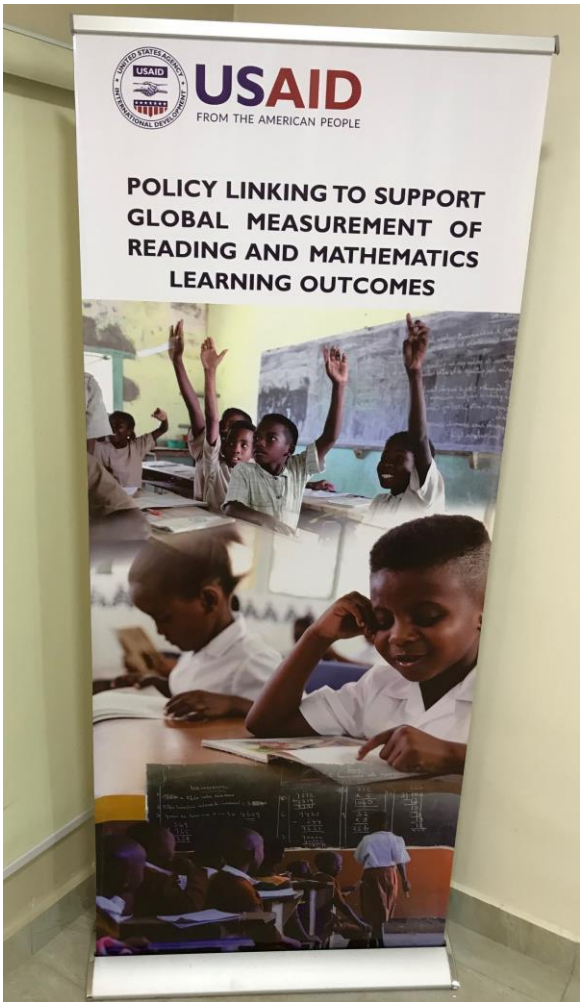
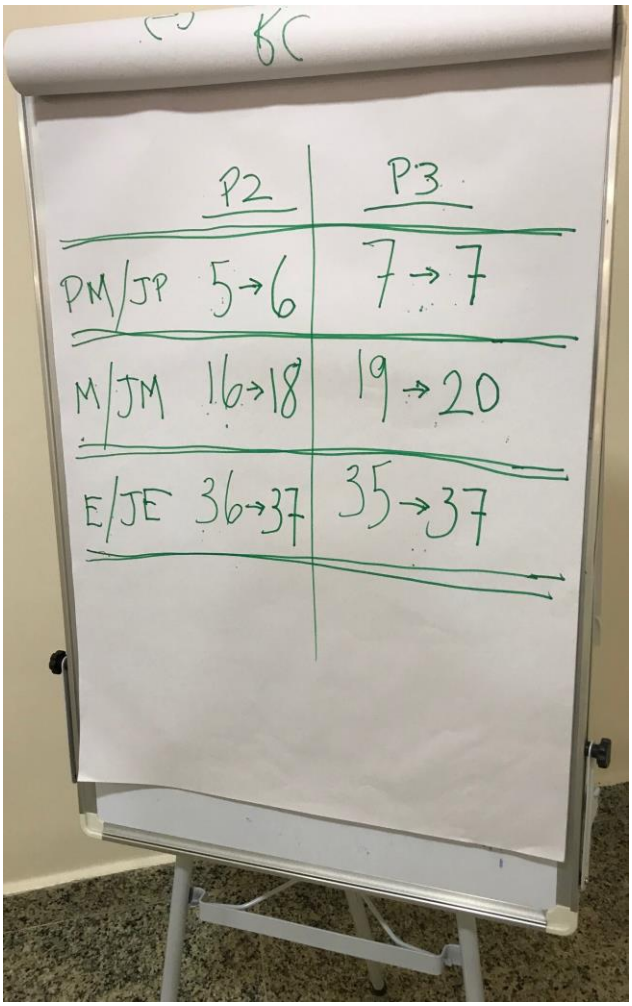
DRAFT Nigeria Policy Linking Pilot Workshop Report: Setting Global Benchmarks for Grades 2 and 3 Early Grade Reading Assessments in Hausa Language





US Agency for International Development (USAID)
Nigerian Federal Ministry of Education (FME)
Universal Basic Education Commission (UBEC)
Nigerian Educational Research and Development Council (NERDC)

March 20, 2020
Management Systems International (MSI)

Acknowledgements

The technical team at Management Systems International (MSI) is grateful for the support provided by several stakeholders and participants for the policy linking pilot workshop in Nigeria.

First, the overall support by officials from the US Agency for International Development (USAID) and its Northern Education Initiative Plus (NEI+) project was invaluable in organizing and launching the workshop.

Second, the administrative support by officials from the Nigerian Federal Ministry of Education (FME) and the Universal Basic Education Committee (UBEC) was instrumental in planning the workshop.

Third, the technical support by officials from the Nigerian Educational Research and Development Council (NERDC) was important in implementing the workshop.

Fourth, the hands-on participation by the panelists – head teachers, teachers, and specialists – from the Northern States and NERDC was critical to obtaining the results from the workshop.

Fifth, the leadership and logistical support by the dedicated personnel – content facilitators, project manager, and logistician -- was key in ensuring the start-to-finish roll-out of the workshop.

We thank all of these stakeholders and participants for making the workshop a success.

Dr. Abdullah Ferdous – Co-Lead Facilitator (MSI)

Dr. Jeff Davis – Co-Lead Facilitator (MSI)

Dr. Chizoba Imoka – Coordinator (MSI)

Table of Contents

Acknowledgements 2

Acronyms and Abbreviations 4

Overview 5

- Development 5
- Piloting 7
- Finalization 7

Preparation 8

- Planning 8
- Assessments 8
- Panelists 9
- Benchmarks 10

Implementation 10

- Task 1 10
- Task 2 14
- Task 3 15

Results 18

- Round 1 19
- Round 2 19
- Consistency 20

Recommendations 21

- Task 1 21
- Task 2 22
- Task 3 22
- Results 23

References 23

Acronyms and Abbreviations

ACER	Australian Council for Educational Research
CAT	Compared, Aggregated, and Tracked
DFID	UK Department for International Development
EGRA	Early Grade Reading Assessment
FME	Federal Ministry of Education
GAML	Global Alliance to Monitor Learning
GPD	Global Proficiency Descriptor
GPL	Global Proficiency Level
GPF	Global Proficiency Framework
IAEG-SDG	Inter-Agency and Expert Group on SDGs
IBE-UNESCO	International Bureau of Education – UNESCO
M&E	Monitoring and Evaluation
MSI	Management Systems International
NERDC	National Council of Educational Research and Training
PLT	Policy Linking Toolkit
PLWG	Policy Linking Working Group
SDG	Sustainable Development Goal
UBEC	Universal Basic Education Commission
UIS	UNESCO Institute for Statistics
UNESCO	UN Educational, Scientific and Cultural Organization
USAID	US Agency for International Development

Overview

The development and piloting of a policy linking method for reporting on Sustainable Development Goal (SDG) Indicator 4.1.1 and US Agency for International Development (USAID) Standard and Supplemental Foreign Assistance Indicators has been in process since 2017. It is anticipated that the method will be finalized by September 2020 for reporting starting in late 2020. The chronology of development, piloting, and finalization below provides an overview.

Development

In September 2015, the SDGs were agreed upon within a resolution adopted in the United Nations General Assembly that featured the 2030 Agenda for Sustainable Development. This included Goal 4.1:

> By 2030, ensure that all girls and boys complete free, equitable, and quality primary and secondary education leading to relevant and effective learning outcomes.

