unesco
Institute for Statistics

Report of the National Assessment of Educational Progress Survey Policy Linking for Measuring Global Learning Outcomes Workshop (June 2021)

Setting Global Benchmarks for Grade 6 English Language and Mathematics in Lesotho

UNESCO
The constitution of the United Nations Educational, Scientific and Cultural Organization (UNESCO) was adopted by 20 countries at the London Conference in November 1945 and entered into effect on 4 November 1946.

The main objective of UNESCO is to contribute to peace and security in the world by promoting collaboration among nations through education, science, culture and communication in order to foster universal respect for justice, the rule of law, and the human rights and fundamental freedoms that are affirmed for the peoples of the world, without distinction of race, sex, language or religion, by the Charter of the United Nations.

To fulfil its mandate, UNESCO performs five principal functions: 1) prospective studies on education, science, culture and communication for tomorrow's world; 2) the advancement, transfer and sharing of knowledge through research, training and teaching activities; 3) standard-setting actions for the preparation and adoption of internal instruments and statutory recommendations; 4) expertise through technical cooperation to Member States for their development policies and projects; and 5) the exchange of specialized information.

UNESCO Institute for Statistics
The UNESCO Institute for Statistics (UIS) is the statistical office of UNESCO and is the UN depository for global statistics in the fields of education, science, technology and innovation, culture and communication. The UIS was established in 1999. It was created to improve UNESCO's statistical programme and to develop and deliver the timely, accurate and policy-relevant statistics needed in today's increasingly complex and rapidly changing social, political and economic environments.

Published in 2022 by:
UNESCO Institute for Statistics
C.P 250 Succursale H
Montréal, Québec H3G 2K8
Canada

Email: uis.tcg@unesco.org
http://www.uis.unesco.org

Acknowledgements

The Policy-Linking methodology is a UNESCO Institute for Statistics (UIS) collaborative project. The CITO International was the technical partner in 2021-2022 and prepared the current report.

This report is based on research funded by the Bill & Melinda Gates Foundation and Educate a Child (EAC) global programme of the Education Above All Foundation (EAA). The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies the donors.

Acknowledgements

The team at Cito is grateful for the support provided by several groups for the policy linking pilot workshop.

First, the organizational support provided by officials at the Examinations Council of Lesotho (ECoL) was critical for the success of the workshops.

Second, the management support provided by officials from the UNESCO Institute for Statistics (UIS) was instrumental in planning and implementing the workshops.

Third, the hands-on support of the local content facilitators was most important for realizing the goal of this workshop.

Finally, the dedication and effort devoted during this week by the panelists were indispensable in establishing the pilot global benchmarks and drawing lessons learned from the workshops.

Sjoerd Crans

Anneke de Graaf

Michel Hop

Sanneke Schouwstra

Angela Verschoor

Table of Contents

Acronyms and Abbreviations

GPD Global Proficiency Descriptor

GPF Global Proficiency Framework

GPL Global Proficiency Level

JE Just Exceeds Minimum Proficiency

JM Just Meets Minimum Proficiency

JP Just Partially Meets Minimum Proficiency

PLT Policy Linking Toolkit

SDG Sustainable Development Goal

SEM Standard Error of Measurement

USAID U.S. Agency for International Development

ECoL The Examinations Council of Lesotho

NAEP The Lesotho National Assessment of Educational Progress Survey

UNESCO Institute for Statistics (UIS)

Glossary of Terms from the Policy Linking Toolkit

Angoff method — A benchmark setting method in which panelists rate items by GPL and then average all panelists' ratings for each GPL to create a benchmark.

Benchmark — The score on an assessment that delineates having met a proficiency level.

Breadth of Alignment — Sufficient coverage of the domains, constructs, and subconstructs in the GPF by at least one assessment item.

Content standards — What content learners are expected to know and be able to do as described in the GPF table on knowledge and skills.

Depth of Alignment — Sufficient coverage of assessment items by the GPF.

Distractor — A set of plausible but incorrect answers to the multiple-choice item on an assessment.

Global Proficiency Descriptor (GPD) — A detailed definition crafted by subject matter experts that clarifies how much of the content described under the statements of knowledge and/or skill(s) in the GPF a learner should be able to demonstrate within a subject at a grade level. These are sometimes called performance standards. Authors have purposefully not used that term, however, as countries have their own performance standards that may differ from global standards for important reasons. The set of GPDs included in the GPF are not meant to be prescriptive in nature but rather to facilitate measurement against SDG 4.1.1.

Global Proficiency Level (GPL) — The four levels of proficiency or performance - below partially meets global minimum proficiency, partially meets global minimum proficiency, meets global minimum proficiency, and exceeds global minimum proficiency - which students can achieve for all targeted grade levels and subject areas. The meets global minimum proficiency level aligns with SDG 4.1.1, and the others allow countries to show progress toward all students meeting or exceeding that level.

Impact data — The data that help panelists understand the consequences of their judgments on the learner population that are subject to application of the benchmarks recommended by the panelists.

Inter-rater consistency — An index that indicates panelists' overall agreement or consensus across all possible pairs of panelists.

Intra-rater consistency — An index that indicates panelists' overall performance in assessing test item difficulty.

Normative information — The distribution of benchmarks set by panelists, with each panelist's location indicated by a code letter or number known only to them.

Performance standards — How much of the content described in statements of knowledge and/or skill(s) (content standards) learners are expected to be able to demonstrate. See also the definition for Global Proficiency Descriptor above.

Policy linking for measuring global learning outcomes — A specific, non-statistical method that uses expert judgment to relate learners' scores on different assessments to global minimum proficiency levels. Policy linking includes processes of alignment and matching between assessments and the GPF and benchmark setting.

Item difficulty statistics — Information on the empirical difficulty of items (i.e., percentage of learners getting an item correct), which gives panelists a rough idea of how their judgments about items compare to actual learner performance.

Standard error of Measurement (SEM) — A statistic that indicates the measurement error associated with a benchmark (panelist judgment).

Statements of knowledge and/or skill(s) — What content learners are expected to know and be able to do for a specific grade and domain, construct, and subconstruct. The statements of knowledge and/or skill(s) are sometimes referred to as content standards. Authors have purposefully not used that term, however, as countries have their own content standards that may differ from global standards for important reasons. The statements of knowledge and/or skill(s) included in the GPF are not meant to be prescriptive in nature but rather to facilitate measurement against SDG 4.1.1.

Statistical linking — Methods that use common persons or common items to relate learners' scores on different assessments. Statistical linking methods include equating, calibration, moderation, and projection.

Stem — The question part of a multiple-choice item on an assessment.

Test-centered method — A family of benchmark-setting methods that make judgments based on a review of assessment material and scoring rubrics; the Angoff method is included in this category.

1. Executive Summary

This document contains the report on the online policy linking workshop that took place from May 31, 2021 until June 5, 2021. The Examinations Council of Lesotho (ECoL) and UNESCO Institute for Statistics (UIS) organized this workshop as a pilot. The objective of the workshop was to set global benchmarks on the 2016 National Assessment of Educational Progress Survey (NAEP) at grade 6 in English and mathematics using a blended policy linking workshop.

This was the first time Lesotho participated in a policy linking workshop. The local participants met physically in a location in Lesotho, whereas the international participants joined via a videoconferencing platform (Zoom). The presence of the international participants was enhanced by excellent facilities provided by the local organizers: microphones, manned camera's, screens and a close cooperation between local and international content facilitators via chat and telephone contact.

The content facilitators and the participants performed their tasks with full dedication and with excellent commitment. They were eager to learn, and at the end of the workshop were grateful for what they had learned and for the opportunity to participate. Consequently, all the activities, from the familiarization at the start to the benchmarking at the end, were carried out with full engagement and with lively and relevant discussions Every step of the process produced important outcomes. The participants gave very positive feedback, both in person and in their evaluation forms.

The closing of the workshop was celebrated with song and dance. The local organizers expressed their hope and belief that this workshop will have a catalyzing effect on the future of Lesotho's education, and a step in the direction of a closer alliance with the global (educational) community.

The participants' work showed that the NAEP for Mathematics is strongly aligned to the Global Proficiency Framework for grade 6, both in depth and breath. English Language is strongly aligned in breath to the Global Proficiency Framework for grade 6 and minimally aligned in depth. Furthermore, the panelists managed to reach almost complete consensus on the matching. The final benchmarks of the panelists show a good consistency, which makes the benchmarks useable for comparing, aggregating, and tracking learning outcomes for the NAEP in Lesotho. The piloting of the policy linking workshop in this blended mode can be considered a success.

2. Background

Policy Linking Overview

In September 2015, Member States of the United Nations formally adopted the 2030 Agenda for Sustainable Development in New York. The agenda contains 17 goals, including a new global education goal (SDG 4). SDG 4 is to ensure inclusive and equitable quality education and promote lifelong learning opportunities for all and has seven targets (UNESCO, 2021). The first target focusses on primary and secondary education (target 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. To monitor progress the indicator 4.1.1 is used: 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 (United Nations, 2021).

To allow countries to use their existing – sub-national, national, and cross-national – assessments to report against Sustainable Development Goal (SDG) 4.1.1, the policy linking methodology was developed (USAID, 2019). Policy linking makes use of a standard-setting methodology (the Angoff approach) to set benchmarks on learning assessments. While it is an existing standard-setting methodology, UIS and its partners have extended its use to help countries set benchmarks using the Global Proficiency Framework (GPF).

Global Proficiency Framework

The Global Proficiency Framework (GPF) describes the global minimum proficiency levels in reading and mathematics that learners are expected to demonstrate at the end of each grade level, from grades one to nine (USAID at all, 2019,2020a, 2020b). The framework was developed by multilateral donors and partners and is based on current national content and assessment frameworks across more than 100 countries. The overarching purpose of the GPF is to provide countries and regional/international assessment organizations with a common reference or scale for reporting progress on indicator 4.1.1 of the SDGs. The four levels outlined in the GPF—Below Partially Meets, Partially Meets, Meets, and Exceeds Global Minimum Proficiency—form a common scale from low to high achievement.

By linking their national assessments to the GPF, countries and donors can compare learning outcomes across language groups in countries as well as across countries and over time, assuming all new assessments are subsequently linked to the GPF.

The policy linking methodology

There are seven stages to policy linking for measuring global learning outcomes that must be completed to facilitate global reporting (USAID at all, 2020c). Countries/assessment agencies and their partners must complete each of these stages for their results to be accepted for reporting against SDG 4.1.1.

1. Initial engagement of a country in which a country makes the decision to move forward with policy linking.
2. Collation of evidence of curriculum and assessment validity and alignment
3. Review of evidence by the 4.1.1 Review Panel
4. Preparation for the policy linking workshop
5. Implementation of the policy linking workshop
6. Review of workshop outcomes by 4.1.1 Review Panel
7. Reporting of the results against SDG 4.1.1

The policy linking methodology is elaborated in the Policy Linking Toolkit, which provides guidance and templates to countries, donors, and partners who conduct policy linking workshops to set global benchmarks[1]. The toolkit and the accompanying Quality Assurance Policy specify the steps to be taken before, during, and following the workshops to ensure consistency and, as a result of comparability of the outcomes. The toolkit covers Stages 4 and 5.

Policy linking workshop

For each assessment, a group of 15 to 20 panelists are invited to participate in the policy linking workshop. The panel should be made up of at least 70 percent master classroom teachers and up to 30 percent non-teachers, preferably curriculum experts. The Policy Linking workshop (USAID at all, 2020c, p.12) begins with a review of the main documents that provide the foundation for the workshop—the GPF and the assessment(s) being linked to the GPF and to SDG 4.1.1. Following this review, facilitators lead panelists through three major tasks:

- Task 1 — The panelists check the alignment between the assessment and the GPF using a standardized procedure. Each panelist indicates the alignment of every item to the GPF.
- Task 2 — The panelists match the assessment items to the appropriate Global Proficiency Level and Global Proficiency Descriptor. Each panelist determines the levels of knowledge and skills required from students to correctly answer each aligned item. The panelists should work in groups to reach consensus
- Task 3 — The panelists set three global benchmarks for each assessment using a standardized method (a modified version of the Angoff methodology) through two rounds of ratings.

The policy linking methodology was piloted in several countries in 2019 and 2020, among which in India, Bangladesh and Nigeria. Also, the ICAN pilot was conducted in 2020. Following these piloting workshops, adjustments were made to the methodology, toolkit, and GPF. Due to the COVID-19 pandemic the piloting was delayed. In 2021 further piloting of the Policy Linking Toolkit will take place in several countries, using remote workshops rather than in-person workshops.

Overview of the National Assessment of Educational Progress Survey (NAEP)

The Examinations Council of Lesotho (ECoL) has been conducting the National Assessment of Educational Progress Survey (NAEP) every two years between 2004 and 2016. Until 2008, the NAEP has been administered in Grade 3 and 6, but since 2010 to Grade 4 and 6 (NAEP 2016, p. xii).

The major objective of the NAEP is to monitor the educational system in Lesotho. The NAEP was designed for the following objectives (NAEP 2016, p. xii):

- To find out what learners know and can do in Literacy and Numeracy
- To determine the actual standards in languages and Mathematics as measured against the curriculum objectives.
- To investigate factors that may be associated with learners' achievement.

[1] http://tcg.uis.unesco.org/policy-linking/

- To provide a basis for further research.

Content and design of the NAEP in grade 6

The NAEP is a low stake system level assessment that summarizes students' achievement at national and district levels. Each student received the same three cognitive tests: Mathematics, Sesotho and English. Apart from the cognitive tests each student also received a learners' questionnaire and a HIV and AIDS questionnaire. Their teachers and principals also received a questionnaire (a teachers' questionnaire, and principals' questionnaire respectively).

The mathematics test tested knowledge with understanding and problem solving in four broad areas: Numbers, Measurement, Shape and Data representation. All 35 Mathematics items were multiple choice items. The Sesotho tests comprised comprehension, language usage and culture, whereas the English tests comprised only comprehension and language usage items. The language tests contained both multiple choice items and open-ended questions. The English test contained in total 29 items divided into three sections: section A (10 items), section B (5 items), section C (14 items).[2]

Sample and data analysis

The sampling design used for NAEP 2016 is a stratified multi-stage sample design (ECoL, 2016, p. 3). For the first stage of sampling, 184 schools were selected. In the second stage, the teachers selected randomly 20 learners using a table with random numbers. In grade 6 data of the English tests were collected from 3136 learners of 181 schools. For mathematics data was collected of 3042 learners of 178 schools. The reporting scale for the National Assessment of Educational Progress Survey (NAEP) 2016 data was based on Classical Test Theory.

[2] Section C was not used in the policy linking workshop, see section "First three policy linking stages", p. 5.

3. Pilot Workshop Preparation

Objective of the workshop

The objective of the workshop was setting global benchmarks on the 2016 National Assessment of Educational Progress Survey (NAEP) at grade 6 in English and mathematics using a remote policy linking workshop. The workshop had a piloting function and should increase the capabilities of ECoL to conduct similar workshops in the future.

