

POLICY LINKING FOR MEASURING GLOBAL LEARNING OUTCOMES TOOLKIT

Linking assessments to the global Proficiency Framework



USAID
FROM THE AMERICAN PEOPLE



UNESCO
INSTITUTE
FOR
STATISTICS



GLOBAL
ALLIANCE
TO MONITOR
LEARNING



POLICY LINKING FOR MEASURING GLOBAL LEARNING OUTCOMES TOOLKIT

LINKING ASSESSMENTS TO THE GLOBAL PROFICIENCY FRAMEWORK

OCTOBER 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.



POLICY LINKING FOR MEASURING GLOBAL LEARNING OUTCOMES TOOLKIT

LINKING ASSESSMENTS TO A GLOBAL PROFICIENCY FRAMEWORK

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

CONTENTS	iii
ACRONYMS	vii
GLOSSARY OF TERMS	ix
ACKNOWLEDGMENTS	xi
CHAPTER I. INTRODUCTION TO POLICY LINKING	1
A. Rationale for Policy Linking.....	1
B. Overview of the Global Proficiency Framework.....	2
C. Overview of Policy Linking	6
D. Policy Linking Stages	7
E. Uses and Benefits of Policy Linking.....	8
F. Using the Policy Linking Toolkit.....	9
CHAPTER II. THE POLICY LINKING METHOD	10
A. Task 1 - Aligning the Assessment to the GPF.....	10
<i>Step 1 - Panelist alignment exercise</i>	10
<i>Step 2 - Facilitator summary of results</i>	12
B. Task 2 - Matching Assessment Items with GPLs and GPDs.....	15
C. Task 3 - The Angoff Method for Setting Benchmarks.....	16
CHAPTER III. PREPARING FOR THE POLICY LINKING WORKSHOP	20
A. Select Workshop Facilitators and Analyst	20
B. Plan Workshop Logistics.....	21
C. Select and Invite Workshop Panelists.....	21
<i>Selecting panelists</i>	21
<i>Inviting panelists and the pre-workshop activity</i>	23
D. Prepare Workshop Materials and Analyses.....	25
<i>Materials that need to be obtained</i>	25
<i>Analysis that needs to be conducted</i>	25
<i>Materials and data that need to be created/adapted</i>	26
E. Train Content Facilitators.....	30
CHAPTER IV. IMPLEMENTING THE POLICY LINKING WORKSHOP	33
A. Workshop Day One.....	34
1. <i>Opening, introductions, logistics, and agenda</i>	34
2. <i>Presentation on the background, objective, and tasks</i>	35
3. <i>Presentation on the GPF</i>	35
4. <i>Presentation on the assessment instrument</i>	36
B. Workshop Day Two	37
5. <i>Presentation on the alignment exercise (Task 1)</i>	37
6. <i>Activity on aligning the assessment(s) with the GPF (Task 1)</i>	37
C. Workshop Day Three.....	38

7. Presentation and discussion of alignment results from Day Two (Task 1)	38
8. Presentation on assessments and the GPLs/GPDs (Task 2)	38
9. Activity on matching the assessments with the GPLs/GPDs (Task 2)	39
D. Workshop Day Four	39
10. Presentation and discussion on the matching results (Task 2).....	39
11. Presentation on global benchmarking (Task 3).....	40
12. Presentation on the Angoff method (Task 3).....	40
13. Activity on Angoff method practice (Task 3).....	41
14. Activity on the Angoff method Round 1 (Task 3).....	42
E. Workshop Day Five.....	42
15. Presentation and discussion of Round 1 results and item difficulty and impact data (Task 3).....	42
16. Presentation on the Angoff method Round 2 (Review) (Task 3).....	43
17. Activity on the Angoff method Round 2 (Task 3).....	43
18. Presentation on the workshop evaluation.....	44
19. Presentation and discussion on the Angoff method Round 2 results.....	44
20. Workshop closing and logistics	45
E. Tips for Hosting Remote Workshops	45
Logistics.....	45
Lead facilitator(s).....	46
Content facilitator training and interaction	46
Pre-sessions.....	46
Discussions.....	47
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.....	50
A. Collect evidence from the workshop	50
B. Submit evidence to UIS.....	50
C. Receive a response back from UIS	51
BIBLIOGRAPHY	53
ANNEXES.....	56
Annex A - Global Proficiency Framework	56
Annex B - Global Minimum Proficiency Levels.....	57
Annex C - Workshop Preparation Checklist.....	58
Annex D - Alignment Rating Form for Task 1	60
Annex E - Workshop Facilitation Slides.....	61
Annex F - Item Rating Forms.....	62
Annex G - Intra- and Inter-Rater Consistency, and Standard Error of Measurement (SEM)	67

Annex H - Workshop Panelist Information	72
Annex I - Sample Invitation Letter for Policy Makers.....	73
Annex J - Sample Invitation Letter for Workshop Panelists	74
Annex K – Sample Explanation for Panelists of Pre-Workshop Activity	75
Annex L - Pre-Workshop Statistics	77
Annex M - Feedback Data Examples and Instructions	78
Annex N - Sample Agenda for an In-Person Workshop	82
Annex O - Sample Agenda for a Remote Workshop.....	84
Annex P - Workshop Evaluation Form.....	90
Annex Q – Content Facilitator Slides	94
Annex R – Benchmark Calculations for the Workshop	95
Annex S - Outline for the Benchmarking Technical Report	97
Annex T - 4.1.1 Review Panel Criteria for Policy Linking Workshop Validity	98
Annex U - Agreement and Consistency Coefficients.....	100
Annex V – Technical Documentation of Workshop Outcomes	102
Annex W - Process Documentation Form	103

TABLES

Table 1. Grade 3 Mathematics Example from the GPF.....	3
Table 2. USAID Foreign Assistance Indicators for Primary-Level Reading and Mathematics	4
Table 3. Policy Linking Stages.....	7
Table 4. Example of Summary Alignment Results for a Grade 3 Assessment by Domain, Construct, and Subconstruct.....	12
Table 5. Mathematics Assessment Alignment Criteria for Grades 1-9	14
Table 6. Reading Assessment Alignment Criteria for Grades 1-9.....	14
Table 7. Item Rating Form for Use with Yes-No Angoff Modification	18
Table 8. Brief Description of the In-Person Workshop Agenda.....	27
Table 9. Discussion Purpose, Do’s, and Don’ts by Task.....	31
Table 10. Summary of Tasks and Activities for the Policy Linking Workshop (Note that day References are for In-Person Workshops).....	33

FIGURES

Figure 1. Education System Alignment.....	5
Figure 2. Example of Comparable Benchmarks on Various Assessments	6
Figure 3. Policy Linking Process and Benefits.....	9
Figure 4. Alignment Scale and Number of Content Standards to Which an Item Aligns	11
Figure 5. Example Alignment of an Item to the GPF with Complete Fit	11
Figure 6. Example Alignment of an Item to the GPF with PARTIAL Fit	12
Figure 7. Example of Matching Items to the GPLs and GPDs.....	16
Figure 8. Item Rating Process for Yes-No Angoff Modification	17
Figure 9. Activities to Prepare for the Policy Linking Workshop.....	20
Figure 10. Composition of Panelists.....	22
Figure 11. Assessment Security Considerations	24
Figure 12. Invitation Adaptations for Remote Workshops	24
Figure 13. Translation of the GPF.....	28
Figure 14. Key Differences between Untimed Assessments (Largely CBAs) and Timed Assessments.....	29
Figure 15. Tips for Facilitators on Opening Presentation	35
Figure 16. Tips for Facilitators on Background Presentation.....	35
Figure 17. Tips for Facilitators on Presentation of the GPF.....	36
Figure 18. Tips for Facilitators on the Assessment Presentation.....	36
Figure 19. Tips for Facilitators on the Alignment Presentation.....	37
Figure 20. Tips for Facilitators on Task 1 – Aligning the Assessment(s) with the GPF	38
Figure 21. Tips for Facilitators on Reviewing the Results of Task 1	39
Figure 22. Tips for Facilitators on the Task 2 Matching Presentation	39
Figure 23. Tips for Facilitators on Overseeing the Task 2 Matching Activity	39
Figure 24. Tips for Facilitators on Review the Task 2 Matching Results.....	40
Figure 25. Tips for Facilitators on the Global Benchmarking Presentation.....	40
Figure 26. Tips for Facilitators on Presenting the Task 3 Angoff Method	41
Figure 27. Tips for Facilitators on the Task 3 Angoff Practice	41
Figure 28. Tips for Facilitators on Overseeing Task 3 – Round 1 Ratings	42
Figure 29. Tips for Facilitators on Sharing Round 1 Results	43
Figure 30. Tips for Facilitators on Presenting Angoff Round 2.....	43
Figure 31. Tips for Facilitators on Overseeing Angoff Round 2 Ratings	44
Figure 32. Tips for Facilitators on Presenting the Evaluation Form	44
Figure 33. Tips for Facilitators on Presenting Final Results.....	45
Figure 34. Tips for Facilitators on Workshop Closing	45

ACRONYMS

AERA	American Educational Research Association
ACER	Australian Council for Educational Research
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
PM _j	Partially Meets Minimum Proficiency
PLT	Policy Linking Toolkit

SDG	Sustainable Development Goal
SEM	Standard of Error Measurement
SME	Subject Matter Expert
USAID	U.S. Agency for International Development

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 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 knowledge and/or skills 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.I.I.

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.

Knowledge and/or skills – What content learners expected to know and be able to do for a specific grade and domain, construct, and subconstruct. Knowledge and/or skills 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 set of knowledge and skills included in the GPF are not meant to be prescriptive in nature but rather to facilitate measurement against SDG 4.I.I.

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 content standards (knowledge and/or skills) 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 of measurement (SEM) – A statistic that indicates the measurement error associated with a benchmark.

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, the UK Foreign Commonwealth and Development Office (FCDO)¹, and the Bill and 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 the USAID Office of Education 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 and 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 UK 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 who supported the pilots in those countries and to the People's Action for Learning (PAL) Network, The Education Partnership (TEP Centre), and Zizi Afrique who 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.

¹ Formerly UK's Department for International Development (DFID)

CHAPTER I. INTRODUCTION TO POLICY LINKING

A. Rationale for Policy Linking

While the number of countries engaging in assessments of learning outcomes 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, international education donors, partners, and other stakeholders need a method for accurately determining how learning outcomes compare both 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 those areas and populations most at 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. The way to address this problem is by linking the different assessments to a common scale. 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.² The UNESCO Institute for Statistics (UIS) has included policy linking in their 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, the Bill and Melinda Gates Foundation,³ ACER, and UNICEF – have demonstrated interest in using or supporting the use of policy linking for setting benchmarks⁴ 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. 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

² 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.

