

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.

ANNEX H—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.

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 I—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 J—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 K—TEMPLATE FOR PANELIST PRE-WORKSHOP ACTIVITY EXPLANATION

This template will need to be edited to fill in all of the bracketed items/information and may also need to be edited if the workshop seeks to only set one benchmark, in which case, facilitators need only include definitions for the “meets” expectation in the below description. Also, you will only want to send panelists information on the grade/subject for which they personally will be setting benchmarks. This means if you are setting benchmarks for reading in grades two and three, for example, you will want to only include information on grade two reading GPDs for grade two panelists. Finally, you will need to attach the assessment or mini-assessment to this description.

Pre-workshop Activity for Teachers: [Activity Name]

Instructions

Each teacher should administer the attached [assessment/mini-assessment] to selected learners in their [Grade, Subject] classrooms.

- 1) Print or write out the assessment. It has [X number of items]. If printed, make sure that the characters/items are large enough for learners to read/see. If written, it needs to be written in the style you use as a teacher.
- 2) Select nine (9) learners, i.e., three (3) who meet the definition provided below for Partially Meets Global Minimum Proficiency, three (3) who meet the definition for Meets Global Minimum Proficiency, and three (3) who meet the definition for Exceeds Global Minimum Proficiency, in your classroom. Write down their names on a separate piece of paper.
 - **[Grade X]:**
 - **Partially Meets Global Minimum Proficiency Learner:** [Include definitions from the GPF to describe the overarching characteristics of learners who fall into this category at the relevant grade in the relevant subject.]
 - **Meets Global Minimum Proficiency Learner:** [See above]
 - **Exceeds Global Minimum Proficiency Learner:** [See above]
- 3) One-on-one, have each learner complete the assessment. [Include additional instructions and steps here following the enumerator/assessment administrator guidance provided for the actual assessment, e.g., should the teachers read the questions/passage out loud or let the learner read it/them quietly or out loud; should the teacher time the learner and ask them to stop after a certain period of time; how should the teacher record learner responses]
- 4) Make sure to mark each assessment with the learner name and the performance level (Partially Meets Global Minimum Proficiency, Meets Global Minimum Proficiency, or Exceeds Global Minimum Proficiency)
- 5) Record which items the learners got correct and which they got wrong.
- 6) Write down the number of questions the learner answered correctly.
- 7) Proceed to the next child.

Should you have any questions about how to administer the attached [assessment/mini-assessment] or to whom it should be administered, please feel free to call or email [Name] at [Phone/Email].

ANNEX L—PRE-WORKSHOP STATISTICS

The data analyst and/or lead facilitator should calculate the following statistics before the policy linking workshop:

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 and the conditional item difficulty levels for learners with 0–25 percent, 26–50 percent, 51–75 percent, and 76–100 percent scores (for workshops where three benchmarks will be set) or for learners with 0–50 percent and 51–100 percent (for workshops where one benchmark will be set) on a given subtask (individually administered) or on an entire test (group administered) using the following steps:

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.
2. Calculate the raw score for each learner by taking the sum of the correct responses to the items.
3. Divide maximum possible score by two or four to calculate score ranges for two or four performance levels, respectively (0–25 percent for partially meets, 26–50 percent for meets, 51–75 percent for exceeds, and 76–100 percent for above exceeds, or 0–50 percent for meets, and 51–100 percent for exceeds).
4. Sort raw scores in ascending order and split the learner item response data file into two or four groups depending on how many benchmarks you will be setting in the workshop. For one benchmark, you will include learners with 0–50 percent scores for meets, and 51–100 percent scores for above meets. For three benchmarks, you will include learners with 0–25 percent scores for partially meets, 26–50 percent for meets, 51–75 percent for exceeds, and 76–100 percent for above exceeds.
5. For each partially meets, meets, exceeds, and above exceeds group, calculate conditional item difficulty level for each item by calculating the proportion of learners who get the item right. This information (conditional item difficulty) will be needed for the post-workshop analysis on intra-rater consistency and inter-rater reliability.

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.

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 those presented in the first and second examples below. Note that for timed assessments, like the EGRA/EGMA, the data analyst will need to create a table on the number of attempted words/items as well, as shown in the third table below.

Table 18: Example Data Distribution Table for Oral Reading Passage

Grade 3 Hausa Oral Reading Fluency—Number of Words Learners Read Correctly in a Minute			
Read Words Correctly	Frequency	Percent	Cumulative Percent
0	649	54.2	54.2
1	14	1.2	55.4
2	11	0.9	56.3
3	8	0.7	56.9
4	13	1.1	58.1
5	13	1.1	59.1
6	11	0.9	60.0

Grade 3 Hausa Oral Reading Fluency—Number of Words Learners <u>Read Correctly</u> in a Minute			
Read Words Correctly	Frequency	Percent	Cumulative Percent
7	11	0.9	61.0
8	8	0.6	61.6
9	15	1.2	62.8
10	4	0.4	63.2
11	10	0.9	64.1
12	14	1.1	65.2
13	10	0.9	66.1
14	17	1.5	67.5
15	11	0.9	68.5
16	9	0.7	69.2
17	18	1.5	70.7
18	10	0.8	71.6
19	10	0.9	72.4
20	10	0.8	73.3
21	15	1.2	74.5
22	16	1.3	75.8
23	14	1.1	77.0
24	5	0.4	77.4
25	11	1.0	78.3
26	10	0.8	79.2
27	7	0.6	79.7
28	15	1.2	80.9
29	10	0.8	81.8
30	10	0.8	82.6
31	14	1.1	83.8
32	15	1.2	85.0
33	18	1.5	86.5
34	52	4.3	90.8
35	110	9.2	100.0
Total	1198	100.0	

