Bangladesh Policy Linking Pilot Workshop Report: Setting Global Benchmarks for Grades 3 and 5 Bangla Language and Mathematics

UNESCO Institute for Statistics (UIS)
Bangladesh Ministry of Primary and Mass Education (MoPME)

December 15, 2019
Management Systems International (MSI)

Acknowledgements

The education team at Management Systems International (MSI) is grateful for the support provided by several groups for the policy linking pilot workshops.

First, the organizational support provided by officials at the Bangladesh Ministry of Primary and Mass Education (MoPME), specifically the Department of Primary Education (DPE), was critical for the success of the workshops.

Second, the management support provided by officials from the UNESCO Institute of Statistics (UIS) in Montreal, the UNESCO regional office in New Delhi, and the UNESCO country office in Dhaka was instrumental in planning and implementing the workshops.

Third, the financial support of the U.K. Department for International Development (DFID) and the Bill & Melinda Gates Foundation (Gates) was essential for operationalizing the workshops.

Fourth, the technical support by UIS and the U.S. Agency for International Development (USAID) was critical in developing the presentations and documents, including the Global Proficiency Framework (GPF) and the Policy Linking Toolkit (PLT). Others collaborating on the technical preparation were DFID, Gates, the World Bank, and numerous other partners.

Fifth, the hands-on support by the panelists – head teachers, teachers, and specialists – from the different divisions in Bangladesh was most important in conducting the policy linking workshops. Their strong engagement and commitment were vital for establishing the pilot global benchmarks and drawing lessons learned from the workshops.

Abdullah Ferdous

Jeff Davis

Sean Kelly

Table of Contents

Acronyms and Abbreviations

DFID	U.K. Department for International Development
DPE	Directorate of Primary Education
EGRA	Early Grade Reading Assessment
GAML	Global Alliance to Monitor Learning
Gates	Bill and Melinda Gates Foundation
GPD	Global Proficiency Descriptor (or Descriptor)
GPL	Global Proficiency Level (or Level)
GPF	Global Proficiency Framework (or Framework)
GPS	Government Primary Schools
IAEG-SDG	Inter-Agency and Expert Group on SDGs
IBE-UNESCO	International Bureau of Education – UNESCO
M&E	Monitoring and Evaluation
MoPME	Ministry of Primary and Mass Education
MSI	Management Systems International
NAC	National Assessment Cell
NAPE	National Academy for Primary Education
NCTB	National Curriculum and Textbook Board
NSA	National Student Assessment
PLT	Policy Linking Toolkit (or Toolkit)
PTI	Primary Teacher Training Institute
RNGPS	Registered Non-Government Primary Schools
SDG	Sustainable Development Goal
UIS	UNESCO Institute for Statistics
UNESCO	U.N. Educational, Scientific and Cultural Organization
USAID	U.S. Agency for International Development

Policy Linking Overview

The development and piloting of a policy linking method for reporting on Sustainable Development Goal (SDG) Indicator 4.1.1 has been in process since 2017. It is anticipated that the method will be finalized by September 2020. The chronology 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:

> 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."

In September 2017, UIS organized a meeting of the Global Alliance to Monitor Learning (GAML) in Hamburg, Germany, to present and resolve issues with reporting on SDG Indicator 4.1.1. The main issue identified by UIS and GAML was setting valid and reliable global benchmarks on the wide variety of national and cross-national assessments. Different benchmarks for each assessment, based on the level of difficulty of those assessments, would allow UIS to use the assessment data sets to calculate the proportions of learners meeting minimum proficiency. At the meeting, 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, UIS and the US Agency for International Development (USAID) 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 bilateral donors). MSI prepared a *Policy Linking Justification Paper* for the workshop, in which it proposed developing a common, non-statistical scale and a step-by-step benchmarking process. The process could be implemented with different assessments for setting global benchmarks that would link assessments to the scale. The group accepted policy linking as a possible method for reporting. They also developed an initial non-statistical scale with four *Global Proficiency Levels* (GPLs or Levels) and labels, along with brief definitions for each level.

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, UIS and USAID co-sponsored two workshops in Washington, DC with 30 subject matter experts in primary school reading and mathematics. The experts adapted and expanded the International Bureau of Education (IBE-UNESCO) global content frameworks drawn from the curriculum and assessment frameworks of over 100 countries. They developed a draft *Global Proficiency Framework* (GPF or Framework) as the common scale for linking different assessments.

In August 2019, the newly-formed Policy Linking Working Group (or Working Group) finalized the draft Framework. It is comprised of four Levels and detailed *Global Proficiency Descriptors* (GPDs or Descriptors) in grades 2 through 6 in reading and mathematics for each level. The Levels are does not meet, partially meets, meets, and exceeds global minimum proficiency. The Descriptors have domains, constructs, subconstructs, and knowledge and skills by grade, subject, and level.

In September 2019, the Policy Linking Working Group completed a draft *Policy Linking Toolkit* (PLT or Toolkit) to provide guidance for workshops to pilot the policy linking method. The Toolkit 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 Framework, 3) matching assessment items with the Levels, 4) setting the global benchmarks (using the Angoff method), 5) calculating quality indicators for the benchmarking, 6) finalizing the results, and 7) writing the technical report. Annexes to the Toolkit include forms for implementing the workshops, formulas for calculating the indicators, and an outline for the technical report.

In October 2019, policy linking was again presented and approved by the IAEG-SDG at a meeting in Addis Ababa, Ethiopia. UIS reported that total of 146 out of 193 countries were committed to providing data for reporting purposes. The combination of the policy linking method and the high percentage of committed countries allowed the IAEG-SDG to advance the classification of SDG Indicator 4.1.1 from Tier II to Tier I for global 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.

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, led 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 January and February 2020, USAID will lead policy linking pilots in Kenya and Nigeria. These workshops will result in setting global benchmarks on Early Grade Reading Assessments (EGRAs) at grade 2 (Kenya) and grades 2 and 3 (Nigeria). There is also the possibility in both countries of setting global benchmarks on national assessments in language and mathematics at the end of upper primary.

Starting in March 2020, there will be additional pilots. The extent of these 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.

