United Nations
Educational, Scientific and
Cultural Organization

UNESCO INSTITUTE FOR STATISTICS

GLOBAL ALLIANCE TO MONITOR LEARNING

SUSTAINABLE DEVELOPMENT GOALS

POLICY LINKING METHOD

LINKING ASSESSMENTS TO GLOBAL STANDARDS

GAML6/REF/1

POLICY LINKING METHOD

LINKING ASSESSMENTS TO GLOBAL STANDARDS

FEBRUARY 2019

This publication was produced for review by the United States Agency for International Development. It was prepared for the Reading and Access Evaluation Project by Management Systems International, A Tetra Tech Company.

POLICY LINKING METHOD

LINKING ASSESSMENTS TO GLOBAL STANDARDS

Contracted under

AID-OAA-M-13-00010

Reading and Access Evaluation

DISCLAIMER

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

CONTENTS

ACRONYMS

COR	Contracting Officer's Representative
E3/ED	Office of Education in the Bureau for Economic Growth, Education and Environment
EGMA	Early Grade Math Assessment
EGRA	Early Grade Reading Assessment
GRN	Global Reading Network
IBE-UNESCO	International Bureau of Education-UNESCO
IRT	Item Response Theory
LLECE	Latin American Laboratory for Assessment of the Quality of Education
MSI	Management Systems International
NAEP	National Assessment of Educational Progress
NAF	National Assessment Framework
NCF	National Curriculum Framework
PASEQ	Program of Analysis of Education Systems of the CONFEMEN
PILNA	Pacific Islands Literacy and Numeracy Assessment
PIRLS	Progress in International Reading Literacy Study
PLD	Performance Level Descriptor
READ	Reinforcing Education Accountability in Development
SACMEQ	Southern and Eastern Africa Consortium for Monitoring Educational Quality
SDG	Sustainable Development Goal
SEA-PLM	Southeast Asia Primary Learning Metrics
SEM	Standard Error of Measurement
SME	Subject-Matter Expert
TIMSS	Trends in International Mathematics and Science Study
UIS	UNESCO Institute of Statistics
UNESCO	United Nations Educational, Scientific and Cultural Organization
USAID	United States Agency for International Development
USG	United States Government

ACKNOWLEDGMENTS

This paper follows the summary report from a three-day measurement workshop on policy linking in August 2018 that was funded and supported by the Office of Education in the Bureau for Economic Growth, Education and Environment (E3/ED) of the United States Agency for International Development (USAID). The Bureau has been extremely supportive of introducing and exploring policy linking as a method for comparing and aggregating results from student assessments within and across countries. The team would like to thank Benjamin Sylla for his leadership as the USAID Contracting Officer's Representative (COR) of the Reading and Access Evaluation Project. The team also extends a special thanks to Melissa Chiappetta and Rebecca Rhodes of the USAID Office of Education for their support and guidance throughout the entire process.

Abdullah Ferdous, Dana Kelly and Jeff Davis of Management Systems International (MSI) authored this paper, with support from Jonathan Weinstock (MSI).

EXECUTIVE SUMMARY

A challenge faced by USAID and other donors of international education projects is identifying where the greatest needs are and how much progress is being made comparatively across countries in improving children's reading and math abilities. The reason is that countries largely use different reading and math assessments. In addition, even when the same assessments are used, the language of the assessment may still differ. This makes fair and accurate comparisons difficult.

The way to address the challenge is by linking the assessments. There are two main types of linking: statistical and non-statistical. Statistical linking is more accurate, but it has greater requirements, i.e., either common students taking different assessments or having common items across assessments (Klein, Hamilton, et al., 1999).[1] Non-statistical linking is less accurate, but it is acceptable when requirements of statistical linking are not met due to issues such as design, logistics, or cost (Buckendahl & Foley, 2015).

Given that the requirements of statistical linking are usually not met in the international context, the U.S. Agency for International Development (USAID) and the Global Reading Network (GRN) sponsored a three-day workshop in August 2018 entitled *Linking Assessments to a Global Standard with Social Moderation*. Workshop participants represented a diverse group of 80 international education stakeholders from education ministries, donor agencies, implementing and evaluation partners, foundations, regional and international assessment programs, universities, and research centers. The workshop focused on developing a mutual understanding of social moderation – also called *policy linking* – as a practical and appropriate non-statistical method for linking student assessments within and across countries. It provided momentum towards the next steps of developing a policy linking toolkit and piloting the method.

This paper bridges the workshop and next steps by presenting a justification for policy linking. It describes the need for linking assessments to facilitate valid comparisons of student learning within and across countries for improved programming and tracking. It includes an overview of different methods used to link assessments and a rationale for choosing policy linking. Procedures and activities in implementing and operationalizing policy linking can be fully explained in the toolkit.

In brief, policy linking is a method in which the link from one assessment to another is the same set of descriptions of levels of student performance. Most often, there are two to four levels. The levels and their descriptions are progressive (or developmental), i.e., with increasing knowledge and skills required for higher levels. The names or labels for the levels reflect this increase, e.g., from "does not meet minimum proficiency" to "meets minimum proficiency." Linking two assessments using the same set of descriptions almost always results in different passing scores for different assessments, e.g., a passing score of 55 on one (easier) assessment that is equal to a passing score of 40 on another (more difficult) assessment. The key is that the difficulty of different assessments is judged in relation to the same set of descriptions of performance levels.

To that end, the first requirement of policy linking is establishing *performance level descriptors* (PLDs) for

[1] The term "common students" refers to the same students taking multiple assessments and "common items" refers to the same items included in multiple assessments. These common students and/or items provide the psychometric links across the assessments for statistical linking to take place. More information is provided in references such as Dorans, Moses, & Eignor (2010) and Kolen & Brennan (2014).

each targeted grade and subject. Each set of PLDs forms a *proficiency scale* that reflects the levels of performance by grade and subject. Since the PLDs and proficiency scales in the international context cover a variety of assessments that can be linked, they should reflect global content frameworks. At present, global PLDs and global proficiency scales do not exist. The UNESCO Institute for Statistics (UIS) and the UNESCO International Bureau of Education (IBE-UNESCO) have made progress in compiling global content frameworks, which offer a solid foundation for developing and finalizing the global PLDs and proficiency scales. Note that establishing the PLDs and proficiency scales needs to take place only once for policy linking, i.e., for all assessments that are linked. (See more information on global PLDs and proficiency scales in the Policy Linking Procedures section starting on page 15.)

A second requirement in policy linking is an internationally-accepted approach for determining the minimum scores for students to demonstrate performance at each level. This process of benchmarking – also called *standard setting* – establishes the minimum scores, or cut scores, on an assessment that correspond to minimum performance in relation to each PLD. Note that standard setting needs to take place multiple times for policy linking, i.e., for each assessment that is linked. (See more information on standard setting in the Policy Linking Procedures section starting on page 17.)

