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

UNESCO Institute for Statistics (UIS)
National Council for Educational Research and Training (NCERT)

December 15, 2019
Central Institute of Educational Technology (CIET) at NCERT
Sri Aurbindo Marg, New Delhi, India
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 National Council of Educational Research and Training (NCERT) was critical for the success of the workshops.

Second, the management support provided by officials from the UNESCO Institute of Statistics (UIS) in Montreal and the UNESCO regional office in New Delhi 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 specialist from UIS and the U.S. Agency for International Development (USAID) was vital 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 from DFID, Gates, the World Bank, and numerous other partners.

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

Abdullah Ferdous

Jeff Davis

Sean Kelly

Table of Contents

Acknowledgements 2

Acronyms and Abbreviations 4

Policy Linking Overview 5

- Development 5
- Piloting 6
- Finalization 6

Pilot Workshop Preparation 7

- Planning 7
- Assessments 7
- Panelists 9
- Benchmarks 9

Pilot Workshop Implementation 10

- Task 1 10
- Task 2 12
- Task 3 13

Pilot Workshop Results 15

- Round 1 16
- Round 2 16
- Consistency 17
- Evaluation 20

Policy Linking Recommendations 21

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

Policy Linking References 23

Acronyms and Abbreviations

CAT	Comparisons, Aggregation, and Tracking
DFID	U.K. Department for International Development
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)
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
NAS	National Achievement Survey
NCERT	National Council of Educational Research and Training
PLT	Policy Linking Toolkit (or Toolkit)
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, such as in Djibouti. 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, 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 measurement experts 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, representatives from 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 three types of information, abbreviated as CAT: 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 and the successful completion of the first pilot workshop in Bangladesh, UIS planned its second pilot workshop in India, with support from MSI, DFID, and Gates. The objective was setting global benchmarks on the 2017 National Achievement Survey (NAS) at grades 3 and 5 in English language and mathematics. The Ministry of Human Resource Development (MHRD) and the National Council of Educational Research and Training (NCERT) approved two four-day workshops at the Central Institute of Educational Technology in the NCERT headquarters, Sri Aurbindo Marg in New Delhi, from November 12 to 21, 2019. MSI assigned two international co-lead facilitators for the workshops. NCERT provided oversight from senior staff members, facilitation from content experts, and support from data analysts and logisticians. The workshops were organized as follows:

Workshop 1: Grade 3 English language and mathematics

Tuesday November 12 to Friday November 15

Workshop 2: Grade 5 English language and mathematics

Monday November 18 to Thursday November 21

For the workshops, the co-lead facilitators prepared three 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 the analysis of the workshop results, including the participants' workshop evaluation data.

Assessments

The NAS has been conducted since 2001. It has been successfully administered over four cycles for grades 3, 5, and 8, along with two cycles for class 10. It measures the achievement of learners relative to the

learning outcomes in the primary school curriculum. The assessments used in this workshop were administered in 2017 to learners in representative samples of schools in grades 3 and 5 English language and mathematics. NAS was also administered in environmental studies (EVS) in grades 3 and 5, as well as in language, mathematics, science, and social science in grade 8. Note that the NAS administered the same assessments in 20 languages, with translation from the original English and Hindi language versions, but this policy linking workshop only used the English language version, with workshop participants coming from states and schools where English is the medium of assessment.

The overall goal of the NAS is to provide systematic assessment results to improve pedagogical processes and learner competencies so that graduates of the system will have essential knowledge and skills for the 21st century. The specific objective of the NAS is to collect and analyze information on the effectiveness of the school system at the different levels, i.e., national, state, and district. Based on disaggregated data and large sample sizes, NCERT produced State Learning Reports (SLRs) and District Report Cards (DRCs) to provide an overview of learning outcomes at those levels. The SLRs and DRCs were shared with stakeholders so that state/district-specific intervention programs could be designed, refined, implemented, and evaluated. NCERT held regional consultation workshops to sensitize the states and districts on using assessment results. This had the additional purpose of helping to gauge the need for developing capacities at the state and district level for interpreting assessment data.

NCERT followed strict procedures in developing and administering the NAS. This included the following:

- Training state and district stakeholders on the assessment procedures;
- Developing a detailed framework for assessing learning competencies;
- Creating guidelines and protocols for the administration of the assessments;
- Building various software templates for data capture, storage, and analysis;
- Sampling representative schools, teachers, and learners at the district level;
- Training data collectors and monitoring data collection at the school level;
- Analyzing data at the different levels and producing the DRCs and SLRs; and
- Developing training packages for data use and addressing learning gaps.