In March 2016, SDG Indicator 4.1.1 was accepted by the UN Statistical Commission for the global indicator framework, with the UNESCO Institute for Statistics (UIS) designated as the custodian for reporting on the indicator. The indicator was developed in collaboration with the USAID and other donor agencies:

> 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) math, by sex."

In September 2017, the Global Alliance to Monitor Learning (GAML) met in Hamburg, Germany to present and resolve issues with reporting on SDG Indicator 4.1.1. The main issue involved setting valid and reliable global benchmarks on a wide variety of national and cross-national assessments. Different benchmarks for each assessment, based on their difficulty level, would allow UIS to calculate the proportions of learners meeting minimum proficiency by country. UIS could then aggregate the (weighted) proportions for global reporting. Management Systems International (MSI) proposed policy linking as a psychometrically acceptable and practical method for setting the global benchmarks on each assessment.

In August 2018, USAID and UIS co-sponsored a workshop in Washington, DC with more than 80 representatives from ministries of education, multilateral and bilateral donors, foundations, assessment organizations, and implementing partners to discuss the feasibility of using policy linking as a method for reporting on SDG Indicator 4.1.1 (as well as on related indicators of other donors, including USAID). MSI prepared a *Policy Linking Justification Paper* for the workshop, in which it proposed developing a common, non-statistical scale with a step-by-step benchmarking process. MSI also prepared and led the technical sessions for the workshop. The representatives accepted policy linking as a method for aggregating and reporting assessment data. They developed an initial scale with four *Global Proficiency Levels* (GPLs) and labels, along with general definitions for each level. The GPLs were 1) does not meet, 2) partially meets, 3) meets, and 4) exceeds global minimum proficiency. Agreement on the GPLs was a first step, as well as a pre-requisite, to moving forward with further development of the policy linking method.

In October 2018, policy linking was presented and approved by the Inter-Agency and Expert Group on SDGs (IAEG-SDG) at a meeting in Stockholm, Sweden as a method for advancing the classification of SDG Indicator 4.1.1 from Tier III to Tier II for global reporting:

> Tier II: Indicator is conceptually clear, has an internationally established methodology and standards are available, but data are not regularly produced by countries.
>
> Tier III: No internationally established methodology or standards are yet available for the indicator, but methodology/standards are being (or will be) developed or tested.

In April and May 2019, USAID and UIS co-sponsored two workshops in Washington, DC with 30 international subject matter experts in primary school reading and mathematics. Again, led by MSI, the experts continued with the development of the policy linking method. This step involved collaborating with the experts to expand on an initiative by the International Bureau of Education (IBE-UNESCO) to develop consensus global content frameworks drawn from the curriculum and assessment frameworks of over 100 countries. The experts reorganized and adapted the global content in the form of draft *Global Proficiency Descriptors* (GPDs) in grades 2 through 6 in reading and mathematics. The GPDs are organized hierarchically by domains, constructs, subconstructs, and knowledge or skill for each grade and subject. Each knowledge or skill is further described in terms of global minimum proficiency for the GPLs. The GPDs, with the GPLs, formed the draft *Global Proficiency Framework* (GPF). It serves as a common non-statistical scale for linking different assessments. (Note that the GPF is available as a separate document and it is explained below, in this report, under Task 1 of the policy linking workshop.)

In August 2019, MSI began collaborating with the newly formed Policy Linking Working Group (PLWG) to review the draft GPF and share information on plans for piloting the method. The PLWG is comprised of representatives from many of the agencies and organizations who participated in the August 2018 workshop, e.g., USAID, UIS, UNESCO, World Bank, DFID, Gates Foundation, ACER, and MSI.

In September 2019, MSI, in collaboration with USAID and the PLWG, led the development of a draft *Policy Linking Toolkit* (PLT) to provide guidance for pilot workshops. The PLT presents a brief rationale for policy linking along with step-by-step guidance on 1) preparing for workshops, 2) checking the alignment between assessments and the GPF, 3) matching assessment items with the GPLs, 4) setting the global benchmarks (using the Angoff method), 5) calculating reliability indicators for the benchmarking, 6) finalizing the results, and 7) writing the technical report. Annexes to the PLT include forms for implementing the workshops, formulas for calculating reliability indicators, and an outline for writing the workshop technical report.

In October 2019, policy linking was again presented and approved by the IAEG-SDG at a meeting in Addis Ababa, Ethiopia as a method for advancing the classification of SDG Indicator 4.1.1 from Tier II to Tier I for global reporting, with the commitment from 146 out of 193 countries to providing data for reporting:

> Tier 1: Indicator is conceptually clear, has an internationally established methodology and standards are available, and data are regularly produced by countries for at least 50 per cent of countries and of the population in every region where the indicator is relevant.

In October 2019, USAID published its Education Reporting Guidance, with support from EnCompass and MSI. Several of the Standard and Supplemental Foreign Assistance Indicators in the guidance are relevant to policy linking and SDG Indicator 4.1.1:

> 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 a 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.
>
> Supp-3 Percent of learners who attain minimum grade-level proficiency in math at the end of grade 2 with USG assistance.
>
> 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 who attain 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 with USG assistance.

The USAID and UIS indicators require student assessment data that can be *compared* and *aggregated* on a global basis, as well as *tracked* over time (abbreviated at CAT). To accomplish this, each national assessment must be linked to a common global reporting scale, as provided in the GPF. The GPF and PLT are referenced in the USAID Education Reporting Guidance as part of the policy linking method for linking the assessments to the GPF, and consequently to each other.

Piloting

In October and November 2019, UIS, with approval from the ministries of education in Bangladesh and India, along with technical support from MSI and financial support from DFID and Gates, funded policy linking pilots in those two countries. The workshops resulted in setting provisional global benchmarks on the grade 3 and 5 Bangladesh and India national assessments in language and mathematics.

In March 2020, USAID funded a policy linking pilot in Nigeria. Another pilot will be held in Kenya. The result of these workshops is setting global benchmarks on Early Grade Reading Assessments (EGRAs) at grade 2 (Kenya) and grades 2 and 3 (Nigeria), along with benchmarks on curriculum-based assessments (CBAs) at grade 3 in language and mathematics in Kenya. There is also the possibility at a later date in both countries of setting global benchmarks on national assessments in language and mathematics at the end of upper primary and lower secondary.

There will be additional USAID-funded pilots, such as in Djibouti. The World Bank is planning pilots in the Gambia and Ghana. The extent of other pilots will depend on the interest level by countries and donor agencies, along with the need to gather information for specific grades, subjects, types of assessments, and geographic areas. All pilots are planned for completion by September 2020.

Finalization

After the pilots, MSI, under USAID funding and with the collaboration of the PLWG, will finalize the GPF and PLT, after which time they will be disseminated by USAID, UIS, and other agencies in late 2020. USAID and UIS plan to hold training sessions and webinars to build capacity for measurement experts who wish to facilitate policy linking workshops. With the tools and training, countries will be able set global benchmarks on their national assessments. Similarly, agencies leading cross-national assessments will have the tools and training to implement policy linking procedures to set their global benchmarks.

In summary, national and cross-national global benchmarks on different assessments will allow USAID, UIS, and other agencies to calculate the percentages of learners achieving a global minimum proficiency level. Through applying the common reporting scale and benchmarking method to the assessments and data sets through policy linking, they will have the information for three national, regional, and global purposes: 1) *comparing* assessment results for drawing lessons learned, 2) *aggregating* assessment results for reporting on indicators, and 3) *tracking* assessment results for measuring progress over time.

The procedures used in the workshop for preparing the sessions, implementing the tasks, and calculating the results are presented in the sections below. Each section concludes with brief comments about the piloting, including suggestions for modifications that can be applied to subsequent pilot workshops. The comments and suggestions from each section are summarized in the final part of this report.