Once a draft toolkit – with global PLDs, proficiency scales, and standard setting processes – has been developed, policy linking should be piloted in selected countries. If the method is deemed practical and appropriate by stakeholders, the toolkit can be finalized so that it provides full guidance on policy linking, with materials to support its application. Education ministries and partners should then be able to link EGRAs, EGMAs and curriculum-based assessments to the global proficiency scales. This should be highly useful to countries that wish to establish internationally valid passing, or minimum proficiency, scores on their assessments. Implementation of policy linking by stakeholders should contribute to publicly-available knowledge bases for monitoring and reporting on global student performance using existing assessments or, in the future, new assessments.

In summary, policy linking is a promising low-cost, relatively rapid, and sufficiently rigorous non-statistical method to compare and aggregate results from different assessments. It can provide a basis for drawing lessons learned across countries to guide future programming and improve learning outcomes. In addition, different agencies that support international basic education should be able to use the policy linking results for global reporting on students who meet minimum proficiency standards by grade and subject.

1. INTRODUCTION

The purpose of this paper is to provide technical support under the 2017 Reinforcing Education Accountability in Development (READ) Act. The act requires developing and using internationally comparable indicators, norms, and methodologies in reporting measurable improvements in literacy, numeracy, and basic skills to the extent practical and appropriate.

Specifically, the aim is to explain and justify social moderation – also called *policy linking* – as a non-statistical method for linking different student assessments. Processes are described for linking assessments to global standards of minimum proficiency by grade level and subject area. Successful implementation of policy linking should allow the USG and other stakeholders to benefit from two kinds of analyses on a global

basis: comparisons and aggregation of assessment results.[2] Such analyses should be possible for assessments by grade and subject both within and across countries.[3]

Following this introductory section, the policy linking paper has four sections. The second section provides an explanation of linking methods, including the advantages and disadvantages of statistical and non-statistical linking. The third section lists steps for implementing policy linking, including writing a policy definition, developing performance level descriptors (PLDs), and setting performance standards. The fourth section covers operationalizing policy linking, with the activities of developing the toolkit, piloting the method, revising the toolkit, and applying the method to different assessments in countries. The fifth section draws conclusions from the paper. A reference section at the end cites articles, books, and presentations for further reading. There is also an annex with information from a recent workshop sponsored by the U.S. Agency for International Development (USAID) and the Global Reading Network (GRN) on policy linking, including draft performance levels/labels and a glossary.

2. OVERVIEW OF LINKING METHODS

Linking is a psychometric approach that makes results, or scores, from one assessment comparable to another. The two main categories of linking are statistical and non-statistical. This section provides an overview of statistical and non-statistical linking, including the different sub-types of statistical linking, and it shows why non-statistical linking is recommended as the preferred option for USAID in making results from different assessments comparable.

Linking methods each have their use, depending on the situation. Technically, statistical linking is preferred since it provides tables for converting each score from one assessment to another. Non-statistical linking usually only provides tables that match up the benchmarks, or *cut score*s, i.e., the corresponding scores on assessments that separate students by performance level. In addition, statistical linking is more accurate since it is based on algorithms, while policy linking is based on judgments. The requirements for statistical linking, however, are much more stringent. Policy linking is useful, and acceptable, when those requirements are not met.

Prior to linking assessments, the proper linking method must be selected according to a set of activities. The first activity is examining the similarity of the assessments in relation to four features. The utility and reasonableness of linking – and the choice of linking method – depends on the extent to which assessments have these features in common (Kolen & Brennan, 2004; Mislevy, 1992).

1. Constructs – Do the assessments measure the similar constructs, such as early grade reading or math?

2. Populations – Are the assessments designed for use with similar populations, such as grade 2 students?

3. Measurement – Do the assessments share common measurement characteristics, such as test format,

[2] Both could be based on the percentages of students in performance levels, such as minimum proficiency.

[3] Assessments in different languages are often used within countries in the same grade level and subject area, such as EGRAs in English and Kiswahili in Kenya (with the same students) or Sindhi and Urdu in Pakistan (with different students). Policy linking provides a method for comparing and aggregating the results from EGRAs both within (for different languages) and across (for Kenya and Pakistan) such countries.

length, and administration procedures?

4. Inferences – Are the assessment scores used to draw similar types of inferences, or make similar decisions, such as meeting minimum proficiency?

After establishing the degree of similarity of the assessments – same, similar, or different – according to these features, the next activity is to select one of five methods to link the assessments: 1) equating, 2) calibration, 3) moderation, 4) projection, or 5) policy linking. The first four methods are categorized as statistical linking since their features lend themselves to statistical processes – requiring either common students or common items – for determining the links between assessments. The last method is categorized as non-statistical linking since it involves judgments to determine the links (Buckendahl & Foley, 2015; Reckase, 2000). The order of the methods reflects their rigor and accuracy. Table 1 below provides information on the linking methods by the degree of similarity of the assessment features.

Table 1: Linking Methods and Assessment Features

Linking Methods	Assessment Features			
	Constructs	**Populations**	**Measurement**	**Inferences**
Equating (Statistical)	Same	Similar	Same	Same
Calibration (Statistical)	Same	Similar	Similar	Same
Moderation (Statistical)	Similar	Same	Similar	Similar
Projection (Statistical)	Similar	Same	Different	Similar
Policy linking (Non-Statistical)	Similar	Similar	Different	Similar

Brief examples of assessments with different features and the corresponding linking method are provided below. The references provided in this paper have detailed information on each of the methods, with policy linking as the only method explained in the subsequent sections.

1. Equating (statistical) – Different students take different versions of a high school graduation test. The versions are constructed based on the same content but with different items – except for a representative subset of common items – due to test security concerns. An equating process, such as fixed anchor calibration, provides a one-to-one correspondence between each score point on the different versions.[4]

2. Calibration (statistical) – All students in a district take a short version of a state assessment with a representative subset of items. The short version is designed to measure performance in a subject area but in less time. The scores can be linked through a calibration process that provides weights to the items on the short version for calculating scores that match to those from the state assessment.

3. Moderation (statistical) – A representative sample of students take two different university entrance examinations. These exams have somewhat similar content but no common items. The scores can be

[4] An example of common person equating would involve a representative sub-sample of students taking multiple assessments. Percentile ranks are calculated for each assessment. Equipercentile equating can then be used to match the scores on the assessments based on the scores associated with same percentile ranks on each assessment.

linked through statistical moderation, i.e., by aligning the score distributions of those same students on the two exams. This alignment process provides comparable scores from the exams.

4. Projection (statistical) – The same students take a multiple-choice test and an essay test in a subject area. Prior to taking the essay test, estimated scores can be projected (or predicted) from the scores on the multiple-choice test and other variables in the population of interest. The projection process takes place using multiple regression, i.e., with an equation comprised of predictor variables and coefficients.