The grades 3 and 5 English language and mathematics assessments had 25 multiple choice (objective) items. The assessments had two forms, with 10 unique items and 5 common items on each form (Table 1). The two forms were administered through spiraling, i.e., with alternate learners taking different forms. The forms were statistically equated through the application of an item response theory (IRT) model. Scores from the 25 items were used to report on the school system at the national, state, and district levels.

Table 1: Number of items and score points per form

		Items			Points
Grade	Subject	Unique	Common	Total	
Grade 3	English Language	10	5	15	15
	Mathematics	10	5	15	15
Grade 5	English Language	10	5	15	15
	Mathematics	10	5	15	15

The sample size for the 2017 NAS was 110,00 schools and 2.2 million learners in the three grade levels from all 36 states in India. The assessments for all learners were conducted on a single day.

Panelists

The DPE invited four groups (or panels) of 18 panelists for grades 3 and 5 English language and mathematics, or a total of 72 panelists. Three panelists were invited from each of the six targeted states for each grade and subject. The invited panelists were comprised of approximately 80 percent head teachers or classroom teachers and 20 percent experts in curriculum, teacher training, and pedagogy. There was near equal gender representation. A total of 58 of the invited panelists participated in the workshops. Fewer panelists participated in the grade 3 workshops due to logistical issues in two states. The number of panelists by grade, subject, and education level is provided below (Table 2).

Table 2: Panelists' background information

Grade	Subject	Total	Education Level	
			< B.A.	≥ B.A.
Grade 3	English Language	13	0	13
	Mathematics	10	1	9
Grade 5	English Language	18	2	16
	Mathematics	17	3	14
Totals	--	58	6	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 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.

Three tasks mentioned above 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 to judge the alignment of the assessments to the Framework. The second task involved training on the Framework and matching items with the Levels and Descriptors to judge the skills and abilities needed by learners to answer the items correctly. The third task involved training on the Angoff method and conducting two rounds of item ratings to set initial and final benchmarks on the four assessments.

NCERT organized the participants and the venue, and provided the 2017 NAS instruments, answer keys, and data sets at the beginning of the workshops. 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 (i.e., partially meets, meets, and exceeds) 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; these data were also entered into pre-programmed spreadsheets to calculate averages by item and category.

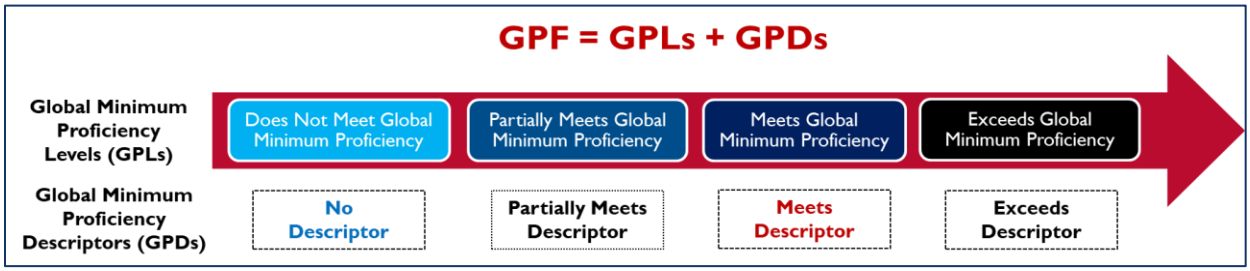
Pilot Workshop Implementation

The pilot workshop involved implementing the three tasks to link the NAS assessments with the Framework by establishing reliable, valid, and fair global benchmarks. Each task is described below, with comments that highlight lessons learned for making improvements to subsequent pilot workshops.

Task 1

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

Figure 1: Global proficiency framework



The graphic shows the two components of the Framework: Global Minimum Proficiency Levels (GPLs, or Levels) and Global Minimum Proficiency Descriptors (GPDs, or Descriptors).

The structure of the Descriptors in the Framework – made up of domains, constructs, subconstructs, and knowledge or skills – was introduced using the following mathematics example (Figure 2).

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 overall knowledge or skills for each subconstruct were identified as content standards, while the knowledge and skills for each construct at the three global minimum proficiency levels were identified as performance standards. In Task 1, the panelists were instructed to focus on the subconstruct(s) and the knowledge or skill associated with each subconstruct.