Preparation

Planning

With the publication of the draft GPF and PLT, as well as the implementation of the policy linking pilot workshops in Bangladesh and India, USAID began preparing for the Nigeria pilot. The objective was setting global benchmarks on the 2018 Early Grade Reading Assessments (EGRAs) in Hausa language at grades 2 and 3, as well as adapting the policy linking method for use with EGRAs. The FME, UBEC, and NERDC approved a four-day workshop in Abuja from Tuesday March 10 to Friday March 13, 2020. MSI assigned two co-lead facilitators and a senior project manager for the workshop. They recruited a workshop coordinator, two Hausa language content facilitators, and a logistician. NEI+ provided advice and support based on their February 2020 workshops to develop national reading frameworks using the GPF.

For the policy linking workshop, the co-lead facilitators prepared three tasks in relation to the GPF: 1) checking the alignment of the assessments with the domains, constructs, and subconstructs, 2) matching the assessment items with the GPLs and GPDs, and 3) implementing the Angoff method to set global benchmarks on the assessments for the GPLs. They also prepared for the analysis of the workshop results, including the alignment and benchmarks, as well as the technical report.

Assessments

According to a study for NEI+ (Evans, 2019) of student assessments in Nigeria, there have been 12 different EGRA administrations in the Northern states between 2010 and 2018. The most recent EGRAs were conducted under USAID's NEI+ and UNICEF/DFID's Reading and Numeracy Activity (RANA). Each of the EGRAs in the table below had large statewide samples of public schools.

Table 1: EGRAs in Northern Nigeria 2016 to 2018

Year	State	Grade	Language	Donor Agency	Implementing Partner
2016	Bauchi, Sokoto	P2, P3	English, Hausa	USAID	Creative (NEI+)
2016	Katsina, Zamfara	P2	Hausa	UNICEF/DFID	FHI 360 (RANA)
2018	Bauchi, Sokoto	P2, P3	English, Hausa	USAID	Creative (NEI+)
2018	Katsina, Zamfara	P1, P2, P3	Hausa	UNICEF/DFID	FHI 360 (RANA)

The assessments used in this workshop were administered by NEI+ for the 2018 midline to P2 and P3 students in Hausa language in samples of schools in Bauchi and Sokoto states. Assessors were drawn from state education institutions, in particular the State Universal Basic Education Boards (SUBEBs) and Local Government Education Authorities (LGEAs). The goals of the assessments were to 1) confirm the appropriateness of the NEI+ approach to improving Hausa reading outcomes, 2) provide insights into potential challenges with implementation, and 3) identify potential limitations of expected outcomes.

The baseline and midline were conducted in random samples of 50 schools in each of Bauchi and Sokoto. The target sample per grade level per school was 12 students. There were 2,330 students in the baseline sample (97 percent of target) and 2,408 students in the midline sample (100 percent of target).

Table 2: EGRA Student Samples

Bauchi P2		Bauchi P3		Sokoto P2		Sokoto P3	
Baseline	Midline	Baseline	Midline	Baseline	Midline	Baseline	Midline
575	602	575	599	597	608	583	599

The EGRAs in Hausa had five subtasks. The same subtasks were administered to the P2 and P3 students. Two of the subtasks were aligned with the GPF – oral reading fluency (ORF) and reading comprehension – but the other three subtasks were not aligned. The 2018 midline tools included an ORF passage with 35 words and five reading comprehension questions, which meant that 40 items from the EGRAs were used for the policy linking. The reading comprehension questions were aligned with parts of the passage and only asked to the student if they had read the corresponding part of the text. Note that the texts and questions in the baseline and midline were different, but they were statistically equated to allow for cross-sectional comparisons over time. The analysis of the baseline and midline for ORF and reading comprehension showed that: 1) the scores were low and 2) they improved (increased).

Table 3: ORF and Reading Comprehension Raw Scores

Subtask	Bauchi P2		Bauchi P3		Sokoto P2		Sokoto P3	
	Baseline	Midline	Baseline	Midline	Baseline	Midline	Baseline	Midline
Oral reading fluency (CWPM)	3.0	7.3	5.1	18.0	1.9	3.9	4.9	10.7
Reading comprehension (out of 5)	0.2	0.5	0.4	1.4	0.1	0.2	0.3	0.8

One reason for the low raw scores for ORF and reading comprehension is the percentage of students with zero scores on the subtasks. The analysis of the baseline and midline zero scores for ORF and reading comprehension showed that: 1) the percentages were high and 2) they improved (declined).

Table 4: ORF and Reading Comprehension Zero Scores

Subtask	Bauchi P2		Bauchi P3		Sokoto P2		Sokoto P3	
	Baseline	Midline	Baseline	Midline	Baseline	Midline	Baseline	Midline
Oral reading fluency (CWPM)	79%	74%	72%	50%	91%	82%	77%	58%
Reading comprehension (out of 5)	89%	82%	81%	60%	94%	90%	86%	69%

P2 benchmarks for Hausa were set by the NEI+ project of 20 CWPM for ORF and 40 percent for reading comprehension. The analysis of the baseline and midline percentages of students achieving or exceeding the benchmarks showed that: 1) the percentages were low and 2) they improved (increased). Note that P2 benchmarks existed for Hausa and not English, while P3 benchmarks existed for English and not Hausa.

Table 5: ORF and Reading Comprehension Percentages Meeting Benchmarks

Subtask	Bauchi P2		Sokoto P2	
	Baseline	Midline	Baseline	Midline
Oral reading fluency (CWPM)	7%	16%	4%	9%
Reading comprehension (out of 5)	7%	14%	3%	5%

Note that the data in this section were presented separately for Bauchi and Sokoto. This followed the data analysis in the technical report. The data in the results section below are combined across states.

Panelists

Following the PLT, MSI set a target of 15 panelists for each of P2 and P3, or a total of 30 panelists. MSI

and UBEC agreed to over-invite these numbers of panelists due to anticipated logistical and other problems with teachers taking leave from their schools and traveling to Abuja for the workshop. UBEC invited 18 panelists per grade level, or a total of 36 panelists, with15 Hausa teachers for each of P2 and P3 (30 total) and three NERDC Hausa curriculum specialists for each grade level (six total).

The invited teachers were from seven Northern States, with concentrations in the intervention states of the NEI+ (Bauchi and Sokoto) and RANA (Katsina and Zamfara) projects. UBEC and MSI made multiple attempts to call each of the panelists to confirm their participation. For the workshop, 13 of the 15 teachers confirmed and participated in P2 while 11 of the 15 teachers confirmed and participated in P3. All six of the Hausa curriculum specialists confirmed and participated, which gave panel sizes of 16 for P2 and 14 for P3. This was close to the original targets and adequate for policy linking.

Table 6: Panelists, including Teachers and Specialists

Panelists	P2		P3	
	Invited	Participated	Invited	Participated
Teachers	15	13	15	11
Specialists	3	3	3	3
Total	18	16	18	14

Benchmarks

In the technical preparations for setting the global benchmarking workshop, MSI's lead facilitators collaborated with NEI+ to obtain the 2016 and 2018 EGRA tools, answer keys, data sets, and reports. MSI produced the training slides, rating forms, and spreadsheets for determining the alignment percentages, benchmarks, and feedback data, i.e., the impact data and reliability estimates (see below). The preparation for benchmarking pilot in Nigeria reflected a variation on the policy linking method used in previous workshops for CBAs. This was necessary due to differences in the formats of the CBAs (in Bangladesh and India) and EGRAs (in Nigeria).

As shown in detail in the implementation section below, the workshop employed a Yes-No variation of the Angoff method (Plake, Buckendahl, & Ferdous, 2005) to set the benchmarks. For policy linking, this involves three tasks, as adapted from a process that is widely accepted for benchmarking workshops (Cizek & Bunch, 2007). In Task 1, the panelists judge the alignment of the assessments in relation to the GPF. In Task 2, they match the assessment items with the GPLs and GPDs based on the skills and abilities needed by learners to answer the items correctly. In Task 3, they use the Angoff method to set initial and final benchmarks on the assessments – i.e., through two rounds of benchmarking – for each grade level.