Table 19: Example Data Distribution Table for a Reading Comprehension Subtask

Grade 3 Hausa Reading Comprehension—Number of Items Learners <u>Answered Correctly</u>			
Score	Frequency	Percent	Cumulative Percent
0	773	64.5	64.5
1	84	7.0	71.5
2	85	7.1	78.6
3	71	5.9	84.5
4	114	9.5	94.0
5	72	6.0	100.0
Total	1198	100.0	

Table 20: Example Data Distribution Table for Timed Reading Assessment

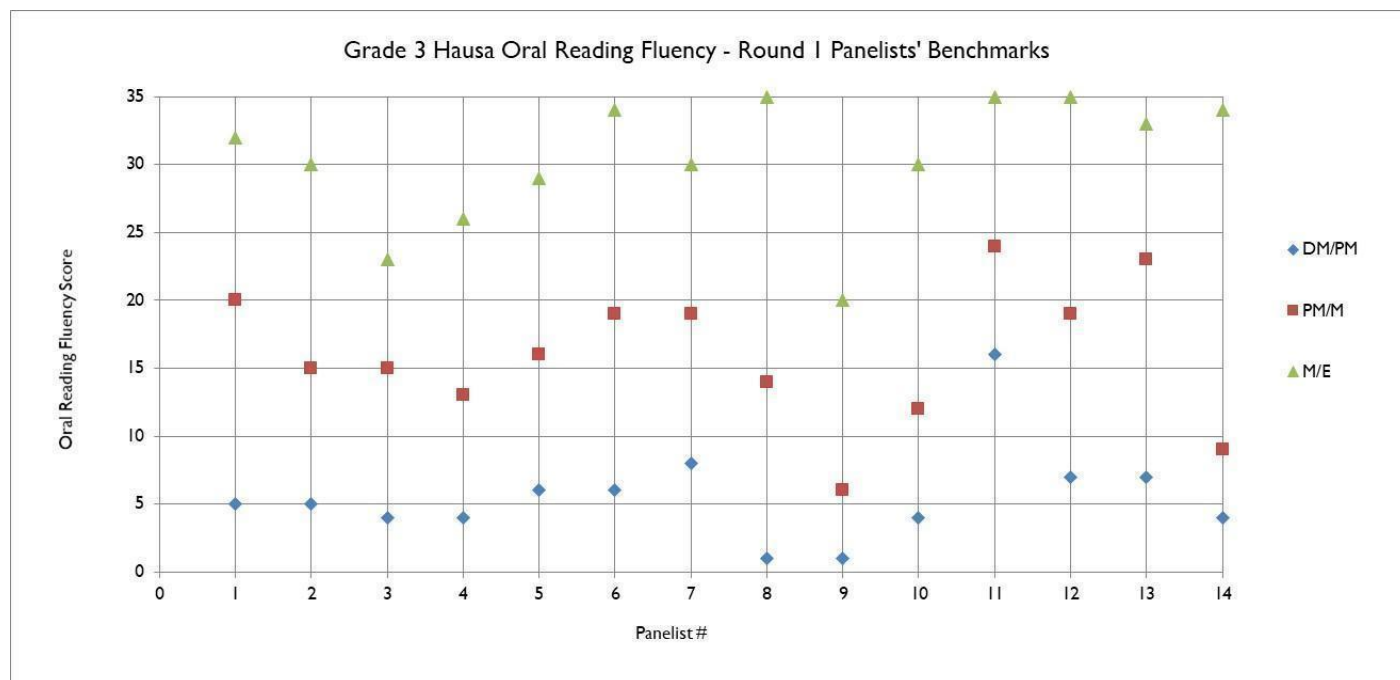
Grade 3 Hausa Oral Reading Fluency—Number of Words Learners Attempted to Read in a Minute			
Attempted Words	Frequency	Percent	Cumulative Percent
7	1	0.1	0.1
8	2	0.2	0.2
9	656	54.8	55.0
10	6	0.5	55.6
11	4	0.3	55.9
12	15	1.2	57.1
13	7	0.6	57.7
14	7	0.6	58.2
15	14	1.2	59.4
16	13	1.1	60.5
17	13	1.1	61.6
18	21	1.8	63.4
19	19	1.6	64.9
20	12	1.0	66.0
21	29	2.4	68.4
22	43	3.6	71.9
23	10	0.8	72.8
24	6	0.5	73.3
25	12	1.0	74.3
26	16	1.4	75.7
27	5	0.4	76.0
28	26	2.1	78.2
29	13	1.1	79.2
30	4	0.3	79.6
31	7	0.6	80.2
33	9	0.7	80.9
34	5	0.4	81.3
35	13	1.1	82.4
Total	211	17.6	100.0
	1198	100.0	

ANNEX M—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.

Figure 38: Example Impact Data Table

Grade 3 Hausa Oral Reading Fluency and Comprehension—Benchmarks			
Categories	Benchmark	Score Range	% of Learners
Below partially meets	N/A	0–6	59.8
Partially meets	7	7–18	10.1
Meets	19	19–34	13.3
Exceeds	35	35–40	16.8
Total			100.0

ANNEX N—SAMPLE AGENDA FOR AN IN-PERSON WORKSHOP

Adaptation Instructions—The project team will need to update the agenda to adjust start/end times, break timing, etc., according to the needs of the country/assessment agency.