The panelists were trained on a three-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;
- 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 associated with only one subconstruct, 2) partial fit was usually associated with more than one subconstruct, and 3) no fit was not associated with any subconstruct.

After the panelists had rated each of the items according to the fit with the subconstructs in the Framework, the co-lead facilitators entered rating totals from each panelist into spreadsheets by subject. 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 grades 3 and 5 English language domains, constructs, and 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	English Language	96%	33%	40%	38%
	Mathematics	92%	100%	85%	79%
Grade 5	English Language	92%	33%	50%	50%
	Mathematics	96%	100%	69%	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 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 facilitators had some concerns about the relatively low number of items per assessment form. While the reasons for having 15 items per form were well understood and made sense given the large sample size and complicated logistics of administering assessments nationwide in India, it would probably be worthwhile to examine the psychometric properties of the assessments prior to the policy linking workshops – i.e., number of items (mostly for validity), internal consistency reliability, and quality of the data set (missing data). Note that the NAS assessments reached high levels of validity and reliability.

Second, with the alignment, the facilitators differentiated between content and performance standards in the Framework and explained the two types of standards to the panelists. However, they did not include the label of Content Standards in the Knowledge or Skill column and Performance Standards in the Global Minimum Proficiency Level column of the Framework. This needs to be done for subsequent workshops.

Third, the facilitators reduced the four-point scale for the item-subconstruct ratings from the previous pilot workshop since it was too detailed, particularly in distinguishing between partial fit and slight fit. They used a three-point scale – complete fit, partial fit, and no fit – which was more appropriate.

Fourth, the facilitators reiterated their observation 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 for the India pilot, 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 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 for each item from the Framework in their test booklets, and they wrote the item numbers in appropriate places in the Framework. This allowed the panelists to cross-reference the items, test booklets, and Framework. 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 facilitators eliminated the part in the Toolkit in which the panelists took the assessments themselves. They used the additional time to deepen their understanding of the assessment items and Framework, especially by matching the items with the Levels in the Framework.

Second, the facilitators instructed the panelists to record the matching information in both the test booklet and Framework. All of the panelists wrote the domain, construct, subconstruct, Level, and Descriptor for each item in their booklets, and then recorded the item number in their Framework. This provided a cross-reference for Task 3 when they did the item ratings.

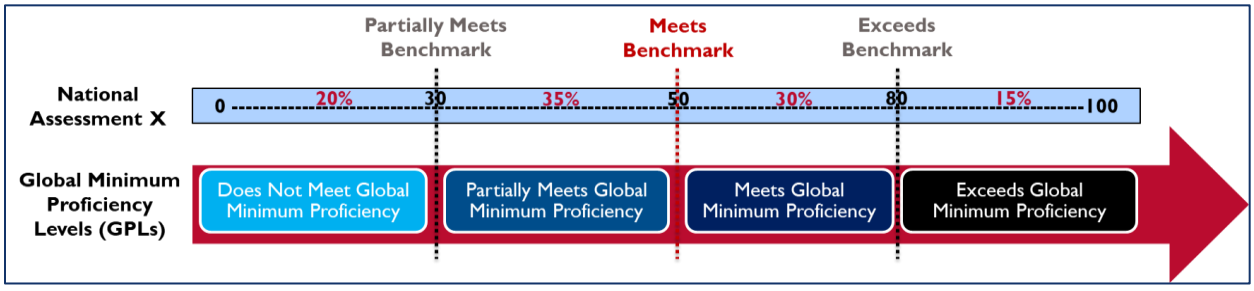
Third, the facilitators had large group discussions after the small group item matching process. It was beneficial for the panelists to gather in a large group – led by the content facilitator – and go through each item and discuss the Levels needed to answer the items correctly. This built more consensus and consistency prior to proceeding with Task 3. Again, it was possible within the timeframe of the workshop since the panelists did take the assessments themselves but rather moved directly to the matching process.