Finalization

In September 2020, the Working Group plans to finalize the Framework and Toolkit will be finalized, at which time it will be disseminated by UIS and USAID. Both organizations plan to hold training sessions

and webinars to build capacity for stakeholders and specialists who would like to implement policy linking. Countries will be able use the method to set global benchmarks on their national assessments for reporting on SDG Indicator 4.1.1. Similarly, organizations responsible for cross-national assessments will be able to follow the same policy linking procedures to set their global benchmarks for reporting.

After September 2020, national and cross-national global benchmarks on different assessments will allow UIS – with support from member countries – to calculate the percentages of learners achieving a global minimum proficiency level. Based on applying the common scale and benchmarking method to the assessments and data sets through policy linking, this will provide information for 1) national, regional, and global *comparisons* of assessment results for drawing lessons learned, 2) global *aggregation* of assessment results for reporting on indicators, and 3) national, regional, and global *tracking* of assessment results for measuring progress over time.

Pilot Workshop Preparation

The workshop preparation, the workshop tasks, and the workshop results are presented in the sections below. Each section concludes with brief comments about what went well and what did not go well with the pilot workshops, along with suggestions for modifications for subsequent pilot workshops. These comments and suggestions are summarized in the final section of this report.

Planning

With the publication of the draft Framework and Toolkit, UIS planned its first pilot workshop in Bangladesh, with support from MSI, DFID, and Gates. The objective was setting global benchmarks on the 2017 National Student Assessment (NSA). The Directorate of Primary Education (DPE) in the Ministry of Primary and Mass Education (MoPME) approved two four-day workshops at the Primary Teacher Training Institute (PTI) in Dhaka from October 21 to 30, 2019. MSI assigned two international co-lead facilitators and recruited two national content facilitators. Lessons learned from these pilots would be summarized by MSI and discussed with UIS. The workshops were organized as follows:

Workshop 1: Grade 3 Bangla language and mathematics

Monday October 21 to Thursday October 24

Workshop 2: Grade 5 Bangla language and mathematics

Sunday October 27 to Wednesday October 30

For the workshop, the co-lead facilitators prepared three pilot tasks: 1) checking the alignment of the assessments with the domains, constructs, and subconstructs in the Framework, 2) matching the assessment items with the Levels and Descriptors in the Framework, and 3) implementing the Angoff method to set global benchmarks on the assessments for each of the Levels. They also prepared for analysis of the workshop results, including the workshop evaluation data.

Assessments

The 2017 NSA was the fourth administration of the assessment. It measures the achievement of learners relative to the learning outcomes in the primary school curriculum. The assessments for this workshop were administered to representative samples in Government Primary Schools (GPS) and Registered Non-Government Primary Schools (RNGPS) in grades 3 and 5 in Bangla language and mathematics.

The NSA serves the dual purposes of 1) informing MoPME activities for classroom instruction, teacher professional development, and curriculum reform, and 2) reporting on domestic and international indicators over time. It is managed by the Monitoring and Evaluation (M&E) Division in the DPE. Technical counterparts are the National Assessment Cell (NAC), the National Curriculum and Textbook Board (NCTB), and the National Academy for Primary Education (NAPE). The assessments are designed by DPE specialists, local consultants, and international consultants.

In 2017, the grades 3 and 5 Bangla language and mathematics assessments had multiple choice (objective) and constructed response (open-ended) items. Each assessment had 30 to 36 multiple choice items and 4 to 5 constructed response items, with 36 to 44 total points (Table 1). Note that some of the constructed response items were scored from 0-1 points while others were scored from 0-2 points.

Table 1: Number of items and score points

Assessment		Items				Total Points
Grade	Subject	Multiple Choice	Constructed Response (0-1)	Constructed Response (0-2)	Total	
Grade 3	Bangla Language	32	4	0	36	36
	Mathematics	30	1	4	35	39
Grade 5	Bangla Language	36	0	4	40	44
	Mathematics	35	1	4	40	44

The sample size for the 2017 NSA was 1,470 schools and 52,547 learners from the eight administrative divisions in Bangladesh. This included 28,402 learners in grade 3 and 24,145 learners in grade 5.

Panelists

The DPE invited four groups (or panels) of 15 panelists for grades 3 and 5 Bangla language and mathematics, or a total of 60 panelists. In each panel, there were 12 head teachers or classroom teachers and three experts in curriculum, teacher training, and pedagogy. The panelists came from each administrative division. There was near equal gender representation. A total of 56 of the invited panelists participated in the workshops. A summary of the panelists' profiles is provided below (Table 2).

Table 2: Panelists' background information

Grade	Subject	Gender		Education Level		Years of Experience	
		Female	Male	< B.A.	≥ B.A.	< 10	≥ 10
Grade 3	Bangla Language	7	8	1	14	0	15
	Mathematics	4	9	0	13	1	12
Grade 5	Bangla Language	7	7	1	13	1	13
	Mathematics	8	6	0	14	1	13
Totals	--	26	30	2	54	3	53

Benchmarks

To set the global benchmarks, the workshops employed a Yes-No variation of the Angoff method (Plake, Buckendahl, & Ferdous, 2005). In this method, panelists are asked to conceptualize minimally proficient

learners – those at or slightly above the benchmarks – from the Framework and estimate how they would perform on each of the assessment items. Using the assessment tools and rating forms, the panelists proceed item by item, making ratings on the items to estimate whether minimally proficient learners in the different Levels would answer each item correctly (yes or no). The number of yes responses by Level are summed and aggregated to yield an individual panelist's benchmark. The benchmarks from all panelists are then averaged to determine the panel's benchmarks.

The three tasks for setting the benchmarks were adapted a process that is widely accepted for benchmarking workshops (Cizek & Bunch, 2007). The first task involved training on policy linking and conducting item-subconstruct ratings. The second task involved training on the Framework and matching items with the Levels and Descriptors. The third task involved training on the Angoff method and conducting two rounds of item ratings.