Implementation

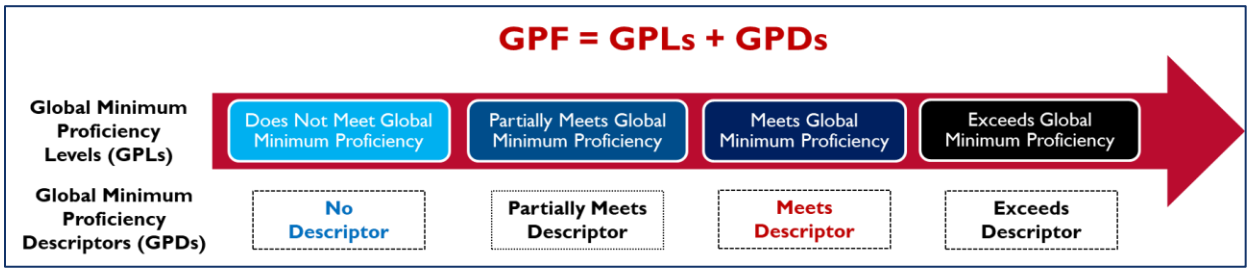
The pilot workshop involved implementing the three tasks outlined above to link the P2 and P3 EGRAs with the GPF through valid and reliable global benchmarks. Each task is described below, followed by comments on lessons learned and the suggestions for subsequent pilot and operational workshops.

Task 1

On the first day of each workshop, the co-lead facilitators provided the panelists, policy makers, and observers with background information on policy linking, including a chronology of the development of

the method in response to the global indicators. The facilitators then provided the panelists with training on the GPF and its role in policy linking. The scale in the GPF was introduced with its the two components: Global Minimum Proficiency Levels (GPLs) and Global Minimum Proficiency Descriptors (GPDs).

Figure 1: Global Proficiency Framework



Furthermore, they explained how two types of standards, i.e., content standards (knowledge or skills) and performance standards (GPLs with GPDs), were integrated into the GPF. For students, the standards were explained as follows:

> Content Standards: WHAT content students are expected to know and be able to do. This is indicated by the appropriate knowledge or skills in the GPF. For example, a student in P3 should be able to identify the main theme of a grade-level passage.
>
> Performance Standards: HOW MUCH content do students need to know and be able to do in relation to the content standards. This is indicated by the appropriate levels (GPLs), with their descriptors (GPDs), in the framework (GPF). For example: A student who meets global minimum proficiency in P3 should be able to identify the main theme of a grade-level passage.

The content standards were the focus for Task 1. The co-lead facilitators provided an excerpt from the content standards matrix in the GPF, which shows the knowledge or skills (content standards) for each domain, construct, and subconstruct by grade level. Note that the performance standards are explained in detail under Task 2.

Figure 2: Content Standards for Reading (Grades 2 to 6)

DOMAIN	CONSTRUCT	SUBCONSTRUCT	KNOWLEDGE OR SKILL (CONTENT STANDARDS)	GR2	GR3	GR4	GR5	GR6
Aural language comprehension	Retrieve information	Understand meaning of words	Understand the meaning of words in text read aloud	x	x			
			Understand how the meaning changes depending upon context		x			
		Identify explicitly stated information	Identify key events, ideas, or characters	x				
			Identify details about key events, ideas, or characters		x			
			Draw basic conclusions and generalizations			x		
Decoding	Precision		Read words accurately	x				
	Fluency		Read words accurately and at a relatively quick pace		x			
			Read texts accurately, at a relatively quick pace, and with some level of prosody			x		
Reading Comprehension	Retrieve information	Understand the meaning of unfamiliar words in grade-level connected text	Use grade 2-level morphological and contextual clues to understand words	x				
			Use grade 3-level morphological and contextual clues to understand words		x			
			Use grade 4-level morphological and contextual clues to understand words			x		
			Use grade 5-level morphological and contextual clues to understand words				x	
			Use grade 6-level morphological, contextual, and syntactical clues					x
		Locate explicitly stated information	Locate prominently-stated information in a single sentence	x				
			Locate prominently-stated information in two consecutive sentences		x			
			Locate prominently-stated information within a single paragraph			x		

Next, the co-lead facilitators provided instructions for the alignment activity, which was to examine whether the EGRAs were suitable for linking to the GPF. This was accomplished by checking the alignment between the ORF and reading comprehension assessment items and the content standards by grade level.

The alignment method in the PLT is two-step process based on a specific and standardized method that is appropriate to policy linking (Frisbie, 2003). This method requires the panelists to make independent and individual judgments on the degree to which 1) the assessment items match with at least one content standard in the GPF (depth) and 2) the domains, constructs, and subconstructs in the GPF are covered by the assessment items (breadth).

The co-lead facilitators trained the panelists on a three-point scale for determining the degree of alignment between the assessment items and the GPF:

Complete Fit (C) signifies that all 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 completely use knowledge of the content standard;

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 knowledge of the content standard;

No Fit (N) signifies that no amount 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 do not use knowledge of the content standard.

The panelists were provided with additional guidelines that 1) complete fit was usually associated with only one content standard (i.e., the knowledge or skill), 2) partial fit was usually associated with more than one content standard, and 3) no fit was not associated with any content standard.

The co-lead facilitators led the panelists through examples from an ORF passage with reading comprehension questions corresponding to complete, partial, and no fit. The panelists then rated each of the EGRA items according to the fit with the content standards in the GPF and entered their results into the item-content standards alignment form.

The co-lead facilitators entered panelists' rating totals into spreadsheets by grade level and analyzed the ratings to examine both parts of the alignment, i.e., for the items (depth) and for the domains, constructs, and subconstructs (breadth). The co-lead facilitators presented an alignment summary in a plenary session based on calculations of the averages of the ratings. Alignment was achieved through either complete or partial fit between the items and the content standards in the GPF.

The pre-determined pilot alignment thresholds were a 75 percent match for the items and a 50 percent match for the domains, constructs, and subconstructs. The item alignment percentages exceeded their threshold. Most of the domains, constructs, and subconstructs exceeded their threshold, with the only exception of the P3 subconstructs. This was due to the development of EGRA reading passage and comprehension questions that was more at the P2 level than the P3 level. Meeting nearly all of the alignment thresholds permitted proceeding with the workshop.

Table 7: Alignment of items with domains, constructs, and subconstructs

Grade Level	Alignment			
	Items	Domains	Constructs	Subconstructs
P2	100%	67%	100%	67%
P3	100%	67%	67%	40%

The completion of Task 1 was a prerequisite for beginning Task 2.

Comments

Task 1 was successful, with the panelists demonstrating that they were able to implement the instructions, i.e., to match up the items with the knowledge or skills (content standards) in the GPF. In addition, the panelists determined that the alignment met all of the pre-determined pilot thresholds for P2 and all but one of the thresholds for P3, thus allowing the workshop to continue. This signified that there was 1) adequate alignment for the item ratings (depth) and 2) sufficient domain coverage to establish content validity (breadth). However, even though the task worked well, the co-lead facilitators had the following comments involving both the implementation of suggestions from the previous pilot workshops and involving new suggestions from this workshop for additional changes to improve the policy linking process.

First, the success of the task provided additional evidence to the co-lead facilitators that the alignment methodology was technically sound. Basing the alignment on an internationally accepted method (Frisbie, 2003), with minor adaptations to the context of policy linking, showed its viability in determining whether the assessments were feasible for policy linking, i.e., benchmarking in relation to the GPF.