Table 21: Sample Agenda for In-Person Workshop

Time	Day 1	Facilitation
08:30–09:00	Registration	Project team
09:00–10:00	Opening, introductions, agenda, and logistics	Government/ assessment agency, donors, and implementing partners (if relevant) as well as lead facilitators
10:00–11:00	Presentation: Background, objective, and proficiency level overview	Lead facilitators
11:00–11:15	Tea break	--
11:15–13:00	Presentation: Overview of the GPF and review of the GPDs	All facilitators
13:00–14:00	Lunch break	--
14:00–14:30	Remaining questions on the GPF	All facilitators
14:30–15:15	Presentation: Overview of the assessment(s)	Content facilitators
15:15–15:30	Tea break	--
15:30–16:30	Presentation: Overview of the assessment(s) continued	Content facilitators
16:30–17:00	Day 1 closing and preview of Day 2	Lead facilitators
Time	Day 2	Facilitation
09:00–09:30	Welcome and review	Lead facilitators
09:30–11:00	Task 1 Presentation: GPF and alignment	Lead facilitators
11:00–11:15	Tea break	--
11:15–12:30	Task 1 Presentation: GPF and alignment continued	Lead facilitators
12:30–13:30	Lunch break	--
13:30–15:15	Task 1 Activity: Alignment of assessment(s) and the GPF	All facilitators
15:15–15:30	Tea break	--
15:30–16:30	Task 1 Activity: Alignment of assessments and the GPF (cont.)	All facilitators
16:30–17:00	Day 2 closing and preview of Day 3	
Time	Day 3	Facilitation
09:00–10:00	Task 1 Presentation: Alignment results	Lead facilitators
10:00–11:00	Task 2 Presentation: Matching assessments and GPDs/GPLs	Lead facilitators
11:00–11:15	Tea break	--
11:15–12:30	Task 2 Activity: Matching assessment items and GPDs/GPLs	All facilitators
12:30–13:30	Lunch break	--
13:30–15:45	Task 2 Activity: Matching assessment items and GPDs/GPLs (cont.)	All facilitators
15:45–16:00	Tea break	--
16:00–17:00	Task 2 Activity: Matching assessment items and GPDs/GPLs (cont.)	All facilitators
Time	Day 4	Facilitation
09:00–10:00	Task 2 Presentation: Matching results	Lead facilitators
10:00–11:00	Task 3 Presentation: Global benchmarking	Lead facilitators
11:00–11:15	Tea break	--
11:15–12:30	Task 3 Presentation: Angoff method	Lead facilitators
12:30–13:30	Lunch break	--
13:30–15:00	Task 3 Activity: Angoff practice	All facilitators
15:00–15:15	Tea break	--
15:15–17:00	Task 3 Activity: Angoff Round 1	All facilitators

Time	Day 1	Facilitation
Time	Day 5	Facilitation
09:00–11:00	Task 3 Presentation: Round 1 results	Lead facilitators
11:00–11:15	Tea break	--
11:15–12:30	Task 3 Activity: Angoff Round 2	All facilitators
12:30–13:30	Lunch break	--
13:30–15:00	Task 3 Activity: Workshop evaluation	All facilitators
15:00–15:45	Task 3 Presentation: Round 2 results	Lead facilitators
15:45–16:00	Tea break	--
16:00–17:00	Closing and logistics	Ministry, USAID (if applicable)

ANNEX O—SAMPLE AGENDA FOR A REMOTE WORKSHOP

Adaptation Instructions—The project team will need to update the agenda to fill in any items in brackets and to adjust comfort break timing, etc., according to the needs of the country/assessment agency. They will also want to establish the actual start times for each of the activities.

Table 22: Example Agenda for Remote Preparation Session 1

(Recommend holding two weeks before the workshop)

Timing	Activity	Facilitator
0–15 minutes	Welcome and introductions	Lead facilitator
15–40 minutes	Overview of policy linking	Lead facilitator
40–55 minutes	Purpose of preparation session	Process facilitator
55–60 minutes	Comfort break	
60–80 minutes	Overview of the GPF	Lead or content facilitator
80–100 minutes	[Grade and Subject] GPF Review	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 23: Example Agenda for Remote Preparation Session 2

(Recommend holding two days after the first preparatory session)

Timing	Activity	Facilitator
0–15 minutes	Welcome and purpose of the preparation session	Lead facilitator
15–30 minutes	Overview of the [assessment name]	Content or lead facilitator
30–55 minutes	Review each item on the [assessment]	Content or lead facilitator
55–60 minutes	Comfort break	
60–100 minutes	Continue reviewing items and discuss [assessment] administration	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 24: Example Agenda for Remote Workshop Session 1

Timing	Activity	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	Lead facilitator

Table 25: Example Agenda for remote Workshop Session 2

Timing	Activity	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)	Lead facilitator
55–60 minutes	Comfort break	
60–90 minutes	Small group discussions on first 5 items ²¹	Content facilitators ^{l21}
90–110 minutes	Plenary discussion on questions that came up in the groups	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 26: Example Agenda for Remote Workshop Session 3

Timing	Activity	Facilitator
0–10 minutes	Welcome and purpose of session 3	Lead facilitator
10–40 minutes	Review inter-session activities and provide clarification	Content facilitator
40–55 minutes	Introduction to Task 2—Matching to GPLs and GPDs	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) ²²	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 27: Example Agenda for Remote Workshop Session 4

Timing	Activity	Facilitator
0–10 minutes	Welcome and purpose of session 4	Lead facilitator
10–40 minutes	Present Angoff methodology and Task 4 and provide clarification	Lead facilitator
40–75 minutes	Small group Angoff ratings using practice items	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)	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)

²¹Each 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.

²² 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 28: Example Agenda for Remote Workshop Session 5

Timing	Activity	Facilitator
0–10 minutes	Welcome and purpose of session 5	Lead facilitator
10–45 minutes	Review and discuss Round 1 ratings in plenary	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	Content facilitator
110–150 minutes	Share and discuss item difficulty and impact data	Lead facilitator
150–180 minutes	Explanation of inter-session activities (reminder of methodology) 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)
- 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 29: Example Agenda for Remote Workshop Session 6

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

ANNEX P—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 30: 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 31: 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 32: 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 33: 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 34: 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 35: 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 Q—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 E**. 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. 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. Below, you will find some additional slides that can be added to the regular workshop slide decks to better orient content facilitators to the key documents, the policy linking process, and their role in the workshop.