³ The Bill and 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.

⁴ A benchmark is a numeric threshold on an assessment that indicates a learner has met a proficiency level.

organization (the National Foundation for Educational Research--NFER) – for this current version. The NFER evaluation of the method, funded by the Bill and 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), 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 (sometimes called content standards)⁵ and how much of that they should be able to demonstrate (referred to in the GPF as global proficiency descriptors, sometimes called performance standards)⁶ that form a common scale for global reporting on learner outcomes in reading and mathematics in Grades 1-9. 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. Note that a country 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 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. Overview of the Global Proficiency Framework

The policy linking method described in this toolkit requires a 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. This is the reason the GPF (See **Annex A**) was created. 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 3 reading GPF, which then allows for comparing, aggregating, and tracking outcomes from those grade 3 reading assessments.

While countries define what knowledge and/or skills learners need to obtain in which grades based on their individual contexts and while they articulate that information through their national curricula, content and assessment frameworks, 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.

The Policy Linking Workshop Group created the GPF, working with 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. 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 (1 through 9). The SMEs reached consensus on the knowledge and skills (sometimes called content standards) and the global performance descriptors (GPDs) (sometimes called performance standards) described in the

⁵ 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 knowledge and skill expectations included in the GPF are not meant to be prescriptive in nature but rather to facilitate measurement against SDG 4.1.1.

⁶ 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.

GPF based on their knowledge of developmental progressions and the UIS' Global Content Framework, which was based on 73 curriculum and assessment frameworks from 25 countries for reading and 115 assessment frameworks from 53 countries for mathematics.^{7,8}

An example from part of the Grade 3 mathematics GPF is shown in **Table I**. It has the domains, constructs, subconstructs, knowledge 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 any learners who do not have the knowledge and/or skills listed in the “partially meets global minimum proficiency” level.

TABLE I. 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, count in and identify the relative magnitude of whole numbers	Count, read, and write whole numbers	Count in whole numbers up to 100. Read and write whole numbers up to 100 in words and in numerals.	Count in whole numbers up to 1000. Read and write whole numbers up to 1000 in words and in numerals.	Count in whole numbers up to 10,000. Read and write whole numbers up to 10,000 in words and in numerals.
			Compare and order whole numbers	Compare and order whole numbers up to 100.	Compare and order whole numbers up to 1000.	Compare and order whole numbers up to 10,000.
			Skip count forwards or backwards	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 <i>(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; or given a picture of 19 shapes, draw 19 more shapes).</i>	Use place-value concepts for tens and ones <i>(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).</i>	Use place-value concepts for hundreds, tens, and ones <i>(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.).</i>	

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 knowledge and/or skills associated with the subconstructs demonstrate what learners need to know and be able to do by grade and subject.

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

- **Below partially meets global minimum proficiency:** Learners lack the basic knowledge and skills for their grade. As a result, they cannot complete the most basic tasks appropriate for their grade.
- **Partially meets global minimum proficiency:** Learners have partial knowledge and skills for their grade. As a result, they can partially complete basic tasks appropriate for their grade.
- **Meets global minimum proficiency:** Learners have sufficient knowledge and skills for their grade. As a result, they can successfully complete basic tasks appropriate for their grade.

⁷ See the previous footnote for a chronology of the development of the GPF.

⁸ See UNESCO (2018a, 2018b) in the references for their 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).

- **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”.

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. 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
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

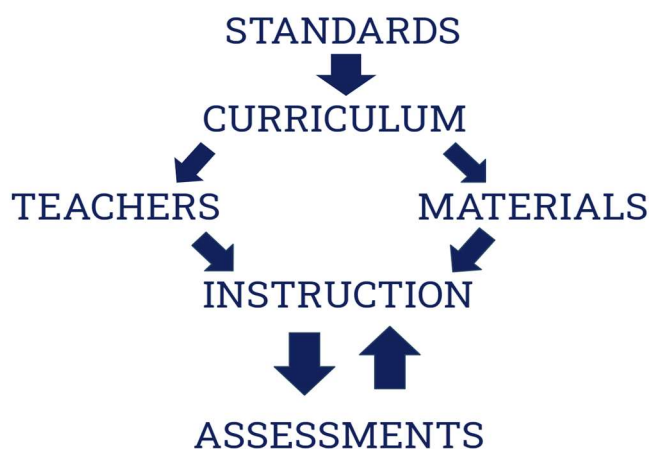
Indicator Number	Indicator Title
Supp-14	Percent of individuals with improved digital literacy skills following participation in USG-assisted programs
Supp-15	Education system strengthened: policy reform
Supp-16	Education system strengthened: data systems strengthened

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 meeting global minimum proficiency will fall into this category) for Grades 1 to 9 in reading and mathematics. They describe how much content learners need to know and be able to do in relation to the defined knowledge and/or skills required by grade and subject. For example, in reading, the GPF says that a learner who meets global minimum proficiency in Grade 3 should be able to identify the general topic in a Grade 3-level continuous text when the topic is prominent but not explicitly stated. In mathematics, a learner who meets global minimum proficiency in Grade 3 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 can elect to use only the “meets” GPL, which is sufficient for reporting on SDG 4.I.I.

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 1**.

FIGURE 1. EDUCATION SYSTEM ALIGNMENT



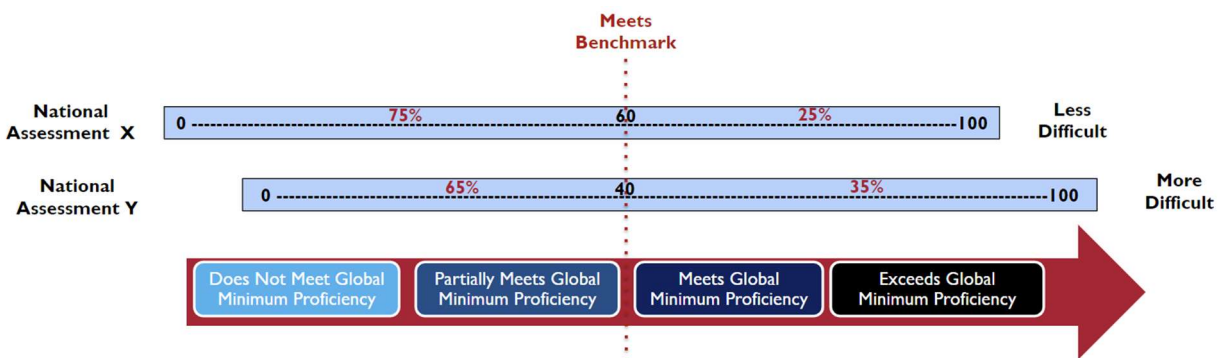
The GPF offers a lens by which countries can examine alignment between the various components of their education system. During the piloting phase for the GPF between September 2019 and July 2020, several countries used it for these purposes.

C. Overview of Policy Linking

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.⁹ 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 2** 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 benchmark of 40 points. In theory, a learner with an ability level of just meeting global minimum proficiency, and 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, e.g., the benchmarks on each assessment that correspond to meeting global minimum proficiency.

FIGURE 2. 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 who understand the performance of learners for specific grades and subjects. They follow the Angoff procedure to 1) examine the country’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.

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

D. 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, and their partners, must complete each of these stages for their results to be accepted for reporting against SDG 4.1.1 and USAID “F” indicators. This toolkit covers Stages 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

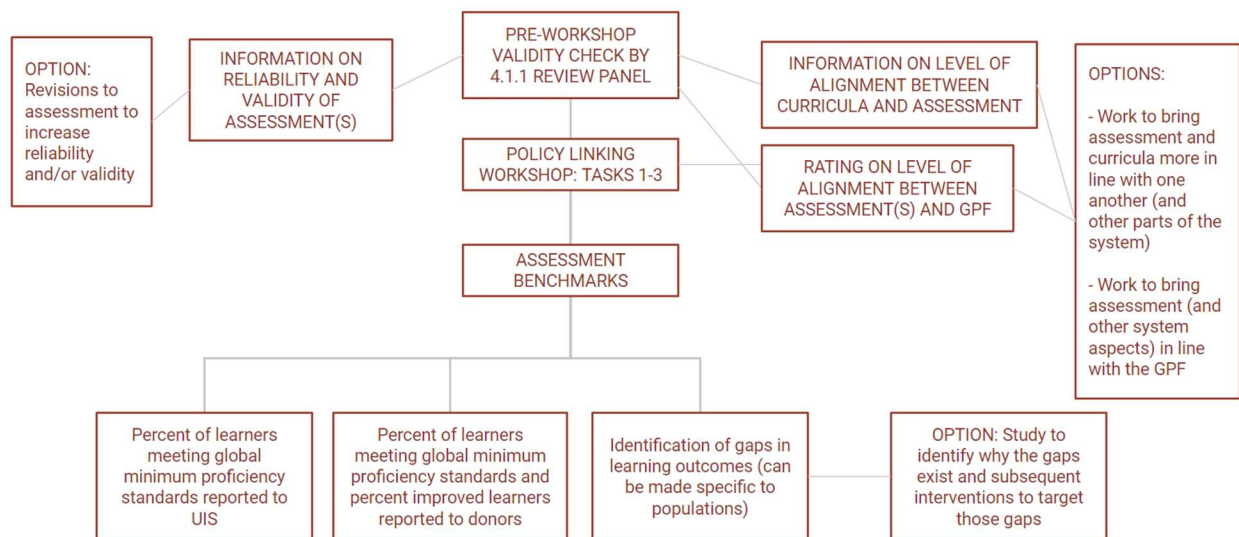
#	Policy Linking Stages	Purpose	Roles/ Responsibilities	Resources (available on UIS website)
1	Initial engagement	For countries to make the decision of whether to move forward with policy linking, either at a national or regional/state level and which assessment(s) they will link to global standards 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 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 quality and alignment	To submit for review by UIS’ 4.1.1 Review Panel to ensure assessments used for global reporting are valid, reliable, & sufficiently aligned to the GPF	Country governments 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, & secure a venue	Country governments with/without support of partners	<ul style="list-style-type: none"> • Policy Linking Toolkit (Chapter 3) • Workshop Preparation Checklist (Annex C - Workshop Preparation Checklist)

#	Policy Linking Stages	Purpose	Roles/ Responsibilities	Resources (available on UIS website)
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 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

E. 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 3** in the second and third stages, the country 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 curriculum and assessment and between its assessment and the GPF. This information might help to inform improvements in country education systems, as described in the GPF section above. Finally, the results of the policy linking workshop should help countries to identify the percentage and profile (assuming the country 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 3. POLICY LINKING PROCESS AND BENEFITS



F. 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, donor representatives, and partners – who will be organizing, funding, and/or implementing the method in their country or region.¹⁰ 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 planned, 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 workshop. Finally, **Chapter VI** presents details on what materials country governments 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.¹¹

The annexes provide all of 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 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.