Second, in spite of the alignment, the co-lead facilitators had concerns about the relatively low number of constructs and subconstructs in the GPF that were covered by the EGRA tool. In other words, there was adequate depth but not enough breadth at the subconstruct level. This was due to two factors: 1) only two of the five subtasks were selected for policy linking due to non-alignment between the other three subtasks and the GPF, and 2) only a limited number of reading comprehension questions were included in the EGRA, all of which aligned with explicit (and none with implicit) comprehension. This is not an issue with the method, but it is with the assessments. The problem could be corrected by 1) designing the EGRAs so that more of the subtasks align with the GPF (e.g., regularly using aural listening comprehension) and 2) expanding the reading comprehension questions, perhaps in conjunction with an additional passage (e.g., allowing for ten instead of five questions). On the other hand, it probably does not make sense to expand the GPF to include more domains since the content covered by the other EGRA subtasks is often pre-reading rather than reading. The subject matter experts who developed the GPF focused on reading comprehension, which is reflected by the large number of constructs and subconstructs in that domain.

Third, and related to the first point, the co-lead facilitators reiterated their observation from previous workshops that the alignment with some of the reading domains – such as aural listening comprehension and decoding – would be difficult for almost any group-administered CBA. It may be useful to consider a CBA at grade 3 with supplemental administration of ORF, which ensuring that reading comprehension is adequately covered by the CBA items. This type of hybrid approach (which is currently being piloted under the USAID-funded project in Lebanon) may have promise for greater breadth, while maintaining adequate depth.

Fourth, with the alignment, the facilitators differentiated between content and performance standards in the GPF and explained the two types of standards to the panelists. Based on a change from the previous pilot to include the label of "content standards" in the knowledge or skill column, the difference between content and performance standards, and their relationship in the GPF, was better understood.

Fifth, the facilitators reduced the four-point scale for the item-content standards ratings from one of the previous pilot workshops since it was too detailed, particularly in distinguishing between partial fit and slight fit. They continued to use a three-point scale – complete fit, partial fit, and no fit – in this workshop, which provided additional evidence that the reduced scale is more appropriate.

Sixth, the co-lead facilitators continued to examine statistical thresholds for alignment. Based on the two initial pilot workshop, reasonable thresholds appear to be at least 75 percent for the item alignment (depth), at least 50 percent for the domain, construct, and subconstruct alignment (breadth). More work

needs to be done with additional pilots, particularly for the EGRAs, and it should be possible to establish such thresholds by the end of the piloting.

Task 2

On the second day of the workshop, the co-lead facilitators began Task 2 by building on Task 1. They extended the previous focus of aligning the assessment items with the knowledge and skills (content standards) to an additional focus of matching the items with the PLDs associated with the different GPLs (performance standards). There were two goals of this activity: 1) to increase the panelists' knowledge of the assessment items and GPF and 2) to identify the GPLs needed to answer each item correctly.

The co-lead facilitators trained the panelists on the GPLs and GPDs. They focused on the minimum knowledge and skills needed by students to answer the items correctly. They focused mostly on the familiarity and complexity of the words for their grade levels.

As part of the training, the co-lead facilitators showed the following GPDs to the panelists as an example of performance standards in reading comprehension for Grade 2.

Figure 3: Performance Standards for Reading (Grade 2)

PARTIALLY MEETS	MEETS	EXCEEDS
READING COMPREHENSION OF SIMPLE, GRADE 2-LEVEL CONNECTED TEXT		
RETRIEVE INFORMATION AT WORD LEVEL Understand in connected text the meaning of unfamiliar words, or of familiar words used in unfamiliar ways (i.e., homophones)		
Identify the meaning of very familiar words but has difficulty identifying the meaning of familiar words when they have regular morphological changes.	Identify the meaning of familiar words, including when they have regular morphological changes.	Identify the meaning of familiar and unfamiliar words.
RETRIEVE INFORMATION AT SENTENCE OR TEXT LEVEL Retrieve prominent information when information is found in a single sentence containing no competing information. The information is generally a response to a 'who, what, when and where" question and the information sought is generally names, facts, or numbers.		
Retrieve explicit pieces of information by direct word matching (e.g., answers the question, 'What is the girl's name?' when the text says, 'The girl's name is Dana.'	Retrieve explicit pieces of information from a single sentence.	Retrieve explicit pieces of information across more than one sentence.

Then, the co-lead facilitators asked the panelists to answer the following questions for each item:

> What knowledge and skills are required to answer the item correctly (aligning with the content standards – which was Task 1)?
>
> What is the lowest GPL – according to the GPDs – that is most appropriate for answering the item correctly (matching with the performance standards – which is Task 2)?

They were provided with the same example of a reading passage and comprehension questions from Task 1. To practice, the panelists were asked to identify the appropriate content standard for a particular comprehension question and then match the question with the appropriate performance standard (GPL).

The panelists then divided into their P2 and P3 panels. Led by content facilitators, they identified the appropriate performance standard for each item and wrote the item next to the GPL in the GPF. If the item matched with more than one content standard – which was usually the case if the alignment had

been rated as a partial fit – the panelists wrote item next to the additional GPL. The panelists discussed the results and reached consensus for matching the items with the appropriate performance standards (GPLs and GPDs).

The completion of Task 2 was a prerequisite for beginning Task 3.

Comments

Task 2 was successful, with the panelists demonstrating that they were sufficiently able to implement the instructions. They matched up the items with the GPLs and their GPDs (performance standards), with reference to the subconstruct(s) and knowledge or skills (content standards). They wrote the information for each item and/or item number in their GPFs. Again, however, even though the task worked well, the co-lead facilitators had the following observations, with implementation of suggestions from the previous pilot workshops and suggestions for additional minor changes to improve the policy linking process.

First, the co-lead facilitators changed the instructions for recording their matching. Since the EGRA items had depth but lacked breadth, it was easier for the panelists to record the matching information only in the GPF. This provided sufficient information for Task 3.

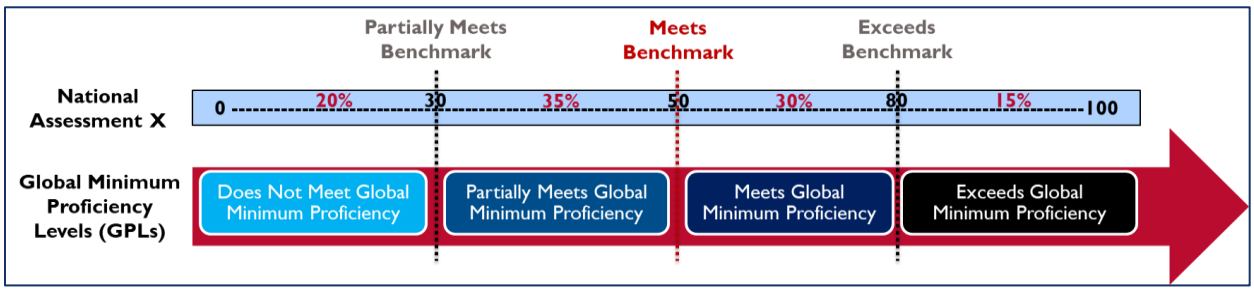
Second, the large group discussions were beneficial. It was useful for the panelists to have the opportunity to go through the matching process with their content facilitators in their panels and discuss their matches. The panelists were able to reach consensus on the GPLs an GPDs (performance standards) appropriate for answering the items correctly. This activity was highly valuable in increasing the panelists' understanding, and consistency, prior to proceeding with the benchmarking in Task 3.

Third, the co-lead facilitators noticed inconsistencies in the two groups and took steps to correct it. The P2 group rated some of the items as requiring lower GPLs and GPDs than the P3 group. The facilitators gathered the content facilitators during the break to discuss the issue, which was then used to reconsider the matching for those items. This situation was apparent since the same test was used for P2 and P3.