ANNEX R—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.

$$PM_j = \sum_{i=1}^l JP_{ij} \quad (1)$$

Equation 2 shows the Meets Minimum Proficiency benchmark for one panelist after Round 1.

$$M_j = PM_j + \sum_{i=l+1}^m JM_{ij} \quad (2)$$

Equation 3 shows the Exceeds Minimum Proficiency benchmark for one panelist after Round 1.

$$E_j = M_j + \sum_{i=m+1}^n JE_{ij} \quad (3)$$

Equation 4 is the Partially Meets Minimum Proficiency benchmark for all panelists after Round 1.

$$P = \frac{1}{z} \sum_{j=1}^z \sum_{i=1}^l PM_{ij}$$

Equation 5 is the Meets Minimum Proficiency benchmark for all panelists after Round 1.

$$M = \frac{1}{z} \sum_{j=1}^z (PM_j + \sum_{i=l+1}^m M_{ij}) \quad (5)$$

Equation 6 is the Exceeds Minimum Proficiency benchmark for all panelists after Round 1.

$$E = \frac{1}{z} \sum_{j=1}^z (M_j + \sum_{i=m+1}^n E_{ij}) \quad (6)$$

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. PM_j for that one panelist is the sum from sub-step 2
- iv. Repeat sub-steps 1 and 2 for each panelist to calculate PM_j for each one

Step 2: Calculate the Meets Minimum Proficiency score (M_j) 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. M_j for that one panelist is the sum from sub-step 2.
- iv. Repeat sub-steps 1 and 2 for each panelist to calculate M_j for each one.

Step 3: Calculate the Exceeds Minimum Proficiency score (E_j) 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. E_j for that one panelist is the sum from sub-step 2.
- iv. Repeat sub-steps 1 and 2 for each panelist to calculate E_j 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 M_j cut scores from the panelists.
- ii. Divide the sum of M_j 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 E_j cut scores from the panelists.
- ii. Divide the sum of E_j cut scores and divide by the total number of panelists.
- iii. This result is a simple average equivalent to E .

ANNEX S—CERTIFICATE OF APPRECIATION TEMPLATE



CERTIFICATE OF APPRECIATION

Presented to:

for serving as a panelist in a
Policy Linking for Measuring Global Learning Outcomes Workshop

LEAD FACILITATOR

DATE



ANNEX T—OUTLINE FOR THE POLICY LINKING TECHNICAL 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. 4.1.1 Review Panel Results
 - a. Criterion 1: Alignment between curriculum, assessment, and GPF
 - b. Criterion 2: Appropriateness of assessment
 - c. Criterion 3: Assessment reliability
4. Standard Setting Methodology
 - a. Selection and Description of Panelists
 - b. Standard Setting Method
 - c. Procedure
 - i. Preparation for the Standard Setting Workshop
 - ii. Conducting Standard Setting Workshop
 - iii. Finalizing the Performance Standards
 - d. Analysis of Round 1 and 2 Ratings
5. Standard Setting Results
 - a. Round 1 Results
 - b. Feedback Data
 - c. Round 2 Results
6. Evaluation of Standard Setting Process
 - a. Procedural Evaluation (Round 1 and 2)
 - b. Internal Evaluation Standard Error of Mean (Round 1 and 2), Inter- and Intra-Panelist Consistency (Round 2), and Agreement and Consistency Coefficients (Round 2)
7. Summary of Results of Criterion 4 for the 4.1.1 Review Panel
8. Conclusions and Recommendations
9. References
10. Annexes
 - a. Method Selection Checklist
 - b. Rating Form
 - c. Evaluation Form
 - d. Frequency Distribution of Learner Test Score
 - e. Difficulty Level of the Test Items
 - f. Other Relevant Documents and Data

ANNEX U—4.1.1 CRITERIA FOR POLICY LINKING WORKSHOP VALIDITY

Table 36: Criterion 4 for Policy Linking Validity

Question	Criteria	Materials
4a)* What was the <i>intra-rater reliability</i> for the second round of ratings?	The <i>intra-rater reliability</i> will vary depending on the number of items on the assessment. The panel will provide guidance on how they determined acceptability.	Countries should provide statistics on intra-rater reliability as well as data that include the scores of each of the raters for both rounds of ratings. Each rater should be assigned a rater number so that their scores can be identified across rounds.
4b)* What was the <i>inter-rater reliability</i> for the second round of ratings?	The <i>inter-rater reliability</i> should be at least .80.	Countries should provide statistics on inter-rater reliability and the scores of each of the raters for both rounds of ratings.
4c)* What was the standard error (SE) for the benchmark at each global proficiency level?	<i>SE</i> should be appropriate for each <i>global proficiency level</i> reported. There is no maximum <i>SE</i> provided in this document, since it will depend on the number of items in the assessment.	Countries should provide the <i>SE</i> and details of how the <i>SE</i> was calculated (either using classical test theory or item response theory) and an explanation of why they believe this to be appropriate given the test features.
4d)* To what extent were the panelists representative of the target population of schools being reported on?	Panelists should be selected to ensure: <ul style="list-style-type: none"> • Gender representation: The panelists must be selected to ensure gender balance, both for the teachers and non-teachers. • Geographical representation: The teachers (and non-teachers, if possible) must be selected to ensure representation from regions, provinces, and/or states. • Ethnic and/or linguistic representation (where applicable): The panel must have diversity that reflects the population; there must be native speakers of assessment languages, as well as classroom teachers who understand learning in second or third languages. • Representation of crisis- and conflict-affected areas. 	Countries should provide an explanation of what criteria they used to select panelists as well as demographic details about each of the panelists and how they meet the requirements listed for this criterion.
4e)* To what extent did the panelists meet the other selection criteria described in the Policy Linking Toolkit?	Panelists should all have: <ul style="list-style-type: none"> • Several years of teaching experience in the grade level for which they are providing ratings (classroom teachers) • Skills in the subject area (all panelists) • Skills in the different languages of instruction and assessment (all panelists) • Knowledge of learners of different proficiency levels, including at least some who would meet the requirements of the Meets Global Minimum Proficiency level and some who would meet the requirements of the Exceeds Global Minimum Proficiency level (all panelists) • Knowledge of the instructional environment (all panelists) • Experience administering the assessment(s) being used for the policy linking workshop 	Countries should provide demographic details about each of the panelists and how they meet the requirements listed under this criterion. Panelists should fill out workshop evaluation forms that include questions about their exposure to the assessment ahead of the workshop and during the workshop, assess their knowledge of the instructional environment, etc.
4f)* To what extent did panelists report understanding the <i>GPE</i> , assessment, and policy linking methodology? And, to what extent did they feel comfortable with their Round 2 evaluations and final <i>benchmarks</i> ?	On a five-point Likert scale, with 1 being strongly disagree, very uncomfortable, etc. and 5 being strongly agree, very comfortable, etc., the average rating for each of these criteria should be 4 or above.	Countries should share all panelist evaluation forms as well as a database of their Likert scale responses and average scores for each of the categories listed in this question.