¹¹ Management Systems International (2019). *Policy linking method: Linking assessments to a global standard*. US Agency for International Development (USAID), Washington, DC.

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 benchmarks¹² for each assessment using a standardized method (a modified version of the Angoff methodology) through two rounds of ratings

Each of these tasks are 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 and 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'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. 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 (roles and responsibilities are described in more detail in below) summarize results of the alignment activity

Step 1 - Panelist alignment exercise

In Step 1, after being given instructions on the tasks and then working through some examples with the facilitators, panelists work independently, going item-by-item to complete the following three sub-steps using the Alignment Rating Form, which can be found in **Annex D** - Alignment Rating Form for Task 1

- 1) Identify the knowledge and/or skills that learners need to answer the item correctly;
- 2) Search through the GPF (using GPF Table 3 to find the domain, construct, and subconstruct that aligns with the knowledge and/or skills needed to answer the item correctly; and
- 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 content standard, i.e., if the learner answers the item correctly, it is because they completely use the knowledge and/or skill(s) of the content standard;

¹² Note that if during Stage 1, 2, or 3, the government decides that they only wish to set a benchmark for the "meets" level or the government or 4.1.1 Review Panel decide 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.

- **Partial Fit (P)** signifies that **part of the content** required to answer the item correctly is contained in the content standard, i.e., if the learner answers the item correctly, it is because they partially use the knowledge and/or skill(s) of the content standard; and
- **No Fit (N)** signifies that **no amount of the content** required to answer the item correctly is contained in the content standards, i.e., if the learner answers the item correctly, it is because they do not use knowledge and/or skill(s) from the content standards.

Further details on the scale appear in **Figure 4** below.

FIGURE 4. ALIGNMENT SCALE AND NUMBER OF CONTENT STANDARDS TO WHICH AN ITEM ALIGNS

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

An example of a “complete fit” item follows in **Figure 5** with Item I from an 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 skills needed to correctly answer this item are contained in this single subconstruct.


FIGURE 5. EXAMPLE ALIGNMENT OF AN ITEM TO THE GPF WITH COMPLETE FIT

<p>I. How is eight hundred and seventy written in standard form?</p> <p>A 807</p> <p>B 870</p> <p>C 817</p> <p>D 871</p>	<p>Domain: Number Knowledge</p> <p>Construct: Whole Numbers</p> <p>Subconstruct: Identify and count whole numbers</p> <p>Knowledge or skill (content standard): Count, read, and write whole numbers to 1000; skip count forwards by twos, fives, tens, and hundreds; and compare and order whole numbers up to 1000</p>
<p>Fit: Complete fit since it only requires the knowledge or skills from a single content standard.</p>	

An example of a “partial fit” item follows in **Figure 6**. The panelist rated this item as a partial fit since to answer this item correctly, a learner would need knowledge or skills from two different content standards.

FIGURE 6. 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 summarize results by taking an average of the number of items that the panelists aligned to each domain, construct, and subconstruct. Note that even though alignment occurs at the knowledge 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 is a partial fit (Note: in these cases, for summary purposes, facilitators should count the domain, construct, and subconstruct that best describes the majority of the knowledge/skills needed to answer the item correctly). An example of summary results for a Grade 3 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	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
SI	Data management	2

A1	Patterns	0
A3	Relations and functions	0
Total		26
Subconstructs		Items
N1.1	Identify, count in, and identify the relative magnitude of whole numbers	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, & 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 will assess both the depth (number of items that have at least a partial fit with at least one content standard 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 is the same as that 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** below. Note that when summarizing results to the subconstruct level, facilitators and/or data analysts will only consider the subconstructs with knowledge and/or skills 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 knowledge or skills that have an “x” listed under the appropriate grade level column in GPF Table 3. For example, **Table 4**, above in this document, only includes the domains, constructs, and subconstructs relevant for Grade 3.

From the below criteria, it is clear that the example Grade 3 assessment described in **Table 4** would be considered “additionally aligned” since it both: 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 3 (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 (min 5 items)
	Subconstructs (breadth):	Items covering at least 50% of the Number subconstructs
Additionally aligned	Domain (depth):	Number (min 5 items) and Measurement and Geometry (min 5 items)
	Subconstructs (breadth):	Items covering at least 50% of the Number, Measurement, and Geometry subconstructs
Strongly aligned	Domain (depth):	Number (min 5 items) and Measurement and Geometry (min 5 items) and Statistics & Probability and Algebra (min 5 items)
	Subconstructs (breadth):	Items covering at least 50% 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 (min 5 items) C (min 5 items)	R (min 5 items)	R (min 5 items)
	Subconstructs (breadth):	Items covering at least 50% of the D and C subconstructs	Items covering at least 50% of the R subconstructs	Items covering at least 50% of the R subconstructs
Additionally aligned	Domain/ Construct (depth):	N/A	N/A	R: B1 (min 5 items) R: B2 (min 5 items)
	Subconstructs (breadth):	N/A	N/A	Items covering at least 50% of the R subconstructs
Strongly aligned	Domain/ Construct (depth):	R (min 5 items)	R: B1 (min 5 items) R: B2 (min 5 items)	R: B1 (min 5 items) R: B2 (min 5 items) R: B3 (min 5 items)
	Subconstructs (breadth):	Items covering at least 50% of the R subconstructs	Items covering at least 50% of the R subconstructs	Items covering at least 50% 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, with support from its partners (if relevant) will need to report out the results of this alignment exercise to the 4.1.1 Review Panel.

B. Task 2 - Matching Assessment Items with GPLs and GPDs

Task 2 builds on the panelists' understanding of the assessment items and the GPF gained through the alignment activity. In this task, panelists are asked to take their alignment work to the next level by matching each item to the appropriate GPL¹³ and GPD in the GPF. They will 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 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 7** 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 will read through the GPDs for each GPL at the grade level (and the lower grades) to determine what GPL(s) and GPD(s) is/are the best match at which grade level. They will 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 knowledge and/or skill (content standard) (as determined in Task 1) and, thus, more than one GPD, the panelist will select the higher of the GPLs since a learner wouldn't 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 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 7 provides an example taken from the Workshop Facilitation Slides included in **Annex E - Workshop Facilitation Slides**. 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.

and then respond to the question, "Who had a pet dog?" This question matches with the knowledge or

¹³ Note that if during Stage 1, 2, or 3, the government decides that they only wish to set a benchmark for the "meets" level or the government or 4.1.1 Review Panel decide 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.

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 grade level of the passage, 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-3-level passage, and the item aligns to the “partially meets global minimum proficiency” level at Grade 3.

FIGURE 7. EXAMPLE OF MATCHING ITEMS TO THE GPLS AND GPDS

<p>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.</p>	
<p>Domain: Reading comprehension</p> <p>Construct: Retrieve information</p> <p>Subconstruct: Locate explicitly stated information</p> <p>Passage grade level: Grade 3</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>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.</p> <p>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.</p> <p>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.</p>

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 itself.

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 knowledge and/or skills correspond with the GPDs would perform on each item. Task 3 relies on the Angoff method for setting benchmarks. The Angoff method is an item-centered method that is appropriate for the various kinds of assessments administered in different countries. With the Yes-No Angoff method, the panelists use an item rating form (see **Annex F** - Item Rating Forms) to rate each of the items on the assessment instruments, using the following four steps:

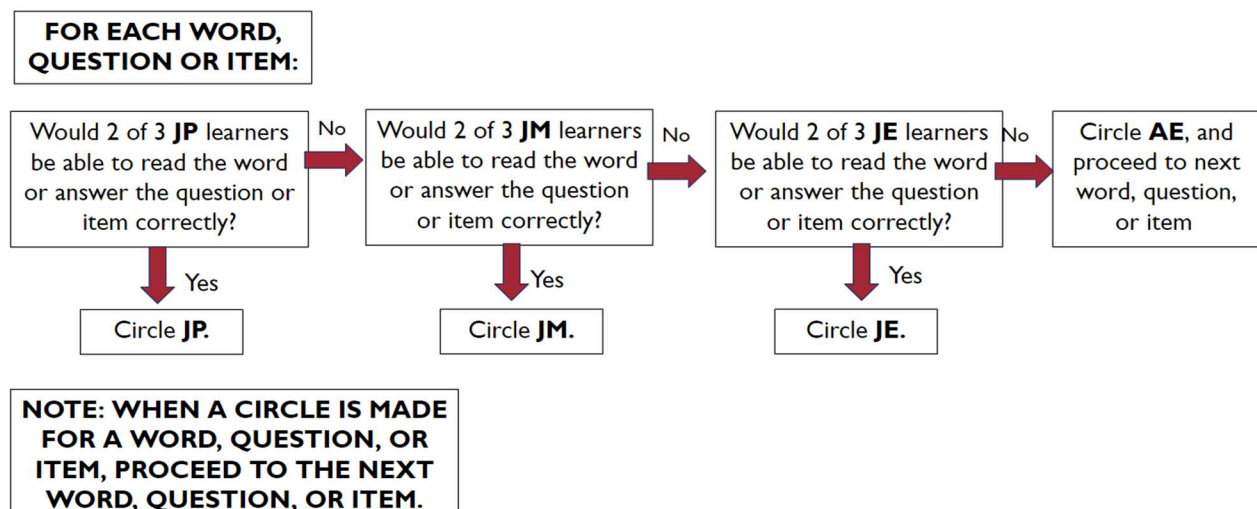
- **Step 1:** Identify or conceptualize three minimally proficient learners at each GPL.¹⁴ Note that minimally proficient learners are those who perform at or just slightly above the GPDs that describe the GPL. Estimate how these learners would perform on each of the assessment items. These learners are called Just Partially Meets (JP), Just Meets (JM), Just Exceeds (JE) learners. As

¹⁴ Note that if during Stage 1, 2, or 3, the government decides that they only wish to set a benchmark for the “meets” level or the government or 4.I.I Review Panel decide the assessment is too short to accommodate three benchmarks at the three main GPLs, then, panelists need only conceptualize learners at the “meets” of JM level.