Task 3

On the third day of the workshop, the panelists received training on setting global benchmarks using the Angoff method, which required making judgments (ratings) on each of the assessment items. The co-lead facilitators presented a hypothetical example of how the benchmarking method would link a national assessment to the GPF, thus allowing for the calculation of the percentages of students attaining minimum proficiency. They showed the following graphic with an example of three benchmarks (30, 50, and 80 points) on a national assessment scale (0-100 points), with percentages of learners in each of the four GPLs (20 percent, 35 percent, 30 percent, and 15 percent).

Figure 4: Example of an assessment and benchmarks



This example was extended to three national assessments of different difficulties, and how this would lead

to different sets of benchmarks for each assessment. The co-lead facilitators showed the panelists how the benchmarking results – when applied to the assessment data sets – could be used for comparing and aggregating assessment results, as well as tracking those results over time. They explained how the system could be used for reporting on the SDG and USAID indicators.

Then the panelists received an introduction to the Angoff benchmarking method and participated in a training session on applying the method to conduct their own item ratings for establishing the benchmarks. The co-lead facilitators provided an opportunity for the panelists to practice rating items, i.e., prior to conducting their ratings for the actual benchmarking.

After the panelists practiced the Angoff method, the co-lead facilitators trained the panelists on the item rating forms and procedures. They explained the rating form – with a section for the panelists to record the ORF results from their classrooms – and how to fill it in.

Figure 5: Item Rating Form

3 JP students: _____

3 JM students: _____

3 JE students: _____

Name of the Panelist: _____

Panelist Code: _____

Rating Form for Setting Grade 3 Oral Language Fluency and Comprehension Benchmarks

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

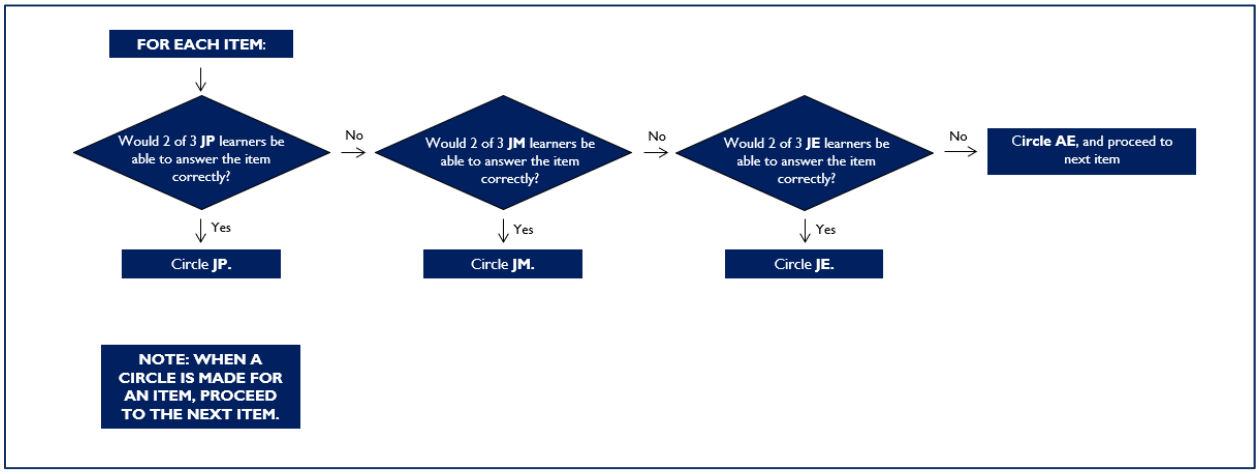
Word No.	Reading Passage (Word)	**Round 1:** No. of **words** learners would attempt to read in a minute			**Round 1** individual and independent ratings				**Round 2:** No. of **words** learners would attempt to read in a minute			**Round 2** individual and independent ratings			
		JP	JM	JE	JP	JM	JE	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

The panelists divided into their two panels and conducted their first round of individual and independent ratings for each of the NSA items. As seen from the sample form, the panelists – individually and independently – were required to give one of four ratings – Just Partially Meets (JP), Just Meets (JM), Just Exceeds (JE), and Above Exceeds (AE) – to each item. The steps in the judgment process are the following:

1. Read each item and reflect on the difficulty of the items, both for ORF and comprehension.
2. Review the matching (Task 2) of the GPL and GPD required for students to answer correctly.
3. Conceptualize three JP, three JM, and three JE students from the GPF and classrooms.
4. Estimate number of words the JP, JM, and JE students would attempt to read in a minute.

5. Reflect on whether two out of three students in each GPL could answer the item correctly.
6. Follow the steps for rating the items, i.e., answer the questions and circle the ratings on the form.

Figure 6: Steps for Rating Items



After conducting the first round of ratings for each of the items, the panelists compiled their ratings to calculate their initial benchmarks. The co-lead facilitators 1) entered the panelists' benchmark data into the spreadsheets, 2) calculated the initial benchmarks for the panels by averaging the benchmarks across the panelists, and 3) produced summaries of the benchmarks. They presented the summaries to the panelists, including the 1) initial benchmarks of each panelist, 2) average benchmarks for the panels, 3) impact data with percentages of scores in GPLs based on the score distributions, and 4) statistics showing the quality of the ratings, including standard errors.

On the fourth day of the workshop, the panelists made their second round of ratings for calculating their final benchmarks. After another review of the initial benchmarks and feedback data, the panelists separated into their panels and revised their item ratings using the same steps as the first round. They were provide guidance that they should 1) focus on item content in relation to the descriptions of the knowledge and skills, 2) consider what students would be able to do given any issues related to testing conditions (i.e., measurement error), and 3) make adjustment to the ratings based on their judgments. After the second round, the co-lead facilitators entered the data, calculated the final benchmarks, and presented the results to the panelists.

Comments

Task 3 was successful, with the panelists able to implement the instructions, i.e., to understand the benchmarking process, conduct the ratings for the items, comprehend the feedback from the first round, and make revisions for the second round. Again, however, even though the task worked well, the co-lead facilitators had the following observations, with implementation of some suggestions from the previous pilot workshops and suggestions for additional minor changes to improve the policy linking process.

First, the matching from Task 2 remained a critical step in Task 3 of the policy linking process. The co-lead facilitators put more emphasis on matching items with the GPLs and GPDs in India than in Bangladesh, and this continued in Nigeria. This included training, panel discussions, and consensus building. The facilitation for this task was a key element. One of the panels in Nigeria had more consistent content facilitation. Once this was observed by the co-lead facilitators, they provided additional support to the other panel. As seen in the results below, there was a reversal for one of the benchmarks, which was then

corrected with increased facilitation support and understanding of the task by the panelists.

Second, the facilitators eliminated the part in the Toolkit in which the panelists took the assessments themselves. For EGRAs, it was better to send the subtasks to the teachers in advance so that they could administer the timed ORF and reading comprehension subtasks to a selection of their students. This allowed them to 1) practice with an EGRA and 2) associate student performance with the GPF.

Third, the co-lead facilitators had moved to additional large group work for Task 3 in Bangladesh and India, which was continued in Nigeria. Normally, there is more individual work in Task 3, but the facilitators found that the large group work was needed to increase the level of understanding for the item ratings and benchmarking. One of the groups in Nigeria stayed late on Day 3 so that they could benefit from additional large group facilitation. Repeated discussions on the GPLs and GPDs appeared to positively affect the reliability of the benchmarks.