The DPE organized the participants and the venue, and provided the NSA 2017 instruments, answer keys, scoring rubrics, and data sets. The co-lead facilitators produced training slides and rating forms based on guidelines in the Toolkit, as well as pre-programmed spreadsheets to calculate benchmarks and feedback data (i.e. location statistics and impact data). Workshop evaluation forms were developed to solicit the panelists' views on the workshop procedures and their own confidence in setting benchmarks.

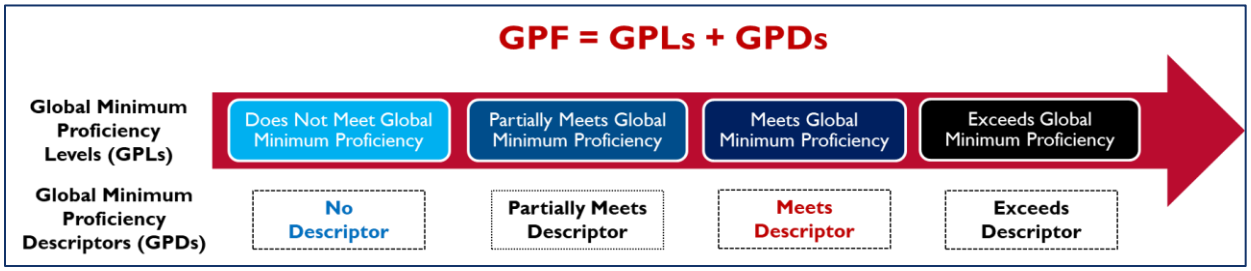
Pilot Workshop Implementation

The pilot workshop involved three tasks to establish reliable, valid, and fair global benchmarks to link the assessments with the Framework: 1) checking the alignment of the assessments with the Framework, 2) matching the assessment items with the Levels and Descriptors, and 3) implementing the Angoff method to set global benchmarks. Each of the three tasks is described below, with comments on lessons learned for improvements that could be implemented in subsequent pilot workshops.

Task 1

On the first day of the workshops, the panelists received training on a method to check the alignment of the assessment items with the content of the Framework (Frisbie, 2003). This required judgments on the degree to which 1) the items matched with the subconstructs in the Framework (alignment depth) and 2) the domains, constructs, and subconstructs in the Framework were covered by the items (alignment breadth). The Framework was introduced using the following graphic (Figure 1).

Figure 1: Global proficiency framework



The graphic shows that the Framework is comprised of two components: Global Minimum Proficiency Levels (GPLs, or Levels) and Global Minimum Proficiency Descriptors (GPDs, or Descriptors). The structure of the Descriptors in the Framework was introduced using the following mathematics example,

with a hierarchy of domains, constructs, subconstructs, knowledge and skills, as well as Descriptors by Level (Figure 2). The knowledge and skills were identified as content standards and the three global minimum proficiency levels were identified as performance standards. The panelists were instructed to focus on the subconstruct(s) and the knowledge or skill associated with each subconstruct.

Figure 2: Example of hierarchy and content

DOMAIN	CONSTRUCT	SUBCONSTRUCT	KNOWLEDGE OR SKILL	GLOBAL MINIMUM PROFICIENCY LEVEL		
				PARTIALLY MEETS	MEETS	EXCEEDS
Number knowledge	Whole number	Identify and count whole numbers	Count, read, and write whole numbers to 1000; skip count forwards by twos, fives, tens, and hundreds.	Count, read, and write whole numbers up to 100.	Count, read, and write whole numbers to 1000; skip count forwards by twos, fives, tens, and hundreds.	Count backwards from 100; skip count backwards using twos, fives, and tens.
		Identify the relative magnitude of whole numbers	Compare and order whole numbers to 100.	Compare and order whole numbers to 20.	Compare and order whole numbers to 100.	Compare and order whole numbers to 1000.
		Represent whole numbers in equivalent ways	Compose and decompose whole numbers to 100; represent whole numbers to 100 concretely, pictorially, and symbolically.	Represent whole numbers to 20 concretely, pictorially, and symbolically.	Compose and decompose whole numbers to 100; represent whole numbers to 100 concretely, pictorially, and symbolically.	Identify the value of a digit based on its place-value position in whole numbers to 1000.

The panelists were trained on a four-point scale for determining the degree of alignment between the assessment items and the Framework:

- Complete Fit (C) signifies that all of the content required to answer the item correctly is contained in the subconstruct, i.e., if the learner answers the item correctly, it is because they completely use knowledge of the subconstruct;
- Partial Fit (P) signifies that part of the content required to answer the item correctly is contained in the subconstruct, i.e., if the learner answers the item correctly, it is because they partially use knowledge of the subconstruct;
- Slight Fit (S) signifies that a slight amount of the content required to answer the item correctly is contained in the subconstruct, i.e., if the learner answers the item correctly, it is because they slightly use knowledge of the subconstruct; and
- No Fit (N) signifies that no amount of the content required to answer the item correctly is contained in the subconstruct, i.e., if the learner answers the item correctly, it is because they do not use knowledge of the subconstruct.

The panelists were provided with guidelines that 1) complete fit was usually with only one subconstruct, 2) partial or slight fit were usually with more than one subconstruct, and 3) no fit was without any subconstruct.

After the panelists had rated each of the items individually and independently according to the fit with the subconstructs in the Framework, the co-lead facilitators entered rating totals from each panelist into a spreadsheet. They 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 facilitators presented a summary based on calculations of the averages of the ratings. Alignment was achieved through either complete or partial fit between the items and the Framework.

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. All of the alignment percentages exceeded these thresholds, except for the near alignment with grade 3 mathematics subconstructs (Table 3). Meeting or exceeding the thresholds allowed the participants to proceed with Task 2 of the pilot linking workshop.

Table 3: Alignment of items with domains, constructs, and subconstructs

Assessment		Alignment (Percentages)			
Grade	Subject	Items	Domains	Constructs	Subconstructs
Grade 3	Bangla Language	89%	67%	80%	75%
	Mathematics	94%	60%	55%	47%
Grade 5	Bangla Language	78%	67%	67%	60%
	Mathematics	100%	80%	70%	57%

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 subconstructs, with reference to the knowledge or skill. They determined that the alignment met draft, pre-determined pilot thresholds, which allowed the workshop to continue since there was 1) adequate alignment to enable item ratings (depth) and 2) sufficient coverage of the framework for process validity (breadth). However, even though the task worked well, the co-lead facilitators suggested the following minor changes to improve the process.