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 the item correctly. The idea is to focus on the item content in relation to the descriptions of knowledge or skills 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 all of this information on their test booklet/assessment instrument during Task 2).
- **Step 3:** Select the lowest GPL, with the associated GPD, for the knowledge and skill needed to answer the item correctly (Note: 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 8** 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). 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 has the knowledge and/or skill 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 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 with the knowledge and skills listed in the appropriate GPL and GPD and determine if they are “reasonably sure” that those learners “would” answer the item correctly.

FIGURE 8. ITEM RATING PROCESS FOR YES-NO ANGOFF MODIFICATION



In completing Step 4, panelists will need to 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% chance OR 2 out of 3 learners (“yes”);** and
- Absolutely positive (“yes”).

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

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

The panelists 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 measurement (SEM) and inter-rater reliability (See **Annex G** - Intra- and Inter-Rater Consistency, and Standard Error of Measurement (SEM) 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 4.1.1 Review Panel at the end of the workshop to inform whether the results of 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 present panelists with:

- **A summary of their ratings** as well as how their individual ratings compare with other panelist ratings. They 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 to 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 to have an idea of the impact of their ratings and benchmarks.

Panelists 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** below:

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 will 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; and
 - 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; and
 - Exceeds = Average of the “exceeds” benchmarks across all panelists.

Note that 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 **Chapter IV** of this toolkit on 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. PREPARING FOR THE POLICY LINKING WORKSHOP

Government officials and donor representatives, if relevant, should have met 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 during Stage I: Initial Engagement. Resources for Stage I are linked in **Table 3** above. One key goal of Stage I is ensuring government buy-in and ownership over the process as well as engagement throughout planning and preparation—with the intention that if the government is not implementing the workshop on their own, following the workshop, they 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—will carry out the five activities shown in **Figure 9**. A detailed checklist of technical and logistical preparations used by the project team, in conjunction with the government officials and donor representatives, can be found in Error! Reference source not found..

FIGURE 9. 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 that 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 policy makers. 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 guiding the panels by following the method, including ensuring that panelists understand the GPF and the assessment content. 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 (reading and/or mathematics) area 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/were implemented. They must be fluent in the language of the 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 Stata and/or SPSS for the assessment data).

Note that it is recommended that recruitment efforts also cover a **national workshop coordinator** and a **national logistician**.

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 E**). The project team should work with relevant government and partner stakeholders to select the appropriate gathering option based on the context, safety of participants, 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 the **Annex C - Workshop Preparation Checklist**.

Finally, in addition to general logistics, during this activity, the project team should agree with the government about ways in which they will continue the engagement with the country government 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 is to actively give key representatives a role in the preparations and execution of the workshop, which will build in-country capacity and permit them to conduct future workshops as needed.

C. Select and Invite Workshop Panelists

Selecting panelists

Use Annex: Annex H

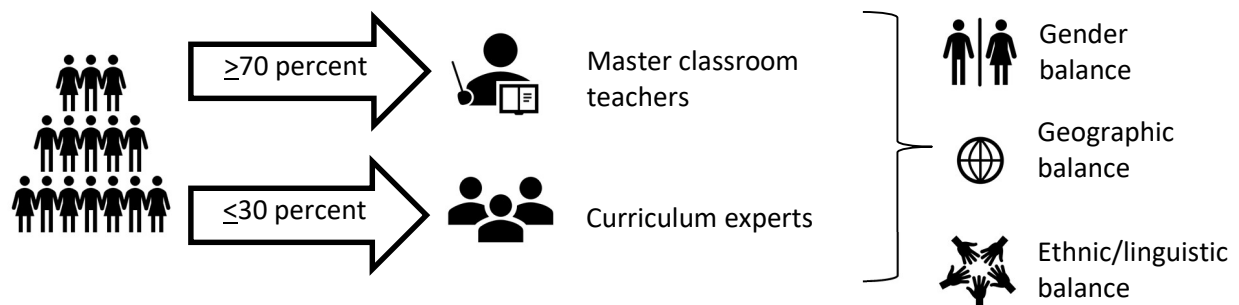
The project team should plan separate panels for each grade, subject, and language of assessment used for policy linking. If multiple assessments are included in a single workshop, e.g., Grade 3 reading and Grade

3 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 (reading or math) knowledge and teaching skills. 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 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 that the process obtains a replicable outcome but is also practical and manageable.¹⁵ As shown in **Figure 10**, the panel should be made up of at least 70 percent master classroom teachers and up to 30 percent non-teachers, preferably curriculum experts

FIGURE 10. COMPOSITION OF PANELISTS



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

- At least 5 years of teaching at or adjacent to the relevant grade level (teachers);
- At least 5 years of teaching experience (curriculum experts);
- Strong skills in the subject (reading or math) area;
- 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; and
- 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

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.

- 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; and
- **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.

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 and submitted with the workshop outcomes using the form included in **Annex H**. This form will give the 4.1.1 Review Panel sufficient data to address the degree to which the panelists meet the criteria.

Inviting panelists and the pre-workshop activity

Use Annex: I, J, and K

Panelists should be invited well in advance of the workshop, at least six weeks is recommended. **Annex I** - Sample Invitation Letter for Policy Makers and **Annex J** - Sample Invitation Letter for Workshop Panelists include draft invitation letters for policy makers 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 a invitation addendum that can be sent to panelists in **Annex K** – Sample Explanation for Panelists of Pre-Workshop Activity.

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/they should be administered to learners ahead of the workshop. The panelist will be asked to select nine learners – three learners 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 – prior to the workshop. 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. **Figure 11** for more information on assessment security.

- For individually administered timed assessments, such as early grade or mathematics assessments (EGRAs or EGMAAs), the sample assessments will include subtasks from reading or mathematics, as appropriate.
- For group 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. Details on invitations for remote workshops are included in **Figure 12** below.

FIGURE 11. 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 amongst 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 and particularly the agency or organization responsible for overseeing the assessment.

Security protocols for policy linking workshops – Assessment security protocols will vary depending on government and 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 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 G 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 12. 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 to each session
- **Information on the preferred hardware for joining** (computers are strongly preferred to allow panelists to see the slides and submit tasks, but smart phones 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 E 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 can describe the assessment and how to administer it as well as provide details on the GPF and how to select learners who fall

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 will need to ensure they have obtained 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 1. Thus, if the facilitators were involved in that stage, they should already have access to all except the starred items (which they will need to request) below.

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** above)*

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 that 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; note that more details on remote workshops are included on **Page 45**).

Analysis that needs to be conducted

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

- a. Item difficulty** - See **Annex L** - Pre-Workshop Statistics for details on how to calculate these statistics using the data from the most recent assessment results.
- b. Data distributions** - See **Annex M** - Feedback Data Examples and Instructions for details on how to prepare this data. The data distributions will show the number and percent of learners who took the assessment that achieved every possible score on the assessment. Note: while this data can be prepared ahead of the workshop, it is not needed until Day 4 when it will form the basis of the analysis of impact information (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) between Angoff rating rounds 1 and 2.

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

Materials and data that need to be created/adapted

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

- a. **Workshop agenda** - Templates included in **Annex J - Sample Invitation Letter for Workshop Panelists**, for in-person workshops, and **Annex O - Sample Agenda for a Remote Workshop**, for remote workshops; these will need to be adapted as described below.
- b. **Panelist IDs** - Need to be assigned on the first day of the workshop and should be confidential between the panelist and the project team.
- c. **Daily attendance sheet** - Needs to be created and tracked during the workshop to ensure each panelist has received all necessary training.
- d. **Relevant grade/subject GPDs**, including the grade below the one being linked. These will be carefully reviewed by the panelists during the workshop including the grade-level for the assessment(s) under consideration and the grade-level below the grade-level of the assessment(s). (The GPF is included in **Annex A**, but 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).
- e. **Facilitation slides** - Details on how to locate the slide templates are included in **Annex E - Workshop Facilitation Slides** for both timed and untimed assessments, but facilitators will need to adapt these; instructions on how to do so are included in the template.
- f. **Alignment rating forms and item rating forms** - **Annex D** - Alignment Rating Form for Task 1 for the alignment form and **Annex F** - Item Rating Forms Annex E - Workshop Facilitation Slides (Examples are included in the annexes, but they may need to be adapted).
- g. **Workshop evaluation forms** – A draft is included in **Annex P - Workshop Evaluation Form**. The project team may wish to add questions to the form and/or turn it into a daily evaluation form.
- h. **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 - Feedback Data Examples and Instructions**).

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

Workshop Agenda

The sample in-person workshop agenda (**Annex N - Sample Agenda for an In-Person Workshop**) 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 country-specific issues. Facilitators should review the agenda, adjust the dates, times for breaks (based on local norms), add in any necessary speeches from government officials, donors, etc. and then send to the government and its partners for their review before finalizing. A brief summary of the five-day in-person workshop agenda is presented below 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 and other sessions (activities, discussions, and feedback) will be panels, preferably led by the content facilitators.

A sample remote workshop agenda (**Annex O - Sample Agenda for a Remote Workshop**) provides a day-by-day list of the remote workshop sessions, time allocations, and facilitation requirements. Note, 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 (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 is optional but 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, including the government and donor agency, 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 (Note: 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, including the government and donor agency, 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 that 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.

Relevant grade/subject GPDs

Annex A - Global Proficiency Framework provides the GPF. 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 the grade-levels as provides a useful summary for policy makers and panelists.

Facilitators should consider with the government whether the two tables, at a minimum, may need to be translated if the language of assessment is not English (See **Figure 13** for details), but facilitators should not make any other changes to the content or language of the GPF.

FIGURE 13. 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. The slides are included in **Annex E** - Workshop Facilitation Slides 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 (called “Timed Assessments” throughout) or untimed assessments (Note: CBAs usually fall into this category as do untimed, group-administered regional and international assessments). Details on the differences between implementing a policy linking workshop for an untimed CBA versus a timed assessment are included in **Figure 14** below.

FIGURE 14. KEY DIFFERENCES BETWEEN UNTIMED ASSESSMENTS (LARGELY CBAS) AND TIMED ASSESSMENTS

- Given test security considerations, facilitators may not be able to send a full CBA or other group-administered, untimed 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 far in the assessment a learner would get within the allotted time and 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** above).