First three policy linking stages

After the engagement of Lesotho, on Wednesday April 21, 2021, Cito joined the meeting between UIS and ECoL. Cito was contracted to facilitate the policy linking workshop and provided the lead facilitator, two content facilitators and a data analyst. After the initial engagement, the country governments or assessment agencies should collate evidence of curriculum and assessment validity and alignment (stage 2 of policy linking) and the 4.1.1. Review Panel should review this collated evidence. However, after the initial engagement of Lesotho, the 4.1.1. Review Panel was not yet in place. “This stage of the process involves the country government sharing standard-, curriculum-, and assessment-related documents (including the most recent round of data) with the project team and examination of those documents by the project team and the 4.1.1 Review Panel to determine whether the assessment(s) meets reliability and validity standards required for a country to proceed with policy linking for reporting global outcomes.” (Policy Linking Toolkit, p. 170). The 4.1.1. Review Panel uses three criteria: Alignment between the assessment and the curriculum, Appropriateness of the assessment for the population, Reliability of the assessment.

As the 4.1.1 Review Panel was not in place, Cito made an initial assessment of whether the assessment(s) meets the standards required to proceed with policy linking. Cito’s content facilitators gave an estimate whether enough items would align. The NAEP English consists of three parts. Section C covers only isolated vocabulary and grammar and has no link to the Global Proficiency Framework. UIS decided to implement the suggestion of Cito to exclude these items from the procedure. For mathematics alignment seemed feasible. Cito could not evaluate the alignment between the NAEP and the curriculum, as we did not receive information about the curriculum.

The implemented sampling procedure, as described in the NAEP 2016 report, ensures that the learners who carried out the assessment are representative of the population against which results are reported. The report about the NAEP does not contain information on the item development and review process, so Cito did not have the information to evaluate appropriateness completely. The reliability of the NAEP was calculated using the data set provided to Cito. The reliability of the English test (without section C) was 0.80 and the reliability of the mathematics test 0.65.

Table 1. Number of grade 6 students in every state

District	Number of grade 6 students participating in NAEP	
	English	Mathematics
BEREA	326	320
BUTHA-BUTHE	185	185
LERIBE	439	509
MAFETENG	383	368
MASERU	551	450
MOHALES HOEK	291	277
MOKHOTLONG	227	223
QACHASNEK	222	216
QUTHING	215	204
THABA TSEKA	297	290
Total	3136	3042

General preparation of the workshop

UNESCO and Cito planned to facilitate the workshop remotely, due to the COVID-19 pandemic. The panelists attended in-person outside the capital and the international team of Cito and UIS attended through Zoom. As ECoL could not provide each panelist with an individual laptop with headset during the workshop, all panelists participated in the remote workshop through a big screen. Two rooms were reserved: one for the plenary meetings and the break-out session for English and one for the break-out sessions of mathematics.

As the panelists did not have a laptop, they worked on paper. For this reason, Cito developed Excel-files for data entry and a two-hour data entry training. After each task the data were entered on location in the developed Excel-files and sent to Cito.

The agenda for a 6-day blended workshop was shared with the stakeholders (ECoL, UNESCO) for suggestions and improvements. The agenda was adapted to the workday in Lesotho and adjusted to allow for data entry. After the funding was in place on Friday 28 of May, the workshop took place in a blended format from Monday 31-05-21 until Saturday 05-06-2021. UIS hosted the workshop using the platform Zoom.

ECoL sought a minimum of two teachers from each district: one teacher of English and one teacher of Mathematics. During the first day of the workshop, ECoL shared the list of panelists. An equal number of female and male teachers were found, in total 24 teachers (see Table 3) and they represented all districts. One teacher of English could not be present during the workshop due to illness.

ECoL expected all teachers to master English sufficiently for understanding all written material and therefore the material (e.g. the Global Proficiency Framework, forms) was not translated. During the workshop two translators (Sesotho) were present.

Table 2. Panelist' background information

	Mathematics	English	Total
District			
Berea	1	1	2
Botha-Bothe	1	1	2
Leribe	2	2	4
Mafeteng	1	1	2
Maseru	2	2	4
Mohale's hoek	1	1	2
Mokhotlong	1	1	2
Qacha's nek	1	1	2
Quthing	1	1	2
Thaba Tseka	1	1	2
Total	12	12	24
Gender			
F	5	7	12
M	7	5	12
Grand Total	12	12	24
Level of education			
Completed 4-year College	5	1	6
Completed Master's Education		4	4
Some College	6	5	11
Some Master's Education	1	2	3
Grand Total	12	12	24

Materials for the workshop and pre-workshop analyses

During the preparation of the workshop, all partners (UIS, ECoL and Cito) followed the week-by-week timeline for the Policy Linking Workshop as described in the UIS Activity plan for Lesotho (see Annex C). All partners strictly followed the timeline, only with respect to the funding the timeline was not met.

Collecting materials and pre-workshop analyses

Before the workshop, ECoL shared the assessments. The panelists were not asked to administer the NAEP to students before the workshop but were asked to take the assessment themselves during the workshop. ECoL also shared the raw data before the workshop and the sampling weights. In preparation for the workshop the distribution of the sum scores was calculated and the p-values using Classical Test Theory (see Appendix F). All students made the same test.

Creating workshop materials

All panelists assembled on location outside the capital. To limit the number of workshop days, a six-day workshop was planned (see the overview in Table 4, in Annex A the complete agenda is presented).

Based on the digital forms, forms were created in a printable format so the panelists could fill out their alignment rating (Figure 10), item rating (Figure 12) and evaluation on paper (Annex B). For the data entry of both subjects, three Excel files were developed: one for the entry of the alignment ratings (separate for Mathematics and Language, see Figure 13), one for the entry of the item ratings (Figure 14) and one for the entry of the evaluation forms (Figure 15).

Cito prepared a package for panelists containing all workshop materials, to be printed on location. The package contained the agenda, the Global Proficiency Framework for Grades 5 to 7, Glossary and acronym list, a handout of the slides of all presentations. Furthermore, the package contained the Alignment rating form, Matching form, Item rating form, and evaluation forms.

Cito adapted the workshop slides to the agenda of Lesotho and their assessment (the NAEP). More importantly, Cito's content facilitators adapted all examples to grade 6. The sample grade 6 items were selected and included in the slides to illustrate the three different tasks and to practice the tasks (alignment, matching, benchmarking).

Table 3. Agenda for a 6-day blended workshop

Day 1—31 May2021	Day 4—3 June2021
Welcome and introductions	Complete Task 2 Matching
Overview Presentation: Policy linking	Task 2 Presentation: Matching results
Overview Presentation: Global Proficiency Framework (GPF)	Task 3 Presentation: Global benchmarking & Angoff
Overview Presentation: National Assessment (NAEP)	Task 3 Activity: Practice and start Angoff ratings
Day 2—1 June2021	**Day 5—4 June2021**
Do the National Assessment & Review GPF	Complete Round 1Angoff ratings
Task 1 Presentation: GPF and alignment	Task 3 Presentation: Round 1 results
Task 1 Activity: Align the National Assessment and the GPF	Task 3 Presentation: Discuss round 1 ratings
Day 3—2June 2021	**Day 6—5 June2021**
Complete Task 1 Alignment	Conduct Angoff ratings Round 2
Task 1 Presentation: Alignment results	Task 3 Presentation: Round 2 results
Task 2 Presentation: Matching NAEP and GPLs	Task 3 Activity: Evaluate workshop
Task 2 Activity: Match NAEP and GPDs/GPLs	Closing and logistics

Training the local content facilitators

During the last week before the workshop, the content facilitator training was held. Cito planned a 5-hour training consisting of 3 different parts for both the local content facilitators for Language and Mathematics:

1. A one-hour introduction into generics and specifics of Policy Linking for both local content facilitators
2. A two-hour interactive session for Language and Mathematics separately focusing on the relevant part of the GPF and on the specific activities of the local content facilitators during the different parts of the workshop (Alignment, Matching and Benchmarking)
3. A 2-hour general rehearsal of the workshop for both Language and Math.

The whole Lesotho team was invited for the introduction (1) and the general rehearsal (3). The interactive sessions were intended for Cito's content facilitators and their local counter parts (Lesotho's content facilitators). It is important that Cito's content facilitator and their counterparts created a good working relationship and understanding of their respective roles during the workshop. In the separate interactive session, they focused on the relevant part of the GPF and

on the specific activities of the local content facilitators during the different parts of the workshop.

The Technical test during the general rehearsal was limited to creating break-out rooms in Zoom. A technical test of all the facilities in the Lesotho venue (big screen, audio, microphones, internet and wide-angle cameras) was not possible. The venue was only available on the workshop days. For this reason, a Technical Test was planned one hour before the registration of the panelists on the first day of the workshop. It proved to be impossible to get all technical equipment to run smoothly at the start of the workshop.

Training for local data entry

As the panelists worked on paper, data entry was needed, and a special 2-hour data entry training was given on the second day of the workshop. On three days (day 3, 5 and 6) data entry had to occur. The panelists handed in their forms at the end of the morning and during lunch time the data had to be entered. As the data had to be analyzed and the results presented that same afternoon, the window for data entry was narrow. During the training the schedule and times for data entry were shown. Next, Cito discussed the steps in data entry and gave a demonstration of data entry for each of the different forms.

The global steps in data entry were:

1. Receive form
    a. Track if each panelist has handed in form (on the tracking form)
    b. Check for errors in the paper forms or data entry and correct errors.
2. Copy the panelists' ratings (as the panelists need their ratings for the next task or round).
3. Data entry in Excel
4. Check if data entry is correct
5. Send all forms to Cito

4. Implementing the blended workshop

Familiarization

Following feedback from other policy linking workshops, the workshop started with a preparation session. After the formal welcome, the first day focused on familiarizing panelists with policy linking, the Global Proficiency Framework and the National Assessment of Educational Progress Survey. The panelists received the printed workshop materials in the venue (such as the Global Proficiency Framework). The materials had only been shared with the panelists digitally before the workshop, because they were travelling from different regions.

During the sessions, the panelists were provided with background information on policy linking, including a chronology of the development of the method in response to the global indicators. UIS asked their regional advisor in Africa to present the panelists with an overview of the Global Proficiency Framework and its role in policy linking. The example of the benchmarks and the proficiency levels is shown in Figure 3.

In the breakout rooms, the content facilitators introduced each of the domains, constructs, subconstructs, statements of knowledge and/or skill(s), and GPLs and GPDs. An example from part of the mathematics GPF is shown in Table 5.

Figure 1. Example of three benchmarks and the global proficiency levels

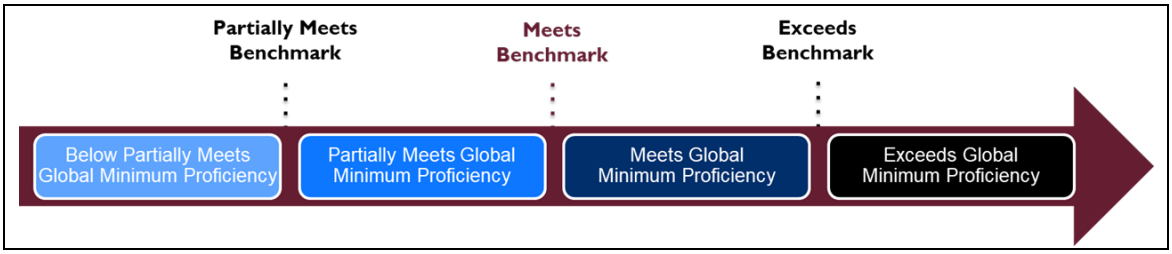


Table 4. Part of the Global Proficiency Framework of Mathematics describing the domain, constructs and subconstructs

Domain		Construct		Subconstruct	
N	Number and operations	N1	Whole numbers	N1.1	Identify and count in whole numbers, and identify their relative magnitude
				N1.2	Represent whole numbers in equivalent ways
				N1.3	Solve operations using whole numbers
				N1.4	Solve real-world problems involving whole numbers
		N2	Fractions	N2.1	Identify and represent fractions using objects, pictures, and symbols, and identify relative magnitude
				N2.2	Solve operations using fractions
				N2.3	Solve real-world problems involving fractions
		N3	Decimals	N3.1	Identify and represent decimals using objects, pictures, and symbols, and identify relative magnitude
				N3.2	Represent decimals in equivalent ways (including fractions and percentages)
				N3.3	Solve operations using decimals
				N3.4	Solve real-world problems involving decimals
		N4	Integers	N4.1	Identify and represent integers using objects, pictures, or symbols, and identify relative magnitude
				N4.2	Solve operations using integers
				N4.3	Solve real-world problems involving integers
		N5	Exponents and roots	N5.1	Identify and represent quantities using exponents and roots, and identify the relative magnitude
				N5.2	Solve operations involving exponents and roots
		N6	Operations across number	N6.1	Solve operations involving integers, fractions, decimals, percentages, and exponents

The day closed with an introduction to the National Assessment of Educational Progress Survey and discussing the first five items of the NAEP in the subject-specific break-out rooms. In the morning of the second day the panelists were asked to study the Global Proficiency Framework and fill-out the NAEP themselves. While answering the items of the NAEP the panelists were

asked to note stumble blocks and aspects of the items that might make the item easy or difficult for Grade 6 students.

Observations

Because the venue was not available prior to the start of the workshop, the technical test took place on the morning of the first day of the workshop, an hour before the start. The number of people present was limited, so not everything could be tested, especially the set-up with the breakout rooms for the two physical rooms at the location.

There were several technical issues during the first day. For example, there was a problem with the sound in the room where the mathematics group was, which was solved by reconnecting again, but this reoccurred during the rest of the workshop.

On this day and during the whole workshop, there was good and frequent contact via WhatsApp chat and telephone between the local and Cito content facilitators. This helped enormously with the technical issues, in the sense that both sides stayed informed, so it was easier to maintain focus when there were problems with connections. The content facilitators used these communication means to confer about content and organizational issues as well.

In both physical rooms, microphones were available, there was a manned camera and a big screen on which the zoom meeting was being projected. This worked very well, given the circumstances This set-up approached being there in-person for the international facilitators.

The presentations about policy linking and about the GPF did not succeed well in engaging the panelists. This is partly understandable, because it is something unfamiliar and complex. A possible other reason is the form of the presentation, which is one-directional.

Familiarization with the GPF is a difficult task, for which the panelists needed a lot of guidance from the content facilitators, both local and international. One complication is that in the presentation preceding the first task, the whole content of the GPF is described, from the key knowledge and skills in the GPF up to the Global proficiency levels (GPL) and Global Proficiency Descriptors (GPD). This mentioning of the GPL and GPD prior to task 1 can be very confusing for panelists, because in the first task (the alignment), the panelists need to focus only on the knowledge and skills required to answer an item correctly.

The length of the NAEP for mathematics and for language was very different: 15 and 35 items respectively. This makes it (more) difficult to keep both groups in synch, to have them both ready for the plenary parts at the same time.

For language, for which the NAEP was in English, it turned out to be difficult for the panelists to relate this to the GPF grade 6. A possible reason for this is that English, although being an official language of Lesotho, is clearly a second language, Sesotho being the first language for more than 90 percent of the population [https://en.wikipedia.org/wiki/Languages_of_Lesotho, retrieved 14/6/2021]. In this workshop, the two effects just mentioned worked in opposite directions, keeping the synching challenge manageable.