First, the facilitators thought that differentiating between content and performance standards in the Framework would promote greater understanding of these standards by the panelists. This would involve including the label of Content Standards in the Knowledge or Skill column and Performance Standards in the Global Minimum Proficiency Level column of the Framework, as well as continuing to explain these standards to the panelists.

Second, the facilitators thought that the four-point scale for the item-subconstruct ratings was too detailed, particularly in distinguishing between partial fit and slight fit. A three-point scale – complete fit, partial fit, and no fit – would be more appropriate.

Third, the facilitators thought that using medians instead of averages would provide a better reflection of the matches between the items and the subconstructs. This modification to the calculations was made during the calculations used in the first and second workshops.

Fourth, the facilitators thought that the alignment between language and some of the reading domains – aural listening comprehension and decoding – would be difficult for almost any group-administered, curriculum-based assessment. No changes were suggested at this time, but it should be reviewed by the subject matter experts who developed the content for the Framework.

Task 2

On the second day of the workshop, the facilitators built on Task 1 by training the panelists on matching the assessment items with the Levels and Descriptors in the Framework. The idea was to 1) increase the

panelists' knowledge of the items and Framework and 2) improve the identification of the Levels corresponding to the items, which would increase the accuracy and consistency of the item ratings in Task 3. The panelists started this task by taking the assessments themselves, making sure that their answers corresponded to the answer keys.

Then, the panelists expanded on the alignment activity by going through each item on their assessments and identifying the Level (performance standard) most appropriate for the item, i.e., in addition to the subconstruct(s) and the knowledge or skill (content standard) from the first day. They had discussions in small groups and focused on the following questions:

- What level of knowledge and skill is required to answer the items correctly?
- What makes an item easy or difficult, e.g. the stem and distractors in addition to the content?
- What is the lowest Level in the Descriptors that is most appropriate for the item?

The panelists wrote the subconstruct and the Level next to each of the items in the test booklet. If the item matched with more than one subconstruct – which was usually the case with partial fit or slight fit – the panelists wrote the additional subconstruct(s) and Level(s) next to the item. The completion of this task 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 Levels (performance standards), with reference to the subconstruct(s) and knowledge or skill (content standards). They wrote the information on their test booklets, and some of the panelists wrote the item numbers on appropriate places in the Framework.

First, the facilitators thought that having the panelists take the assessments themselves took up time that could have been devoted towards more efficiently increasing their understanding the assessment items and the Framework. This would mean eliminating taking the assessments and increasing the amount of time to match up the items with the Levels and Descriptors.

Second, the facilitators thought that information from the matching process could be recorded on both the test booklet and Framework, as was done by some panelists. All of the panelists could write the domain, construct, subconstruct, Level, and Descriptor for each item in their booklets, and then record the item number in their Framework. This would provide a better reference for Task 3 – by having the necessary information recorded on the two source documents – when they do the item ratings.

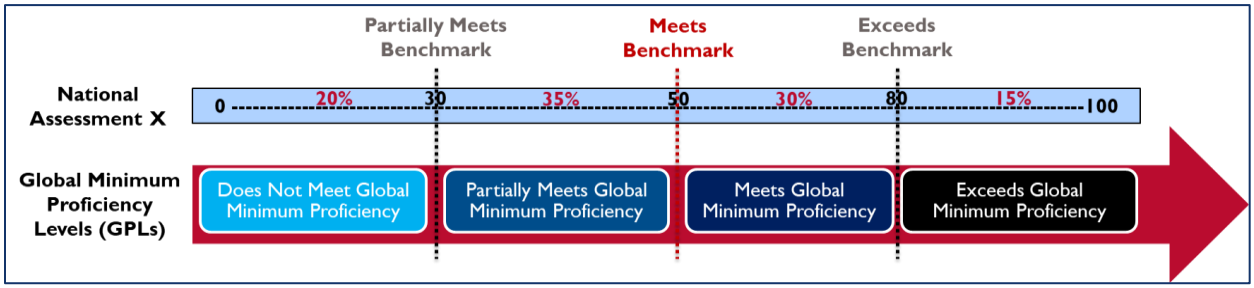
Third, the facilitators thought that having large group discussions after the small group item matching process would increase understanding of the Levels. It could be beneficial for the panelists to gather in a large group – led by the content facilitator – to go through each item and discuss the Levels needed to answer the items correctly. This would build more consensus prior to proceeding with Task 3. It would be possible within the timeframe of the workshop since the panelists would not be taking the assessments themselves.

Fourth, the facilitators thought that more emphasis could be placed on the item construction as a factor in item difficulty. For example, having a more difficult stem or distractors would increase the difficulty for the learners and would require a higher Level within the same subconstruct and knowledge or skill. It would be beneficial to spend more time on issues around item construction, including presenting an example of an item with the same stem but different distractors, and the subsequent effect of those distractors on item difficulty.

Task 3

On the third and fourth days of the workshop, the panelists received training on implementing the Angoff method to set global benchmarks. The facilitators showed the panelists how the benchmarking method would link the NSA to the Framework. The following graphic (Figure 3) showed a hypothetical example of the three benchmarks (30, 50, and 80 points) on a national assessment scale (0-100 points), with percentages of learners in each of the four Levels (20 percent, 35 percent, 30 percent, and 15 percent).

Figure 3: Example of an assessment and benchmarks



Then the panelists participated in a training session on conducting item ratings using the Angoff benchmarking method. The panelists separated into their panels to rate practice items in order to build on their understanding of the knowledge and skills needed to answer assessment items correctly.

After the panelists practiced on applying the Yes-No variation of Angoff, the co-lead facilitators trained the panelists on the item rating forms and procedures. The panelists divided into their two panels and conducted their first round of individual and independent ratings for each of the NSA items. One of four ratings – Just Partially Meets (JP), Just Meets (JM), Just Exceeds (JE), and Above Exceeds (AE) – was given by each panelist for each item. The steps in the judgment process are shown below (Figures 4 and 5).