Fourth, the facilitators placed more emphasis on the item construction as a factor in item difficulty. In particular, they instructed the panelists that 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. The facilitators presented and discussed an example of an item with the same stem but different sets of distractors, and the subsequent effect of those sets of 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

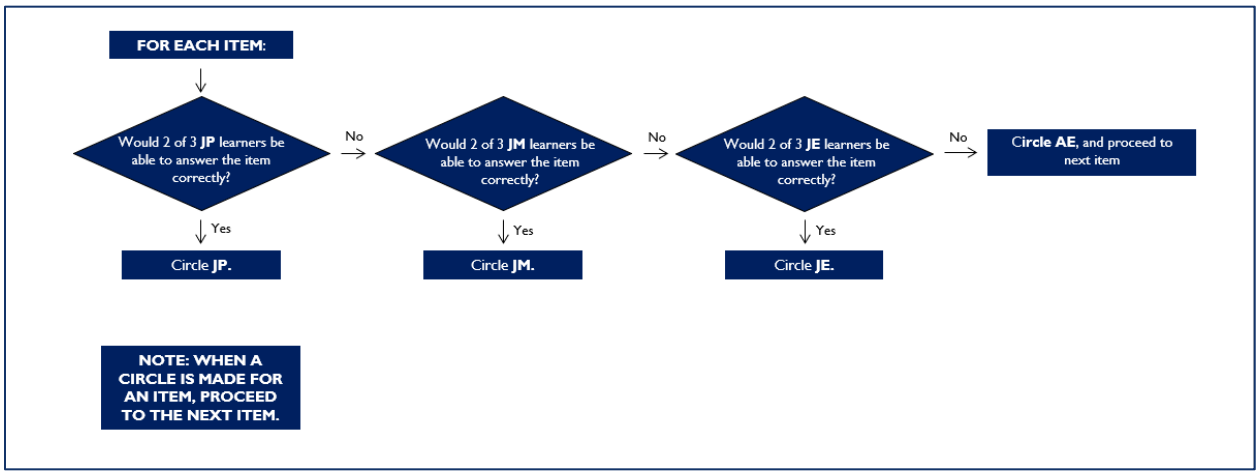


Then the panelists participated in a training session on conducting item ratings using the Angoff benchmarking method. The panelists separated into their panels to rate practice items in order to build on their understanding of the knowledge and skills needed to answer assessment items correctly.

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 (Figure 4).

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



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 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 facilitators thought that the level of matching from Task 2 was better than in the previous workshop. Part of the reason was that there were only 15 items on the assessment form, which allowed the panelists to concentrate on fewer items. This was reflected in the lower standard error estimates after the first round.

Second, the facilitators implemented additional large group work during Task 3. Prior to conducting Round 2, they led a session with each panel to again go through each item and discuss the matching with the Levels in the Framework. Repeated discussions on the Descriptors and Levels appear to positively influence the reliability of the benchmarks between the panelists.

Third, the facilitators spent more training time spent on calculating benchmarks to promote better understanding the benchmarking numbers. The panelists were instructed to total their JP, JM, and JE columns and then calculate each of their three benchmarks. This helped with making revisions during the second round. Otherwise, the panelists did not have a solid sense of what happens to the benchmarks if they change the ratings of items.

Fourth, the facilitators thought that the workshops continued as good starting points towards establishing thresholds, specifically for the alignment and the consistency. 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 subconstruct alignment (breadth), less than 1.00 for standard errors (with some variation, i.e., this would apply to an average assessment with a length of 30 items), and 0.70 for inter-and intra-rater reliability. More work needs to be done with additional pilots, and it should be possible to establish such 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 after Rounds 1 and 2. For the English language and mathematics panels at each grade level, 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.

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 English language had the highest percentage of scores in the combined meets or exceeds (60.6 percent) categories, while grade 3 mathematics had the lowest (36.5 percent).

Table 3: Round 1 benchmarks by grade and subject

Level	Benchmarks (in points)			
	Grade 3		Grade 5	
	English Language	Mathematics	English Language	Mathematics
Does Not Meet	--	--	--	--
Partially Meets	4	6	3	3
Meets	10	12	9	9
Exceeds	13	14	14	13

Table 4: Round 1 score ranges by grade and subject

Level	Score Ranges (in points)			
	Grade 3		Grade 5	
	English Language	Mathematics	English Language	Mathematics
Does Not Meet	0-3	0-5	0-2	0-2
Partially Meets	4-9	6-11	3-8	3-8
Meets	10-12	12-13	9-13	9-12
Exceeds	13-15	14-15	14-15	13-15

Table 5: Round 1 impact data by grade and subject

Level	Impact Data (in percentages)			
	Grade 3		Grade 5	
	English Language	Mathematics	English Language	Mathematics
Does Not Meet	3.4%	14.7%	4.1%	6.2%
Partially Meets	36.0%	48.8%	51.0%	55.0%
Meets	28.9%	18.9%	40.2%	27.8%
Exceeds	31.7%	17.6%	4.7%	11.0%

Round 2

After providing the results from the initial benchmarks in Round 1 to the panelists and conducting the Round 2 ratings, the co-lead 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.