5. Policy linking (non-statistical) – Different students at a similar grade level from multiple countries take their own national assessments in the same subject area. The minimum passing scores on the assessments can be linked to each other by mapping them to a common proficiency scale. The mapping process can be implemented through performance standard setting involving judgments by subject-matter experts (SMEs).

The two general categories of statistical and non-statistical linking methods have advantages and disadvantages, which are summarized in Table 2 below.

Table 2: Advantages and Disadvantages of Statistical and Non-Statistical Linking

Advantages	Disadvantages
Statistical Linking (Equating, Calibration, Moderation, Projection)	
• Provides stronger and more precise linking relationships between assessment results. • Allows the interchangeable use of scores from different assessments. • Provides error estimates to evaluate the strength of linking relationships. • Allows for technical reporting with statistical evidence of linking accuracy.	• Can require careful selection of representative sub-samples of common students or items. • Can require complex test construction by a team of psychometricians and SMEs. • Can require strict security measures so that common items are not released. • Can require large sample sizes for advanced psychometric analyses.
Non-Statistical Linking (Policy Linking)	
• Uses existing assessments since there are not strict requirements for assessment design. • Has less stringent requirements for commonalities in the assessment features. • Avoids requiring common persons or items to put assessment results onto the same scale. • Involves simpler analyses and does not require large sample sizes.	• Does not provide as strong of a linking relationship between assessment results. • Does not give statistical error estimates (standard setting errors are estimated). • Does not allow interchangeable use of scores since linking only takes place at cut scores. • Does require added collection of validity evidence to judge the procedure's accuracy.

Based on the assessment features associated with each linking method, policy linking is the most practical and appropriate option for linking different assessments to global definitions of proficiency with existing assessments from different countries when common students do not take multiple assessments across countries and common items are not shared across assessments. In these cases, without special studies, statistical linking is generally not an option.[5]

[5] After the fact, linking different assessments using a statistical method can be done if a special study is conducted that involves re-administering the assessments to common students. These types of studies are generally not feasible since they involve added time, cost, logistics, and expertise, and they can involve government or other approvals.

3. POLICY LINKING PROCEDURE

After examining the options and selecting policy linking as a way forward, the next step in policy linking is to develop global proficiency scales – by grade and subject – and link the assessments to the scales. In this way, the assessments are also linked to each other, at the cut scores, through policy linking.

The procedure has six steps, carried out in two phases. In Phase I, descriptions of performance levels are defined, and proficiency scales are established. This phase is implemented *only once globally* by grade and subject. In Phase II, an appropriate standard setting method is applied to establish the cut score(s) on an assessment that corresponds to the levels of performance. This phase is implemented *for each assessment*.

The phases and steps are as follows.

Phase I (completed once globally):

1. Define the content frameworks, i.e., the grade-appropriate knowledge and skills by subject area that students are expected to learn before they go to the next grade.

2. Determine the performance levels and labels, i.e., the number of performance levels and the labels for each of those levels.

3. Write the policy definition, i.e., the general statements of what students in each performance level are expected to know and be able to do in relation to the content frameworks.

4. Develop the PLDs, the detail on what students in different performance levels are expected to demonstrate in each grade and subject in relation to the content frameworks.

Phase II (completed for each country or assessment):

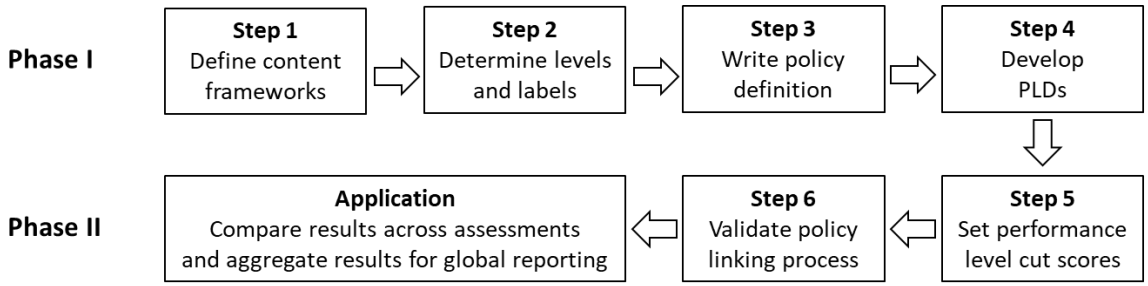
5. Set the performance level cut scores on each assessment by engaging local educators as judges to apply the PLDs to their assessments, i.e., standard setting.

6. Validate the policy linking process through procedural evidence, e.g., criteria for selection of judges and analysis of inter-judge consistency.

The two phases and six steps for developing the global proficiency scales and setting the cut scores on assessments are shown in Figure 1 and elaborated below. Four scales (in Phase I) need to be developed i.e., grade 2 and end of primary in reading and math. Each assessment – by grade and subject – goes through the process of setting cut scores and validating the process. It should then be possible for the results from policy linking, i.e., the percentages of student scores at each performance level, to be used to compare the outcomes from different assessments and aggregate these outcomes for global reporting.[6]

[6] Note that the cut scores from individual assessments are applied to their score distributions to calculate the percentages of scores in each performance level. The percentages from different assessments can be compared and aggregated (after weighting) for lessons learned and global reporting. In addition, the percentages can be applied to the number of beneficiaries to calculate a "count" of students reaching minimum proficiency.

Figure 1: Steps in Policy Linking



Step 1: Define the Content Frameworks (Phase I)

The first step of Phase I involves defining the content frameworks. Policy linking requires universal frameworks that specify what students should know and be able to do in the targeted grades and subjects. Here are examples from grade 2 reading and math in the U.S. (National Governors Association, 2010):

Grade 2 Reading:
2.3: Know and apply grade-level phonics and word analysis skills in decoding words.
2.4: Read with sufficient accuracy and fluency to support comprehension.

Grade 2 Math:
2.2: Know from memory all sums of two one-digit numbers.
2.3: Read and write numbers to 1,000 using base-ten numerals and number names.

IBE-UNESCO developed global content frameworks as a basis for reporting on the Sustainable Development Goals (SDGs) 4.1.1.a and 4.1.1.b of minimum proficiency in reading and math at grades 2/3 and the end of primary, respectively. They reviewed 73 national assessment frameworks (NAFs) and national curriculum frameworks (NCFs) from 25 countries for reading, along with 115 NAFs and NCFs from 53 countries for math (IBE-UNESCO and UIS, 2017 and 2018). The assessment frameworks included EGRAs and EGMAs. They summarized the reading and math knowledge and skills in the assessment and curriculum frameworks at those grades.