ANNEX V—AGREEMENT AND CONSISTENCY COEFFICIENTS

Following the workshop, the data analyst should conduct reliability analysis using Subkoviak’s method. Subkoviak’s method estimates an agreement coefficient and a consistency coefficient using a reliability estimate for the total test scores and absolute value of Z.

$$Z = (\text{benchmark for the test} - \text{mean observed test score} - 0.5) / \text{Standard deviation of observed test score}$$

Absolute values of Z are used to obtain the estimates of the agreement coefficient and consistency coefficient from lookup tables.

Suppose an assessment of 50 items was administered to a sample of learners, that the sample mean and standard deviation were 35.5 and 7.0 respectively, that a benchmark of 30 was used to make meeting or not meeting global minimum proficiency decisions, and total score reliability was 0.80. In this case, the calculated value of Z is $[(30 - 35.5 - 0.5)/7] = -0.86$. Using **Table 37**, the agreement coefficient is found by locating the intersection of the row containing the absolute value of Z (0.86) and the column containing the reliability of 0.80. The agreement coefficient in this case is 0.86 (between 0.85 and 0.87), indicating that a high proportion of consistency decisions would be expected. When reliability statistics and Z scores are not in increments of .10, the data analyst should round to the nearest .10; so, a reliability statistic of .73 would become .70 and a Z score of .86 would become .90, for instance.

Table 37: Approximate Value of Agreement Coefficient using Absolute Value and Reliability Coefficient

z	r								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.00	0.53	0.56	0.60	0.63	0.67	0.70	0.75	0.80	0.86
0.10	0.53	0.57	0.60	0.63	0.67	0.71	0.75	0.80	0.86
0.20	0.54	0.57	0.61	0.64	0.67	0.71	0.75	0.80	0.86
0.30	0.56	0.59	0.62	0.65	0.68	0.72	0.76	0.80	0.86
0.40	0.58	0.60	0.63	0.66	0.69	0.73	0.77	0.81	0.87
0.50	0.60	0.62	0.65	0.68	0.71	0.74	0.78	0.82	0.87
0.60	0.62	0.65	0.67	0.70	0.73	0.76	0.79	0.83	0.88
0.70	0.65	0.67	0.70	0.72	0.75	0.77	0.80	0.84	0.89
0.80	0.68	0.70	0.72	0.74	0.77	0.79	0.82	0.85	0.90
0.90	0.71	0.73	0.75	0.77	0.79	0.81	0.84	0.87	0.90
1.00	0.75	0.76	0.77	0.77	0.81	0.83	0.85	0.88	0.91
1.10	0.78	0.79	0.80	0.81	0.83	0.85	0.87	0.89	0.92
1.20	0.80	0.81	0.82	0.84	0.85	0.86	0.88	0.90	0.93
1.30	0.83	0.84	0.85	0.86	0.87	0.88	0.90	0.91	0.94
1.40	0.86	0.86	0.87	0.88	0.89	0.90	0.91	0.93	0.95
1.50	0.88	0.88	0.89	0.90	0.90	0.91	0.92	0.94	0.95
1.60	0.90	0.90	0.91	0.91	0.92	0.93	0.93	0.95	0.96
1.70	0.92	0.92	0.92	0.93	0.93	0.94	0.95	0.95	0.97
1.80	0.93	0.93	0.94	0.94	0.94	0.95	0.95	0.96	0.97
1.90	0.95	0.95	0.95	0.95	0.95	0.96	0.96	0.97	0.98
2.00	0.96	0.96	0.96	0.96	0.96	0.97	0.97	0.97	0.98

Source: Subkoviak, 1988; Brown, 1989.

The corrected decision consistency coefficient agreement is found by locating the intersection of the same value of Z and test reliability coefficient. **Table 38** reveals that the consistency coefficient is 0.56 out of a possible 0.71, indicating the assessment procedure is adding only modestly to consistency in decision making.

Table 38: Approximate Value of Consistency Coefficient using Absolute Value and Reliability Coefficient

z	r								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.00	0.06	0.13	0.19	0.26	0.33	0.41	0.49	0.59	0.71
0.10	0.06	0.13	0.19	0.26	0.33	0.41	0.49	0.59	0.71
0.20	0.06	0.13	0.19	0.26	0.33	0.41	0.49	0.59	0.71
0.30	0.06	0.12	0.19	0.26	0.33	0.40	0.49	0.59	0.71
0.40	0.06	0.12	0.19	0.25	0.32	0.40	0.48	0.58	0.71
0.50	0.06	0.12	0.18	0.25	0.32	0.40	0.48	0.58	0.70
0.60	0.06	0.12	0.18	0.24	0.31	0.39	0.47	0.57	0.70
0.70	0.05	0.11	0.17	0.24	0.31	0.38	0.47	0.57	0.70
0.80	0.05	0.11	0.17	0.23	0.30	0.37	0.46	0.56	0.69
0.90	0.05	0.10	0.16	0.22	0.29	0.36	0.45	0.55	0.68
1.00	0.05	0.10	0.15	0.21	0.28	0.35	0.44	0.54	0.68
1.10	0.04	0.09	0.14	0.20	0.27	0.34	0.43	0.53	0.67
1.20	0.04	0.08	0.14	0.19	0.26	0.33	0.42	0.52	0.66
1.30	0.04	0.08	0.13	0.18	0.25	0.32	0.41	0.51	0.65
1.40	0.03	0.07	0.12	0.17	0.23	0.31	0.39	0.50	0.64
1.50	0.03	0.07	0.11	0.16	0.22	0.29	0.38	0.49	0.63
1.60	0.03	0.06	0.10	0.15	0.21	0.28	0.37	0.47	0.62
1.70	0.02	0.05	0.09	0.14	0.20	0.27	0.35	0.46	0.61
1.80	0.02	0.05	0.08	0.13	0.18	0.25	0.34	0.45	0.60
1.90	0.02	0.04	0.08	0.12	0.17	0.24	0.32	0.43	0.59
2.00	0.02	0.04	0.07	0.11	0.16	0.22	0.31	0.42	0.50