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.

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 - Alignment Rating Form for Task 1)** – These will be used for the panelists’ ratings of the alignment between the assessments and the GPF.
- **Item rating forms (Annex F - Item Rating Forms)** – These will be used for the panelists’ ratings of each assessment item in relation to the GPLs and GPDs.

The annexes include examples from timed assessment (in this case EGRA/EGMA) and CBA alignment and item rating forms. 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 forms is pretty basic, 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 - Item Rating Forms** with details on how to choose and adapt the forms.

Workshop evaluation forms

Panelists should fill out an evaluation at the minimum 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 - Workshop Evaluation Form** includes the minimum evaluation questions that must 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 level on the orientation, training, and process. The results will provide evidence of the panelists’ confidence with their judgments, as well as seek additional comments on the policy linking experience. The results will be included in the workshop report and presented to the 4.I.I Review Panel and government/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 teams opt 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. (Note: 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** - Feedback Data Examples and Instructions. 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 data will need to be generated 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/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)
- Relevant grade/subject GPDs from the GPF
- Assessment instrument (if assessment security protocols allow for it; see Error! Reference source not found. 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** – Content Facilitator Slides. 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 dos and don'ts of facilitating discussions during and following completion of each of the tasks (Note: the same rules apply to answering panelists questions and facilitating practice ratings), as shown in **Table 9** below.

TABLE 9. DISCUSSION PURPOSE, DO'S, AND DON'TS BY TASK

Task	Discussion Purpose	Dos	Don'ts
<p>Task 1 - Assessment and GPF alignment (panelists work independently)</p>	<p>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</p>	<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 subconstruct alignment 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 subconstruct
<p>Task 2 - Matching the assessment items with the GPLs and GPDs (panelists work together in groups)</p>	<p>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</p>	<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 skills 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
<p>Task 3, Round 1 - Rating the items using the Angoff method (panelists work independently)</p>	<p>To ensure panelists understood the task. One way to check this is by asking 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.”</p> <p>And, to give the panelists an opportunity to talk about disagreements on ratings, as this might</p>	<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 	<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 aligned X item to X subconstruct (Note - panelist ratings are supposed to be confidential, which is why they are presented to the group by panelist

Task	Discussion Purpose	Dos	Don'ts
	inform some panelists' Round 2 rating decisions.	panelists who rated no to explain why and vice-versa <ul style="list-style-type: none"> ● Encourage panelists to consider the item difficulty and impact data and decide if that affects their Round 2 judgements 	number rather than name <ul style="list-style-type: none"> ● 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. 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 benchmarks¹⁶ for each assessment using a standardized method (a modified version of the Angoff Procedure)

Table 10 below 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 E** at the end of this Chapter). 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 on **Page 29** above). 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 (NOTE THAT 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
		6. Activity on aligning the assessments with the GPF
	Day 3	7. Presentation and discussion on the alignment results

¹⁶ Note that if during Stage 1, 2, or 3, the government decides that they only wish to set a benchmark for the “meets” level or the government or 4.I.I Review Panel decide 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
Task 2		8. Presentation on the assessments and the GPLs/GPDs
		9. Activity on matching the assessments with the GPLs/GPDs
Task 3	Day 4	10. Presentation and discussion on the matching results
		11. Presentation on global benchmarking
		12. Presentation on the Angoff method
		13. Activity on Angoff practice
	Day 5	14. Activity on Angoff Round 1
		15. Presentation and discussion of the Round 1 results
		16. Presentation on Angoff Round 2
		17. Activity on Angoff Round 2
		18. Presentation on and completion of the 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. Note that there are references to the facilitation slides for the opening and presentations. There are two sets of slides:

- Group-administered assessments untimed assessments with multiple choice (MC) and constructed response (CR) items, namely CBAs (164 slides).
- Individually administered assessments with timed subtasks, namely EGRA/EGMAs (166 slides); and

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

A. Workshop Day One

I. Opening, introductions, logistics, and agenda

Materials: Facilitation slides, panelist workshop packets (See Page 30 above)

Slides: I-11 (CBA and Timed Assessments)

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 15. TIPS FOR FACILITATORS ON OPENING PRESENTATION

Government officials, 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: 12-28 (CBA and Timed 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 16. TIPS FOR FACILITATORS ON BACKGROUND PRESENTATION

When introducing the GPF and PLT, provide context for the workshop by giving brief background and describing future activities. Use the graphic with the GPF scale, including the four proficiency levels and 3 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: 29-40 (CBA and Timed Assessments)

In this presentation, you will introduce the GPF, including introducing each of the domains, constructs, subconstructs, knowledge and skills covered by the subconstructs, 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, ask panelists to give examples of items that might be used to measure the performance standard described in the GPD, etc.

FIGURE 17. 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: 42-45 (CBA and Timed Assessments) *(Note: you will need to create additional slides for this presentation; the recommendation is one slide per assessment item or pair of 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, etc.). 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” descriptor perform; what about “meets” learners and exceeds learners?) 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 18. 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 meeting, meeting, or exceeding 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 I)

Materials: Facilitation slides, panelist workshop packets

Slides: 46-65 (CBA and Timed Assessments)

In this presentation, you will revisit the GPF, specifically the subconstructs and the knowledge and skills learners need to have to meet the standards described in the subconstructs. You will describe the three-step process panelists will engage in to check the alignment of the assessments with the knowledge and skills described by the GPF (See **Page 10** above 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 (See Slides 57-59) and sample math items included in the CBA slides (See Slides 57-59) that you can use for this purpose, or you can select/develop your own. Note, 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** above.

FIGURE 19. 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 knowledge and/or skills and the GPDs (content and performance standards). Go carefully through the examples and each of the two steps and sub-steps described in Section on Task I beginning on Page 10. 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 I)

Materials: Facilitation slides, panelist workshop packets

Slides: 66-70 (CBA and Timed Assessments)

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 the alignment of the assessment items with the GPF subconstructs. You will explain to the panelists that alignment is conducted between the items and the GPF knowledge and/or skills and at the end, there must be sufficient breadth and depth of alignment for policy linking

to work well.

FIGURE 20. 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-knowledge and/or skill ratings, and submit their form to the content facilitators for analysis by the lead facilitators or data analyst. Panelists should only be aligning to knowledge and/or skills for the relevant grade level, as depicted by the “x’s” 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: 71-77 (CBA and Timed Assessments) *(Note: it is recommended that you create additional slides for this presentation, including one slide per item where there was significant disagreement amongst panelists on the knowledge and/or skills that the item aligned to)*

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 items and alignment ratings where there was a considerable amount of disagreement between panelists on the knowledge and/or skills that the item aligned to. Tips on facilitating this discussion are included in **Table 9** above in the Content Facilitator Training Section.

FIGURE 21. 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 subconstructs in the GPF, namely that the assumption will be that globally 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 assessments and the GPLs/GPDs (Task 2)

Materials: Facilitation slides, panelist workshop packets

Slides: 78-87 (CBA and Timed Assessments)

In this presentation, you will build on the alignment conducted during Task 1 (to the knowledge or skills, also called content standards) 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 above on **Pages 15 and 16** 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 (See Slides 83-86) and sample math items included in the Untimed Assessment slides (See Slides 83-86) that you can use for this purpose, or you can select/develop your own.

FIGURE 22. TIPS FOR FACILITATORS ON THE TASK 2 MATCHING PRESENTATION

Remind panelists that this activity builds on the understanding of the CBA 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 knowledge and/or skill(s) needed 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 knowledge and/or skills from both are required to correctly answer the item.

9. Activity on matching the assessments with the GPLs/GPDs (Task 2)

Materials: Facilitation slides, panelist workshop packets

Slides: 88-95 (Untimed Assessments and Timed Assessments)

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 knowledge and skills 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 23. TIPS FOR FACILITATORS ON OVERSEEING THE TASK 2 MATCHING ACTIVITY

Make sure that 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: 96-104 (Untimed Assessments and Timed 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/how those were resolved, etc. 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 to indicate challenges that may arise early on (e.g., if a grade 3 panel matches an item to a lower grade level than the grade 2 panel). Additional tips on facilitating this discussion are included in **Table 9** above.

FIGURE 24. TIPS FOR FACILITATORS ON REVIEW 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: 105-112 (Untimed Assessments and Timed 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 (See 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 (See Slide 107). You will cover the third graphic in the presentation (See Slide 108) with the percentages of learners in the GPLs (categories) from the assessment data sets, which is used for comparisons, aggregation, and tracking (CAT) on SDG 4.1.1 and USAID indicators.

FIGURE 25. 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: 113-127 (Untimed Assessments); 113-130 (Timed Assessments)

In this presentation, you will explain the standardized process for setting benchmarks using the Yes-No version of the Angoff method (See **Pages 16-19** above). 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 the panels.

FIGURE 26. 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 knowledge and/or skills required to answer the item correctly. A flowchart for the ratings and examples is provided for the panelists in the slides and in **Figure 8** above, along with ratings tips.

13. Activity on Angoff method practice (Task 3)

Materials: Facilitation slides, panelist workshop packets

Slides: 128-133 (Untimed Assessments); 130-135 (Timed Assessments)

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 (See Slides 130-133) and sample math items included in the Untimed Assessments slides (See Slides 132-135) that you can use for this purpose, or you can select/develop your own. Note, sample items should not be too similar to the actual assessment items that panelists will rates, 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 27. 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 knowledge and skills (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: 134-136 (Untimed Assessments); 136-138 (Timed Assessments)

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, which are then averaged as benchmarks for the panels. Panelists will complete their Round 1 ratings individually but can ask one-on-one questions of facilitators during the process.

FIGURE 28. 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 knowledge or skills from the GPF and the GPDs.

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: 137-150 (Untimed Assessments); 139-152 (Timed Assessments)

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) 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 18** above 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** above for tips on how to run this discussion.

FIGURE 29. 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 help panelists better understand the different rating options to better inform their Round 2 ratings. Tips for this discussion are included in **Table 9** above. 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 is 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 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: 151-154 (Untimed Assessments); 153-156 (Timed Assessments)

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. You will tell the panelists that they should use Round 1 as a starting point for making their Round 2 revisions.