1. Re-read each item, including the stem and the options (correct answer and distractors).
2. Match the items with the Descriptors (from Task 2) required for learners to answer correctly.
3. Conceptualize three Just Partially Meets, three Just Meets, and three Just Exceeds learners.
4. Follow the steps for rating the items, i.e., answering the questions and circling the ratings.

Figure 4: Steps for rating multiple choice items

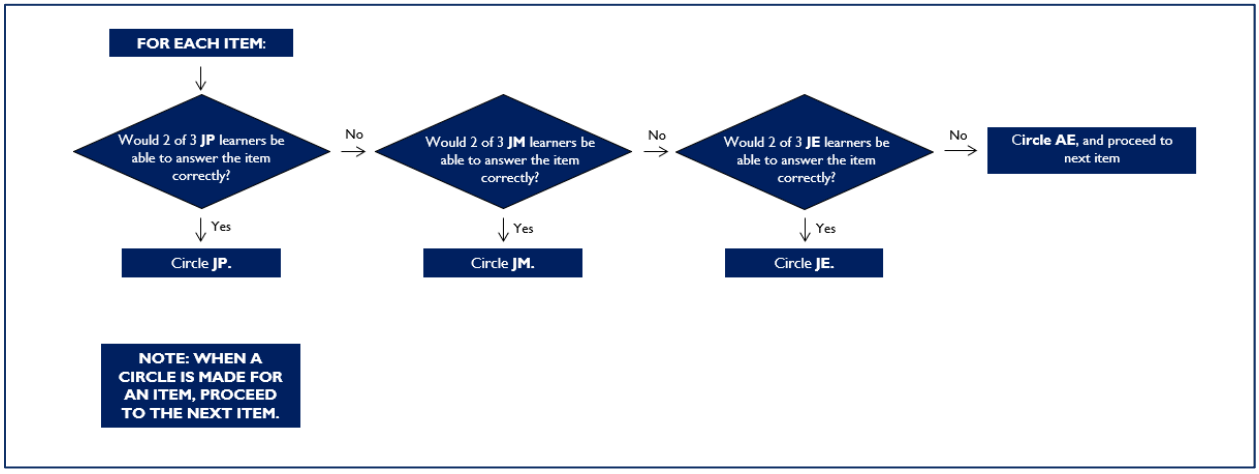
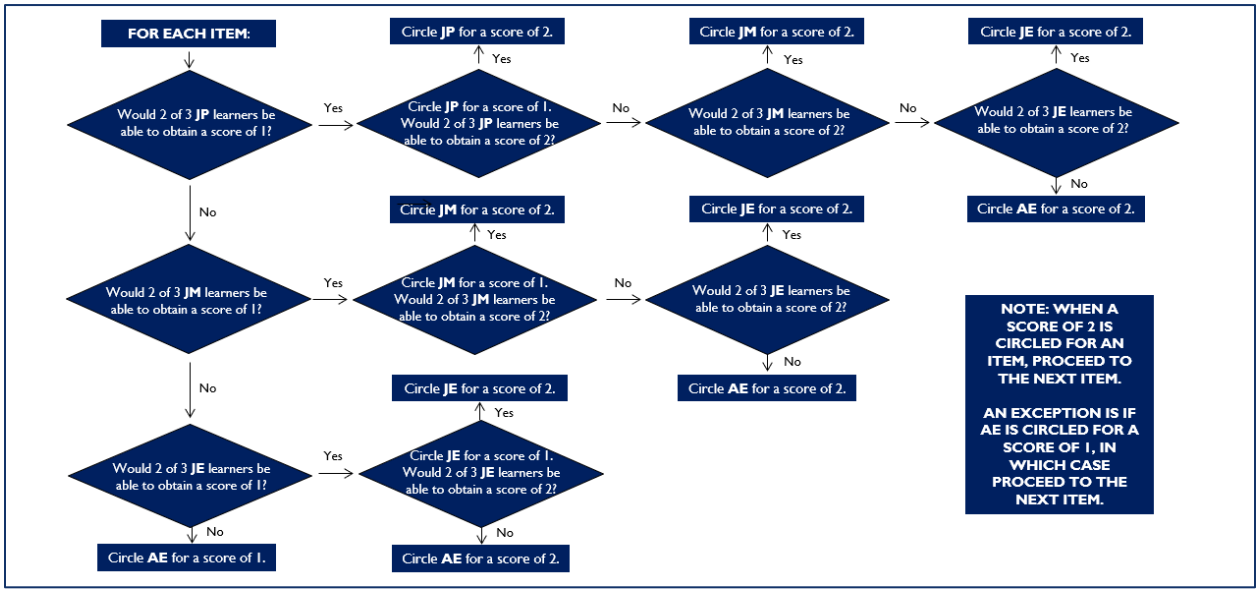


Figure 5: Steps for rating constructed response items



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

The fourth day of the workshop involved the second round of ratings, or the 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 the item construction, including the stem and distractors, 3) consider what learners would be able to do given any issues related to measurement error, and 4) 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.

Finally, the panelists completed workshop evaluation forms. The evaluation had six sections: 1) Framework Training, 2) Assessment Training, 3) Policy Linking Training, 4) Round 1 Ratings, 5) Round 2 Ratings, and 6) Overall Evaluation. The first three sections focused on guidance for the alignment, matching, and rating activities. The next two sections focused on the two rounds of ratings. The last section focused on the panelists' opinions on the organization and facilitation of the workshop.

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 multiple choice and constructed response items), comprehend the feedback from the first round, and make revisions for the second round. The facilitators suggested the following minor changes to improve the process.

First, the facilitators thought that the level of matching from Task 2 was not strong enough, thus causing some inconsistencies in the item ratings. This was reflected in the standard error statistics after the first round. The changes suggested under Task 2 should help with this. However, it might be necessary to go through the items one-by-one – as a group – prior to the first round of item ratings. Repeated discussions on the Descriptors and Levels appear to positively influence the reliability of the benchmarks between the panelists.