Fourth, the extra time in large groups in Task 3 was needed since the EGRA-based benchmarking method was different and more complicated than the CBA-based method. The difference was the timed tasks in EGRA, which led to conditionalities for items. For instance, the panelists had to estimate the number of words in the ORF passage that a student would reach, depending on the knowledge and skills associated with their GPL. This meant that the panelists would only rate the items based on projections of students in particular levels who would attempt the items. Furthermore, the panelists only rated the comprehension questions that the students were projected to receive, depending on whether they attempted to read the part of the passage associated with the question. The CBA-based method does not have this feature since it is not a timed test.

Fifth, the facilitators continued with training the panelists to calculate their own benchmarks rather than waiting for the data entry and calculation by the facilitators. The panelists were instructed to total their JP, JM, and JE columns and then calculate each of their three benchmarks. This training and subsequent calculation promoting better understanding on the part of the panelists for the benchmarking numbers. This also helped with making revisions during Round 2, since the panelists could see the difference in their benchmarks depending on their item ratings.

Sixth, the facilitators continued to examine statistical thresholds for reliability. Based on the two initial pilot workshop, reasonable thresholds appear to be less than 1.00 for standard errors (with some variation depending on the number of items in the assessment), and 0.70 for inter-and intra-rater reliability. More work needs to be done to test these thresholds through additional pilots, which should result in establishing firm thresholds by the end of the piloting. The thresholds are useful indicators of reliability after Rounds 1 and 2.

Results

The co-lead facilitators analyzed the panelists' ratings after Rounds 1 and 2. For P2 and P3, this included calculating the following: 1) benchmarks, 2) score ranges, 3) impact data (using the score distributions), and 4) consistency of the results. All analyses are presented by round, except for the location statistics, which are only presented for Round 2.

Note again that the P2 and P3 EGRA tools were the same, with 35 passage reading words and five reading comprehension questions, for a total of 40 points. In general, the results should show a progression from P2 to P3, i.e., higher benchmarks at each GPL for P3 than P2 (since the same subtasks should be easier for P3 students).

Round 1

The co-facilitators produced summary tables and graphs from Round 1, which showed the initial benchmarks, score ranges, and impact data. The impact data examined the percentages of scores in the different GPLs. All analyses were conducted for P2 and P3.

The benchmarks showed progression from P2 to P3, except for the "exceeds" benchmark, which had a reversal, with the P2 benchmark higher than the P3 benchmark. Impact data showed 12 percent of students meeting global minimum proficiency at P2 and 30 percent at P3.

Table 8: Round 1 Benchmarks

Grade	Benchmarks (in points)			
	Does Not Meet	Partially Meets	Meets	Exceeds
P2	--	5	16	36
P3	--	7	19	35

Table 9: Round 1 Score Ranges

Grade	Benchmark Ranges (in points)			
	Does Not Meet	Partially Meets	Meets	Exceeds
P2	0-4	5-15	16-35	36-40
P3	0-6	7-18	19-34	35-40

Table 10: Round 1 Impact Data

Grade	Impact Data by Proficiency Level (in percentages)			
	Does Not Meet	Partially Meets	Meets	Exceeds
P2	81.0%	7.0%	8.1%	3.9%
P3	59.8%	10.1%	13.3%	16.8%

Round 2

After providing the results from the initial benchmarks in Round 1 to the panelists and conducting the Round 2 ratings, the facilitators produced summary tables and graphs from Round 2, which showed the final benchmarks, score ranges, and impact data. Again, the impact data examined the percentages of scores in the different GPLs. All analyses were conducted for P2 and P3.

Most of the benchmarks increased from Round 1 to Round 2. The reversal at the "exceeds" level was changed by the panelists so that the two grades had the same benchmark. Impact data showed 11 percent of students meeting global minimum proficiency at P2 and 29 percent at P3.

Table 11: Round 2 Benchmarks

Grade	Benchmarks (in points)			
	Does Not Meet	Partially Meets	Meets	Exceeds
P2	--	6	18	37
P3	--	7	20	37

Table 12: Round 2 Score Ranges

Grade	Benchmark Ranges (in points)			
	Does Not Meet	Partially Meets	Meets	Exceeds
P2	0-5	6-17	18-36	37-40
P3	0-6	7-19	20-36	37-40

Table 13: Round 2 Impact Data

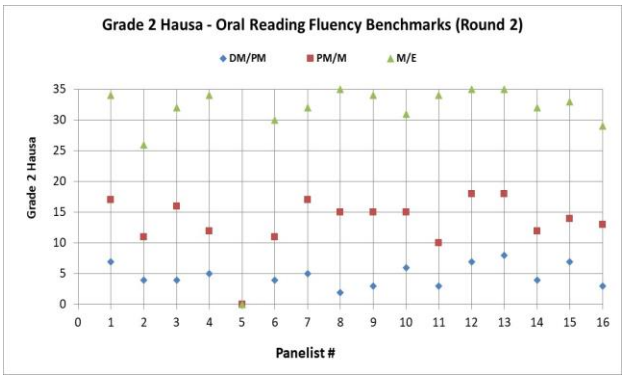
Grade	Impact Data by Proficiency Level (in percentages)			
	Does Not Meet	Partially Meets	Meets	Exceeds
P2	81.8%	7.2%	7.4%	3.6%
P3	59.8%	11.2%	14.4%	14.6%

Consistency

Feedback data were provided on the consistency in, or reliability of, the panelists' ratings. The feedback data included location statistics and standard errors of measurement (SEM).

The location statistics are provided only for the final benchmarks. They showed strong consistency in the panelists' ratings for both grades and subjects, with some outliers in the ORF benchmarks.

Figure 7: Location statistics for P2



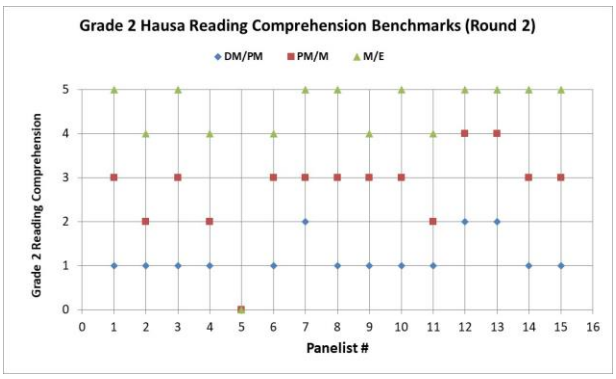
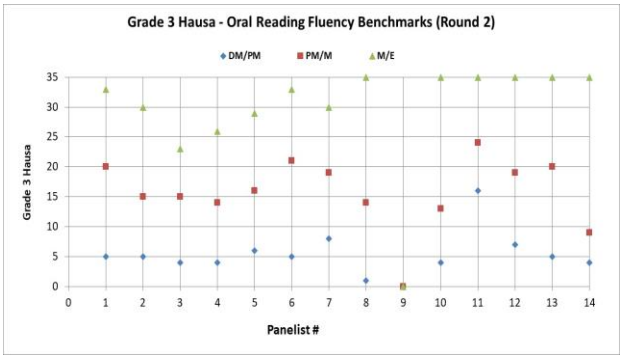
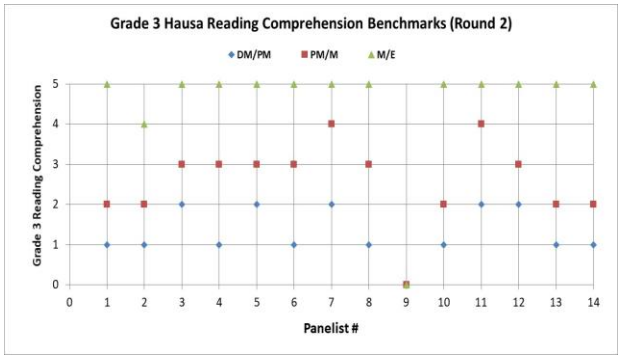


Figure 8: Location statistics for P3





The SEMs were calculated at the benchmark level. Values of less than 1.00 indicate substantial agreement between the panelists in their estimated ratings for the benchmarks.