Task 1: Alignment

The following days, the panelists were asked to work individually in the morning while the local content facilitators were present and, in the afternoon, the sessions contained presentations by facilitators and activities for panelists to complete in groups. The panelists had to execute three tasks during the workshop:

- Task 1 — Rate the alignment between the NAEP and the GPF

- Task 2 — Match the NAEP items to the appropriate Global Proficiency Level and Global Proficiency Descriptor.
- Task 3 — Set three global benchmarks for the NAEP

On the afternoon of the second day of the workshop, the panelists received an introduction to their first task: aligning the National Assessment of Educational Progress Survey to the Global Proficiency Framework (GPF). Alignment is important, because it ensures there are enough items in the assessment that measure the knowledge and/or skill(s) depicted in the GPF for policy linking to work. The purpose of the alignment task was to ensure panelists have fully understood the GPF and to allow them to identify which statements of knowledge and/or skill(s) describe the knowledge and/or skill(s) required of children to answer assessment items correctly.

The alignment method in the policy linking toolkit is a two-step process based on a specific and standardized method that is appropriate to policy linking (Frisbie, 2003). In the first step, panelists independently rate the alignment between the NAEP items and GPF knowledge and/or skill(s) statement(s) and in the second step the facilitators compile and summarize the ratings to check the alignment between the assessments and the GPF.

In the break-out rooms, the content facilitators started to practice together with the panelists in conducting item-statement of knowledge and/or skill(s) ratings with sample items. The content facilitators trained the panelists to rate each item using a scale of Complete Fit, Partial Fit, and No Fit as follows:

- Complete Fit (C) signifies that all content required to answer the item correctly is contained in the statement of knowledge and/or skill(s), i.e., if the learner answers the item correctly, it is because they completely use the knowledge and/or skill(s) described in the statement.
- Partial Fit (P) signifies that part of the content required to answer the item correctly is contained in the statement of knowledge and/or skills, i.e., if the learner answers the item correctly, it is because they partially use knowledge and/or skill(s) described in the statement.
- No Fit (N) signifies that no amount of the content required to answer the item correctly is contained in the statements of knowledge and/or skill(s), i.e., if the learner answers the item correctly, it is because they do not use knowledge and/or skill(s) described in the GPF.

The panelists were provided with additional guidelines that 1) complete fit was usually associated with only one statement in the GPF, 2) partial fit was usually associated with more than one statement of knowledge and/or skill(s), and 3) no fit was not associated with any one statement of knowledge and/or skill(s) in the GPF.

The next morning, panelists were asked to work individually and independently to rate the alignment between each NAEP item and the GPF knowledge and/or skill(s) statements. They had to start with the first item and proceed item-by-item and find the GPF knowledge and/or skill(s) statements that align (if any) with the knowledge or skill(s) needed to answer the item correctly. They were asked to record their ratings on the alignment rating form which they received by email (see Annex B). After they completed the alignment rating, they had to send their rating form to an email address created exclusively for this workshop.

After the panelists sent their alignment forms on day 3, the lead facilitator completed the second step. All alignment ratings forms were merged into one file, checked and analyzed.

All results were summarized at the subconstruct level. Only the subconstructs were considered with knowledge and/or skill(s) expected at the grade level for which alignment was being conducted (grade 6). The data analyst took the average of the number of items that the

panelists aligned to each grade 6 subconstruct, construct and domain. Each item was counted only once (even if it was a partial fit), non-fitting items were not counted towards alignment.

Alignment English

All results were summarized at the subconstruct level. Only the subconstructs were considered with knowledge and/or skill(s) expected at the grade level for which alignment was being conducted (grade 6). The data analyst took the average of the number of items that the panelists aligned to each grade 6 subconstruct, construct and domain. Each item was counted only once (even if it was a partial fit), non-fitting items were not counted towards alignment.

Averaging the panelists' ratings, we see that all 15 items (on average) aligned to Reading comprehension. At least 5 items were aligned to Retrieve information, but less than 5 items to Interpret information (on average 2,5). The NAEP English is therefore minimally aligned in depth rather than strongly aligned (see Table 6).

We see that on average all subconstructs of Reading comprehension are covered (see in Table 19 in Annex D). The NAEP English assessment was therefore strongly aligned in breadth (see the criteria in Table 6). We do see that 10 out of 15 items are aligned to R1.2 Retrieve explicit information in a grade-level text by direct- or close-word matching.

Table 5. Reading Alignment Criteria for Grades 1–9

Level of Alignment	Category	Grade 1–2 Criteria	Grade 3–6 Criteria Grade	Grade 7–9 Criteria
Minimally Aligned	Domain/Construct (depth):	D (minimum five items) C (minimum five items)	R (minimum five items)	R (minimum five items)
	Subconstructs (breadth):	Items covering at least 50 percent of the D and C subconstructs	Items covering at least 50 percent of the R subconstructs	Items covering at least 50 percent of the R subconstructs
Additionally Aligned	Domain/Construct (depth):	N/A	N/A	R: R1 (minimum 5 items) R: R2 (minimum 5 items)
	Subconstructs (breadth):	N/A	N/A	Items covering at least 50 percent of the R subconstructs
Strongly Aligned	Domain/Construct (depth):	R (minimum five items)	R: B1 (minimum 5 items) R: B2 (minimum 5 items)	R: R1 (minimum 5 items) R: R2 (minimum 5 items) R: R3 (minimum five items)
	Subconstructs (breadth):	Items covering at least 50 percent of the R subconstructs	Items covering at least 50 percent of the R subconstructs	Items covering at least 50 percent of the R subconstructs

Key:
D—Decoding
C—Comprehension of spoken or signed language
R—Reading comprehension
R1—Retrieve information
R2—Interpret information
R3—Reflect on information

Alignment NAEP Mathematics

"When summarizing results to the subconstruct level, facilitators and/or data analysts should only consider the subconstructs with knowledge and/or skill(s) expected at the grade level for which alignment is being conducted. " (PLT, p. 15). Averaging the panelists' ratings, on average 28 of the 34 items, aligned to grade 6 subconstructs. One item was excluded from the ratings, because information was missing from the item (item 29). In the Global Proficiency Framework 24 subconstructs are mentioned for grade 6 and the NAEP covered 18 of those subconstructs (an average of >0.5, see Table 20 in Annex D). In breadth the NAEP is strongly aligned to the Global Proficiency Framework for Grade 6 as the items covered more than 50% of all grade 6 subconstructs.

The NAEP Mathematics items covered all five domains and all 12 constructs for grade 6. According to the new criteria in the Policy Linking Toolkit, for strong alignment in Depth at least 5 items should align to the domain Number and Operations, at least 5 items to Measurement and Geometry and at least 5 items to Statistics and Probability and Algebra (see Table 7). On average 12.1 items covered the domain of Number and Operations, 13.3 items the domains Measurement and Geometry, and 5.4 items the domains Statistics and Probability and Algebra. For this reason, the NAEP is also strongly aligned to the GPF in depth.

Table 6. Mathematics Alignment Criteria for Grades 1–9

Level of Alignment	Category	Criteria
Minimally Aligned	Domain/Construct (depth):	Number (minimum five items)
	Subconstructs (breadth):	Items covering at least 50 percent of the Number and Operations subconstructs
Additionally Aligned	Domain/Construct (depth): Subconstructs (breadth):	Number (minimum 5 items) and Measurement and Geometry (minimum 5 items) Items covering at least 50 percent of the Number, Measurement, and Geometry subconstructs
Strongly Aligned	Domain/Construct (depth):	Number (minimum five items) and Measurement and Geometry (minimum five items) and Statistics and Probability and Algebra (minimum five items)
	Subconstructs (breadth):	Items covering at least 50 percent of all subconstructs

Observations

From the alignment task onwards, the language group and mathematics group stayed in their separate physical rooms, coming together digitally via the Zoom platform for the plenary activities. This worked well, except for one moment where the sound problem in the math room could not be fixed quickly. This was circumvented by the math group physically joining the language group in the other room just for this moment.

In the plenary presentation on alignment, examples were presented of the three types of fit, but only for mathematics. It would have helped the language panelists if there would also have been similar examples for language in the presentation.

Although the working language of the workshop was English, the panelists benefitted greatly from being assisted by the local content facilitators in Sesotho from time to time. Such interventions/discussions were then summarized and communicated to the international content facilitator either by the interpreter or by the local content facilitator themselves.

The manual filling in of the alignment forms went smoothly, as well as the data entry process by the data entry persons in both groups.

The addition of codes for the knowledge or skill statements is a big improvement. However, in the mathematics GPF some inconsistencies were still found.

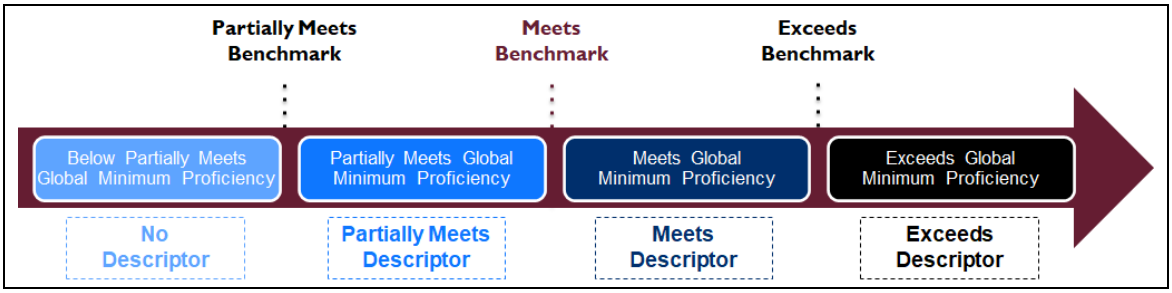
Table 7. The new knowledge or skill codes for mathematics.

Domain	Construct	Subconstruct		Knowledge or Skill
	N1 Whole numbers	N1.1	Identify and count in whole numbers, and identify their relative magnitude	**N1.1.1** - Count, read, and write whole numbers
				N1.1.2 - Compare and order whole numbers
				N1.1.3 - Skip count forwards or backwards
		N1.2	Represent whole numbers in equivalent ways	**N1.2.1** - Determine or identify the equivalency between whole numbers represented as objects, pictures, and numerals
				N1.2.2 - Use place-value concepts
				N1.2.3 - Round whole numbers
		N1.3	Solve operations using whole numbers	**N1.3.1** - Add and subtract whole numbers
				N1.3.2 - Find the double or half of a set of objects
				N1.3.3 - Multiply and divide whole numbers
				N1.3.4 - Demonstrate fluency with basic addition and subtraction facts
				N1.3.5 - Demonstrate fluency with basic multiplication and division facts
				N1.3.6 - Identify factors and multiples of whole numbers
				N1.3.7 - Perform calculations involving two or more operations on whole numbers
		N1.4	Solve real-world problems involving whole numbers	**N1.4.1** - Solve real-world problems involving the addition and subtraction of whole numbers, including with measurement and currency units
				N1.4.2 - Solve real-world problems involving the multiplication and division of whole numbers, including with measurement and currency units

Task 2: Matching

On the third day, after the panelists completed task 1, they received training for the next task: Matching the NAEP items with the Global proficiency levels and descriptors. Task 2 builds on the panelists' understanding of the items and GPF gained through the alignment activity. The purpose of Task 2 is to further narrow down the expectations of learners measured by each assessment item. The panelists should identify the descriptors (GPDs) of global minimum proficiency that match with the items.

Figure 2. Global Proficiency Levels (GPLs) and Global Proficiency Descriptors (GPDs) in the Global Proficiency Framework



A Global Proficiency Descriptor is a detailed definition crafted by subject matter experts that clarifies how much of the content described under the statements of knowledge and/or skill(s) in the Global Proficiency Framework a learner should be able to demonstrate within a subject at a grade level. The Global Proficiency Descriptors (GPD) describes the minimum proficiency for the Global Proficiency Levels (GPLs), i.e., the minimum knowledge or skill(s) necessary for classification into each GPL (by grade and subject), see Figure 4.

The Global Proficiency Descriptors are organized by domain, construct and subconstruct, with descriptors for each subconstruct. In Table 9 an example is displayed of Global Proficiency Descriptors for the three GPLs (partially meets, meets and exceed global minimum proficiency). For mathematics full consensus was reached, even though for one item this took a lot of

discussion. Apart from the item in which information was missing (item 29), the panelist agreed that another two items did not align or match with the Global Proficiency Framework (item 3 and item 10). Also, for English full consensus was reached, but for 3 of the 15 items this took a lot of discussion.

Table 8. Example of the Global Proficiency Descriptors for three Proficiency Levels.

G1: PROPERTIES OF SHAPES AND FIGURES					
G1.1: Differentiate shapes and figures by their attributes					
G1.1.2_P	Recognize and name three-dimensional figures by their attributes *(e.g., faces, edges, vertices).*	G1.1.2_M	Identify parallel and perpendicular sides of shapes.	G1.1.2_E	N/A
G1.1.3_M	N/A	G1.1.3_M	N/A	G1.1.3_E	Use the defining attributes (i.e., type of angle, parallel and perpendicular lines) of complex two-dimensional shapes to classify them.
G1.1.5_P	Recognize and name types of triangles *(e.g., isosceles, scalene, equilateral, and right angle).*	G1.1.5_M	Recognize and name types of quadrilaterals *(e.g., parallelogram; trapezium, etc.).*	G1.1.5_E	N/A
G1.1.7_P	Recognize types of angles by their magnitude *(e.g., right, straight, acute, obtuse).*	G1.1.7_M	N/A	G1.1.7_E	Estimate the size of angles by comparing to reference/benchmark angles *(e.g., estimate the size of a given angle with reference to the fact that it is smaller than a right angle and larger than 45°).*

Observations

In the two subject groups, the matching activity was carried out slightly differently. In the mathematics group, the whole group proceeded together item by item, the discussion being facilitated by both content facilitators. The language group worked first in subgroups, trying to reach consensus there first, and then brought together these results in the whole group, where further discussion ensued when the subgroups had differing opinions. Here as well the content facilitators had an important facilitating part.

There was confusion as to which grades were applicable, both for the alignment and the matching task. In the end, the math group used for alignment: if you can find a knowledge or skill statement *in any grade* → fit. For matching: find a descriptor *in grade 5-7*, if it is not in grade 6 → translate the level (according to the diagonal pattern that exists in the GPD for math), *even if the subconstruct is N/A in grade 6*.

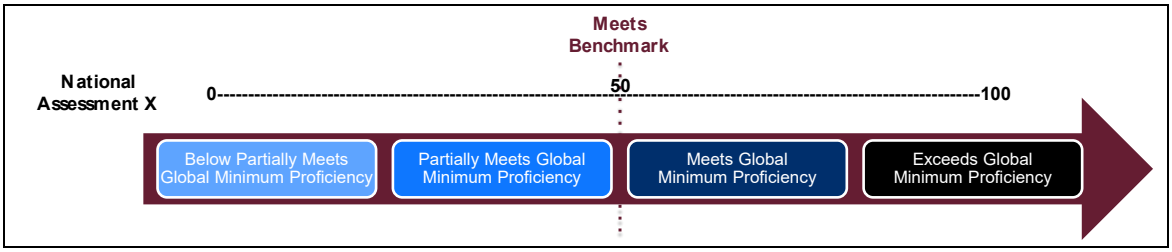
The conclusions from the matching task were recorded by both content facilitators and at the end of the task compared as a check.

The matching task took longer than scheduled, but not longer than expected.