Second, the facilitators thought that more training time spent on calculating benchmarks would be helpful in understanding the benchmarking numbers. The panelists should be instructed on totaling their JP, JM, and JE columns and then calculating each of the three benchmarks. This would also help with making revisions during the second round. Sometimes the panelists did not have a solid sense of what happens to the benchmarks if they change the ratings of items.

Third, the facilitators thought that the workshops were good starting points towards establishing thresholds, specifically for the alignment and the consistency. Based on this initial pilot workshop, reasonable thresholds appear to be at least 75 percent for the item alignment (depth), at least 50 percent for the subconstruct alignment (breadth), less than 1.00 for standard errors (on test of average length, or 30 items), and 0.70 for inter-and intra-rater reliability. It should be possible to establish such reasonable thresholds by the end of the piloting. The thresholds would be useful as indicators both during workshops – after Round 1 – and of the workshop results – after Round 2.

Pilot Workshop Results

The co-lead facilitators analyzed the panelists' ratings and benchmarks after Rounds 1 and 2. This included calculating the following for the panels: 1) benchmarks, 2) score ranges, 3) impact data (using the score distributions), and 4) consistency of the results. All analyses are presented by round, with the exception of the consistency statistics, which are provided for both rounds together.

Round 1

The co-facilitators produced summary tables and graphs from the first round, which showed the initial benchmarks, score ranges, and impact data for each Level (Tables 3, 4, and 5). The impact data examined the percentages of scores in the different Levels. All analyses were conducted by grade and subject.

The impact data were variable in Round 1. Grade 3 Bangla language had the highest percentage of scores in meets or exceeds (80.3 percent), while grade 5 mathematics had the lowest (37.3 percent).

Table 3: Round 1 benchmarks by grade and subject

Level	Benchmarks (in points)			
	Grade 3		Grade 5	
	Bangla Language	Mathematics	Bangla Language	Mathematics
Does Not Meet	--	--	--	--
Partially Meets	8	10	13	7
Meets	16	23	29	26
Exceeds	32	34	41	39

Table 4: Round 1 score ranges by grade and subject

Level	Score Ranges (in points)			
	Grade 3		Grade 5	
	Bangla Language	Mathematics	Bangla Language	Mathematics
Does Not Meet	0-7	0-9	0-12	0-6
Partially Meets	8-15	10-22	13-28	7-25
Meets	16-31	23-33	29-40	26-38
Exceeds	32-36	34-39	41-44	39-44

Table 5: Round 1 impact data by grade and subject

Level	Impact Data (in percentages)			
	Grade 3		Grade 5	
	Bangla Language	Mathematics	Bangla Language	Mathematics
Does Not Meet	2.6%	9.7%	2.8%	2.0%
Partially Meets	12.1%	38.6%	39.7%	50.7%
Meets	70.1%	35.8%	55.6%	37.2%
Exceeds	10.2%	15.9%	1.9%	10.1%

Round 2

After providing the results from the initial benchmarks in Round 1 to the panelists and conducting the Round 2 ratings, the co-facilitators produced a parallel set of summary tables and graphs with final benchmarks, score ranges, and impact data for each Level (Tables 6, 7, and 8). Again, all analyses were conducted by grade and subject.

The impact data were less variable in Round 2. Grade 3 Bangla language had the highest percentage of scores in meets or exceeds (77.8 percent), while grade 5 mathematics had the lowest (47.3 percent).

Table 6: Round 2 benchmarks by grade and subject

Level	Benchmarks (in points)			
	Grade 3		Grade 5	
	Bangla Language	Mathematics	Bangla Language	Mathematics
Does Not Meet	--	--	--	--
Partially Meets	10	6	12	8
Meets	19	18	29	26
Exceeds	32	35	42	40

Table 7: Round 2 score ranges by grade and subject

Level	Score Ranges (in points)			
	Grade 3		Grade 5	
	Bangla Language	Mathematics	Bangla Language	Mathematics
Does Not Meet	0-9	0-5	0-11	0-7
Partially Meets	10-18	6-17	12-28	8-25
Meets	19-31	18-34	29-41	26-39
Exceeds	32-36	35-39	42-44	40-44

Table 8: Round 2 impact data by grade and subject

Level	Impact Data (in percentages)			
	Grade 3		Grade 5	
	Bangla Language	Mathematics	Bangla Language	Mathematics
Does Not Meet	4.4%	2.4%	2.1%	3.2%
Partially Meets	17.8%	30.2%	45.3%	49.5%
Meets	62.6%	54.6%	51.8%	39.8%
Exceeds	15.2%	12.8%	0.8%	7.5%

Consistency

Feedback data were provided on the consistency in panelists' ratings. The feedback data included location statistics (Figures 6, 7, 8, and 9), standard errors, and inter- and intra-rater reliability (Tables 9 and 10).

Figure 6: Round 1 location statistics for grade 3 Bangla language and mathematics

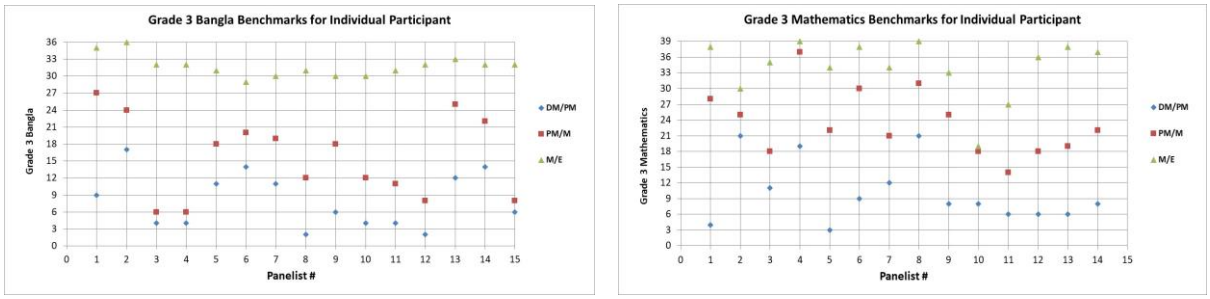


Figure 7: Round 2 location statistics for grade 3 Bangla language and mathematics