Source: Subkoviak, 1988; Brown, 1989.

ANNEX W—POLICY LINKING PROCESS DOCUMENTATION FORM

INSTRUCTIONS

Thank you for helping the UNESCO Institute of Statistics (UIS) and the United States Agency for International Development (USAID) document the policy linking for measuring global learning outcomes process. This documentation will help us refine the methodology, the Policy Linking Toolkit, the Criteria for Policy Linking Validity, and the Global Proficiency Framework, and keep track of lessons learned along the way. We ask that you use this form to document your experience with this process, including any challenges you face, ways in which you addressed those challenges, and any implications this may have for the process in other countries or with other assessments and/or revisions that should be made to the toolkit.

Please complete each stage tab as it is being completed.

During the workshop implementation (Stage 4) please attempt to complete the short questionnaire for each day at the end of that day, while the information is still fresh in your mind. Please provide as many details as you can.

A. COVER SHEET

1. Country(ies)		
2. Grades, Subjects, AND Languages Selected for Policy Linking		
Grade(s)	Subject(s)	Language(s) of Assessment
3. Policy Linking Implementation Team		
Stage	Stage lead name and organization	Contact information for the lead
Stage 1		
Stage 2		
Stage 3 (Please provide information on the person who submitted and will receive feedback from the 4.1.1 Review Panel)		
Stage 4		
Stage 5		
Stage 6		
Stage 7 (Please provide information on the person who submitted and will receive feedback from the 4.1.1 Review Panel)		

STAGE 1: INITIAL COUNTRY ENGAGEMENT

This stage involves the initial engagement with the relevant ministry around the objectives, key stakeholders, and planning of the policy linking work. These discussions include decisions around which grade levels, subjects, and languages will be covered as well as roles and responsibilities, budget, and gathering of necessary data, assessment instruments, and details on the assessment methodology and curriculum framework that will be used. If the policy linking work will be completed entirely by the country government, without support from partners, this stage is still relevant, but it mostly involves making decisions about the assessments, grade levels, and languages of assessment that will be linked to SDG 4.1.1 as well as how many benchmarks the government wishes to set.

<p>Country Engagement Description (Describe the process that was followed, who reached out to whom and why, and how communication occurred (e.g., sharing overview documents, webinar, meeting).)</p>	
<p>Did the government decide to engage external partners/facilitators? Why or why not?</p>	
<p>Country engagement successes/high points (Describe what went well)</p>	
<p>Challenges faced during country engagement (Describe any challenges that emerged)</p>	
<p>Strategies used to address challenges (Explain any strategies that were used to address these challenges as well as any materials or training that may be helpful in the future to mitigate such challenges)</p>	
<p>Implications for policy linking process/toolkit/training, the GPF, and/or the Criteria for Policy Linking Validity (Following this experience, please explain any implications for the policy linking process, training, toolkit, and related documents, such as the GPF and Criteria for Policy Linking Validity)</p>	

Summary Notes and Recommendations

STAGE 2 AND 3, CRITERION 1: ALIGNMENT BETWEEN THE ASSESSMENT, THE ASSESSMENT FRAMEWORK, AND THE CURRICULUM

This stage of the process involves the country government sharing standard-, curriculum-, and assessment-related documents (including the most recent round of data) with the project team and examination of those documents by the project team and the 4.1.1 Review Panel to determine whether the assessment(s) meets reliability and validity standards required for a country to proceed with policy linking for reporting global outcomes. More details about each of the below criteria are included in the Criteria for Policy Linking Validity. While many of the Criteria for Policy Linking Validity are yes/no questions, please provide details to support each response, e.g., what type of sampling was used or how was the assessment adapted for special educational needs and disabilities (SEND) learners.

Describe the process and the outcome of aligning the curriculum and national standards with the GPF	General comments about the process:	
	Were materials provided/obtained on time to/by the project team?	
	Were all necessary materials provided to the policy linking team?	
	Criteria for Policy Linking Validity 1a) Are the content expectations for the grade/subject clearly defined in the country's curriculum? <i>Please provide details both on the project team's judgement and the official judgement of the 4.1.1 Review Panel.</i>	
	1b) Is the content domain for the assessment clearly defined? Please provide details both on the project team's judgement and the official judgement of the 4.1.1 Review Panel.	
	1c) Do the items in the assessment appropriately sample from the assessment content domain such that the assessment can be considered a comprehensive assessment of reading or mathematics as defined in the assessment framework? <i>Please provide details both on the project team's judgement and the official judgement of the 4.1.1 Review Panel.</i>	
	1d)* Is the assessment aligned with the GPF? Please provide details both on the project team's judgement and the official judgement of the 4.1.1 Review Panel.	
	2a)* Is there evidence that the items in the assessment have been reviewed qualitatively and/or quantitatively to determine their validity? Please provide details both on the project team's judgement and the official judgement of the 4.1.1 Review Panel.	
	2b) Have the items been reviewed to ensure fairness to all relevant subgroups of the population, including students with SEND? Please provide details both on the project team's judgement and the official judgement of the 4.1.1 Review Panel.	
	2c)* Is the cohort that took the assessment representative of the population against which results will be reported? Please provide details both on the project team's judgement and the official judgement of the 4.1.1 Review Panel.	
	2d) If a sample is used, is the sample appropriately powered to detect reasonable differences over time? Please provide details both on the project team's judgement and the official judgement of the 4.1.1 Review Panel.	
	3a)* Is the value of coefficient alpha[1] (see definition above) for the grade-level subject assessment greater than or equal to 0.7? Please provide details both on the project team's judgement and the official judgement of the 4.1.1 Review Panel.	