Task 3: Benchmarking

On the fourth day the panelists received training in setting global benchmarks using the Angoff method. The facilitator first presented a hypothetical example of how the benchmarking method would link a national assessment to the GPF, thus allowing for the calculation of the percentages of students attaining minimum proficiency (see Figure 5). This example was extended to three national assessments of different difficulties, and how this would lead to a different benchmark for each assessment. The facilitators discussed how the benchmarking results – when applied to the assessment data sets – could be used for comparing and aggregating assessment results, as well as tracking those results over time.

Figure 3. Example of an assessment and a benchmark



The panelists then received an introduction to their third task: setting benchmarks with the Angoff benchmarking method. The lead facilitator emphasized that the ratings for task 3 should be individual and independent and that, in contrast to task 2, consensus on the rating is not needed, even though consistency is desired.

The benchmarks represent the panel's estimates of scores that a minimally proficient learner at each level would obtain on the assessment. The panelists were asked to rate the items using the following steps:

Step 1: Identify and/or conceptualize three Just Partially Meets (JP), three Just Meets (JM), and three Just Exceeds (JE) learners based on an understanding of the GPF.

Step 2: Carefully read the first item on the assessment and, building from Task 1, consider the knowledge and/or skill(s) required to answer the item correctly. Consider what makes the item easy or difficult (e.g., the wording of the item stem and the strength of the incorrect options, or distractors) and what kind of errors may be possible or reasonable.

Step 3: Building from Task 2, select the domain, construct, subconstruct, knowledge or skill, and GPLs/GPDs in the GPF that are most relevant for the item.

Step 4: Based on an understanding of Steps 1–3, follow this procedure (displayed in Figure 6): Ask whether minimally proficient JP learners would be able to answer the item correctly, i.e., are you reasonably sure (≥ 67 percent chance, or 2 out of the 3 JP learners)?

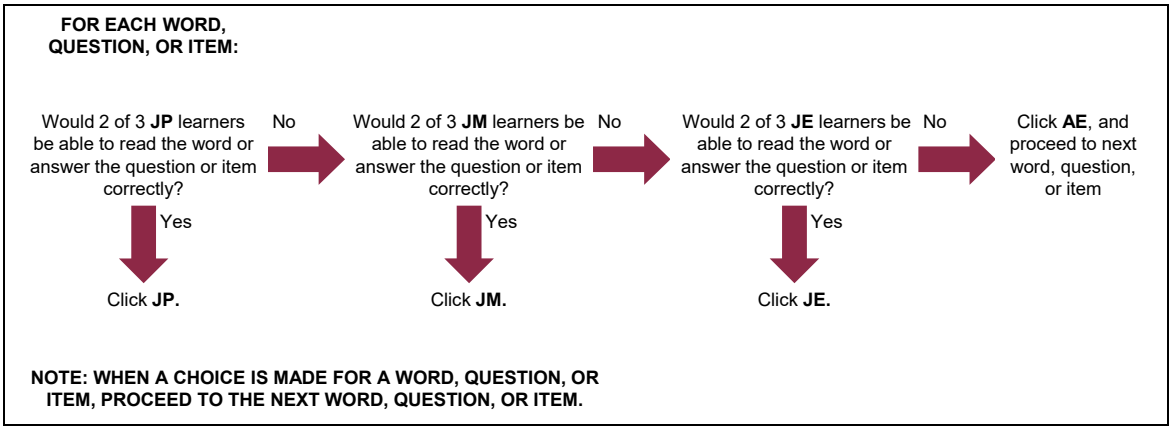
- If "yes," place an "X" under JP and proceed to the next item.
- If "no," ask whether minimally proficient JM learners would be able to answer the item correctly?
    - If "yes," place an "X" under JM and proceed to the next item.
    - If "no," ask whether minimally proficient JE learners would be able to answer the item correctly?
        - If "yes," place an "X" under JE and proceed to the next item.
        - If "no," place an "X" under AE and proceed to the next item.

The global benchmarks are calculated based on the total ratings by each panelist and the averages across all the panelists.

Round 1

After practicing with the benchmarking, the panelists continued with the first round of Item Rating on the fifth day. Again, the panelists were asked to conduct the ratings individually and independently. They were asked to focus on the item content in relation to the statements of knowledge and/or skill(s) in the GPF and take into considerations the difficulty of the item. To obtain realistic ratings, the panelists should consider what a learner *would* answer at the respective GPL, rather than what a learner *should* answer.

Figure 4. Steps for Rating Items



After the panelists conducted their first ratings in the morning of the fifth day, they handed in their forms to the persons responsible for data entry. They kept track of the forms sent and checked whether:

- The panelist rated all items
- The panelist had filled in the ID at the top (rather than the name, or missing)

Once all the forms were entered, the data entry file was sent to Cito and the data analysis could start. The data-analysts performed the analyses and compiled a report to give feedback to the panelists during the workshop. In the report the following was contained:

- Per item the average rating, the minimum, maximum, and standard deviation of the ratings.
- A list of sum scores of panelists ratings for the three benchmarks
- A plot of anonymous ratings (referred to as location statistics in the policy linking toolkit)
- The p-values as calculated prior to the workshop
- The benchmarks of the panel, containing for each minimum proficiency level the benchmark, the score range and the estimated percentages of learners in the category.
- The intra- and inter-rater consistency

The lead facilitator presented the preliminary results of Round 1. The content facilitators then facilitated an item-wise discussion. The content facilitators focused during the discussion on those items where panelists strongly disagreed. The facilitators invited the panelists to share their views during the discussion.

Round 2

During the morning of the last day, the panelists conducted the second rating using the same procedure. After the panelists conducted their second ratings in the morning of the sixth day, they handed their forms to the data entry persons. Like the day before, they tracked the submission of the forms and checked the forms. After the data entry, the file was sent to Cito. While the panelists filled out a short questionnaire, the data analyst analyzed the ratings. In the afternoon, the lead facilitator shared the results with the panelists.

Observations

As expected, the conceptualization of three Just Partially Meets (JP), three Just Meets (JM), and three Just Exceeds (JE) learners based on an understanding of the GPF provided to be the most challenging part of the workshop for the panelists. This is not because something went

wrong, but because this is inherently difficult. First, to “switch off” your own intuitions and knowledge based on your own experience in your own country and on the country’s curriculum, and instead building a picture of a JP “global learner” based on all the descriptors of the Partially Meets level. Secondly, to decide what it means, based on this picture, that this learner is *just* in the PM level: which tasks at Below Partially Meets level is such a learner able to carry out and which tasks at PM level? And the same for the other two levels. And then to apply this to the actual items on the NAEP.

All content facilitators showed thoroughness in their support, both in assisting the panelists in understanding the benchmarking task and in facilitating the discussion between round 1 and round 2. All panelists showed great commitment to do a good job.

The filling in of the forms, by the panelists and by the data entry persons, went as smoothly as it did with the alignment task. With some effort, the bottleneck that is the analysis after round 1 was successfully negotiated but this remains a risk.

The benchmarking task took shorter than scheduled, but not shorter than expected.

Workshop evaluation

Near the end of the sixth day, after returning the Round 2 ratings, all panelists were asked to share their opinion about the workshop. Their evaluations are completely anonymous. They were informed that their opinion was important to improve the workshop and to evaluate the validity and reliability of the standard setting process. The panelists had about one hour to answer the questions about:

a) The training on the Global Proficiency Framework
b) The training on the National Assessment of Educational Progress Survey
c) The training on the alignment methodology
d) The training on the matching methodology
e) The training on the benchmark-setting (Angoff) methodology
f) Benchmark Round 2 evaluation
g) Overall evaluation

The questions included are presented in the policy linking toolkit (see also Annex F). As the panelists worked on paper, a paper-based version of the questionnaire (originally in Microsoft Forms) was made. The evaluation consists of Likert-type scales and open-ended questions on the panelists’ satisfaction with the orientation, training, and process.

5. Results of the benchmarking

Round 1

The data analyst and lead facilitator produced summary tables and graphs from the first round, which showed the initial benchmarks, score ranges, and impact data for each Minimum Proficiency Level (see Table 10 and Table 11). In the plenary room the panelists were presented with anonymous normative information on the panelists ratings (see Figure 7 and Figure 8). We saw that the ratings of panelists varied considerably, especially for the lowest benchmark (Partially meets). We also see a clear ceiling effect with English. Exceeds is at the maximum (15) for all panelists.

Figure 5. Anonymous information on the panelists' ratings of English Round 1

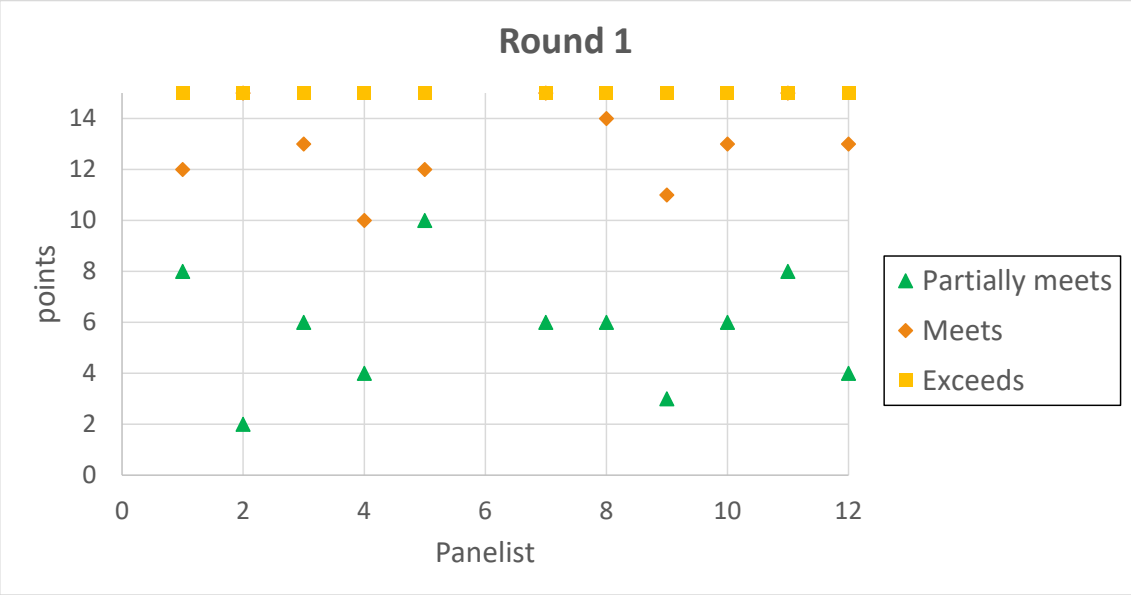
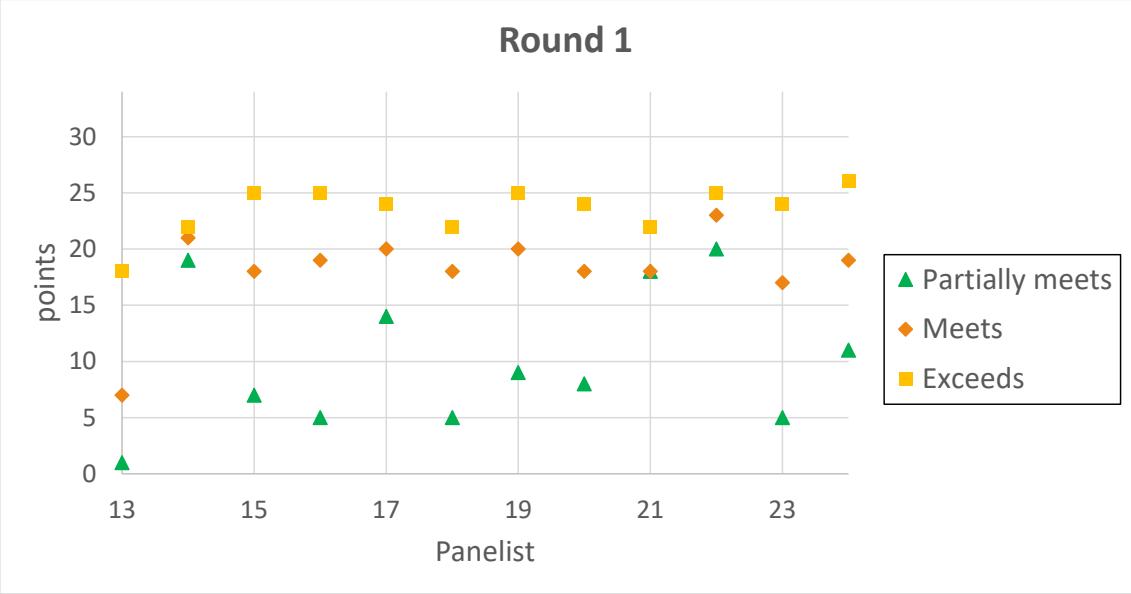


Figure 6. Anonymous information on the panelists' ratings of Mathematics Round 1



After round 1 the benchmark was calculated as the average of the panelists' benchmarks. The average benchmark was rounded down, as stipulated in the policy linking toolkit. For English, the impact information shows 31.8% of the learners would fall in the Below Partially Meets Proficiency level and more than half (61.4%) in the Partially Meets level (see Table 10).

For Mathematics, the impact information shows that only eleven percent (11.5%) would fall in the Below Partially Meets Proficiency level and more than half (65.2%) would fall in the Partially Meets level (see Table 11) using round 1 benchmarks.

Table 9. Round 1 benchmarks, score range and impact for English

Minimum Proficiency Levels	Round 1 Benchmark	Score Range	Percentage of Learners		
			Male	Female	Total
Below Partially Meets	N/A	0 - 4	40.1%	24.8%	31.8%
Partially Meets	5.73	5 - 12	55.3%	66.6%	61.4%
Meets	13.00	13 - 14	3.6%	7.4%	5.7%
Exceeds	15.00	15 - 15	0.9%	1.2%	1.1%

Table 10. Round 1 benchmarks, score range and impact for Mathematics

Minimum Proficiency Levels	Round 1 Benchmark	Score Range	Percentage of Learners		
			Male	Female	Total
Below Partially Meets	N/A	0 - 9	13.8%	9.7%	11.5%
Partially Meets	10.17	10 - 17	64.9%	65.3%	65.2%
Meets	18.17	18 - 22	17.6%	20.7%	19.3%
Exceeds	23.5	23 - 34	3.6%	4.3%	4.0%

Round 2

After providing the results from the initial benchmarks in Round 1 to the panelists, the panelists discussed the items. They focused on items for which the ratings differed a lot. After the discussion the panelists individually conducted the Round 2 ratings and submitted their forms. The data analyst produced a parallel set of summary tables and graphs with final benchmarks.

We see that in Round 2 the ratings of panelists varied less than in Round 1, especially for Mathematics (Figure 7 and Figure 8).