The SEMs of the P2 ORF ratings at each benchmark from Rounds 1 and 2 were below the provisional threshold of 1.00. However, the SEMs of the P3 ORF ratings at each benchmark were slightly above the threshold. The co-lead facilitators believe that this was likely due to better alignment of the EGRAs with P2 than P3. Note that there were some issues with the facilitation of the P3 panel, but this was corrected.

The SEMs of the P2 and P3 reading comprehension ratings showed consistency for both Rounds 1 and 2 at each benchmark.

Table 14: SEM by Round

Grade (Subtask)	SEMs by Benchmark					
	Partially Meets		Meets		Exceeds	
	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2
P2 (ORF)	0.43	0.49	0.75	0.71	0.88	0.68
P2 (Comprehension)	0.07	0.12	0.16	0.17	0.16	0.14
P3 (ORF)	1.01	1.02	1.41	1.17	1.30	1.15
P3 (Comprehension)	0.14	0.15	0.21	0.21	0.17	0.08

Recommendations

These policy linking recommendations reflect many of the comments from Tasks 1, 2, and 3. There are also comments about the presentation of the results.

Task 1

First, the alignment method is closer to finalization. The internationally accepted method (Frisbie, 2003) showed its viability with the EGRAs in Nigeria – as it had with the CBAs in Bangladesh and India – in determining whether the assessments are feasible for policy linking. It still needs additional piloting, particularly with EGRAs and EGMAs, but it is working as planned, with minor adaptations that have been made, such as establishing reliability thresholds that are appropriate in the international context.

Second, the alignment issues with EGRAs became more apparent during the workshop. This related to both the limited number of subtasks that align with the GPF and the limited number of reading comprehension questions in a typical EGRA. Rather than make changes in the policy linking method, including the GPF, the recommendation is to reflect on ways to correct the content coverage in EGRAs, which were originally designed as non-curriculum-based assessments. For instance, one possibility is to use aural listening comprehension more often as a subtask. Another possibility is to use a hybrid approach to assessing reading – with a combination of aligned EGRA subtasks and CBA items – would have more construct validity, as identified by international reading subject matter experts.

Third, differentiating between content and performance standards in the GPF contributed to greater understanding of the GPF by the panelists. It appeared to help that an explanation of the two types of standards was included in the facilitation as well as in the labels of the GPF. The recommendation is to continue with this approach

Fourth, the pilot showed that the three-point scale for the item-content standards ratings was more efficient than the original four-point scale, which was too detailed, particularly in distinguishing between

partial fit and slight fit. The three-point scale – complete fit, partial fit, and no fit – is recommended for future pilots as the more appropriate scale for alignment.

Fifth, the workshops provided more information about the alignment thresholds, particularly as they apply to the EGRAs. Based on these initial pilot workshops, reasonable thresholds appear to be 75 percent for the item alignment (depth), 50 percent for the domain, construct, and subconstruct alignment (breadth), The recommendation is to continue with these thresholds in subsequent workshops.

Task 2

First, the information from the matching process was recorded on both the test booklet and GPF in the previous pilots. However, given the extensive depth and lack of breadth for the EGRA items, it was easier for the panelists to record the matching information only in the GPF. The recommend is to record the matching information on both the test booklet and the GPF for CBAs but only in the GPF for EGRAs.

Second, focusing on large group discussions for most of Task 2 was useful in promoting understanding of the GPLs and GPDs (performance standards) on the part of the panelists. It was also better for the panelists to work with their content facilitators to consider the GPLs needed for answering the items correctly. This activity was highly valuable in increasing the panelists' understanding, and consistency, prior to proceeding with the benchmarking in Task 3.

Third, having the co-lead facilitators closely collaborate with both content facilitators during Task 2 was important in establishing consistency in the application of the matching method. In addition, the communication between the content facilitators can be important. The other pilots did global benchmarking for different subject areas at the same time, but this workshop had the same subject area, and even the same assessment, across the panels. This made the collaboration more important. The recommendation is to encourage more collaboration when the subject area is the same for different panels within the same workshop. It can still be useful across subjects, but it is clearly important within subjects.

Task 3

First, the level of matching from Task 2 continued the improvement trend from the India workshop due to the application of the method and increased time for the activity. This paid dividends in Task 3 by reducing the inconsistencies in the item ratings. The standard error of measurement estimates were generally low, except for the anomaly of using the same EGRA for two grade levels, and observing that the subtasks were more appropriate for one grade level than for another. The recommendation is to continue with the method of having adequate time to go through the items one-by-one – as a group – prior to the first and second rounds of item ratings. Repeated discussions on the GPLs and GPDs appeared to have a positive influence on the reliability of the benchmarks of the panelists.

Second, replacing the part in the PLT in which the panelists took the assessments themselves by having the panelists administer the assessments to a selection of their students was highly useful for the EGRAs. The recommendation is to send the subtasks to the teachers in advance so that they can administer the timed ORF and reading comprehension subtasks prior to the workshop. This allowed them to 1) practice with an EGRA, including the timed ORF subtask and 2) associate student performance with the GPF.

Third, additional training on calculating benchmarks helped the panelists in understanding the benchmarking numbers. They were instructed on totaling their JP, JM, and JE columns to calculate the benchmarks. This also helped with making revisions during Round 2 due to seeing the influence of the ratings on the benchmarks. The recommendation is to continue with this training in subsequent pilots.

Fourth, the workshops provided more information about the reliability or consistency thresholds. Based on these initial pilot workshops, a reasonable threshold appears to be less than 1.00 for standard errors on a 30- to 40-item test. The recommendation is to continue with the thresholds in subsequent pilots.

Results

First, as with the other workshops, the presentations of the data and the pre-programmed spreadsheets were improved for this pilot, as were the graphics. The recommendation is to continue with this improvement, though it is important to note that different spreadsheets are needed for each workshop, depending on the type of assessment, e.g., CBA or EGRA. The graphics are similar from one assessment type to another.

Second, the presentation of the data between benchmarking Rounds 1 and 2 had difficulties. The panelists did not seem to understand the data as well as they had in previous workshops. The recommendation is to improve the tailoring of the data presentations to the audience, though sometimes this is difficult to determine in advance. On the other hand, the spent on understanding the calculations of the benchmarks helped in making revisions during Round 2. This recommendation is to continue with this process in subsequent workshops.

References

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

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

Cizek, G.J. & Bunch, M.B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publishing.

Evans, N. (2019). *Overview of reading assessments in Nigeria: 2011 to 2018*. Washington, DC: Northern Education Initiative Plus.

Ferdous, A., Evans, N., & Davis, J. (2019). *Global proficiency framework for reading and mathematics: Grades 2 to 6*. Washington, DC: US Agency for International Development.

Ferdous, A., Kelly, D., & Davis, J. (2019). *Policy linking method: Linking assessments to global standards*. Washington, DC: US Agency for International Development.

Ferdous, A., Kelly, S., Davis, J., & Watson, C. *Policy linking toolkit: Linking assessments to a global proficiency framework*. Washington, DC: US Agency for International Development.

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.

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

Northern Education Initiative Plus (2016). *Early grade reading assessment baseline report*. Abuja, Nigeria: NEI+ (for USAID).

Northern Education Initiative Plus (2018). *Early grade reading assessment midline report*. Abuja, Nigeria: NEI+ (for USAID).

Plake, B.S., Buckendahl, C., & Ferdous, A.A. (2005). *Setting multiple performance standards using the Yes/No Method: An alternative item mapping method*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

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

USAID (2019). *Education reporting guidance*. Washington, DC: USAID (https://www.edu-links.org/sites/default/files/media/file/Education-Reporting-Guidance-2019.10.16-508_Final.pdf).