There was little difference in the impact data, with only an increase of one point for the grade 5 English language meets benchmark. Grade 3 English language had the highest percentage of scores in meets or exceeds (60.6 percent), while grade 5 English language had the lowest (35.6 percent).

Table 6: Round 2 benchmarks by grade and subject

Level	Benchmarks (in points)			
	Grade 3		Grade 5	
	English Language	Mathematics	English Language	Mathematics
Does Not Meet	--	--	--	--
Partially Meets	5	5	3	2
Meets	10	12	10	9
Exceeds	13	14	14	13

Table 7: Round 2 score ranges by grade and subject

Level	Score Ranges (in points)			
	Grade 3		Grade 5	
	English Language	Mathematics	English Language	Mathematics
Does Not Meet	0-4	0-4	0-2	0-1
Partially Meets	5-9	6-11	3-9	2-8
Meets	10-12	12-13	10-13	9-12
Exceeds	13-15	14-15	14-15	13-15

Table 8: Round 2 impact data by grade and subject

Level	Impact Data (in percentages)			
	Grade 3		Grade 5	
	English Language	Mathematics	English Language	Mathematics
Does Not Meet	6.7%	9.5%	4.1%	2.4%
Partially Meets	32.7%	54.0%	60.3%	58.8%
Meets	28.9%	18.9%	30.9%	27.8%
Exceeds	31.7%	17.6%	4.7%	11.0%

Consistency

Feedback data were provided on the consistency in panelists' ratings. The feedback data included location statistics (Figures 5, 6, 7, and 8), standard errors of measurement (SEM), and inter-rater consistency (IRC) and intra-rater reliability (IRR) (Tables 9 and 10).

The location statistics are provided only for the final benchmarks. They showed strong consistency in the panelists' ratings for both grades and subjects, particularly for grade 5. The benchmarks for partially meets

(blue diamonds) benchmarks showed no overlap while the meets (green squares) and exceeds (blue triangles) benchmarks showed some overlap, except for grade 5 English language.

Note that the last panelist in each figure shows the average for all panelists. Also, as shown in Table 2, the number of panelists varied with each grade and subject. The lowest number of panelists was ten (grade 3 mathematics), but this was adequate for reliable benchmarks.

Figure 5: Location statistics for grade 3 English language (13 panelists)

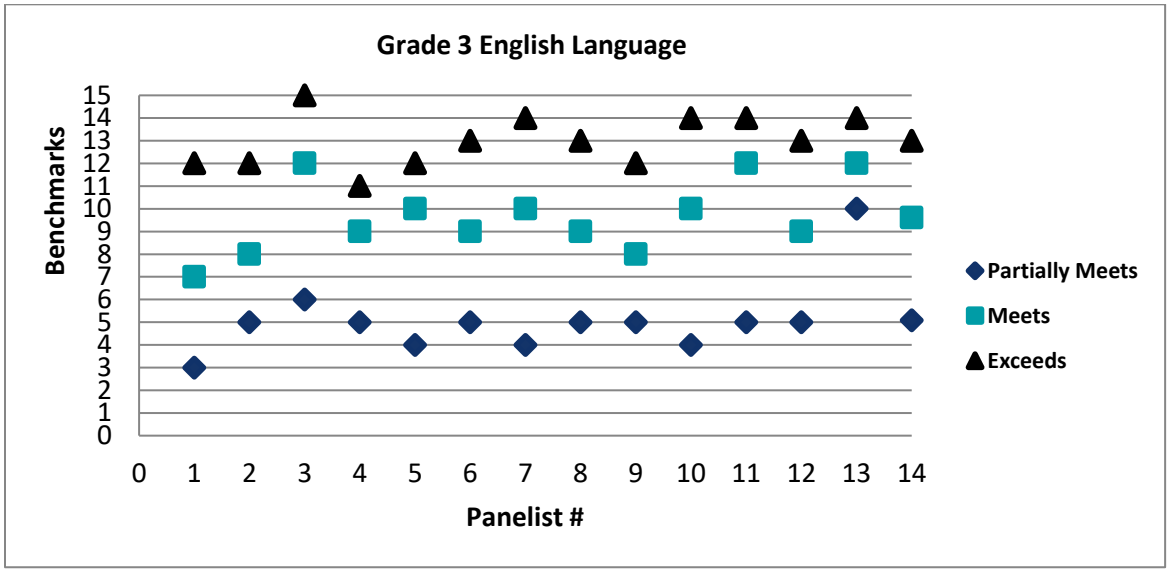


Figure 6: Location statistics for grade 3 mathematics (10 panelists)