Since the targeted grades for SDG reporting correspond with those of the USAID F-indicators and self-reliance metrics, IBE-UNESCO's mapping can be used as a starting point for defining the global reading and math content standards for policy linking at those grades. SMEs can review the IBE-UNESCO content frameworks and then gather at a workshop to determine the final frameworks for policy linking.

Step 2: Determine the Performance Levels and Labels (Phase I)

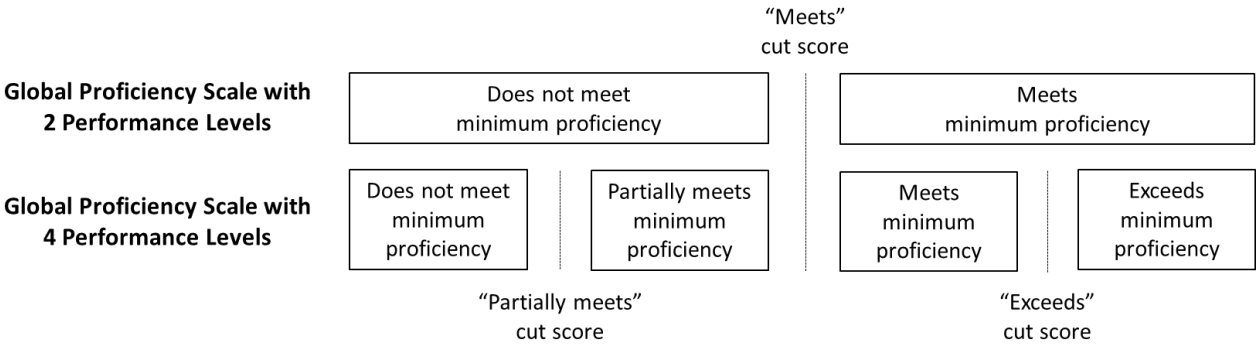
Next, education policy makers and practitioners need to determine the number of performance levels and their labels for use on the global proficiency scales.[7] Typically, at least two performance levels but no more than four levels are used (Perie, 2008). After determining the number of levels, the next task is to name

[7] Again, note that the global proficiency scales are qualitative, i.e., defined through descriptions of performance at each level. The cut scores are quantitative and separate the score distributions for assessments into those same levels. The scores in those performance levels are named by the labels and described by the descriptors.

the levels, i.e., to label them. There are no clear-cut guidelines on this. However, experts recommend choosing labels that relate to the purpose of making supportable inferences from the classifications (Cizek & Bunch, 2007).

An example of two levels, allowing for USAID (and UIS) reporting, would be "does not meet minimum proficiency" and "meets minimum proficiency." An example of three levels, providing more description of progression towards minimum proficiency, would add an intermediary level of "partially meets minimum proficiency." An example of four levels would add a higher level for "exceeds minimum proficiency" that surpasses the meets proficiency level. Using these levels and labels, proficiency scales with two levels (one cut score) and four levels (three cut scores) are shown in Figure 2. Note that the names of the cut scores refer to the lowest point in the higher level.

Figure 2: Performance Levels and Cut Scores



Here are other examples of four levels and labels, based on existing assessment programs. These could also be considered for the global proficiency scales:

- below basic / basic / *proficient* / advanced;
- does not meet the standard / approaches the standard / *meets the standard* / exceeds the standard; and
- level 1 / level 2 / *level 3* / level 4.

The levels and labels in italics are considered the "desired" level of student performance that policymakers expect all students to achieve.

Step 3: Write the Policy Definition (Phase 1)

Using the levels/labels, the policy makers and practitioners need to write a generic definition, i.e., broad statements of student performance at each level. The definition is not linked to specific content. It provides overall information on performance at each level. It is particularly useful when reporting on the results from multiple assessments, i.e., for different grades and subjects, since all scores from all grades and subjects can be broadly interpreted using the same generic policy definition. The definition outlines the progression of knowledge and skills with the same level of rigor at each performance level. It is written for all performance levels except for the lowest level.

Examples of policy definitions from U.S. national and statewide assessment programs are provided in Table

3. Each of the three programs has four performance levels, though the definitions are only written for the upper three levels.[8] Note that the "proficient" or "meets the standard" level usually has more description, since it is the minimum, or desired, level of performance for all students. The labels correspond to two of the examples in Step 2 above.

Table 3: Illustrative Policy Definitions for Performance Levels

Assessment	Performance Levels		
National Assessment of Educational Progress (NAEP)	**Basic:** Partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.	**Proficient:** Solid academic performance at each grade. Students have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations and analytical skills appropriate to the subject matter.	**Advanced:** Superior performance beyond proficient.
Pennsylvania Statewide Testing Program	**Basic:** Marginal academic performance, work approaching, but not yet reaching, satisfactory performance, indicating partial understanding and limited display of the skills included in the academic standards.	**Proficient:** Satisfactory academic performance indicating a solid understanding and adequate display of the skills included in the academic standards.	**Advanced:** Superior academic performance indicating an in-depth understanding and exemplary display of the skills included in the academic standards.
Arizona Statewide Testing Program	**Approaches the Standard:** This level denotes understanding of the knowledge and application of the skills that are fundamental for proficiency in the standards.	**Meets the Standard:** This level denotes demonstration of solid academic performance on challenging subject matter reflected by the standards. This includes knowledge of subject matter, application of such knowledge to real-world situations and content-relevant analytical skills.	**Exceeds the Standard:** This level denotes demonstration of superior academic performance evidenced by achievement substantially beyond the expected goal of all students.

In writing a policy definition, experts strongly recommend that the words can clearly and concisely distinguish between the levels. The definitions should state the degree of knowledge and skills expected of students at each level (Perie, 2008). The policy definition is the backbone of the full descriptions by grade level and subject area in Step 3 below.

Experts recommend a one-day workshop with the policy makers and practitioners to determine the levels/labels and write the policy definition. During the August 2018 USAID and GRN-sponsored workshop, there was an activity in which the participants suggested draft performance levels/labels and a

[8] For instance, the policy definition for the lowest levels of Below Basic or Does Not Meet the Standard could be included but they would be stated as "does not reach basic" or "does not approach the standard." On EGRAs, the lowest level is often reserved for non-readers, i.e., those students who are unable to read a word of written text.

draft policy definition. This provides a starting point for Steps 2 and 3 when operationalizing of policy linking. The results from the workshop activity are provided in an annex.

Step 4: Develop the Performance Level Descriptors (Phase 1)

In the final step of Phase I, practitioners need to develop full descriptions of the performance levels by grade and subject. These descriptions identify the knowledge and skills required by students to achieve each performance level. They provide stakeholders – governments, donors, partners, schools, and communities – with essential information on what students at each performance level are expected to know and be able to do, as well as what knowledge and skills are necessary to reach the next performance level.