Describe the process and the outcome of aligning the curriculum and national standards with the GPF	3b(i)* For paper-and-pencil assessments that contain selected response items, how has the scoring been quality assured to ensure appropriate scores for each student? <i>Please provide details both on the project team's judgement and the official judgement of the 4.1.1 Review Panel.</i>	
	3b(ii)* For paper-and-pencil assessments that contain constructed response items and/or oral assessments with selected response and/or constructed response items, how have those responsible for scoring been quality assured to ensure consistency of scoring (inter-rater reliability)? Please provide details both on the project team's judgement and the official judgement of the 4.1.1 Review Panel.	
	3c) Is there any additional evidence relating to the reliability of the assessment? Please provide details both on the project team's judgement and the official judgement of the 4.1.1 Review Panel.	
	Where there any challenges with Stages 2 and 3? How were they addressed?	
	What concerns, if any, did you have about the reliability/validity of the assessment?	
	What level of alignment did the project team and the 4.1.1 Review Panel, respectively, find between the assessment and the GPF? If the ratings differed, why?	
	What was the final recommendation and grade provided by the 4.1.1 Review Panel? Was there a report prepared on the results of Stages 2 and 3? If yes, please submit that report with this form.	
	General comments about the outcomes of the pre-workshop reliability and validity process:	
What recommendations would you make to the GPF based on this exercise? Please provide a justification for the proposed changes and how they might affect the validity of the policy linking results.		
Based on the alignment exercise, do you think the country should move forward with policy linking? Why/why not?		
Recommendations for changes in the policy linking process/toolkit to improve the process:		
Recommendations for changes in the Criteria on Policy Linking Validity to improve the process:		

Summary Notes and Recommendations

STAGE 4: PREPARATION FOR THE POLICY LINKING WORKSHOP

(IF APPROVAL RECEIVED FROM UIS FOLLOWING STAGE 3 TO PROCEED)

This stage involves all of the technical and logistical planning for the in-country workshop, including facilitator and panelist selection and invitations, workshop logistical planning, and materials preparation.

Facilitator Selection/Training	General comments about the process:	
	Please describe the qualifications of the lead facilitators and how they were selected and trained (if necessary)	
	Please describe the qualifications of the content facilitators and how they were selected and trained	
	Can the process/training be improved? If yes, how?	
Workshop Logistics	Will the workshop be held completely remotely (with everyone remote), partly remotely (with lead facilitators remote and panelists and, potentially, content facilitator in person), or in-person? Why was the decision made in the way it was?	
	If the workshop will be led by non-government officials, how will government officials be engaged and trained on the methodology for the future?	
	What logistical challenges arose and how were they mitigated? What recommendations, if any, do you have to improve workshop logistical planning for future workshops?	

Panelist Selection	How many panelists were invited to participate?			
	Please describe how the panelists were selected and the sampling criteria used			
	Can the process be improved? If yes, how?			
	Please include the number of panelists selected for participation of the following groups:			
		PANEL 1	PANEL 2	PANEL 3
	Men			
	Women			
	Public School Teacher			
	Private School Teacher			
	Curriculum Expert			
	Works in Rural Area			
	Works in Urban Area			
	Taught learners with disabilities			
	Disability expert			
	Experience giving assessment			
	Geographical Area 1			
	Geographical Area 2			
	Geographical Area 3			
	Geographical Area 4			
Geographical Area 5				
Geographical Area 6				
Geographical Area 7				
Material Preparation	What challenges, if any, arose with collecting and developing materials?			
	What additional information/materials would help future facilitators with the materials preparation process?			
	What changes, if any, do you recommend to the toolkit, GPF, slides, or other materials that would improve usability or the process in any other way?			
Overall	Are there any logistical details that you missed in the planning process? What were they, and how might this issue be avoided in the future?			

Summary Notes and Recommendations

STAGE 5: IN-COUNTRY WORKSHOP

This stage involves implementation of the five-day workshop in country. Please include details about all activities undertaken, panelist attendance and participation, progress toward daily objectives, and key outcomes of the workshop. Also, please try to complete the questions for the relevant day at the end of that day. The Overall Questions should be addressed on Day 5, if possible, and the workshop outcomes can be entered once analysis is completed after the workshop.

Workshop dates:

Please include the attendance sheets, with details on all participants, their positions, and roles in the workshop, as an attachment to this form.

WORKSHOP DAY-BY-DAY

These questions were written in an attempt to align with activities meant to be completed each day (for in-person workshops). For remote workshops, the timing will vary. If some activities are not completed on the day listed below, questions related to those activities should be answered whenever the activities are completed (even if on a different day). If this happens, please note the day that the question was answered so that we can update the form to match the appropriate timing moving forward.