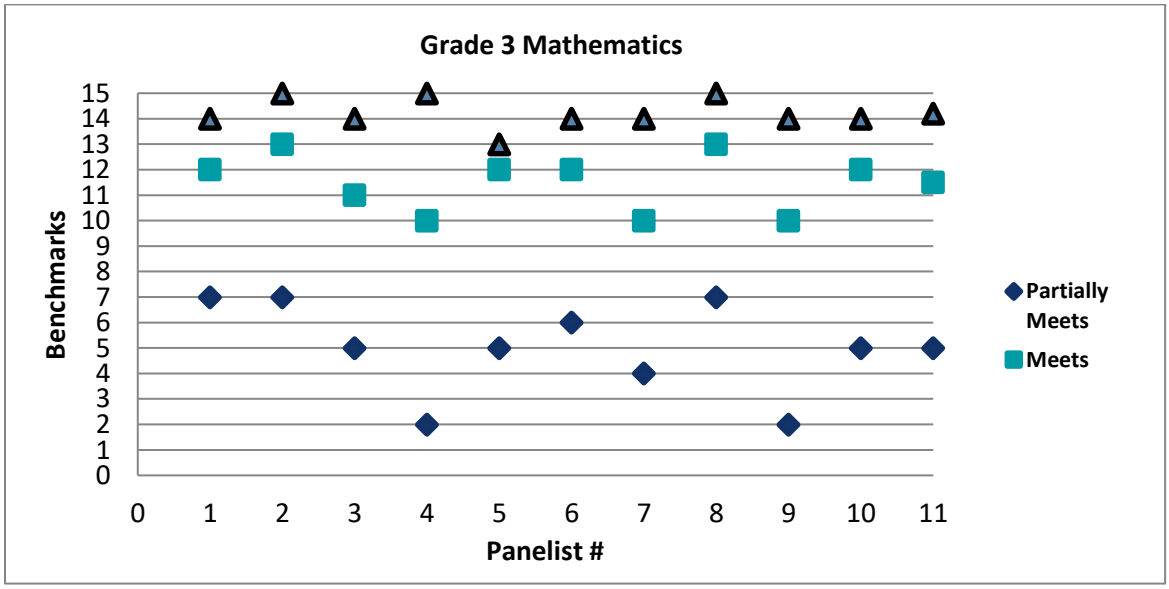


Figure 7: Location statistics for grade 5 English language (18 panelists)