To develop full descriptions, a PLD writing workshop should occur with a group of SMEs who have experience teaching the subject and working with a variety of student populations. In addition, SMEs should have demonstrated knowledge and practice related to the content frameworks. Typically, five to ten SMEs per subject and grade participate in a three- to four-day workshop (Perie, 2008). The SMEs start with the policy definition and expand it to cover in the specific knowledge and skills at each performance level by grade and subject. The PLDs should reflect the content frameworks defined in Step 1. The SMEs should carefully consider the development or progression of students within a grade. Moreover, it is expected that the SMEs reach consensus on the descriptions.

Table 4 shows Grade 6 English PLDs adapted from a U.S. statewide assessment program. They illustrate the level of detail required and desired progression across levels. The PLD writing workshop, as with the illustrative policy definitions, should reference multiple examples of PLDs by grade and subject.

Table 4: Illustrative Performance Level Descriptors

Performance Level Descriptors (PLDs)
Partially meets minimum proficiency. A student performing at this level demonstrates limited comprehension of grade-level literary and informational texts and may use textual evidence to summarize and/or analyze a text. The student inconsistently analyzes how an element of literature or informational text develops and influences the text. The student may determine a central idea in an informational text. The student may determine how the author uses organization, structure, form, text features, figurative language and/or word choice to achieve a purpose. The student determines the point of view in a text. The student provides an incomplete comparison between texts in different forms or genres. The student may identify the development of an argument and may evaluate the author's claims and evidence in a text. The student may use context and word structure to determine the meanings of words, may interpret figurative language and may understand some word meanings. In writing, the student inconsistently uses reasoning and evidence to develop an argumentative/informational essay on a topic for an intended audience. The student organizes a narrative using limited narrative techniques. The student writes a text-dependent analysis essay that responds to a text or texts and demonstrates a weak analysis that may include inadequate evidence to support its intended purpose. The student may use transitions. The student recognizes and demonstrates a partial command of the conventions of standard English grammar, usage and mechanics.
Meets minimum proficiency. A student performing at this level demonstrates comprehension of grade-level literary and informational texts by using textual evidence to summarize and/or analyze a text. The student analyzes how an element of literature or informational text develops and influences the text. The student determines a central idea in an informational text. The student determines how the author uses organization, structure, form, text features, figurative language and/or word choice to achieve a purpose. The student

Performance Level Descriptors (PLDs)
determines the effectiveness of point of view in a text. The student compares and contrasts texts in different forms or genres. The student traces the development of an argument and evaluates the author's claims and evidence in a text. The student uses context and word structure to determine the meanings of words, interprets figurative language and understands nuances in word meanings. In writing, the student uses logical reasoning and relevant evidence to develop an organized argumentative/informational essay on a topic in a formal style for an intended audience. The student organizes a narrative with a controlling point, using precise words, phrases and narrative techniques. The student writes a text-dependent analysis essay that responds to a text or texts and demonstrates an organized analysis that cites textual evidence to support its intended purpose. The student uses a variety of appropriate transitional words, phrases and clauses. The student recognizes and demonstrates a command of the conventions of standard English grammar, usage and mechanics to convey ideas precisely and for effect.
Exceeds minimum proficiency. A student performing at this level demonstrates thorough comprehension of grade-level literary and informational texts by using key textual evidence to effectively summarize and/or analyze a text. The student thoroughly analyzes how an element of literature or informational text develops and influences the text. The student determines a central idea in an informational text. The student determines how the author uses organization, structure, form, text features, figurative language and/or word choice to achieve a purpose. The student determines the effectiveness of point of view in a text. The student thoroughly contrasts texts in different forms or genres. The student traces the development of an argument and thoroughly evaluates the author's claims and evidence in a text. The student uses context and word structure to determine the meanings of words, interprets figurative language and understands nuances in word meanings. In writing, the student uses logical reasoning and substantive evidence to develop a cohesive argumentative/informational essay on a topic in a formal style for an intended audience. The student thoroughly organizes a narrative with a controlling point, using precise words, phrases and narrative techniques. The student writes a text-dependent analysis essay that responds to a text or texts and demonstrates an organized and thorough analysis that cites substantial and relevant evidence to support its intended purpose. The student uses a variety of appropriate transitional words, phrases and clauses. The student recognizes and demonstrates a thorough command of the conventions of Standard English grammar, usage and mechanics to convey ideas precisely and for effect.

The PLDs should comprise the global proficiency scales by grade and subject as shown in Figure 2 above. They should comprise the content of the qualitative scales by defining the progression of knowledge and skills that students are expected to demonstrate from the lower to the upper performance levels.[9] The next step shows how the global proficiency scales form the foundation for setting the quantitative cut scores for each assessment, and how applying the assessments to the scales allows for comparing and aggregating results both within and across countries.

Step 5: Set the Performance Level Cut Scores (Phase II)

In this first step of Phase II, stakeholders in the countries set the performance level cut scores on their assessments based on the PLDs, i.e., the global proficiency scales, by grade and subject. This establishes a link between each assessment and the relevant global proficiency scale by setting cut scores for each performance level on the assessments. Again, setting the cut scores takes place for each assessment.

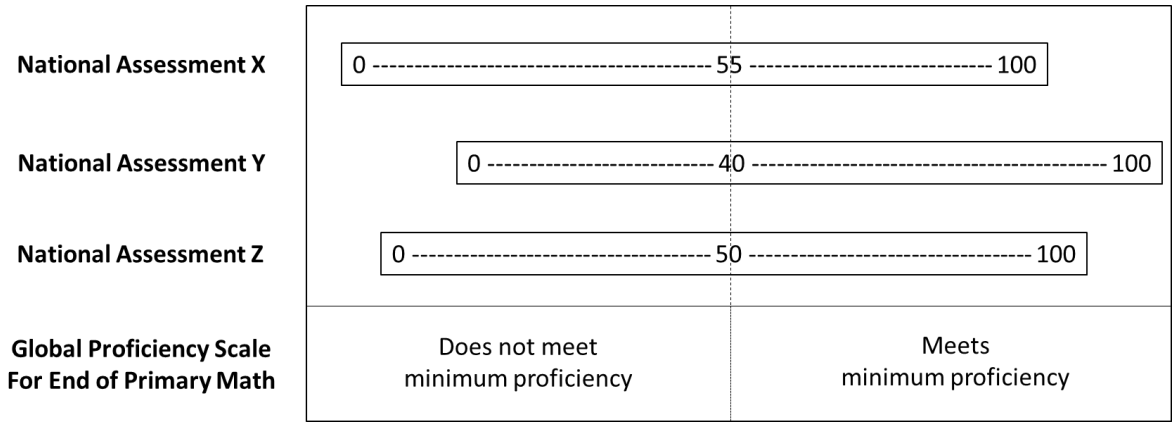
As an example, Figure 3 shows the hypothetical results from setting cut scores through policy linking for three math assessments: X, Y, and Z. These could be national assessments in three different countries for

[9] Note that the Phase I steps only need to be implemented once, i.e., to establish the global proficiency scales. The following Phase II steps need to be implemented for each assessment.

math at the end of primary school. For simplicity, it can be assumed that the three assessments are placed on a scale with only two performance levels: does not meet minimum proficiency and meets minimum proficiency. In Step 4, assume for purposes of this example that there is an end-of-primary math PLD for meets minimum proficiency, which gives us the qualitative global proficiency scale for the grade and subject. It is further assumed that the scores for all three assessments are converted to a 0-100 scale. Independent, country-based panels of SMEs for the three assessments make judgments on the cut scores that correspond to the global proficiency scale, i.e., what score a student needs to attain to meet the definition of minimum proficiency.