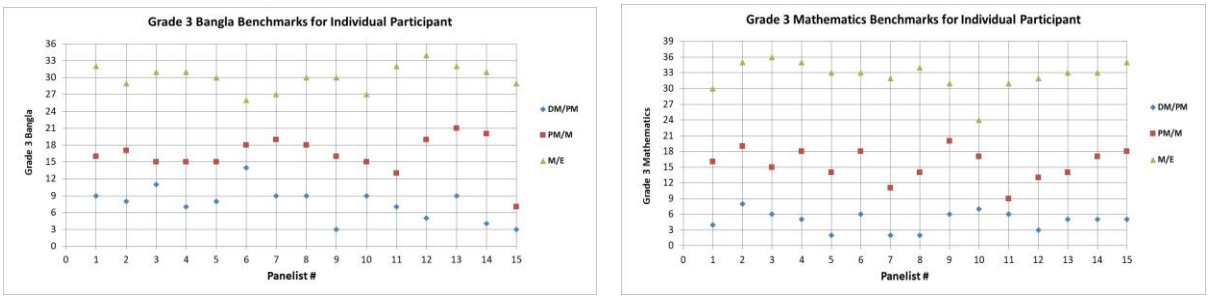


Figure 8: Round 1 location statistics for grade 5 Bangla language and mathematics

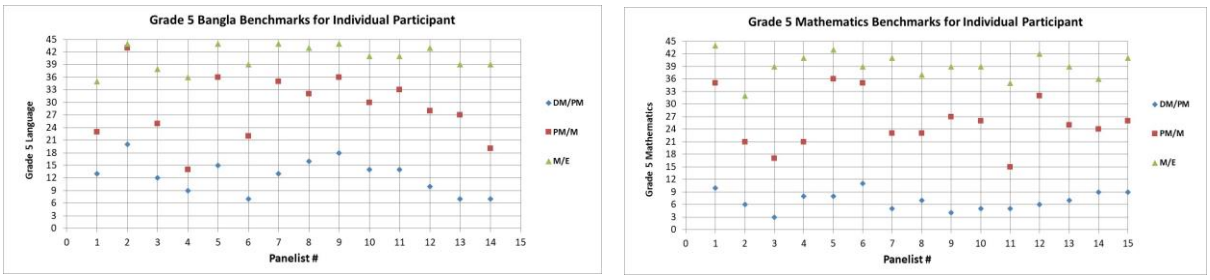
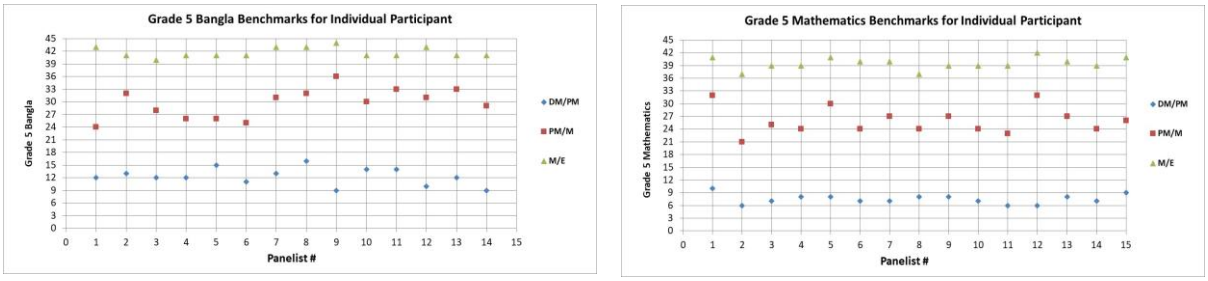


Figure 9: Round 2 location statistics for grade 5 Bangla language and mathematics



The location statistics showed that the consistency in the panelists' ratings improved substantially from Round 1 to Round 2 for both grades and subjects. The partially meets (green triangles), meets (red squares), and exceeds (blue dots) global minimum proficiency benchmarks in Round 2 also showed less overlap and more separation.

The other three reliability statistics were calculated as follows, with provisional thresholds:

- Standard error of measurement (SEM) is calculated at the benchmark level. Values of less than 1.00 (on a test of average length, or 30 items) indicates substantial agreement between the panelists in their estimated benchmarks.
- Inter-rater consistency (IRC) is calculated at the item level. Values of 0.70 or greater indicate substantial agreement between the panelists in their item ratings.
- Intra-rater reliability (IRR) is also calculated at the item level. A high value indicates high consistency. Values of 0.70 or greater indicate substantial consistency in the panelists' item ratings.

Most of the SEM of the benchmarks from Round 1 were generally above the provisional threshold of 1.00, particularly for the Meets benchmark. All of the SEM were below 1.00 in Round 2 (Table 9).

Table 9: SEM by grade, subject, and round

Statistic	Grade 3				Grade 5			
	Bangla Language		Mathematics		Bangla Language		Mathematics	
	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2
Partial Meets	1.30	0.80	1.67	0.50	1.14	0.58	0.61	0.30
Meets	1.92	0.91	1.75	0.82	2.16	0.98	1.71	0.86
Exceeds	0.50	0.58	1.54	0.78	0.86	0.33	0.84	0.38

Most of the IRC and IRR statistics were around the threshold of 0.70 in both Rounds 1 and 2. There was little or no improvement from Round 1 to Round 2 for either statistic (Table 10).

Table 10: IRC and IRR by grade, subject, and round

Statistic	Grade 3				Grade 5			
	Bangla Language		Mathematics		Bangla Language		Mathematics	
	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2
IRC	0.75	0.79	0.68	0.73	0.75	0.78	0.71	0.71
IRR	0.77	0.76	0.82	0.83	0.81	0.81	0.84	0.84

Evaluation

The evaluation results from the panelists were reasonably high, with averages between 3.1 and 3.4 out of a maximum of 4.0 (minimum 1.0). They were similar by grade and subject (Table 11).