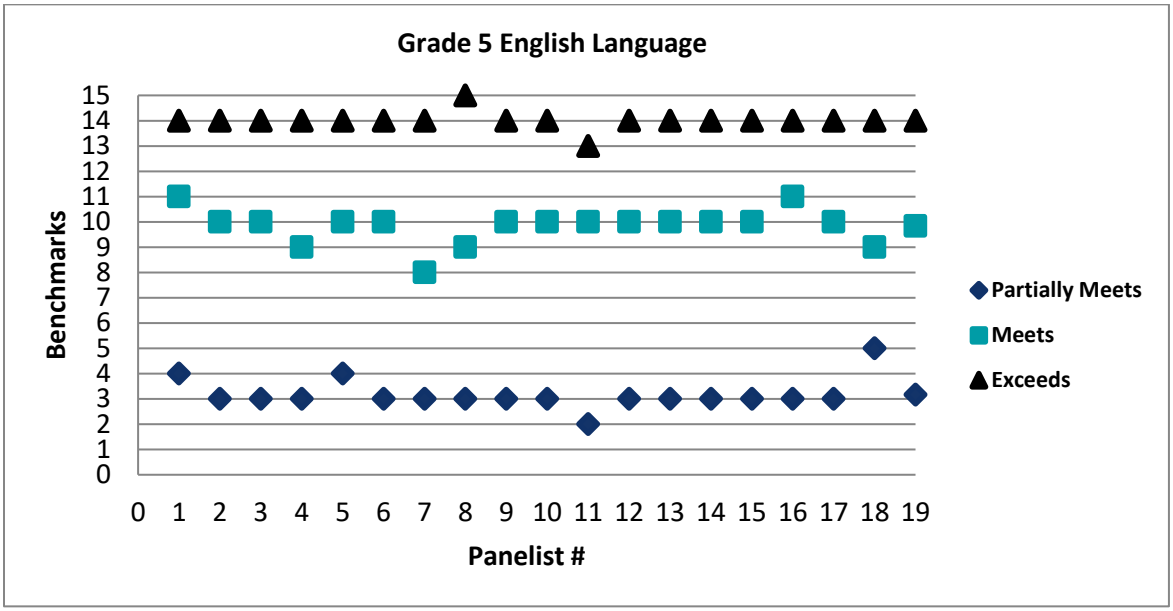
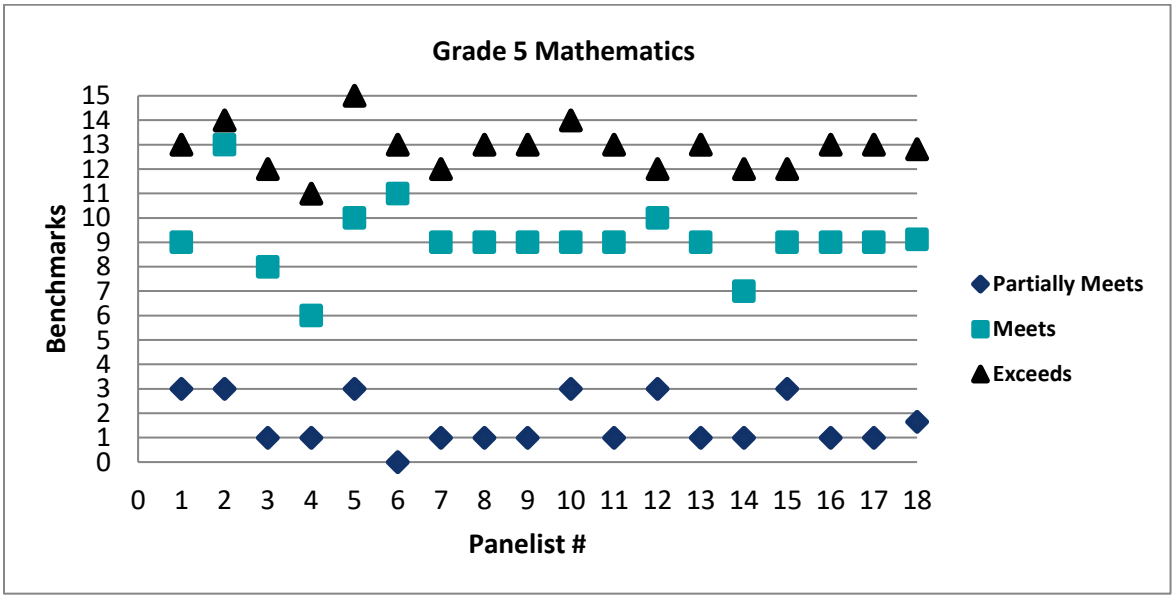


Figure 8: Location statistics for grade 5 mathematics (17 panelists)



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

- Standard error of measurement (SEM) was calculated at the benchmark level. Values of less than 1.00 (on an assessment of average length, or 30 items) indicate substantial agreement between the panelists in their estimated benchmarks.
- Inter-rater consistency (IRC) was 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) was also calculated at the item level. Values of 0.70 or greater indicate substantial consistency within the panelists in their item ratings.

All of the SEM of the benchmarks from Round 1 were below the provisional threshold of 1.00. Most of the SEM improved from Round 1 to Round 2 (Table 9).

Table 9: SEM by grade, subject, and round

SEM	Grade 3				Grade 5			
	English Language		Mathematics		English Language		Mathematics	
	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2
Partially Meets	0.62	0.48	0.61	0.63	0.51	0.15	0.68	0.26
Meets	0.75	0.46	0.37	0.39	0.69	0.17	0.63	0.37
Exceeds	0.56	0.33	0.14	0.21	0.31	0.08	0.42	0.24

All of the IRC and IRR statistics were above the threshold of 0.70 in both Rounds 1 and 2. Both statistics stayed about the same from Round 1 to Round 2, with slight increases in Round 2 (Table 10).