Figure 7. Anonymous information on the panelists' ratings of English Round 2

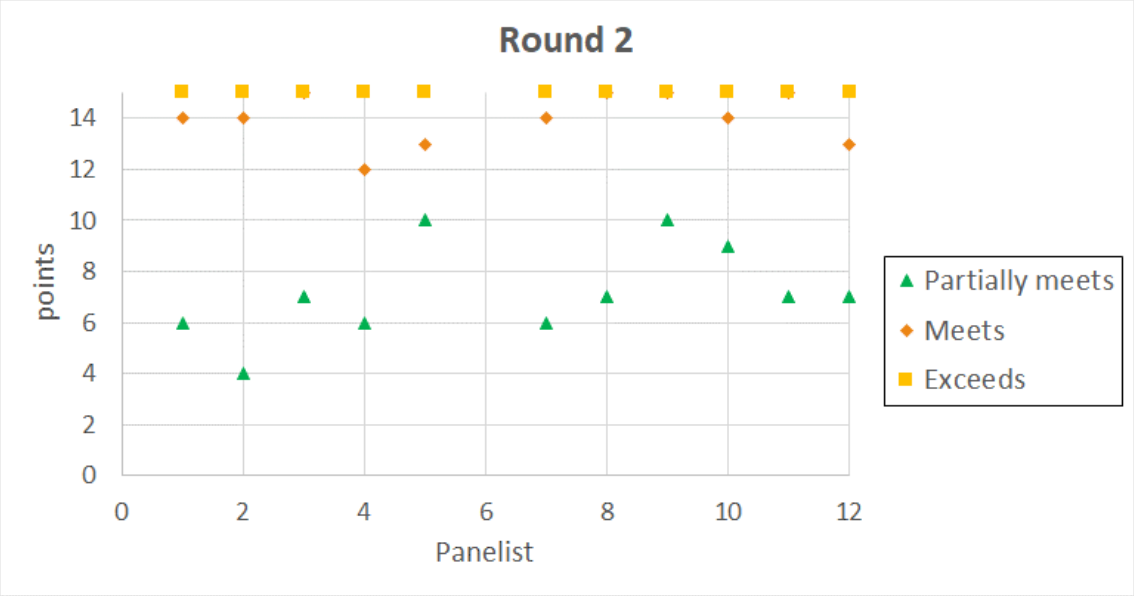
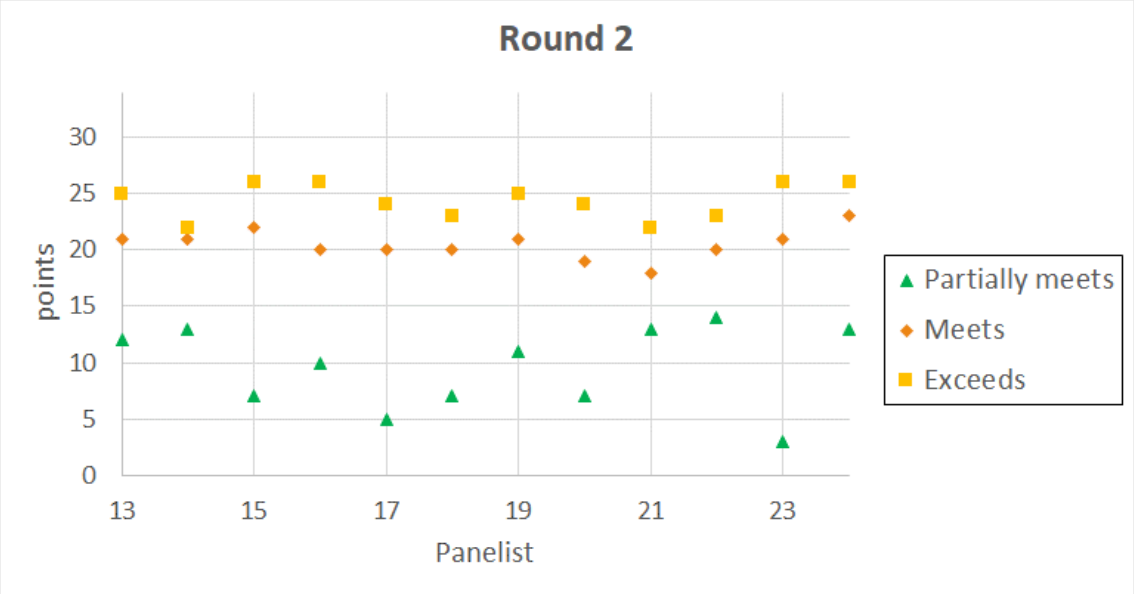


Figure 8. Anonymous information on the panelist's ratings of Mathematics Round 2



For English, the results show that in Round 2 almost half of the learners (49.9%) fall in the Below Partially Meets level (see Table 12). Also, almost half of the students (46.7%) fall in the Partially Meets level. Only 2.3% fall in the Meets level and only 1.1% in the Exceeds level. The benchmarks were set higher in round 2 than in round 1. The Below Partially Meets benchmark was set lower in round 2 than in round 1 and the Meets and Exceeds benchmarks higher (see Table 13). Consequently, after round 2 a higher percentage of learners falls in the Below Partially meets proficiency level. The Exceeds benchmark is set at the top of the scale, which is a clear ceiling effect.

Table 11. Round 2 benchmarks, score range and impact for English

Minimum Proficiency Levels	Round 2 Benchmark	Score Range	Percentage of Learners		
			Male	Female	Total
Below Partially Meets	N/A	0 - 6	58.3%	42.8%	49.9%
Partially Meets	7.18	7 - 13	39.3%	53.0%	46.7%
Meets	14.00	14 - 14	1.4%	3.0%	2.3%
Exceeds	15.00	15 - 15	0.9%	1.2%	1.1%

Table 12. Comparison of Round 1 benchmarks and Round 2 benchmarks for English

Minimum Proficiency Levels	Round 1 Benchmark	Round 1 Percentage of Learners	Round 2 Benchmark	Round 2 Percentage of Learners
Below Partially Meets	N/A	31.8%	N/A	49.9%
Partially Meets	5.73	61.4%	7.18	46.7%
Meets	13.00	5.7%	14.00	2.3%
Exceeds	15.00	1.1%	15.00	1.1%

For Mathematics, we see that in Round 2 the Partially Meets benchmark was set at a lower score and the Meets and Exceeds benchmarks slightly higher (see Table 15). Consequently, after round 2 a higher percentage of learners falls in the Partially Meets proficiency level. A lower percentage of learners than was the case in Round 1 fall in the other levels. Only 6.5% of the learners fall in the Below Partially Meets level (Table 14). More than three quarters of the students (81.4%) fall in the Partially Meets level, 9.4% in the Meets level and 2.7% in the Exceeds level.

Table 13. Round 2 benchmarks, score range and impact for Mathematics with 34 items

Minimum Proficiency Levels	Round 2 Benchmark	Score Range	Percentage of Learners		
			Male	Female	Total
Below Partially Meets	N/A	0 - 8	7.6	5.6%	6.5%
Partially Meets	9.58	9 - 19	81.4%	81.4%	81.4%
Meets	20.50	20 - 23	8.7%	10.0%	9.4%
Exceeds	24.33	24 - 34	2.3%	3.0%	2.7%

Table 14. Comparison of Round 1 benchmarks and Round 2 benchmarks for Mathematics with 34 items

Minimum Proficiency Levels	Round 1 Benchmark	Percentage of Learners	Round 2 Benchmark	Percentage of Learners
Below Partially Meets	N/A	11.5%	N/A	6.5%
Partially Meets	10.17	65.2%	9.58	81.4%
Meets	18.17	19.3%	20.5	9.4%
Exceeds	23.5	4.0%	24.33	2.7%

6. Evaluation of the Standard Setting Process

Internal Evaluation SEM, Panelist Consistency and Panelists' Agreement

In addition to calculating benchmarks and impact data, the Policy Linking Toolkit also requires calculating measures of consistency and presenting evaluation feedback results. These measures of consistency are reported in Table 16 and Table 17.

As shown in Table 16, the Standard Error of Measurement (SEM), which measures how much panelists' benchmarks are spread around a "true" benchmark, was in both rounds under 1.0 for both English with 15 items and under 2.00 for Mathematics with 34 items. The results show that the Standard Error of Measurement is smaller for the Exceeds benchmarks. This is a consequence of a ceiling effect for this benchmark. For English, all panelists have put the Exceeds benchmark at the maximum sum score (15) (see the previous section).

Table 15. Standard Error of Measurement by Round

	SEM by Benchmark					
	Round 1			Round 2		
Subjects	Partially Meets	Meets	Exceeds	Partially Meets	Meets	Exceeds
English	0.71	0.50	0	0.55	0.30	0
Mathematics	1.81	1.12	0.63	1.05	0.38	0.45

The results show that the inter-consistency for both English and Mathematics was higher in Round 2 than in Round 1. The inter-rater consistency index evaluates the panelists' overall agreement or consensus across all possible pairs of panelists. Inter-rater consistency is calculated at the item level and for the entire assessment. The value ranges between 0 and 1. According to the Policy Linking Toolkit values of 0.80 or greater are desirable, as they indicate substantial agreement between the panelists. Both for English and Mathematics the interrater consistency was above the 0.80 (see Table 17).

The intra-rater consistency index evaluates the panelists' overall consistency in estimating item difficulty. Intra-rater consistency is calculated for each panelist across all items on the assessment. The value ranges between 0 and 1. A lower value indicates high consistency and a higher value indicates low consistency. We see that the intra-rater consistency is quite high (given the scale of 0 to 1): above .75.

Table 16. Inter-rater consistency and intra-rater consistency by subject and round

	Round 1		Round 2	
Subjects	Inter-Rater Consistency	Intra-Rater Consistency	Inter-Rater Consistency	Intra-Rater Consistency
English	0.83	0.81	0.87	0.78
Mathematics	0.83	0.77	0.88	0.78

Procedural Evaluation

All panelists shared their opinion about the workshop through a questionnaire (see Annex F). The panelists indicated on a five-point scale (Strongly Disagree-Disagree-Neutral-Agree-Strongly Agree) how strongly they agreed with several statements about six aspects of the workshop. On average, we see that the respondents were positive about the workshop. All six aspects received an average score above 4 (on a scale of 1 to 5). The overall evaluation shows

that the panelists are overall very positive: 4.44 on a scale of 1 to 5 (the neutral category has been added to the scale, which was missing in the example in the Policy Linking Toolkit).

Table 17. Workshop evaluation results

Part of the workshop	Scale	Number of statements	Average scale score	Standard deviation of scale score	N
The training on the Global Proficiency Framework	1-5	8	4.65	0.32	23
The training on the National Assessment of Educational Progress Survey[3]	1-5	5	4.47	0.42	23
The training on the alignment methodology	1-5	5	4.53	0.39	23
The training on the matching methodology	1-5	5	4.41	0.61	23
The training on the benchmark-setting (Angoff) methodology[4]	1-5	10	4.50	0.42	23
Benchmark Round 2 evaluation	1-5	8	4.28	0.46	23
Overall evaluation	1-5	3	4.44	0.55	23

[3] One question was left out because the question was not applicable: "Administering the assessment helped me to understand how minimally proficient learners would perform on the assessment (this is only applicable if the panelists were able to assess learners ahead of the workshop").
[4] One question was missing on the paper-based form "I was able to follow the instructions and complete the Round 1 form accurately".

7. Summary of results of criterion 4 for the 4.1.1 Review Panel

The results of the policy linking workshop in Lesotho are summarized in Table 19 and Table 20. In the policy linking toolkit (Annex U, p. 164) six criteria are mentioned for the validity of policy linking workshop. The evaluation of the validity is based on the intra-rater and inter-rater reliability, the standard error of measurement, the representativeness of the panel and panelists' understanding of the procedures.

The 4.1.1 Review Panel will review the workshop outcomes (PLT, p. 52) and make a recommendation whether the policy linking has been carried out appropriately and the reported outcomes are validated. If not, more evidence might be required, or the workshop needs to be rerun because the policy linking was not carried out appropriately and/or outcomes cannot be validated. The 4.1.1 Review Panel will also provide a grade for the adequacy of the policy linking workshop. If four of the six criteria are met, two of which must be criteria b and c (inter-rater reliability and SE), the grade will be "Good". If all six criteria are met, the grade will be "Excellent".

For English (Table 19), the intra-rater and inter-rater reliability meet the requirements. The standard error of measurement is low. However, the third benchmark ("Exceeds") might not be valid. All panelists put the Exceeds benchmark at the maximum, so there is no variation and a clear ceiling effect (even though this is not mentioned as a criterium). The panel has good gender representation and a good geographical representation. The panelists are all teachers in English, but their experience is unknown. The panelists rated their understanding of the GPF, assessment, and policy linking methodology above 4 and they felt on average comfortable with their Round 2 evaluations and final benchmarks. The adequacy of the policy linking workshop for English in Lesotho can be considered to be good.

For mathematics (Table 20), the intra-rater and inter-rater reliability meet the requirements. The standard error of measurement is low. The panel has good gender representation and a good geographical representation. The panelists are all teachers in mathematics, but their experience is unknown. The panelist rated their understanding of the GPF, assessment, and policy linking methodology above 4 and they felt on average comfortable with their Round 2 evaluations and final benchmarks. The adequacy of the policy linking workshop for mathematics in Lesotho can be considered to be good.

Table 18. Summary of Results for Criteria for Policy Linking Validity English Grade 6

Question		Criteria	Response
a)	What was the intra-rater reliability for the second round of ratings?	The intra-rater reliability will vary depending on the number of items on the assessment. The panel will provide guidance on how they determined acceptability.	0.78
b)	What was the inter-rater reliability for the second round of ratings?	The inter-rater reliability should be at least .80.	0.87
c)	What was the Standard Error of Measurement (SEM) at each global proficiency level?	SEM should be appropriate for each global proficiency level reported. There is no maximum SEM provided in this document, since it will depend on the number of items in the assessment.	Number of items: 15 0.55 (Partially Meets) 0.30 (Meets) 0.00 (Exceeds)
d)	To what extent were the panelists representative of the target population of schools being reported on?	Panelists should be selected to ensure: • Gender representation – The panelists must be selected to ensure gender balance, both for the teachers and non-teachers. • Geographical representation – The teachers (and non-teachers, if possible) must be selected to ensure representation from regions, provinces, and/or states. • Ethnic and/or linguistic representation (where applicable) • Representation of crisis-and-conflict-affected areas.	• 58% female, 42% male • From each district one or two teachers (8% or 17%)
e)	To what extent did the panelists meet the other selection criteria described in the Policy Linking Toolkit?	Panelists should all have: • Several years of teaching experience in the grade level for which they are providing ratings (classroom teachers) • Skills in the subject area (all panelists) • Skills in the different languages of instruction and assessment (all panelists) • Knowledge of learners of different proficiency levels, including at least some who would meet the requirements of the meets minimum proficiency level and some who would meet the requirements of the exceeds minimum proficiency level (all panelists) • Knowledge of the instructional environment (all panelists) • Experience administering the assessment(s) being used for the policy linking workshop.	• 100% teachers of English • 42% completed either college or master education

f)	To what extent did panelists report understanding the GPF, assessment, and policy linking methodology? And, to what extent did they feel comfortable with their Round 2 evaluations and final benchmarks?	On a five-point Likert scale, with 1 being strongly disagree, very uncomfortable, etc. and 5 being strongly agree, very comfortable, etc., the average rating for each of these criteria should be 4 or above.	**GPF** • I understand the purpose of the GPF - 4,82 • I understand the relationship between domains, constructs, subconstructs, knowledge and skills, and GPDs - 4,91 • The GPDs were clear and easy to understand - 4,45 **NAEP** • I understand the purpose of the assessment - 4,82 • I understand the constructs assessed in the assessment - 4,82 • I understand how the assessment is administered - 4,55 **Alignment** • I understand the purpose of alignment - 4,73 • I understand the alignment methodology - 4,45 • I understand the difference between no fit, partial fit, and complete fit - 4,73 **Matching** • I understand the purpose of matching - 4,64 • I understand the matching methodology - 4,55 • I understand how the alignment activity links to the matching activity - 4,64 **Benchmarking methodology** • I understand the process I need to follow to complete the benchmarking exercise - 4,8 • I understand how the benchmarking methodology links to the steps on alignment and matching - 4,73 • I understand the difficulty level of the assessment items - 4,64 **Benchmark round 2** • I understand the data on others' ratings - 4,56 • I understand the item difficulty data and how it relates to this process - 4,73 • I understand the impact data and how it relates to this process - 4,36 **Comfortable with Round 2** • How comfortable are you with your final performance predictions? - 4,5