In this example, Figure 3 shows that Assessment X is judged as less difficult. The cut score is 55, which means that a student needs a score of 55 and above to meet minimum proficiency. Assessment Y is judged as more difficult, so a score of 40 or above meets minimum proficiency. Assessment Z is judged as medium difficult, so a score of 50 or above meets minimum proficiency. The process for making the judgments for setting the performance level cut scores – or standard setting – is explained below.[10]

Figure 3: Example of Performance Level Cut Scores



Experts recommend using an internationally accepted test-centered (Angoff, 1971), examinee-centered (Kahl, Crockett, DePascale, & Rindfleisch, 1995), or item-mapping (Lewis, Green, Mitzel, & Patz, 1999) standard setting method to establish performance level cut scores. The choice of the standard setting method depends on the measurement characteristics of the assessments (Cizek & Bunch, 2007). Table 5 shows some of the most common standard setting methods, with the types of methods and the associated measurement characteristics. Methods such as Angoff are especially useful when it is important that the passing score represents the standard of a large and diverse group of test-takers (Livingston & Zieky, 1982). A modified version of the Angoff method has been used to set cut scores on EGRAs (Plake, Ferdous, & Buckendahl, 2005; Ferdous, Simon, & Davis, 2017).

[10] The higher or lower placement of a performance level cut score for an assessment does not necessarily mean that a country with a higher cut score, for instance, has a lower percentage of students who meet minimum proficiency than a country with a lower cut score. The percentages depend on the cut scores and the score distributions associated with those assessments. Of course, the higher or lower placement of a cut score for an assessment changes the performance level percentages for that country.

Table 5: Standard Setting Methods and Measurement Characteristics

Methods	Types of Methods	Measurement
Angoff and Modified Angoff	Test-centered, i.e., having the judges estimate the probabilities of students answering items correctly	Yes/no or multiple-choice items
Borderline and Contrasting Groups	Examinee-centered, i.e., having the judges observe the students when they take the assessment	Open-response items
Bookmark	Item-mapping, i.e., having the judges place the cut scores on a list of items ordered by difficulty	Multiple-choice or open-response items

For each assessment, experts recommend that 10–15 SMEs gather for a three- or four-day standard setting workshop. If the workshop involves multiple subject areas, e.g., reading and math, then twice as many SMEs are needed. The SMEs for each assessment are selected based on a consistent process for each assessment. For instance, with a group of 12 SMEs, it might be helpful to include three people from each of four sub-groups – assessment specialists, classroom teachers, teacher trainers, and curriculum specialists – to comprise the group of judges. In addition to these job titles, a certain level of education may be required along with a minimum number of years of experience and geographic/cultural/linguistic diversity. The policy-linking toolkit should include more details on selection of SMEs.

The workshop should begin with a plenary training of the SMEs on the PLDs, global proficiency scales, and standard setting method. This should be conducted by the lead facilitator, who is usually a psychometrician or other well-qualified assessment specialist. If there are multiple assessments, each group needs a leader. As described below, the workshop proceeds through two rounds of standard setting, with feedback and plenary sessions after each round. Detailed information on this process should be provided in the toolkit.

In Round 1, the SMEs provide independent judgments about each assessment to set their initial cut scores. For instance, with a test-centered method, the SMEs make judgments about each item, or subtask, on the test to set their initial cut scores based on their understanding of the PLDs and the student populations. The Modified Angoff method requires the SMEs to indicate whether they thought a student meeting minimum proficiency would answer the item correctly (Plake, Ferdous, & Buckendahl, 2005).

After collecting each SMEs' Round 1 judgments, an analyst compiles the results to estimate the group-recommended Round 1 cut scores. The SMEs are provided feedback on the cut scores, i.e., the individual SME cut score locations, or relative positions, of their judgments. The implications of the cut score placement on the percentages of scores in the performance levels are also provided.

In Round 2, the SMEs have an opportunity to reconsider their Round 1 judgments and to identify any errors or misconceptions about the process of setting the cut scores. The Round 2 judgments are considered more accurate than those from Round 1, as SMEs may have had inconsistent understanding of PLDs, minimum proficiency, or the judgmental procedure (Hambleton, Impara, et al., 2000; Ferdous & Plake, 2003; Ferdous & Buckendahl, 2013).

The analyst again compiles the Round 2 judgments to calculate cut scores. As in Round 1, the implications of the cut score placements are provided. In standard setting, it is not necessary to have consensus on the judgments, but rather the SMEs' cut scores are averaged to determine the recommendations. The combined Round 2 judgments by the SMEs become the recommended cut scores for the assessment.

Step 6: Validate the Policy Linking Process (Phase II)

In the final step of policy linking, experts validate the process in each country through evidence on the standard setting procedure and consistency of the SMEs' judgments. For procedural evidence, selecting and training the SMEs, understanding of PLDs, and collecting and processing ratings data should be well documented. For internal consistency evidence, the precision of estimated cut scores, inter-judge consistency and intra-judge consistency should be estimated (Chang, 1999; Ferdous & Plake, 2005).[11]

Once the cut scores for a country are set, it should be noted that subsequent assessments in the same countries should be linked to the assessment used for the policy linking. Otherwise, the classifications of students into the performance categories, such as the percentage of students reaching minimum proficiency, would depend on the level of difficulty of these later assessments. These linkages can be implemented in a country using an appropriate linking method, preferably equating.

4. OPERATIONALIZING POLICY LINKING

This section provides a list of activities that need to be conducted to implement the steps in policy linking. Operationalizing the process is divided into three groups of activities: toolkit development and revision, country-level implementation, and global-level implementation.

Toolkit development and revision:

The toolkit needs to be developed, piloted, finalized, and disseminated according to the six activities below. The development includes producing the policy definition and PLDs for inclusion in the toolkit. The dissemination can also include training for facilitators of policy linking workshops so that they can validly and reliably implement the process.