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

Statistic	Grade 3				Grade 5			
	English Language		Mathematics		English Language		Mathematics	
	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2
IRC	0.74	0.78	0.81	0.78	0.75	0.91	0.73	0.85
IRR	0.83	0.84	0.86	0.89	0.87	0.86	0.84	0.87

Evaluation

The evaluation results from the panelists were high, with averages between 3.2 and 3.8 out of a maximum of 4.0 (minimum 1.0). They were similar by grade and subject, with the overall evaluation averages of between 3.4 and 3.7 (Table 11).

Table 11: Workshop evaluation results by category

Category	Average Ratings (maximum of 4.0)			
	Grade 3		Grade 5	
	English language	Mathematics	English language	Mathematics
Framework Training	3.5	3.3	3.4	3.9
Assessment Training	3.4	3.2	3.2	3.6
Policy Linking Training	3.5	3.3	3.3	3.8
Round 1 Item Ratings	3.2	3.3	3.3	3.5
Round 2 Item Ratings	3.3	3.2	3.3	3.4
Overall Evaluation	3.5	3.4	3.5	3.7

Several 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, met pre-determined draft thresholds in both Rounds 1 and 2. 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 somewhat improved from the first pilot workshops but still need revisions. As with the first workshops, a constraint was the amount of preparation time, 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 took more time with discussions on the feedback between Rounds 1 and 2. As mentioned above, some of this time was spent on understanding the calculations of the benchmarks, which then helped with making revisions during the second round. However, again, some of this additional time was possible due to the relatively low number of items on the assessments (which NCERT said would likely be augmented in the next assessment cycle).

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, policy linking would benefit from initial analyses of assessment quality, in addition to alignment with the Framework. Rapid examination of the psychometric properties of the assessments prior to the policy linking workshops – i.e., number of items (mostly for validity), internal consistency reliability, and quality of the data set (missing data) – would be helpful in determining the suitability of the assessments for policy linking.

Second, differentiating between content and performance standards in the Framework contributed to would greater understanding of these standards by the panelists. However, it would help to include 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.

Third, the three-point scale for the item-subconstruct ratings worked better than the 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 – would be more appropriate.

Fourth, there was an ongoing concern that 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, it was useful to change from having the panelists take the assessments themselves to devoting more time towards increasing their understanding the assessment items and the Framework. This meant 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 was recorded on both the test booklet and Framework. All of the panelists wrote the domain, construct, subconstruct, Level, and Descriptor for each item in their booklets, and then recorded the item number in their Framework. This provided a better reference for Task 3 – by having the necessary information recorded on the two source documents – when they do the item ratings. However, it would be helpful to do this activity on assessments with more items to judge whether it could be completed within the time allocated.

Third, having large group discussions after the small group item matching process increased the panelists' understanding of the Levels. It was 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 built more consensus prior to proceeding with Task 3. Again, it was possible within the timeframe of the workshop since the panelists did not take the assessments themselves.

Fourth, placing more emphasis on item construction as a factor in item difficulty was useful. An example presented having a more difficult stem or distractors to increase the difficulty for the learners and requiring a higher Level within the same subconstruct and knowledge or skill. Perhaps it would be beneficial to spend more time on issues around item construction, including presenting examples from both reading and mathematics 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 stronger than in the previous workshops, thus causing fewer inconsistencies in the item ratings. This was reflected in the standard error statistics after Rounds 1 and 2. It was helpful 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 Descriptors and Levels appeared to positively influence the reliability of the benchmarks between the panelists.

Second, more training time spent on calculating benchmarks was helpful in understanding the benchmarking numbers. The panelists were instructed on totaling their JP, JM, and JE columns and then calculating each of the three benchmarks. This also helped with making revisions during the second round. The panelists had a better sense of what happens to the benchmarks if they change the ratings of items.

Third, the workshops added more information on starting points towards establishing thresholds, specifically for the alignment and the consistency. Based on these initial pilot workshops, 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, as with the first workshops, 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 was given to discussions on the feedback between Rounds 1 and 2. As mentioned above, some of this time was spent on understanding the calculations of the benchmarks, which then helped in making revisions during the second round. This process – including the discussions on matching the assessment items with the subconstructs and the Levels in the Framework – should be continued in subsequent workshops.

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.