Table 11: Workshop evaluation results by category

Category	Average Ratings (maximum of 4.0)			
	Grade 3		Grade 5	
	Bangla language	Mathematics	Bangla language	Mathematics
Framework Training	3.5	3.3	3.3	3.4
Assessment Training	3.3	3.3	3.3	3.3
Policy Linking Training	3.4	3.2	3.4	3.4
Round 1 Item Ratings	3.2	3.2	3.4	3.4
Round 2 Item Ratings	3.3	3.3	3.1	3.1
Overall Evaluation	3.4	3.3	3.4	3.3

A few comments were received from the panelists. They focused on the importance of the training, the need for additional workshops, and the lack of adequate time for certain activities.

Comments

The pilot workshop results showed that the panelists able to implement the instructions and make judgments with adequate consistency. The panelists and MoPME observers thought that the impact data from Round 2 were reasonable. The statistical analyses of indicators, i.e., the standard errors and inter-rater consistency indicators, did not meet pre-determined draft thresholds in Round 1, but this improved to adequate levels in Round 2. Otherwise, these comments are similar to those from Task 3. The panelists reported having satisfaction with the training and confidence in their ratings and benchmarks. The facilitators suggested the following minor changes to improve the process.

First, the facilitators thought that the presentations of the data were adequate but could be improved. A constraint was the amount of preparation time for the workshop, since spreadsheets need to be pre-programmed using the test format from the country. Additional graphics, which can be included in the spreadsheets given more advance time, would be helpful in communicating the results.

Second, the facilitators thought that more time should be given to discussions on the feedback between Rounds 1 and 2. As mentioned above, some of this time could be spent on understanding the calculations of the benchmarks, which would then help making revisions during the second round.

Policy Linking Recommendations

These policy linking recommendations are summarized from the comments in the Tasks 1, 2, and 3 sections, and in the Results section.

Task 1

First, differentiating between content and performance standards in the Framework would promote greater understanding of these standards by the panelists. This would involve including the label of Content Standards in the Knowledge or Skill column and Performance Standards in the Global Minimum Proficiency Level column of the Framework, as well as continuing to explain these standards to the panelists.

Second, the four-point scale for the item-subconstruct ratings was too detailed, particularly in distinguishing between partial fit and slight fit. A three-point scale – complete fit, partial fit, and no fit – would be more appropriate.

Third, using medians instead of averages would provide a better reflection of the matches between the items and the subconstructs. This modification to the calculations was made during the calculations used in the first and second workshops.

Fourth, the alignment between language and some of the reading domains – aural listening comprehension and decoding – would be difficult for almost any group-administered, curriculum-based assessment. No changes were suggested at this time, but it should be reviewed by the subject matter experts who developed the content for the Framework.

Task 2

First, having the panelists take the assessments themselves took up time that could have been devoted towards more efficiently increasing their understanding the assessment items and the Framework. This would mean eliminating taking the assessments and increasing the amount of time to match up the items with the Levels and Descriptors.

Second, the information from the matching process could be recorded on both the test booklet and Framework, as was done by some panelists. All of the panelists could write the domain, construct, subconstruct, Level, and Descriptor for each item in their booklets, and then record the item number in their Framework. This would provide a better reference for Task 3 – by having the necessary information recorded on the two source documents – when they do the item ratings.

Third, having large group discussions after the small group item matching process would increase understanding of the Levels. It could be beneficial for the panelists to gather in a large group – led by the content facilitator – to go through each item and discuss the Levels needed to answer the items correctly. This would build more consensus prior to proceeding with Task 3. It would be possible within the timeframe of the workshop since the panelists would not be taking the assessments themselves.

Fourth, more emphasis could be placed on the item construction as a factor in item difficulty. For example, having a more difficult stem or distractors would increase the difficulty for the learners and would require

a higher Level within the same subconstruct and knowledge or skill. It would be beneficial to spend more time on issues around item construction, including presenting an example of an item with the same stem but different distractors, and the subsequent effect of those distractors on item difficulty.

Task 3

First, the level of matching from Task 2 was not strong enough, thus causing some inconsistencies in the item ratings. This was reflected in the standard error statistics after the first round. The changes suggested under Task 2 should help with this. However, it might be necessary to go through the items one-by-one – as a group – prior to the first round of item ratings. Repeated discussions on the Descriptors and Levels appear to positively influence the reliability of the benchmarks between the panelists.

Second, more training time spent on calculating benchmarks would be helpful in understanding of benchmarking numbers. The panelists should be instructed on totaling their JP, JM, and JE columns and then calculating each of the three benchmarks. This would also help with making revisions during the second round. Sometimes the panelists did not have a solid sense of what happens to the benchmarks if they change the ratings of items.

Third, the workshops were good starting points towards establishing thresholds, specifically for the alignment and the consistency. Based on this initial pilot workshop, reasonable thresholds appear to be about 75 percent for the item alignment (depth), 50 percent for the subconstruct alignment (breadth), less than 1.00 for standard errors on a 30-item test, and 0.70 for inter- and intra-rater reliability. It should be possible to establish reasonable levels for these kinds of thresholds in subsequent pilots. The thresholds would be useful both during the workshops – particularly after Round 1 – and also as quality control checks on the workshop results – after Round 2.

Results

First, the presentations of the data were adequate but could be improved. A constraint was the amount of preparation time for the workshop, since spreadsheets need to be pre-programmed using the test format from the country. Additional graphics, which can be included in the spreadsheets given more advance time, would be helpful in communicating the results.

Second, more time should be given to discussions on the feedback between Rounds 1 and 2. As mentioned above, some of this time could be spent on understanding the calculations of the benchmarks, which would then help making revisions during the second round.

Policy Linking References

American Institutes for Research (2016). *Bangladesh national student assessment 2015 grades 3 and 5: Draft technical report*. Washington, DC: American Institutes for Research.

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.

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

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

Ferdous, A., Kelly, S., Davis, J., & Watson, C. (2019). *Policy linking toolkit: Linking assessments to a global proficiency framework*. Washington, DC: U.S. 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.

Ministry of Primary and Mass Education. (2018). *National student assessment 2017 grades 3 and 5*. Dhaka, Bangladesh: Directorate of Primary Education.

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.