1. Organizing a workshop to produce the policy definition and the PLDs (that comprise the global proficiency scale) for each grade and subject.

2. Developing the toolkit so that it is ready for piloting, with guidance and materials, including forms and handouts;

3. Piloting the toolkit in two countries, with EGRA/EGMA at grades 2 and a national assessment for reading and math at the end of primary;

4. Revising the toolkit based on the findings from the piloting and conducting a review process with selected stakeholders;

5. Posting the toolkit on appropriate websites, since the same toolkit can be used for policy linking on all assessments; and

6. Conducting training sessions for facilitators of policy linking workshops from interested countries

[11] These internal consistency statistics should be evaluated against those from other policy linking workshop and, optimally, against standards. However, at this time, standards have not been developed.

(optional).

Country-level implementation:

Once the toolkit is developed, countries that choose to use policy linking can implement the guidelines by conducting the following two activities, i.e., the steps in Phase II (cut scores and validation) along with a country-level application activity:

1. Applying policy linking to existing assessments for setting performance level cut scores; and

2. Interpreting assessment results based on those cut scores for reporting on indicators.

Global-level implementation:

The assessment results based on the new performance level cut scores from within and across countries can then be applied by stakeholders in two global-level activities:

1. Comparing the assessment results within and across countries for drawing lessons learned; and

2. Aggregating the assessment results for reporting on global indicators.

The toolkit should include step-by-step guidance on implementing policy linking. It should also contain agendas, glossaries, activities, slides for presentations, videos (optional), judgment forms, data entry templates, data analysis tools, evaluation forms and reporting templates. Finally, it should include finalized policy definitions and PLDs for the global proficiency scales.

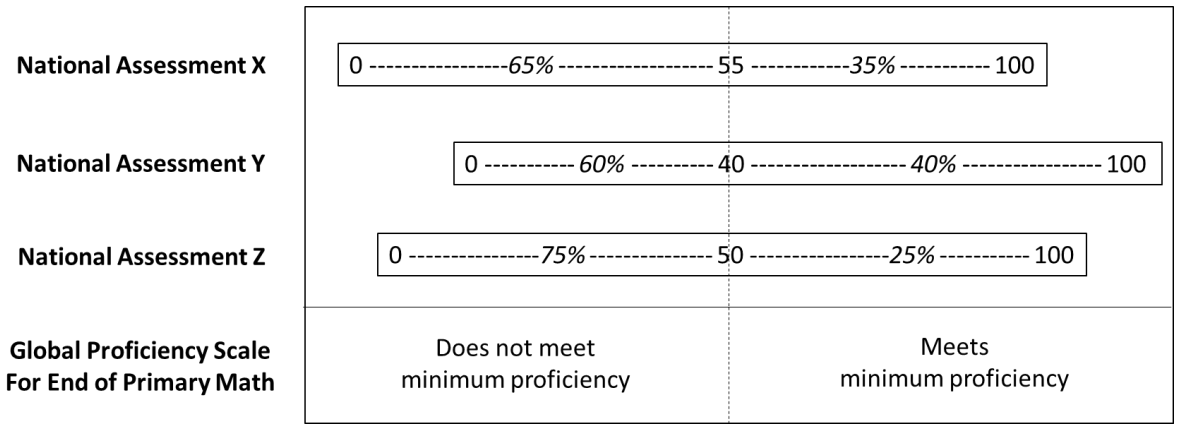
A recommendation is to pilot test the policy linking method in two countries. The countries can be chosen based on criteria that may include 1) recent administration of EGRA and EGMA at grade 2, 2) recent administration of national assessments in reading and math at the end of primary school, 3) availability of the assessment instruments, data files, and technical reports, and 4) willingness of the USAID Mission and host-country education ministry to cooperate on and participate in the piloting. We suggest that some of the same experts facilitate each of the pilots so that they can more readily apply information gathered from the first pilot to the second pilot as well as accelerate the process of revising the toolkit.

Another recommendation is to pilot-test results and lessons learned then be presented and discussed with the governments, the donor community, implementing partners, and other stakeholders. Based on the discussions, the toolkit should be revised and posted in the public domain. The toolkit should include this justification paper, along with practical, detailed guidance and materials for implementing workshop activities, interpreting workshop results, and using results for improved student learning. As mentioned above, training workshops can also be organized for education ministries and partners on implementing the policy linking approach. The training could include presenting the contents and procedures in the toolkit and providing simulation exercises. It may likely increase the validity and reliability of the cut scores and inferences from policy linking.

Once policy linking is implemented in a critical mass of countries, the assessment results can be compared since they should be linked to their global proficiency scales. The cut scores and the score distributions for each assessment can be used to determine the percentages of student scores by performance level.

For instance, in Figure 4, national assessment X with a cut score of 55 might result in 35 percent of students in the meets minimum proficiency level. Other possible percentages are shown for countries with assessments Y and Z. These percentages can also be weighted and aggregated for global reporting.

Figure 4: Example of Performance Level Percentages



Finally, the percentages can be converted into counts when the percentages are multiplied by the number of beneficiaries to calculate the number of students in each level. As with the percentages, these counts can be aggregated as well.

5. CONCLUSIONS

Several key conclusions that can be drawn from the paper are listed below.

1. Policy linking has potential for helping USAID address technical issues in comparing and aggregating student assessment results both within and across countries. The method can be applied to indicators of minimum grade-level proficiency in reading and math at the end of grade 2 and the end of primary school.

2. Results from policy linking can help fulfill USAID's aim of "strengthening its ability to draw cross-country data comparisons and share lessons learned and best practices." The method makes it possible to further capitalize on USAID investments to spread promising interventions to other countries.

3. Statistical linking methods are not technically or practically viable for use with current assessments since they require the same students to take multiple assessments or the same items across assessments and can thus be costly. Language issues are a related factor that inhibit the use of statistical linking.

4. Basing linking on policy definitions and PLDs through a non-statistical, judgmental approach creates a psychometrically acceptable and practical pathway to link assessments. It is applicable in situations where different assessments are used for similar purposes.

5. Since the number of performance levels determines score reporting, there need to be at least two levels, i.e., meets minimum proficiency and does not meet minimum proficiency, for USAID (and UIS) reporting. Adding one or two levels would permit more description of progress towards proficiency.

6. PLDs are based on content frameworks and provide detail on what students should know and be able to do by grade and subject. Following work by UIS, these need to be developed. The PLDs form the basis for global proficiency scales and performance level cut scores on assessments.

7. After developing PLDs by grade and subject, setting cut scores on assessments for performance levels can be accomplished using an internationally accepted approach. Different methods are appropriate for this purpose, depending on the characteristics of the assessments.

8. The policy linking process is validated by documenting the workshop procedures and estimating the consistency of the cut scores. Statistics such as inter-judge consistency should be calculated and evaluated across workshops, with the objective of developing standards for them.