FIGURE 30. TIPS FOR FACILITATORS ON PRESENTING ANGOFF ROUND 2

The panelists need to realize that their ratings should change from Round 1 to Round 2 based on an increased level of understanding, both for the panelists themselves and for the panels. This should lead the panelists to become both self-sufficient and 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: 155-157 (Untimed Assessments); 157-159 (Timed Assessments)

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 31. TIPS FOR FACILITATORS ON OVERSEEING ANGOFF ROUND 2 RATINGS

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

18. Presentation on the workshop evaluation

Materials: Facilitation slides, panelist workshop packets

Slides: 158 (Untimed Assessments); 160 (Timed Assessments)

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, a break, etc; if that is the case, the 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. You should be sure to emphasize to the panelists that the evaluations are confidential and that you will not know who rated what; so, they are strongly encouraged to share their honest feedback. This information will inform future workshops.

FIGURE 32. 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: 159-161 (Untimed Assessments); 161-163 (Timed Assessments)

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 format. 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** above.

FIGURE 33. 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: 162-164 (Untimed Assessments), 164-166 (Timed 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. Complete any final logistics and take a group photo, if appropriate.

FIGURE 34. 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.

E. Tips for Hosting Remote Workshops

Tips for hosting remote workshops follow based on the first 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 sub-section on **Panelist Packets** in **Chapter III** above for details).
- Ensure panelists are able to join via a laptop (strongly preferred) or smartphone so that they can see slides and submit tasks. Allow panelists to submit tasks either as soft copies, photos/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), etc.
- 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 invitations 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; 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-workshops 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 security 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 the by "meets" GPDs as part of their inter-session activity.

Discussions

One major disadvantage of remote workshops is that panelists don't 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 panelist 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/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 to 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 re-join the call if they have questions.

CHAPTER V. DOCUMENTING THE WORKSHOP OUTCOMES

A. Production of the technical documentation (after the workshop is completed)

Materials: NA

Slides: NA

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 S - Outline for the Benchmarking Technical Report** 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 – Benchmark Calculations** for the Workshop) for details on how to calculate these); panelist ratings and impact data (See **Annex M - Feedback Data Examples and Instructions**); statistics, such as intra-rater and inter-rater consistency indices and the SEM (See **Annex G - Intra- and Inter-Rater Consistency, and Standard Error of Measurement (SEM)** 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 SEM reports 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 SEM is calculated at the benchmark level. High SEM values (more than two score points) indicate a lack of consistency in panelists' estimated benchmarks, and low values (less than one score point) indicate a high level of consistency in panelists' estimated benchmarks. **Annex G** provides the formal equations and steps for calculating it.

Results of the panelists' workshop evaluations (See **Annex P - Workshop Evaluation Form** for the evaluation form and **Annex T - 4.I.I Review Panel Criteria for Policy Linking Workshop Validity** for details on how this information should be summarized and presented to the 4.I.I Review Panel) provide evidence of how well the policy linking method was implemented and to what extent

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

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

panelists understood, applied, and had confidence in their benchmarks. 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 U - Agreement and Consistency Coefficients**.

Technical documentation (See **Annex V – Technical Documentation of Workshop Outcomes**) for a report template) should be provided to the donor agency (if relevant) and the government (who 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 - Process Documentation Form** to help inform updates to the Toolkit and/or GPF.

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; Brennan & Wan, 2004 for methods on calculating classification accuracy and consistency. Subkoviak's method in **Annex P** is computationally straightforward.

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 **CPLV** document.

A. Collect evidence from the workshop

Materials: NA

Slides: NA

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. The submission of information 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: NA

Slides: NA

Resources: CPLV

UIS has quarterly submission deadlines: March 31, June 30, September 31, 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 their evidence according to the timeline indicated in **Table 7** below.

TABLE 7: TIMELINE FOR SUBMITTING RESULTS TO UIS & RECEIVING RESPONSES

Submission of Documents for Stage 3	Decision from the CPLV and UNESCO (Stage 4)	Policy Linking Workshop (Stage 5)	Submission of Documents for Stage 6	Decision from CPLV and UNESCO (Stage 7)
January	March 31	April – June	By June 30	September 31
February	March 31	April – June	By June 30	September 31
March	March 31	April – June	By June 30	September 31
April	June 30	July – Sept.	By Sept. 31	December 31

Submission of Documents for Stage 3	Decision from the CPLV and UNESCO (Stage 4)	Policy Linking Workshop (Stage 5)	Submission of Documents for Stage 6	Decision from CPLV and UNESCO (Stage 7)
May	June 30	July – Sept.	By Sept. 31	December 31
June	June 30	July – Sept.	By Sept. 31	December 31
July	September 31	Oct. – Dec.	By Dec. 31	March 31
August	September 31	Oct. – Dec.	By Dec. 31	March 31
September	September 31	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 their congressional reporting, along with reporting requirements in terms of quality. Project teams should check with their Contracting Officer’s Representative (COR) as USAID to determine the appropriate timeline for submission of results.

C. Receive a response back from UIS

Materials: NA

Slides: NA

Resources: Annex T - 4.1.1 Review Panel Criteria for Policy Linking Workshop Validity 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 in 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 SEM).
- 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 their recommendation, including stipulating any additional documentation that must be submitted before they can recommend validated outcomes. UIS will share the outcomes with the government and confirm next/final steps.

Once the outcomes of policy linking have been validated by the 4.1.1 CPLV and accepted by UNESCO-UIS, the government can submit the data for reporting against SDG 4.1.1 and/or USIAD'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 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 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 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

- 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, DC: 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, DC: 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*. US Agency for International Development (USAID), Washington, DC.
- 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 - Global Proficiency Framework

The Global Proficiency Framework is a separate document that is included in the toolkit package.

Annex B - Global Minimum Proficiency Levels

Does not meet 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

Activity	Responsible	Deadline	✓	Comments
1. Workshop memo				
a. Update background & methodology				
b. Update on methodology				
c. List participants				
d. Describe logistics				
e. Develop agenda				
2. Participant lists/database				
a. Finalize teacher panelist list				
b. Finalize specialist panelist list				
c. Finalize government/policy-maker list				
d. Finalize international observer list				
3. Reimbursements				
a. Finalize amounts for teacher panelists				
b. Finalize amounts for specialist panelists				
c. Finalize amounts for government				
4. Invitations				
a. Prepare invitations for teacher panelists				
b. Prepare invitations for specialist panelists				
c. Prepare invitations for government				
d. Prepare invitations for observers				
5. Materials				
a. Translate reading GPF into local language				
b. Develop practice passages/questions				
c. Finalize ratings forms and print				
d. Print assessment instruments				
e. Finalize facilitation slides and print				
f. Purchase stationery, name tags, and banner				
g. Finalize daily attendance forms and print				
6. Practice assessments (for panelists)				
a. Prepare instructions				
b. Send to panelists				
7. Wire/cash transfers (to country)				
a. Transfer to account in country				
b. Receive by account in country				
c. Withdraw from bank for payments				
8. Wire/cash transfers (within country)				
a. Transfer to participants				
b. Receive by participants				
9. Conference hotel				

a. Reserve rooms for participants				
b. Pay for rooms				
c. Reserve venue including breakout				
d. Pay for venue				
e. Inspect and arrange venue				
10. National consultants				
a. Recruit candidates				
b. Vet candidates				
c. Onboard consultants				
11. Transportation				
a. Determine transport modes for panelists				
b. Purchase panelist air tickets (if needed)				

Coordinator: _____

Logistician: _____

Annex D - Alignment Rating Form for Task I

These columns are only required where there is partial fit. You can use these to record any other domains, constructs and subconstructs that relate to the item

Question	Domain	Construct reference	Subconstruct reference	Knowledge or skill	Fit	Domain	Construct reference	Subconstruct reference	Fit	Knowledge or skill
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); you will also find math examples items in these slides that can be used for either type of assessment workshop.
- 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.

The slides can be found online.

Annex F - Item Rating Forms

Sample Form I. Assessment with 20 objective (multiple choice) items:

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).

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. Assessment with 5 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).

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. 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, assign 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).

ORAL READING PASSAGE 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	dadɪ	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															

ORAL READING COMPREHENSION IN HAUSA

Item no.	Condition	Questions	Round 1 individual and independent ratings				Round 2 individual and independent ratings			
			JP	JM	JE	AE	JP	JM	JE	AE
1	≤ 9 words attempted	Su waye abokan juna? {Kande da Delu}								
2	≤ 18 words attempted	Ina suke tafiya kullum? {Makaranta}								
3	≤ 22 words attempted	Me Kande ta zo da shi daga makaranta? {Aiki}								
4	≤ 28 words attempted	Wa ya taimaka wa Kande? {Delu}								
5	≤ 35 words attempted	Me ya faru a ajin su Kande? {Kande ta Samu yabo/ yabo}								
Total										

Annex G - Intra- and Inter-Rater Consistency, and Standard Error of Measurement (SEM)

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; the lower number indicates high consistency and higher 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 a 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)
- 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% 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% 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% scores on the subtask or the entire test

P_{ij} = 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% scores on the subtask or the entire test

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%, 26-50%, 51-75%, and 76-100% 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 gets the item right;
- ii. Calculate raw score for each learner by taking sum of correct responses to the items;
- iii. Divide maximum possible score by four to calculate score ranges for four categories (0-25% for partially meets, 26-50% for meets, 51-75% for exceeds, and 76-100% for above exceeds);
- iv. Sort raw score in ascending order, and split learner item response data file into four groups by including learners with 0-25% scores for partially meets, 26-50% for meets, 51-75% for exceeds, and 76-100% for above exceeds;
- v. For each partially meets, meets, exceeds, and above exceeds group, 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 values $|P_{ijk} - P_{ie}|$ 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) 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 I 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,

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

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

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

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)
- R_{bi} = Panelist b 's response to item i ; $k = 1, 2, 3, 4$ (1= partially meets, 4=above exceeds)
- d = Maximum absolute possible difference between two judges' ratings.

As 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 possible maximum absolute difference is 3.

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 1: Calculate the total possible inter-rater inconsistency.

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

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

- i. Take the absolute value of the difference in ratings between each panelist;
- ii. Add together all of the absolute values. The result is the total observed inter-rater inconsistency for 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 of Measurement (SEM)

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

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

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

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

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 Measurement

Calculate the SEM 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 SEM for that benchmark.
5. Repeat steps 1 through 4 as necessary for each benchmark.