Table 19. Summary of Results for Criteria for Policy Linking Validity Mathematics Grade 6

Question		Criteria	Response
g)	What was the intra-rater reliability for the second round of ratings?	The intra-rater reliability will vary depending on the number of items on the assessment. The panel will provide guidance on how they determined acceptability.	0.78
h)	What was the inter-rater reliability for the second round of ratings?	The inter-rater reliability should be at least .80.	0.88
i)	What was the Standard Error of Measurement (SEM) at each global proficiency level?	SEM should be appropriate for each global proficiency level reported. There is no maximum SEM provided in this document, since it will depend on the number of items in the assessment.	Number of items: 34 1.81 (Partially Meets) 1.12 (Meets) 0.63 (Exceeds)
j)	To what extent were the panelists representative of the target population of schools being reported on?	Panelists should be selected to ensure: • Gender representation – The panelists must be selected to ensure gender balance, both for the teachers and non-teachers. • Geographical representation – The teachers (and non-teachers, if possible) must be selected to ensure representation from regions, provinces, and/or states. • Ethnic and/or linguistic representation (where applicable) • Representation of crisis-and-conflict-affected areas.	• 42% female, 58% male • From each district one or two teachers (8% or 17%)
k)	To what extent did the panelists meet the other selection criteria described in the Policy Linking Toolkit?	Panelists should all have: • Several years of teaching experience in the grade level for which they are providing ratings (classroom teachers) • Skills in the subject area (all panelists) • Skills in the different languages of instruction and assessment (all panelists) • Knowledge of learners of different proficiency levels, including at least some who would meet the requirements of the meets minimum proficiency level and some who would meet the requirements of the exceeds minimum proficiency level (all panelists) • Knowledge of the instructional environment (all panelists) • Experience administering the assessment(s) being used for the policy linking workshop.	• 100% teachers of Mathematics • 42% completed either college or master education

l)	To what extent did panelists report understanding the GPF, assessment, and policy linking methodology? And, to what extent did they feel comfortable with their Round 2 evaluations and final benchmarks?	On a five-point Likert scale, with 1 being strongly disagree, very uncomfortable, etc. and 5 being strongly agree, very comfortable, etc., the average rating for each of these criteria should be 4 or above.	**GPF** • I understand the purpose of the GPF - 4,42 • I understand the relationship between domains, constructs, subconstructs, knowledge and skills, and GPDs - 4,92 • The GPDs were clear and easy to understand - 4,08 **NAEP** • I understand the purpose of the assessment - 4,67 • I understand the constructs assessed in the assessment - 4,5 • I understand how the assessment is administered - 4,25 **Alignment** • I understand the purpose of alignment - 4,67 • I understand the alignment methodology - 4,5 • I understand the difference between no fit, partial fit, and complete fit - 4,67 **Matching** • I understand the purpose of matching - 4,58 • I understand the matching methodology - 4,42 • I understand how the alignment activity links to the matching activity - 4,42 **Benchmarking methodology** • I understand the process I need to follow to complete the benchmarking exercise - 4,67 • I understand how the benchmarking methodology links to the steps on alignment and matching - 4,42 • I understand the difficulty level of the assessment items - 4,25 **Benchmark round 2** • I understand the data on others' ratings - 4,5 • I understand the item difficulty data and how it relates to this process - 4,33 • I understand the impact data and how it relates to this process - 4 **Comfortable with Round 2** • How comfortable are you with your final performance predictions? - 4,08

8. Conclusions and Recommendations

Due to the travel restrictions of COVID-19, the international facilitators hosted the workshop using a videoconferencing platform (Zoom). The participants met in person in one single location with two rooms. For many of the participants, this was the first time they participated in an international workshop and the first time using a videoconferencing platform.

After getting used to this mode the first day, the participants engaged in lively discussion regarding the alignment of the NAEP items with the Global Proficiency Framework, the matching and the Item ratings. The participants performed their tasks with dedication. Every step of the process produced important outcomes. The participants gave very positive feedback, both in person and in their evaluation forms. In this respect the piloting of the policy linking workshop in this blended mode can be considered a success.

The participants' work showed that the NAEP for English is in breadth strongly aligned to the Global Proficiency Framework and minimally aligned in depth. Mathematics is both in depth and breadth strongly aligned to the Global Proficiency Framework for grade 6. Furthermore, the panelists managed to reach complete consensus on the matching both for English and for mathematics. The final benchmarks of the panelists show a good consistency, which makes the benchmarks useable for comparing, aggregating, and tracking learning outcomes for the NAEP in Lesotho.

Recommendations

Based on Cito's observations during the workshop, several lessons can be drawn that are useful for coming workshops that are conducted in a blended mode such as was used for this workshop.

Workshop Preparation

Collecting workshop materials and pre-workshop analyses

- In the policy linking toolkit, the materials to be collected, such as the assessment instrument and the data file, are clearly described. The UIS activity plan ensured the workshop materials were exchanged in a timely manner.
- It is important that the Review Panel 4.1.1 is in place. We found that the reliability of the mathematics test is rather low, and some items of English did not fit with the Global Proficiency Framework. To ensure the reliability of the results of the workshop, an independent panel needs to evaluate before the workshop whether the assessment(s) meets the standards required to proceed with policy linking.

Creating workshop materials

- A technical test should be held well in advance of the workshop. A technical test with all locations and participants will also make clear in advance if back-up material or equipment is needed (e.g. the WhatsApp contact) and to troubleshoot any technology issues.
- A list of participants with their contact details should be available at least a week ahead of the workshop. The contact details and demographic information can be checked, and a panelist ID can be provided individually. This also would allow inviting panelists to a technical test and providing them with the Global Proficiency Framework prior to the workshop.
- It would be much easier for the panelists to familiarize themselves with the Global Proficiency Framework and to execute the tasks, if they received key documentation in

the form of a hand-out translated in their own language, especially the Global Proficiency Framework, but also the presentations.
- Working with two physical break-out rooms (and a digital plenary room) worked well. It prevented a lot of confusion, which often occurs when people participate for the first time in Zoom and work with digital break-out rooms. If participants participate individually, from their own device, they need some instruction and practice time when participating for the first time. This instruction can be given during a Technical test or, for example, during the registration on the first day of the workshop.

Training the local content facilitators

- The local content facilitators received a more intensive training and participated in a general rehearsal. Conducting a rehearsal with the local content facilitators helped in raising the awareness of the goals of the workshop and of the tasks panelists must perform.
- The local content facilitators should also receive the (translated) Global Proficiency Framework well ahead of the workshop.

Implementing the blended workshop

- To facilitate the sessions and discussions, it is essential that everything is translated (from English to the local language and vice versa). The presence of an interpreter (in addition to the local content facilitator) should be planned for all sessions.
- A three-week workshop as is described in the policy linking toolkit is the preferred option. The schedule in the six-day blended workshop is very tight and forms a risk for the quality of the results. In a six-day workshop, there is very little room for adapting to unforeseen circumstances or solving technical problems, such as occurred during the first day.
- In a blended workshop, more time is needed for collecting, checking, merging, analyzing and reporting the results of the alignment and two Rounds of Item rating. The process of collecting and checking the forms locally and doing data-entry locally, made the process much smoother.
- When conducting a blended workshop with panelists in-person in one location, the set-up as used in this workshop is recommended: subject groups in separate physical rooms, one digital break-out room for one of the groups, good audio and video facilities such as microphones, cameras and screens. Also, a good and frequent contact between local and international content facilitators, for example via WhatsApp and/or telephone.

Familiarization

The familiarization phase is new in the policy linking toolkit. We feel the familiarization is an important addition.

- The agency or governmental organization that has created the assessment, is best suited to give a presentation about the assessment, instead of the lead facilitator.
- The presentations, both plenary and in the subgroups, should be more pedagogically informed, with suitable involvement of the panelists: more practicing than presenting. This to enhance engagement of the panelists and to avoid them feeling overwhelmed.
- The presentations should take the starting point of the panelists more into account. The panelists seem to have difficulty with the many acronyms and technical words. A didactical approach can help in making the slides clearer and less word-based aiming at more language independent information. A translation of the slides would help as well.

- The two plenary starting presentations/activities on the first day: Overview of the policy linking and Overview of the GPF should be given by an experienced trainer with in-depth knowledge of policy linking and of the GPF.
- Perform the familiarization of the GPF in two steps: up to and including the knowledge or skill statements before the Alignment task, and the GPD and GPL between the Alignment task and the Matching task. This avoids possible confusion by the panelists and a possible overload of information on the first day.
- In conducting a workshop for more subjects and/or grades, it would be helpful if the assessments for the different groups were of similar length.
- In selecting a national assessment on language for policy linking, take particular care to select an assessment for which the language in the assessment is the first language of the learners taking the assessment.

Task 1: Alignment

- In the plenary presentation on alignment, also provide examples for the three types of alignment for languages,
- The remaining inconsistencies in the mathematics GPF should be repaired.
- The panelists should focus on knowledge or skill statements, not whether it is the appropriate grade.

Task 2: Matching

- Evaluate the different ways the matching activity was carried out in the two subject groups –first find consensus in subgroups or work immediately with the whole group – and choose one of them (or another) for future workshops.
- Give clearer instructions in the PLT on how to deal with items that match with a descriptor from a grade other than the one under consideration.
- Perform an extra check by letting both the local and the international content facilitator administer the conclusions and comparing afterwards.
- Schedule more time for the matching task, especially for the consensus discussions.

Task 3: Benchmarking

- Take particular care to spend enough time and effort on the conceptualization of JP, JM and JE learners.
- In this conceptualization, distinguish clearly between the hypothetical learner fitting the Global Proficiency Descriptors for a Global Proficiency Level and the actual learners in the country: these latter ones may not be representative for the former ones, because of different choices made in the curriculum or specific circumstances in the country for example. Therefore, be careful with the interpretation of p-values of items as indicative of ‘global’ difficulty.
- Schedule less time for the Benchmarking task, without compromising the effort needed to conceptualize JP, JM and JE learners.
- Schedule sufficient time for the data entry and analysis, both after round 1 and after round 2.

9. References

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.) Educational Measurement (2nd ed.). Washington, DC.: American Council on Education.

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

Examinations Council of Lesotho (2016). Lesotho National Assessment of Educational Progress grade 4 and 6: The 2016 survey report. Lesotho.

UNESCO. (2021, March). SDG 4: Education. https://en.unesco.org/gem-report/sdg-goal-4.

United Nations (2021, March). Sustainable development Goals. *Global indicator framework adopted by the General Assembly (A/RES/71/313), annual refinements contained in E/CN.3/2018/2 (Annex II), E/CN.3/2019/2 (Annex II), and 2020 Comprehensive Review changes (Annex II) and annual refinements (Annex III) contained in E/CN.3/2020/2.* https://unstats.un.org/sdgs/indicators/Global%20Indicator%20Framework%20after%202020%20review_Eng.pdf

USAID (2019). *Policy Linking Method: Linking assessments to global standards. Draft paper.* Downloaded 26/3/2021 from https://www.edu-links.org/sites/default/files/media/file/Final%20Policy%20Linking%20Justification%20Paper_03062019.pdf

USAID, UNESCO Institute for Statistics (UIS), UK's Foreign, Commonwealth, and Development Office (FCDO), The Bill and Melinda Gates Foundation, Australian Council for Education Research (ACER), World Bank (2019). *Global Proficiency Framework: Reading and Mathematics*. Downloaded from https://www.edu-links.org/resources/global-proficiency-framework-reading-and-mathematics.

USAID, UNESCO Institute for Statistics (UIS), UK's Foreign, Commonwealth, and Development Office (FCDO), The Bill and Melinda Gates Foundation, Australian Council for Education Research (ACER), World Bank (2020a). *Global Proficiency Framework for Mathematics Grades 1 to 9*. Downloaded from https://www.edu-links.org/sites/default/files/media/file/GPF_Math_Final_Jan19.pdf

USAID, UNESCO Institute for Statistics (UIS), UK's Foreign, Commonwealth, and Development Office (FCDO), The Bill and Melinda Gates Foundation, Australian Council for Education Research (ACER), World Bank (2020b). *Global Proficiency Framework for Reading Grades 1 to 9*. Downloaded from https://www.edu-links.org/sites/default/files/media/file/GPF_Reading_Final_Dec23.pdf

USAID, World Bank, UNESCO Institute for Statistics (UIS), UK's Foreign, Commonwealth, and Development Office (FCDO), Australian Council for Education Research (ACER), MSI (2020c). Policy Linking for Measuring Global Learning Outcomes Toolkit: Linking Assessments to the Global Proficiency Framework. Downloaded from https://www.edu-links.org/sites/default/files/media/file/Policy_Linking_for_Measuring_Global_Learning_Outcomes_Final.pdf.

Warm T.A. (1989). Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika*, *54*, 427-450.