9. The international education community has expressed support for policy linking. Momentum exists for proceeding to the next steps of developing a policy linking toolkit, piloting the method, and posting the toolkit so that stakeholders can implement it for their assessments.

10. It may be advisable to provide training to countries and partners who wish to implement policy linking for assessments. Training can include presenting the contents and procedures in the toolkit and providing simulation exercises. This may likely increase the validity and reliability of the cut scores and inferences.

REFERENCES

Angoff, W.H. (1971). Scales, norms and equivalent scores. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508–600). Washington D.C.: American Council on Education.

Beck, M. (2003). *Standard setting: If it is science, it's sociology and linguistics, not psychometrics*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Buckendahl, C.W. & Foley, B.P. (2015). *Policy linking as cut score moderation: Considerations for practice.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

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

Dorans, N.J., Moses, T.P., & Eignor, D.R. (2010). *Principles and practices of test score equating*. Princeton, NJ: ETS.

Ferdous, A.A. & Buckendahl, C.W. (2013). Evaluating panelists' standard setting perceptions in a developing nation. *International Journal of Testing, 13(1),* 4–18.

Ferdous, A.A. & Plake, B.S. (2003). Understanding the factors that influence decisions of panelists in a standard setting study. *Applied Measurement in Education, 18(3),* 257–267.

Ferdous, A.A. & Plake, B.S. (2005). *A mathematical formulation for computing inter-panelist inconsistency for*

Body of Work, Bookmark, and Yes/No Variation of Angoff Method. Paper presented for the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Ferdous, A.A., Simon, G., & Davis, J. (2017). *Setting reading fluency and comprehension benchmarks in Lebanon*. Beirut, Lebanon: QITABI.

Hambleton, R.K., Impara, J., Mehrens, W., & Plake, B.S. (2000). *Psychometric review of Maryland School Performance Assessment Program*. Baltimore, MD: MSDOE.

Kahl, S.R., Crockett, T.J., DePascale, C.A., & Rindfleisch, S.L. (1995). *Setting standards for performance levels using the student-based constructed response method*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Klein, S.P., Hamilton, L.S., McCaffrey, D.F., & Stecher, B.M. (1999). What do test scores in Texas tell us? *Education Policy Analysis Archives*, 8(49), 1-22.

Kolen, M.J. & Brennan, R.L. (2004). *Test equating, linking, and scaling: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Lewis, D.M., Green, D.R., Mitzel, H.C., & Patz, R.J. (1999). *The bookmark standard setting procedure*. Monterey, CA: McGraw-Hill.

Livingston, S.A. & Zieky, M.J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: ETS.

Mislevy, R.J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: ETS, Policy Information Center.

Mullis, I.V.S., Martin, M.O., Goh, S., & Cotter, K. (Eds.) (2016). *TIMSS 2015 encyclopedia: Education policy and curriculum in mathematics and science.* Boston, MA: TIMSS & PIRLS International Study Center.

National Governors Association Center for Best Practices (2010). *Common core standards*. Washington, DC: NGACBP.

Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practices, 27(4),* 15–29.

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

Reckase, M.D. (2000). *The evaluation of the NAEP achievement levels setting process: A summary of the research and development efforts conducted by ACT*. Iowa City, IA: ACT, Inc.

ANNEX

Performance Levels

Groups of participants in the USAID and GRN workshop developed draft performance levels/labels and policy definitions as an activity. The combined results are presented below. They can provide a starting point for the initial steps in operationalizing policy linking.

Note that most of the groups selected four performance levels, which became five levels when the results were combined. Also, some of the levels have multiple labels and some of the definitions have repetition.

Beginner
(No definition)

Below minimum proficiency / Partially proficient / Below proficiency
- Denotes partial mastery of the knowledge and skills that are prerequisite for work at a grade level
- Partially achieves standard at the relevant grade level
- Partially completes straightforward grade-appropriate tasks
- Insufficient knowledge to permit application

Minimum proficiency / Basic
- Exhibits foundational skills required to complete grade-appropriate academic tasks
- Denotes foundational knowledge and skills that permit meaningful engagement by subject and grade
- Denotes mastery of the knowledge and skills that are prerequisite for work at a given grade
- Successfully completes straightforward grade-appropriate tasks
- Sufficient knowledge to permit basic application in limited contexts

Proficient
- Achieves standard at the relevant grade level (minimum proficiency)
- Reliably and independently exhibits skills required to complete grade-level tasks
- Sufficient knowledge to permit general application in a variety of contexts

Advanced
- Successfully completes complex grade-appropriate tasks

Glossary

The workshop participants were provided with the following glossary of assessment and linking terms.

Anchor item also **common or linking item**
Common set of an assessment items administered in combination with non-anchor items with the aim of establishing equivalence in difficulty between different assessments and/or subtasks.

Assessment
Systematic process of planning, developing, and gathering information for making inferences. An assessment instrument is used to systematically collect, assess, analyze, and interpret student learning.

Benchmarks
Clear, identifiable points of reference on a continuum of achievement measurement.

Calibration
Linking method that puts an assessment's items and students onto the same scale.

Content standards
Goal statement identifying the knowledge and/or skills to be measured by an assessment instrument.

Criterion-referenced
Scores interpreted with respect to standards.

Cutoff score / Cut score
The minimum pre-established performance level or score a student must attain to be assigned a pre-determined level of achievement.

Equating
Linking method that sets up a common scale for different assessment's forms and/or subtasks. Equating can be based on a common person or item design.

Learning outcome
Direct, specific, and observable measure of student knowledge and/or skill.

Linking
Method that relates an assessment and/or subtask's scores to another assessment and/or subtask's scores.

Norm-referenced
Scores interpreted with respect to scores of other students.

Performance standard
Established level of knowledge and/or skill achievement. Can be linked to a specific cutoff score.

Performance level descriptor (PLD)
Description that outlines the knowledge and skills that students must perform at a given level.

Policy definition (of PLDs)
Statements that assert policymakers' position on the desired levels of performance across content areas and grade levels. They do not refer to specific cutoff scores.

Projection
Linking method that predicts scores on one assessment from those on another assessment.

Raw score
Total number correct responses of an assessment and/or subtask.

Scale score
Score based on some transformation(s) applied to the raw scores. Used to report scores from different versions of an assessment on the same scale.

Standard setting
Process of setting a performance cut score or cut scores.

Social moderation also **policy linking**
Linking method that obtains a concordance table of comparable scores by matching assessments by direct subjective judgement. Matches levels of performance on different assessments directly to one another.

Statistical moderation
Linking method that obtains a concordance table of comparable scores by mapping two assessments' score distributions.

United States Agency for International Development
1300 Pennsylvania Avenue, NW
Washington, D.C. 20004