Annex H - Workshop Panelist Information

Grade and subject for which panelist will serve as a panelist:

Subject Group: 1) Reading

2) Mathematics

Grade level: _____

Name: _____

Occupation: _____

Address: _____

Email: _____

Cell Number: _____

Gender: 1) Female

2) Male

Ethnicity (if relevant): _____

Education Level: _____

Years of Experience/Expertise: _____

Professional Organization/Affiliation: _____

Prior Training(s) in Reading/Mathematics: 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) Use for Classroom Instruction (for teachers only): _____

Annex I - Sample Invitation Letter for Policy Makers

Note that this includes sample letters and a sample pre-workshop assessment. All should be modified depending on the context.

February 27, 2020

[Name]
Executive Secretary
Nigerian Educational Research and Development Council (NERDC)
Sheda, Abuja, Nigeria.

Invitation to a Policy Linking Workshop

In pursuit of the Sustainable Development Goals on education (SDG 4.1.1), Nigeria has been chosen as a pilot country to test out a global reporting method called *Policy Linking*. This method allows countries to determine whether its learners are reaching global minimum proficiency in reading and mathematics. USAID is using similar indicators for its global reporting.

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

The Hausa Early Grade Reading Assessment (EGRA) will be used to pilot the Policy Linking methodology for early primary assessments in Nigeria. There will be two panels, one for P2 EGRA and one for P3 EGRA. The teachers will be guided through a systematic process that involves reviewing assessment materials and setting benchmarks: for the Primary 2 and 3 Hausa early grade reading assessments (EGRA).

Up to four (4) administrators from NERDC are invited to participate as observers. It 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 that policy linking can be scaled up in Nigeria to assessments for other grade levels, subject areas, and languages. The workshop will take place **Tuesday March 10 to Friday March 13, 2020**. **Registration will be at 8:30am on March 10.**

Activity Name	Arrival Date	Departure Date	Venue
Workshop to set global benchmarks using Hausa EGRA	Tuesday, March 10, 2020 Registration at 8:30am	Friday, March 13, 2020 Last session ends by 4:00pm	Hawthorne Suites by Wyndham at I Uke St, Garki, Abuja

Transportation expenses relating to the participation of the administrators will be covered by Management Systems International (MSI) under contract with the US Agency for International Development (USAID). In addition, lunch and refreshments will be served.

If you have questions or require further clarifications, please contact [Name] via phone [number]. Please kindly confirm your participation by Tuesday March 3, 2020. Your participation in this workshop is crucial and we look forward to collaborating with you.

Sincerely, [Name and Title]

Annex J - Sample Invitation Letter for Workshop Panelists

February 26, 2020

Teachers and Curriculum Specialists

Dear Sir/Madam,

Invitation to a Policy Linking Workshop

In pursuit of the Sustainable Development Goals on education, Nigeria has been chosen as a pilot country to test out a global reporting method called *Policy Linking*. This method allows countries to determine whether its learners are reaching global minimum proficiency in reading and mathematics.

Through Policy Linking, countries will link their national assessments to a common global reporting scale using benchmarks. Setting the benchmarks requires judgments by panels of teachers.

The Hausa Early Grade Reading Assessment (EGRA) will be used to pilot the Policy Linking methodology in Nigeria. Accordingly, 30 Hausa Primary 2 and 3 teachers – 15 P2 and 15 P3 – are required to participate on two assessment panels. These teachers will come from the following states:

Bauchi = 10 (5 P2 and 5 P3)

Sokoto = 10 (5 P2 and 5 P3)

Zamfara = 10 (5 P2 and 5 P3)

Total = 30 (15 P2 and 15 P3)

The teachers will be guided through a process to set the benchmarks. Participation in the workshop will provide a valuable learning opportunity for the selected teachers. As one of those teachers, you are invited to this five-day workshop, which will be held from **Monday March 9 to Friday March 13**. Teachers will arrive and depart from Abuja as follows:

Activity Name	Arrival Date	Departure Date	Venue
Workshop to set global benchmarks on learner performance using Hausa EGRA	Sunday, March 8, 2020 (all participants)	Saturday, March 14, 2020 (Sokoto and Zamfara) Sunday, March 15, 2020 (Bauchi)	TBD in Abuja (information will be communicated as soon as possible)

Note that all teachers will arrive in Abuja by plane. The arrival date is Sunday March 8th due to flight schedules from Bauchi and Sokoto.

All expenses relating to your participation will be fully covered by Management Systems International (MSI) under contract with the US Agency for International Development (USAID). This includes transportation, accommodation, feeding at the workshop, and per diem (daily allowance).

If you have questions or require further clarifications, please contact [Name] via phone [Telephone number]. Due to the need to confirm plane tickets, please kindly confirm your participation on or before 10am on March 2, 2020.

Your participation in this workshop is crucial and we look forward to you joining us.

Sincerely, [Name and Title]

Annex K – Sample Explanation for Panelists of Pre-Workshop Activity

Pre-workshop Activity for Teachers: Rapid Reading Assessment (RRA)

Instructions

Each teacher should administer this passage to selected learners in their P2 or P3 classrooms.

- 1) Print or write out the reading passage. It has 35 words. If printed, make sure that the characters are large enough. 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 2:**
 - **Partially Meets Global Minimum Proficiency Learner:** Learners who can say or sign accurately some words in a grade 2-level continuous text, generally very common and simple words.
 - **Meets Global Minimum Proficiency Learner:** Learners who can say or sign accurately a grade 2-level continuous text with few errors (e.g., no more than 10 percent of the words in the text).
 - **Exceeds Global Minimum Proficiency Learner:** Learners who can say or sign accurately a grade 2-level continuous text with no errors.
 - **Grade 3:**
 - **Partially Meets Global Minimum Proficiency Learner:** Learners who can say or sign accurately a grade 3-level continuous text, at a pace that is slow by country standards for fluency for the language in which the assessment is administered (e.g., often word-by-word).
 - **Meets Global Minimum Proficiency Learner:** Learners who can 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 Global Minimum Proficiency Learner:** Learners who can 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.
- 3) One-by-one, have each learner read the passage on the paper. Allow one minute on your watch.
- 4) As the learner reads, count the following:
 - the number of words s/he has attempted to read in a minute
 - the number of words s/he has read correctly in a minute.
- 5) Write both numbers – attempted words and correctly read words -- on the paper next to the learner’s name. Note: Please bring this paper to the workshop.
- 6) If the child cannot read any of the words on the first line correctly, then discontinue the test.
- 7) If a child hesitates or stops on a word for 3 seconds, say “ci gaba.”
- 8) After s/he has finished reading as much of the text as possible, ask the comprehension questions.
- 9) Only ask the comprehension questions that correspond to the text that s/he has read.
- 10) If the learner does not give a response right away, silently count to 10 and ask the next question, or stop altogether if that is the last question.
- 11) Write down the number of questions that the learner answered correctly.
- 12) Proceed to the next child.

13) Have the teacher complete the rapid reading assessment for all nine (9) learners. It should take no longer than 30-45 minutes of your time altogether.

Reading passage	# of words	Comprehension questions
Kande da abokiyarta Delu sukan tafi Makaranta tare kullum.	9	1. Su waye abokan juna? [Kande da Delu]
Wata rana Kande ta zo da aiki daga makaranta.	18	2. Ina suke tafiya kullum? [Makaranta]
Delu ta taimaka mata.	22	3. Me Kande ta zo da shi daga makaranta? [Aiki]
Kande ta samu yabo a ajinsu.	28	4. Wa ya taimaka wa Kande? [Delu]
Kande da Delu Sun ji dafi sosai.	35	5. Me ya faru a ajin su Kande? [Kande ta Samu yabo/ yabo]

Annex L - Pre-Workshop Statistics

The Data Analyst and/or Lead Facilitator should calculate the following statistics before the policy linking workshop:

Item difficulty

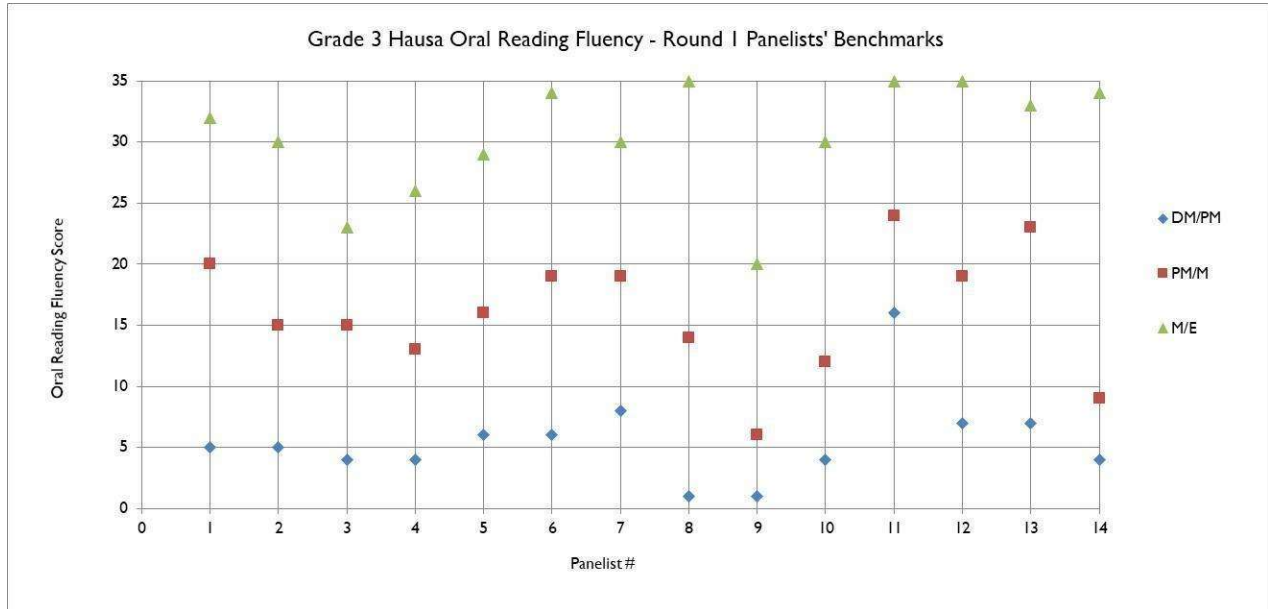
Item difficulty informs facilitators and panelists on how difficult an item is based on how learners performed on the item in the most recent iteration of the assessment. The data analyst should calculate the empirical item difficulty level (i_e) and conditional item difficulty levels (i_{jk}) for learners with 0-25%, 26-50%, 51-75%, and 76-100% scores 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;
2. Calculate the raw score for each learner by taking sum of correct responses to the items;
3. Divide maximum possible score by four to calculate score ranges for four categories (0-25% for partially meets, 26-50% for meets, 51-75% for exceeds, and 76-100% for above exceeds);
4. Sort raw score in ascending order, and split learner item response data file into four groups by including learners with 0-25% scores for partially meets, 26-50% for meets, 51-75% for exceeds, and 76-100% for above exceeds;
5. For each partially meets, meets, exceeds, and above exceeds group, calculate conditional item difficulty level (i_{jk}) for each item by calculating taking the proportion of learners who gets the item right.