DAY 1	1. How many panelists showed up, and how many were missing? Reasons why, if known?	
	2. How well did the welcomes and introductions go? Please provide details about what went well and what did not. What might be improved for future workshops?	
	3. How did the overview of the GPF go? What worked well, and what did not?	
	4. How did the overview of the assessment go? What worked well, and what did not?	
	5. Did you have the panelists practice taking the assessment during the session? Why or why not? If you did, would you recommend it for other workshops?	
	6. How many of the panelists were able to assess learners ahead of the workshop using the instrument being linked? If some were not able to, why not?	
	7. How did that pre-workshop assessment process go? Did the panelists understand it? Do their reported scores make sense? Does it seem that they understood how to select the learners to assess? What might have been improved, if anything?	
	8. General reflection/feedback on the process (e.g., allocation of time for each activity, slides, facilitation).	
	9. What were the major points of discussion during the day?	
	10. What recommendations do you have for the policy linking process/toolkit based on the results of Day 1?	
	11. What recommendations do you have for the Global Proficiency Framework based on today?	

DAY 2	1. How many panelists showed up, and how many were missing? Reasons why, if known?	
	2. How well did the alignment of the assessment with the GPF go? Did the panelists understand the task? Please provide details about what went well and what did not.	
	3. General reflection/feedback on the process (e.g., allocation of time for each activity, slides, facilitation)	
	4. Please provide information on the results of alignment (including a table that shows the panelists' alignment of each PLD with each assessment item) and discrepancies with pre-workshop alignment.	
	5. How clear were panelists on the difference between complete and partial alignment? Please provide details.	
	6. If linking a reading assessment, how clear were panelists on the purpose of examining the grade-level of the text? How did that process or reviewing text complexity go?	
	7. What issues around alignment came up, if any?	
	8. What were the major points of discussion/confusion that came up during the alignment process?	
	9. What recommendations do you have for the policy linking process/toolkit based on the alignment process?	
	10. What recommendations do you have for the Global Proficiency Framework based on the alignment process?	
	11. What recommendations do you have for the 4.1.1 Criteria for Policy Linking Validity based on the alignment process?	
DAY 3	1. How many panelists showed up, and how many were missing? Reasons why, if known?	
	2. How well did the introduction/discussion of the GPDs go? Please provide details about what went well and what did not.	
	3. How well did the introduction to the matching process go? Please provide details about what went well and what did not.	
	4. General reflection/feedback on the process (e.g. allocation of time for each activity, slides, facilitation).	
	5. What major points of conversation/ questions/confusion came up during the matching process?	
	6. In your opinion, do participants have a good understanding of the GPDs? Clear understanding of the GPDs in their context? Clear understanding of what a minimally proficient student is? Understanding of how the assessment is conducted? Understanding of the assessment item difficulty?	
	7. What other challenges/issues came up, if any?	
	8. What recommendations do you have for the policy linking process/toolkit based on the matching process?	
	9. What recommendations do you have for the Global Proficiency Framework based on the matching process?	

DAY 4	1. How many panelists showed up, and how many were missing? Reasons why, if known?	
	2. How well did the presentation on setting global benchmarks go? Please provide details about what went well and what did not.	
	3. How well did the presentation of the Angoff rating method and rating practice items go? Please provide details about what went well and what did not.	
	4. General reflection/feedback on the process (e.g., allocation of time for each activity, slides, facilitation).	
	5. What major points of conversation/ questions/confusion came up during the discussion of benchmarks? To what extent do ALL panelists appear to be engaged, contributing, and asking questions?	
	6. What major points of conversation came up during the discussion of the Angoff method and during practice item rating?	
	7. In your opinion, do participants have a good understanding of the purpose of setting benchmarks? Clear understanding of the Angoff method? Why or why not?	
	8. To what extent do you feel panelists understood that they needed to consider only those learners who meet the expectations listed for the relative performance level when making judgements?	
	9. What other challenges/issues came up, if any?	
	10. What recommendations do you have for the policy linking process/toolkit based on today?	
	12. What recommendations do you have for the Global Proficiency Framework based on today?	
DAY 5	1. How many panelists showed up, and how many were missing? Reasons why, if known?	
	2. How did the discussion on Round 1 judgements go? Were the panelists surprised? What major points of conversation came up as a result of differences in judgements?	
	3. To what extent did you review each item for which there were differing judgements? Why or why not?	
	4. How did panelists react to the information presented on item difficulty, the location statistics, and the impact information? To what extent do you think panelists understood the data? In what ways do you think this information affected their Round 2 ratings?	
	5. General reflection/feedback on the process (e.g., allocation of time for each activity, slides, facilitation).	
	6. What major points of conversation/ questions/confusion came up during the discussion of Round 1 ratings and other feedback data?	
	7. What major points of conversation/ questions/confusion came up during the presentation on Round 2 ratings?	
	8. In your opinion, by the end of the workshop, did participants have a good understanding of the policy linking process? The GPF? The assessment and the level of item difficulty? Why or why not?	
	9. What other challenges/issues came up, if any?	

	10. What recommendations do you have for the policy linking process/toolkit based on today?	
	11. What recommendations do you have for the Global Proficiency Framework based on today?	
Overall feedback	1. Did any unexpected logistical challenges arise during the workshop? What recommendations do you have to avoid similar issues in future workshops?	
	2. To what extent did you find it useful/not useful to kick off the workshop with the in-depth overview of the GPF/assessment before diving into the tasks?	
	3. How useful/not useful was the pre-workshop exercise? Why?	
	4. What general reflections do you have on the workshop?	

Workshop Results

	Question	Response	Notes
Workshop Results Statistics	1. What was the intra-rater reliability for the first and second rounds of ratings? Please add any comments on this.		
	2. What was the inter-rater reliability for the first and second rounds of ratings? Please add any comments on this.		
	3. What was the standard error of measurement (SE) at each benchmark level for each round? Please add any comments on this.		
	4. To what extent were the final panelists in attendance representative of the target population of schools being reported on?		
	5. To what extent did the panelists who showed up meet the other selection criteria described in the Policy Linking Toolkit?		
Overall Outcomes	6. To what extent did panelists report understanding the GPF, assessment, and policy linking methodology? And, to what extent did they feel comfortable with their round 2 evaluations and final benchmarks?		
	7. How confident are you in the final benchmarks provided by the panelists?		

Summary Notes and Recommendations