10. Annexes

Annex A: Agenda for the blended 6-day workshop

LESOTHO POLICY LINKING WORKSHOP
FOR REPORTING ON SDG 4.1
May 31st – June 5th, 2021

Overview

Day	Time	Activity
Monday, May 31	9:00 – 17:30	Welcome, Introductions, Overview of policy linking, Overview of the Global Proficiency Framework (GPF), Overview of the NAEP, Review of the NAEP items
Tuesday, June 1	9:00 – 17:30	Taking the NAEP, Review of the GPF, Discussing NAEP and GPF, Introduction to Task 1: Alignment, Practice Alignment
Wednesday, June 2	9:00 – 17:45	Complete Task 1, Introduction to Task 2: Matching, Practice Matching, Review Task 1 Results
Thursday, June 3	9:00 – 18:30	Complete Task 2, Review Task 2 Results and Discussion, Introduction to Task 3: Benchmarking, Practice Benchmarking, Facilitator-Panelist Consultations
Friday, June 4	9:00 – 18:30	Complete Round 1 Benchmark Ratings, Review and Discuss Round 1 Benchmark Ratings, Review and Discuss Item Difficulty and Impact Data, Presentation on Task for Round 2, Facilitator-Panelist Consultations
Saturday, June 5	9:00 – 17:15	Complete Round 2 Ratings, Present and Discuss Round 2 Ratings and Workshop Outcomes, Presentation of Certificates, Closing

LESOTHO POLICY LINKING WORKSHOP FOR REPORTING ON SDG 4.1

May 31st – June 5th, 2021

Monday, May 31, 2021

Start	End	Activity	Facilitation
9:00	09:30	Registration	Project team
09:30	10:45	Welcome and introductions	ECOL, UIS
10:45	11:00	Morning tea break	
11:00	11:45	Presentation: Overview of policy linking	UIS
11:45	12:45	Presentation: Overview of the GPF	UIS
12:45	13:45	Lunch break	
13:45	15:15	GPF Review	Content facilitators
15:15	15:30	Afternoon tea break	
15:30	16:15	Presentation: Overview of the NAEP	ECOL
16:15	17:15	Review NAEP items	Content facilitators
17:15	17:30	Explanation of individual work next day & closing	Content facilitators

LESOTHO POLICY LINKING WORKSHOP FOR REPORTING ON SDG 4.1

May 31st – June 5th, 2021

Tuesday, June 1, 2021

Start	End	Activity	Facilitation
9:00	9:30	Introduction of Day 2 and solving issues of Day 1	Lead facilitator
9:30	10:15	Taking the NAEP	Content facilitators
10:15	10:45	Review GPF and identify any elements that are still unclear	Content facilitators
10:45	11:00	Morning tea break	
11:00	12:45	Review GPF and identify any elements that are still unclear (Continued)	Content facilitators
12:45	13:45	Lunch break	
13:45	14:45	Discussion of taking the NAEP and reviewing GPF	Content facilitators
14:45	15:30	Task 1 Presentation: GPF and alignment	Lead facilitator
15:30	15:45	Afternoon tea break	
15:45	16:30	Task 1: Small group discussions on first 5 items	Content facilitators
16:30	17:15	Task 1: Plenary discussion on questions that came up in the groups	Content facilitators
17:15	17:30	Explanation of individual work next day & closing	Content facilitators

LESOTHO POLICY LINKING WORKSHOP FOR REPORTING ON SDG 4.1

May 31st – June 5th, 2021

Wednesday, June 2, 2021

Start	End	Activity	Facilitation
9:00	09:15	Welcome and purpose of session 3	Lead facilitator
9:15	10:45	Task 1: Alignment of NAEP and the GPF	Content facilitators
10:45	11:00	Morning tea break	
11:00	12:45	Task 1: Alignment of NAEP and the GPF (cont.)	Content facilitators
12:45	13:45	Lunch break	
13:45	15:45	Task 2 Presentation: Matching NAEPs and GPDs/GPLs	Content facilitators
15:45	16:00	Afternoon tea break	
16:00	16:45	Task 2 Activity: Matching NAEP items and GPDs/GPLs	Content facilitators
16:45	17:30	Task 1 Presentation: Alignment results	Lead facilitator
17:30	17:45	Explanation of individual work next day & closing	Lead facilitator

LESOTHO POLICY LINKING WORKSHOP FOR REPORTING ON SDG 4.1

May 31st – June 5th, 2021

Thursday, June 3, 2021

Start	End	Activity	Facilitation
9:00 -	09:15	Welcome and purpose of session 4	Lead facilitator
9:15 -	10:45	Task 2: Small groups complete Task 2 together	Content facilitators
10:45 -	11:00	Morning tea break	
11:00 -	12:45	Task 2 Plenary discussion: Matching NAEP items and GPDs/GPLs and results of matching	Content facilitators
12:45 -	13:45	Lunch break	
13:45 -	14:15	Task 3 Presentation: Global benchmarking	Lead facilitator
14:15 -	15:00	Task 3 Presentation: Angoff method	Lead facilitator
15:00 -	15:30	Task 3 Presentation: Angoff practice	Content facilitators
15:30 -	15:45	Afternoon tea break	
15:45 -	16:30	Task 3: Plenary discussion of questions that arose in small groups	Content facilitators
16:30 -	17:15	Task 3a Activity: Angoff Round 1	Content facilitators
17:15 -	17:30	Explanation of individual work next day & closing	Content facilitators
17:30 -	18:30	Consultation hour in which panelists can consult the content facilitator	Content facilitators

LESOTHO POLICY LINKING WORKSHOP FOR REPORTING ON SDG 4.1

May 31st – June 5th, 2021

Friday, June 4, 2021

Start	End	Activity	Facilitation
9:00	09:15	Welcome and purpose of session 5	Lead facilitator
9:15	10:45	Task 3a: Complete Round 1 ratings on all remaining items	Content facilitators
10:45	11:00	Morning tea break	
11:00	12:45	Task 3a: Complete Round 1 ratings on all remaining items (continued)	Content facilitators
12:45	13:45	Lunch break	
13:45	15:30	Task 3a: Review and discus Round 1 ratings in plenary	All facilitators
15:30	15:45	Afternoon tea break	
15:45	16:45	Task 3a: Review Round 1 ratings in small groups, going through each item where there was disagreement	Content facilitators
16:45	17:15	Task 3a: Share and discuss item difficulty and impact data	Content facilitators
17:15	17:30	Explanation of individual work next day & closing	Content facilitators
17:30	18:30	Consultation hour in which panelists of each state can consult the content facilitator	Content facilitators

LESOTHO POLICY LINKING WORKSHOP FOR REPORTING ON SDG 4.1

May 31st – June 5th, 2021

Saturday, June 5, 2021

Start	End	Activity	Facilitation
9:00	09:15	Welcome and purpose of session 6	Lead facilitator
9:15	10:45	Task 3b: Complete Task 3 Activity Angoff Round 2	Content facilitators
10:45	11:00	Morning tea break	
11:00	12:45	Task 3b: Complete Task 3 Activity Angoff Round 2 (continued)	Content facilitators
12:45	13:45	Lunch break	
13:45	14:45	Workshop evaluation	Individual
14:45	15:30	Task 3b Presentation: Round 2 results	Lead facilitator
15:30	15:45	Afternoon tea break	
15:45	16:45	Discuss outcomes and final panelist questions	Lead facilitator
16:45	17:15	Closing and logistics	ECOL, UIS

Annex B: Example of the forms

Figure 9. Alignment rating form (English) for paper-based rating

Panelist ID	

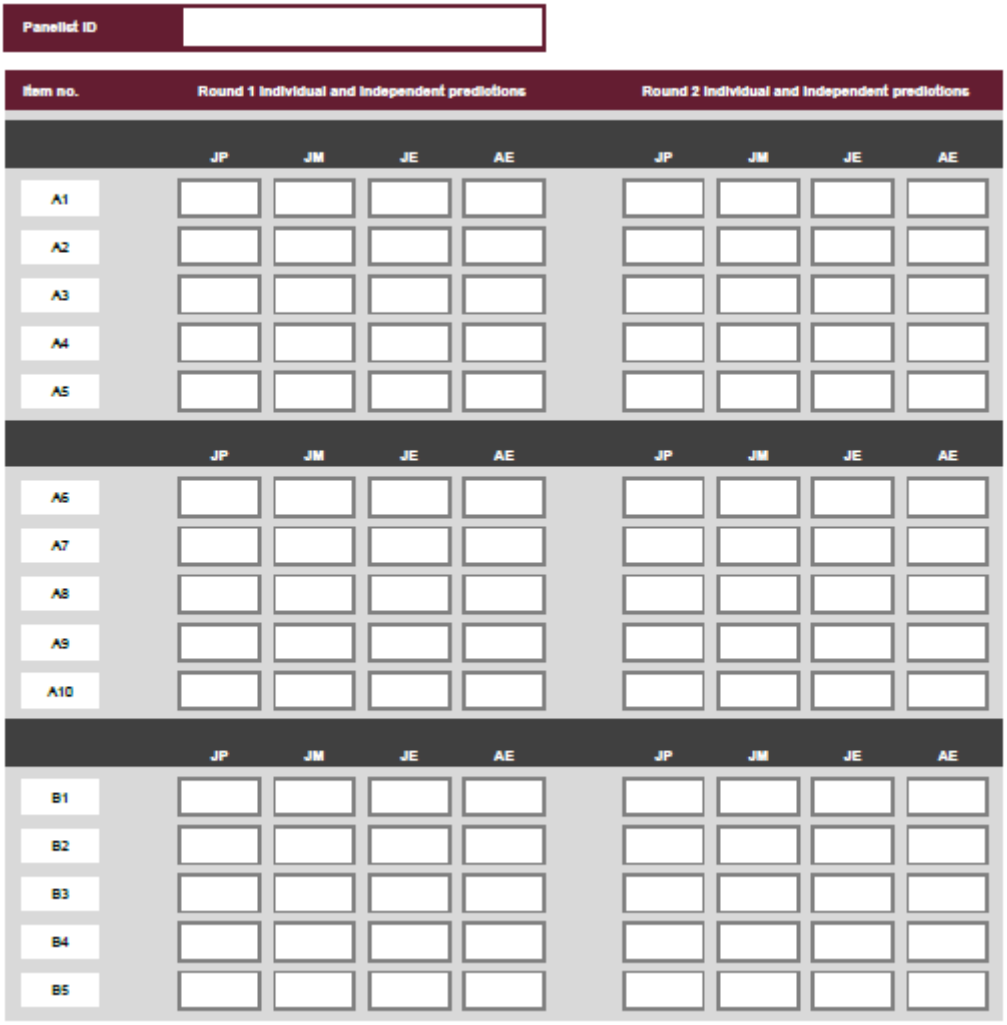
						In case of partial fit (record other domains, constructs and subconstructs that relate to the item)				
Question	Domain	Construct reference	Subconstruct reference	Knowledge or skill	Fit	Domain	Construct reference	Subconstruct reference	Knowledge or skill	Fit
A1										
A2										
A3										
A4										
A5										
A6										
A7										
A8										
A9										
A10										
B1										
B2										
B3										
B4										
B5										

Figure 10. Matching form for the local content facilitator (English)

Panelist ID	

Question	Domain	Construct reference	Subconstruct reference	Knowledge or skill	Fit	Lowest GPL	Difficulty	Consensus
A1								
A2								
A3								
A4								
A5								
A6								
A7								
A8								
A9								
A10								
B1								
B2								
B3								
B4								
B5								

Figure 11. Item rating form (English) for paper-based rating



Panelist ID	

Item no.	Round 1 individual and independent predictions				Round 2 individual and independent predictions			
	JP	JM	JE	AE	JP	JM	JE	AE
A1								
A2								
A3								
A4								
A5								
	JP	JM	JE	AE	JP	JM	JE	AE
A6								
A7								
A8								
A9								
A10								
	JP	JM	JE	AE	JP	JM	JE	AE
B1								
B2								
B3								
B4								
B5								

Figure 12. Data entry file for Alignment rating results (English)

	Panelist 1				Panelist 2				Panelist 3			
	Knowledge or skill	Fit	Knowledge or skill	Fit	Knowledge or skill	Fit	Knowledge or skill	Fit	Knowledge or skill	Fit	Knowledge or skill	Fit
A1												
A2												
A3												
A4												
A5												
A6												
A7												
A8												
A9												
A10												
B1												
B2												
B3												
B4												
B5												

Figure 13. Data entry file for Item rating results

Panelist nr	1	1	2	2	3	3	4	4
PID								
Round	1	2	1	2	1	2	1	2
	0		0		0		0	
Question	Round1	Round2	Round1	Round2	Round1	Round2	Round1	Round2
A1								
A2								
A3								
A4								
A5								
A6								
A7								
A8								
A9								
A10								
B1								
B2								
B3								
B4								
B5								

Figure 14. Data entry file for the Evaluation form

	TRAINING ON THE GLOBAL PROFICIENCY FRAMEWORK							
Response Number 1. PIN	2a. I understand the purpose of the GPF	2b. I understand the relationship between domains, constructs, subconstructs, knowledge and skills, and GPDs	2c. The GPDs were clear and easy to understand	2d. The discussion of the GPDs helped me understand what is expected of learners in Mathematics/Language at the end of grade 8	2e. The practical exercise using the GPDs was useful to improve my understanding	2f. There was an equal opportunity for everyone to contribute their ideas and opinions	2g. There was an equal opportunity for everyone to ask questions	2h. The amount of time spent on the GPD training was sufficient
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								

Annex C: UIS Activity plan

WEEK-BY-WEEK TIMELINE FOR LESOTHO PL WORKSHOP
Country, UIS, and Cito Tasks

Number	Activity	Role/Responsibility	Workshop Format for which Step is Relevant	Task Complete?	Date Complete
Week of March 21 - 27					
1	Decide on which assessment, grade level, and language to focus	Country with support from UIS/Cito	Both		
2	Decide what format the workshop will take (all remote or hybrid with participants gathering in one or multiple places) and the timing of the workshop	Country with support from UIS/Cito	Both		
Week of March 28 - April 3					
3	Start cost estimation	Country with support from UIS	Both		
4	Draft Activity Plan for engagement	UIS	Both		
5	Draft Non-Disclosure Agreement (NDA)	UIS and Country	Both		
6	Tailor the GPF to the relevant grades/subjects so that it can be translated	UIS	Both		
Week of April 4 - April 10					
7	Identify local Content Facilitators	Country	Both		
8	Identify interpreters (if relevant)	Country	Both		
9	Identify logistician (if needed)	Country	Both		
10	Identify other potential costs for the workshop, including phone/internet cards, transportation, lodging, per diems, meals, water, and materials during the workshop (see budget template)	Country	Both		
11	Review draft Activity Plan and provide any feedback	Country	Both		
12	UIS and Cito complete Non-Disclosure Agreements (NDAs)	UIS and Cito	Both		
Week of April 11 - April 17					
13	Submit budget to UIS	Country	Both		
14	Send assessment instruments to UIS/Cito	Country	Both		
15	Send data to UIS/Cito	Country	Both		
16	Begin to translate GPF into local language, if necessary and back-translate to check quality	Country	Both		
17	Decide on remote conferencing service for workshop	All	Both		
18	Draft agenda	Cito	Both		
Week of April 18 - 24					
19	Provide feedback on draft agenda	Country	Both		
20	Provide Ministry logo for certificates and banner (the latter only for hybrid workshops) and determine who from the Ministry will sign	Country	Both		
21	Identify panelists (both teachers and content specialists), including collecting their contact information; ensure panel is representative	Country	Both		
22	Draft certificates and banner	UIS	Both		
23	Finalize agenda	Cito	Both		
24	Draft workshop slides, including example items, and rating forms to send to UIS and the Country for review	Cito	Both		
25	Analyze data to produce data distributions, item difficulty data, etc.	Cito	Both		
Week of April 25 - May 1					
26	Identify and secure physical space for workshop	Country	Hybrid		
27	Invite panelists	Country, UIS, or Cito - depending on country's preference	Both		
28	Identify and invite any workshop observers - from other donors, Ministries, etc.	Country with support from UIS/Cito	Both		
29	Provide feedback on certificate and banner	Country	Both		
30	Review workshop slides, including example items, and rating forms and send feedback to Cito	UIS and Country	Both		
Week of May 2 - 8					
31	Reserve hotel rooms for panelists, if needed	Country	Hybrid		
32	Finalize contracts with local Content Facilitators, interpreters, and logistician (the latter two, if applicable)	UIS and Country	Both		
33	Finalize MOU with country based on approved budget	UIS	Both		
34	Identify modality for fund tranfer/expense coverage between UIS/Country	UIS and Country	Both		
35	Finalize certificates and banners	UIS	Both		
36	Finalize item rating forms and slides based on UIS feedback	Cito	Both		
37	Make logistical arrangements for content facilitator training	Cito	Both		
Week of May 9 - 15					
38	Determine what food/refreshments will be provided to participants and procure	Country	Hybrid		
39	Arrange or procure materials, such as notebooks, pens, flipcharts, folders, name tags/tents	Country	Hybrid		
40	Finalize slides for content facilitator training	Cito	Both		
41	Finalize the agenda (with any last-minute changes), acronym list, glossary, assessment, GPF, revaluation forms, certificates, banners, daily attendance forms, and any other documents	Cito	Both		
Week of May 16 - 22					
42	Confirm panelist participation	Country	Both		
43	Translate slides, forms, and any other documents for panelists	Country	Both		
44	Assign panelist IDs	Cito	Both		
45	Meet with Content Facilitators	Cito	Both		
Week of May 23-29					
46	Prepare funds to disperse to participants for per diems, travel, etc.	Country	Hybrid		
47	Distribute panelist IDs	Country	Remote		
48	Print the agenda, acronym list, glossary, assessment, GPF, rating forms, evaluation forms, slides with notes fields, certificates, banners, daily attendance forms, and any other documents	Country	Both		
49	Distribute the agenda, acronym list, glossary, assessment, GPF, rating forms, evaluation forms, slides with notes fields, certificates, banners, and any other documents	Country	Remote		
50	Inspect venue to plan for workshop, locations of breakout rooms, and to test remote access (if applicable, e.g., if not a government facility)	Country	Hybrid		
51	Train Content Facilitators	Cito	Both		
52	Remote platform testing with panelists or venue to make sure are participants can access the platform and don't need technical support	All	Both		
Week of May 31-June 5: Workshop					