Annex M - Feedback Data Examples and Instructions

Normative information (sometimes called location statistics)

After each round of ratings, the data analyst will create a graph like the one below 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.



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 percentage of learners who took the assessment that received an overall score of zero through the highest score possible on the assessment. They will use that information to prepare a table like those presented in the second and third 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 first table below.

Grade 3 Hausa Oral Reading Fluency - No. of Words Learners <u>Attempted</u> 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	

Grade 3 Hausa Oral Reading Fluency - No. of Words Learners <u>Read Correctly</u> 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
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	

Grade 3 Hausa Reading Comprehension - No. of Items Learners Who Answered Correctly

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	

Impact information

To generate the impact information, the data analyst will 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.

Grade 3 Hausa Oral Reading Fluency and Comprehension - Benchmarks with No Standard Error of Measurement		
Categories	Score Range	% of Learners
Does not meet	0-6	59.8
Partially meets	7-18	10.1
Meets	19-34	13.3
Exceeds	35-40	16.8
Total		100.0

Annex N - Sample Agenda for an In-Person Workshop

Time	Day 1	Facilitation
08:30 – 09:00	Registration	Project team
09:00 – 10:00	Opening, introductions, agenda, and logistics	Government, donors, and IPs (if relevant) as well as lead facilitators
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 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: Assessments and GPDs/GPLs	Lead facilitators
11:00 – 11:15	Tea break	--
11:15 – 12:30	Task 2 Activity: Match between assessments and GPDs/GPLs	All facilitators

12:30 – 13:30	Lunch break	--
13:30 – 15:45	Task 2 Activity: Match between assessments and GPDs/GPLs (cont.)	All facilitators
15:45 – 16:00	Tea break	--
16:00 – 17:00	Task 2 Activity: Match between assessments 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 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	MOE, USAID

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. They will also want to establish the actual start times for each of the activities.

Preparation session I – <Date and start time; recommend holding two weeks before workshop session I>

Timing	Activity	Facilitator
0-15 mins	Welcome and introductions	Lead facilitator
15-40 mins	Overview of policy linking	Lead facilitator
40-55 mins	Purpose of preparation session	Process facilitator
55-60 mins	Comfort break	
60-80 mins	Overview of the GPF	Lead or content facilitator
80-100 mins	<Grade and Subject> GPF Review	Lead or content facilitator
100-110 mins	Explanation of inter-session activities	Lead facilitator
110-120 mins	Closing remarks	Lead facilitator

Panelist inter-session activities:

- Review <grade and subject> GPF, and identify any elements that are unclear (submit 1 week prior to workshop)

Preparation session 2 – <Date and start time; recommend holding two days after preparation session 1>

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

Panelist inter-session activities:

- Administer the <assessment> to 3 learners (from the appropriate grade/age group for each GPL)

Workshop session I – <Date and start time>

Timing	Activity	Facilitator
0-10 mins	Welcome and purpose of session I	Lead facilitator
10-55 mins	Review GPF activity and provide clarification	Content or lead facilitator
55-60 mins	Comfort break	
60-105 mins	Discussion of <assessment> administration activity	Content or lead facilitator
105-120 mins	Evaluation approach and completion of evaluation I	Lead facilitator

Workshop session 2 – <Date and start time>

Timing	Activity	Facilitator
0-10 mins	Welcome and purpose of session 2	Lead facilitator
10-20 mins	Address any concerns raised in evaluation 1	Content or lead facilitator
20-55 mins	Introduction to alignment task (Task 1)	Lead facilitator
55-60 mins	Comfort break	
60-90 mins	Small group discussions on first 5 items ²¹	Content facilitators ^[2]
90-110 mins	Plenary discussion on questions that came up in the groups	Lead facilitator
110-120 mins	Explanation of inter-session activities and close	Lead facilitator

Panelist inter-session activities:

- Complete Task 1 - alignment review on all remaining items (submit 4 hours after session)
- Complete evaluation 2 (submit with alignment review)

²¹ Each small group will have a content facilitator; recommend lead facilitator(s) stay out of the small groups so that the small groups can identify what questions they have and bring them back to the plenary.

Workshop session 3 – <Date and start time>

Timing	Activity	Facilitator
0-10 mins	Welcome and purpose of session 3	Lead facilitator
10-40 mins	Review inter-session activities and provide clarification	Content facilitator
40-55 mins	Introduction to Task 2 – Matching to the GPLs and GPDs	Lead facilitator
55-120 mins	Practice with Task 2	Lead facilitator
120-130 mins	Comfort break	
130-230 mins	Small groups complete Task 2 together (groups organized by grade/subject/language) ²²	Content facilitator
230-240 mins	Explanation of inter-session activities and close	Lead facilitator

Panelist inter-session activities:

- Complete evaluation 3 (submit 1 hour after close of session)

²² Ibid.

Workshop session 4 – <Date and start time>

Timing	Activity	Facilitator
0-10 mins	Welcome and purpose of session 4	Lead facilitator
10-40 mins	Present Angoff methodology and Task 4 and provide clarification	Lead facilitator
40-75 mins	Small group Angoff ratings using practice items	Content or lead facilitator
75-80 mins	Comfort break	
80-100 mins	Plenary discussion of questions that arose in small groups	Lead facilitator
100-110 mins	Start Round 1 ratings (raise questions that come up)	Independent work
110-120 mins	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)
- Complete Round 1 ratings on all remaining items (submit 4 hours after close of session or 1 hour after one-on-one meeting with a lead facilitators, whichever comes later)
- Complete evaluation 4 (submit with Round 1 ratings)

Workshop session 5 – <Date and start time>

Timing	Activity	Facilitator
0-10 mins	Welcome and purpose of session 5	Lead facilitator
10-45 mins	Review and discuss Round 1 ratings in plenary	Content facilitator
45-50 mins	Comfort break	
50-110 mins	Review Round 1 ratings in small groups (organized by grade/subject/language), going through each item where there was disagreement	Content facilitator
110-150 mins	Share and discuss item-difficulty and impact data	Lead facilitator
150-180 mins	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 4 hours after close of session or 1 hour after one-on-one meeting with a lead facilitators, whichever comes later)
- Complete evaluation 5

Workshop session 6 – <Date and start time>

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

Annex P - Workshop Evaluation Form

Part I: Training on Global Proficiency Descriptors

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.

GPD training	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I understand the purpose of the 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 and to ask questions					
The amount of time spent on the GPD training was sufficient					

Do you have any additional comments on the GPD training?

Part II: Training on the assessment and policy linking method

Today you have been trained on the assessment on which we are undertaking the policy linking and the policy linking 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.

Assessment training	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I understand the purpose of the assessment					
I understand the constructs assessed in the assessment					
Administering the assessment helped me to understand how minimally proficient learners would perform on the assessment (Note: Not applicable for Group Administered assessment)					
The amount of time spent on the assessment training was sufficient					

Do you have any additional comments on the assessment training?

Policy linking training	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I understand the process I need to follow to complete the policy linking exercise					
I understand the difficulty level of the assessment items					
The discussion of the procedure was sufficient to allow me to feel confident in making decisions					
The practice exercise helped me to understand what I need to do					
There was an equal opportunity for everyone to contribute their ideas and opinions and to ask questions					
The amount of time spent on the policy linking method training was sufficient					

Do you have any additional comments on the policy linking training?

Part III: Round I evaluation

During Round I, you were asked to predict whether minimally proficient learners would be able to answer the questions correctly.

Round 1	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I am confident about the performance predictions I made during Round 1					
I was able to follow the instructions and complete the Round 1 form accurately					
I was given sufficient time to complete the Round 1 performance predictions					

Do you have any additional comments on Round I?

Part IV: Round 2 evaluation

During Round 2, you were given actual performance information and data about the impact of using the Round 1 results. You then were asked to give revised performance predictions. Please select the best answer below.

Round 2	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I understand the data on others' ratings, the item difficulty data, impact data, etc.					
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					
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

Coming soon.

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. The first three equations 1-3 are used to calculate benchmarks for each panelist and the last three equations 4-6 are used to calculate benchmarks recommended by the panel. For these equations, i indicates the items or words, j indicates panelists, l indicates number of item or words attempted by JP, m indicates number of items or words attempted by JM, and n indicates number of items or words attempted words by JE.

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 the 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 - Outline for the Benchmarking 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. Quality Assurance Process 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. Conclusions and Recommendations
8. References
9. 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 T - 4.1.1 Review Panel Criteria for Policy Linking Workshop Validity

Question	Criteria	Materials
a. 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 his/her scores can be identified across rounds.
b. 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.
c. What was the <i>Standard Error of Measurement</i> (SEM) at each <i>global proficiency level</i> ?	<i>SEM</i> should be appropriate for each <i>global proficiency level</i> reported. There is no maximum <i>SEM</i> provided in this document, since it will depend on the number of items in the assessment.	Countries should provide the <i>SEM</i> and details of how the <i>SEM</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.
d. 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.
e. To what extent did the panelists meet the other selection criteria described in the Policy	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 	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

<p>Linking Toolkit?</p>	<p>(classroom teachers)</p> <ul style="list-style-type: none"> ● 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 minimum proficiency level and some who would meet the requirements of the exceeds minimum proficiency level (all panelists) ● Knowledge of the instructional environment (all panelists) ● Experience administering the assessment(s) being used for the policy linking workshop. 	<p>include questions about their exposure to the assessment ahead of the workshop and during the workshop, assess their knowledge of the instructional environment, etc.</p>
<p>f. 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>?</p>	<p>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.</p>	<p>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.</p>

Annex U - Agreement and Consistency Coefficients

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.

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

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 2**, 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.

TABLE 2. 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. The table reveals that the consistency coefficient is 0.56, indicating

the assessment procedure is adding only modestly to consistency in decision making.

TABLE 3. 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 V – Technical Documentation of Workshop Outcomes

Coming soon.

Annex W - Process Documentation Form

Coming soon.