Annex D: Alignment of the NAEP items with the domains, constructs and subconstructs

Table 20. English: Number of items aligned to each grade 6 domain, construct and subconstructs

Domain	Items
D Decoding	0.0
R Reading comprehension	15.0
Total	15.0
Construct	**Items**
D1 Precision	0.0
D2 Fluency	0.0
R1 Retrieve information	11.5
R2 Interpret information	2.5
R3 Reflect on information	1.0
Total	15.0
Subconstruct	**Items**
D1.1 Identify symbol-sound/fingerspelling and/or symbol-morpheme correspondences	0.0
D1.2 Decode isolated words	0.0
D2.1 Say or sign a grade-level continuous text at pace and with accuracy	0.0
R1.1 Recognize the meaning of common grade-level words	0.8
R1.2 Retrieve explicit information in a grade-level text by direct- or close-word matching	10.0
R1.3 Retrieve explicit information in a grade-level text by synonymous matching	0.7
R2.1 Identify the meaning of unknown words and expressions in a grade-level text	0.8
R2.2 Make inferences in a grade-level text	1.4
R2.3 Identify the main and secondary ideas in a grade-level text	0.3
R3.1 Identify the purpose and audience of a text	0.4
R3.2 Evaluate a text with justification	0.4
R3.3 Evaluate the status of claims made in a text	0.3
Total	15.0

Table 21. Mathematics: Number of items aligned to each grade 6 domain, construct and subconstructs

Domain	Items
N Number and operations	12.1
M Measurement	5.7
G Geometry	7.6
S Statistics and probability	3.5
A Algebra	1.9
Total	30.8
Construct	**Items**
N1 Whole numbers	9.2
N2 Fractions	1.2
N3 Decimals	1.8
M1 Length, weight, capacity, volume, area, and perimeter	2.6
M2 Time	1.8
G1 Properties of shapes and figures	5.5
G2 Spatial visualizations	1.3
G3 Position and direction	0.8
S1 Data management	2.6
S2 Chance and probability	0.9
A1 Patterns	1.2
A3 Relations and functions	0.8
Total	30.8
Subconstruct	**Items**
N1.1 Identify and count in whole numbers, and identify their relative magnitude	1.8
N1.2 Represent whole numbers in equivalent ways	1.3
N1.3 Solve operations using whole numbers	5.5
N1.4 Solve real-world problems involving whole numbers	0.6
N2.1 Identify and represent fractions using objects, pictures, and symbols, and identify relative magnitude	0.0
N2.2 Solve operations using fractions	0.5
N2.3 Solve real-world problems involving fractions	0.7
N3.1 Identify and represent decimals using objects, pictures, and symbols, and identify relative magnitude	0.4
N3.2 Represent decimals in equivalent ways (including fractions and percentages)	0.6
N3.3 Solve operations using decimals	0.4
N3.4 Solve real-world problems involving decimals	0.3
M1.1 Use non-standard and standard units to measure, compare, and order	1.0
M1.2 Solve problems involving measurement	1.6
M2.1 Tell time	1.0
M2.2 Solve problems involving time	0.8
M3.1 Use different currency units to create amounts	1.3
G1.1 Recognize and describe shapes and figures	5.5
G2.1 Compose and decompose shapes and figures	1.3
G3.1 Describe the position and direction of objects in space	0.8

S1.1 Retrieve and interpret data presented in displays	2.6
S2.1 Describe the likelihood of events in different ways	0.9
A1.1 Recognize, describe, extend, and generate patterns	1.2
A3.1 Solve problems involving variation (ratio, proportion, and percentage)	0.6
A3.2 Demonstrate an understanding of equivalency	0.2
Total	30.8

Annex E. Difficulty Level of the Items

Table 22. P-value and Item-Total correlation of the NAEP English items

Question	N	P-value	P0-25	P26-50	P51-75	P76-100	Rit
QuestionA1	3136	0.60	0.26	0.61	0.75	0.88	0.31
QuestionA2	3136	0.49	0.16	0.41	0.68	0.90	0.40
QuestionA3	3136	0.54	0.11	0.52	0.76	0.95	0.46
QuestionA4	3136	0.38	0.08	0.29	0.55	0.84	0.40
QuestionA5	3136	0.22	0.03	0.13	0.30	0.72	0.39
QuestionA6	3136	0.42	0.06	0.33	0.62	0.93	0.47
QuestionA7	3136	0.34	0.05	0.21	0.50	0.90	0.47
QuestionA8	3136	0.16	0.01	0.05	0.22	0.65	0.41
QuestionA9	3136	0.58	0.21	0.50	0.83	0.99	0.46
QuestionA10	3136	0.60	0.19	0.53	0.85	0.99	0.48
QuestionB1	3136	0.30	0.08	0.23	0.43	0.70	0.32
QuestionB2	3136	0.44	0.24	0.43	0.51	0.71	0.15
QuestionB3	3136	0.43	0.05	0.32	0.67	0.90	0.48
QuestionB4	3136	0.60	0.21	0.53	0.85	0.98	0.46
QuestionB5	3136	0.55	0.12	0.48	0.81	0.94	0.49

Table 23. P-value and Item-Total correlation of the NAEP mathematics items

Question	N	P-value	P0-25	P26-50	P51-75	P76-100	Rit
Question1	3040	0.97	0.86	0.97	1.00	1.00	0.13
Question2	3040	0.75	0.37	0.72	0.93	1.00	0.26
Question3	3040	0.37	0.17	0.33	0.54	0.94	0.15
Question4	3040	0.55	0.21	0.51	0.75	0.94	0.24
Question5	3040	0.26	0.23	0.28	0.20	0.50	-0.17
Question6	3040	0.23	0.06	0.18	0.41	0.89	0.21
Question7	3040	0.16	0.03	0.11	0.32	0.94	0.22
Question8	3040	0.69	0.25	0.64	0.94	1.00	0.34
Question9	3040	0.25	0.09	0.19	0.44	0.89	0.19
Question10	3040	0.47	0.23	0.40	0.70	0.78	0.20
Question11	3040	0.43	0.19	0.39	0.58	0.94	0.18
Question12	3040	0.35	0.19	0.34	0.41	0.83	0.03
Question13	3040	0.59	0.23	0.52	0.86	1.00	0.30
Question14	3040	0.29	0.11	0.26	0.40	0.72	0.11
Question15	3040	0.41	0.11	0.34	0.70	0.89	0.29
Question16	3040	0.38	0.20	0.33	0.56	1.00	0.13
Question17	3040	0.68	0.26	0.64	0.88	0.89	0.26
Question18	3040	0.28	0.11	0.23	0.45	0.89	0.20
Question19	3040	0.10	0.07	0.09	0.11	0.72	0.02
Question20	3040	0.40	0.07	0.32	0.72	0.89	0.35
Question21	3040	0.51	0.23	0.45	0.76	0.94	0.25
Question22	3040	0.20	0.10	0.18	0.27	0.78	0.07
Question23	3040	0.42	0.27	0.42	0.45	0.78	-0.01
Question24	3040	0.59	0.22	0.53	0.87	1.00	0.32
Question25	3040	0.34	0.15	0.29	0.50	0.94	0.17
Question26	3040	0.13	0.11	0.13	0.11	0.44	-0.08
Question27	3040	0.73	0.42	0.70	0.89	1.00	0.24
Question28	3040	0.62	0.28	0.57	0.85	0.89	0.26
Question29	3040	0.20	0.14	0.17	0.30	0.22	0.06
Question30	3040	0.09	0.01	0.05	0.20	0.33	0.21
Question31	3040	0.54	0.18	0.49	0.78	0.94	0.28
Question32	3040	0.24	0.14	0.21	0.33	0.44	0.06
Question33	3040	0.74	0.33	0.70	0.93	1.00	0.30
Question34	3040	0.46	0.20	0.40	0.70	0.94	0.23
Question35	3040	0.23	0.12	0.17	0.40	0.89	0.21

Annex F. Questions and instructions in the Evaluation form of the workshop

EVALUATION OF THE WORKSHOP

We kindly ask you to share your opinion about the policy linking workshop. Please complete this short questionnaire inquiring about your experience. Your answers will be used to improve the workshop and the training. Your feedback will not be shared widely except as part of an aggregation (average) of all panelists ratings or reflect on your participation in the workshop. Your feedback will also not be attributed to you.

1. PIN

TRAINING ON THE GLOBAL PROFICIENCY FRAMEWORK

During the first and second day of the workshop, you have been trained on the Global Proficiency Descriptors (GPDs). Please read the following statements carefully and place a mark in that category indicating your level of agreement.

2. GPD training	**Strongly disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly agree**
I understand the purpose of the GPF					
I understand the relationship between domains, constructs, subconstructs, knowledge and skills, and GPDs					
The GPDs were clear and easy to understand					
The discussion of the GPDs helped me understand what is expected of learners in Mathematics/Language at the end of grade 6					
The practical exercise using the GPDs was useful to improve my understanding					
There was an equal opportunity for everyone to contribute their ideas and opinions					
There was an equal opportunity for everyone to ask questions					
The amount of time spent on the GPD training was sufficient					

3. Please describe in your own terms what the purpose of the GPF is and what the GPDs tell you.
4. Please list any questions or areas of confusion you have about the GPF.
5. Please list any tips/requests for facilitators that would make the training work better for you.

TRAINING ON THE NAEP

During the first and second day of the workshop, you have been trained on the assessment(s) that we will use for policy linking. Please read the following statements carefully and place a tick in each category to indicate the degree to which you agree with each statement.

6. Assessment training	**Strongly disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly agree**

I understand the purpose of the assessment					
I understand the constructs assessed in the assessment					
I understand how the assessment is administered					
I feel I have a good sense of how minimally proficient learners would perform on the assessment					
The amount of time spent on the assessment training was sufficient					

7. Please list any questions you have about the assessment(s).
8. Please list any tips/requests for facilitators that would make the training work better for you.

TRAINING ON ALIGNMENT METHODOLOGY

The second and third day, you have been trained on the alignment methodology. Please read the following statements carefully and place a tick in each category to indicate the degree to which you agree with each statement.

9. Alignment training	**Strongly disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly agree**
I understand the purpose of alignment					
I understand the alignment methodology					
I understand the difference between no fit, partial fit, and complete fit					
I feel confident with my alignment ratings					
The amount of time spent on the alignment training was sufficient					

10. Please list any questions or areas of confusion you have about the alignment methodology/process.
11. Please list any tips/requests for facilitators that would make the training work better for you.

TRAINING ON MATCHING METHODOLOGY

During the third and fourth day, you have been trained on the matching methodology. Please read the following statements carefully and place a tick in each category to indicate the degree to which you agree with each statement.

12. Alignment training	**Strongly disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly agree**
I understand the purpose of matching					
I understand the matching methodology					
I understand how the alignment activity links to the matching activity					
I agree with the group consensus on the GPLs and GPDs to which we aligned each item (expand below if not)					
The amount of time spent on the matching training was sufficient					

13. Please describe any group decisions on matching with which you don't agree and why.
14. Please list any questions or areas of confusion you have about the matching methodology/process.
15. Please list any tips/requests for facilitators that would make the training work better for you.

TRAINING ON THE BENCHMARK-SETTING (ANGOFF) METHODOLOGY

During the fourth and fifth day, you have been trained on the benchmark-setting methodology. Please read the following statements carefully and place a tick in each category to indicate the degree to which you agree with each statement.

16. Policy linking training	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I understand the process I need to follow to complete the benchmarking exercise					
I understand how the benchmarking methodology links to the steps on alignment and matching					
I understand the difficulty level of the assessment items					
The discussion of the procedure was sufficient to allow me to feel confident in the methodology					
I understand how my ratings will result in a final benchmark					
There was an equal opportunity for everyone to contribute their ideas and opinions					
There was an equal opportunity for everyone to ask questions					
The amount of time spent on the policy linking method training was sufficient					
I feel confident in my Round 1 ratings					
I was given sufficient time to complete the Round 1 performance predictions[5]					

17. Please describe the benchmarking methodology in your own terms.
18. Please list any questions or areas of confusion you have about the benchmarking methodology/process.
19. Please list any tips/requests for facilitators that would make the training work better for you.

BENCHMARK ROUND 2 EVALUATION

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

20. Round 2	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I understand the data on others' ratings					
I understand the item difficulty data and how it relates to this process					
I understand the impact data and how it relates to this process					
I am confident about the performance predictions I made during Round 2					

[5] Additional question on request of observers. This question is not included in the reported evaluation to keep evaluations comparable across countries.

My performance predictions were influenced by the information showing the ratings of other panelists					
My performance predictions were influenced by the item difficulty data showing the actual performance of learners on the assessment					
My performance predictions were influenced by the impact information showing the outcomes for the sample of learners					
I was given sufficient time to complete the Round 2 performance predictions					

21. Do you have any additional comments on Round 2?

OVERALL EVALUATION

22. How comfortable are you with your final performance predictions?
    a) Very uncomfortable
    b) Somewhat uncomfortable
    c) Neutral[6]
    d) Fairly comfortable
    e) Very comfortable
23. If you marked either of the uncomfortable options, please explain why.
24. Overall, how would you rate the success of the policy linking workshop?
    a) Totally Successful
    b) Successful
    c) Neutral[7]
    d) Unsuccessful
    e) Totally Unsuccessful
25. How would you rate the organization of the workshop?
    a) Totally Successful
    b) Successful
    c) Neutral[8]
    d) Unsuccessful
    e) Totally Unsuccessful
26. Please provide any comments you feel would be helpful to us in planning future policy linking workshops.

Thank you for your participation in the workshop.

[6] Added the Neutral on request of UIS project leader
[7] Added the Neutral on request of UIS project leader
[8] Added the Neutral on request of UIS project leader