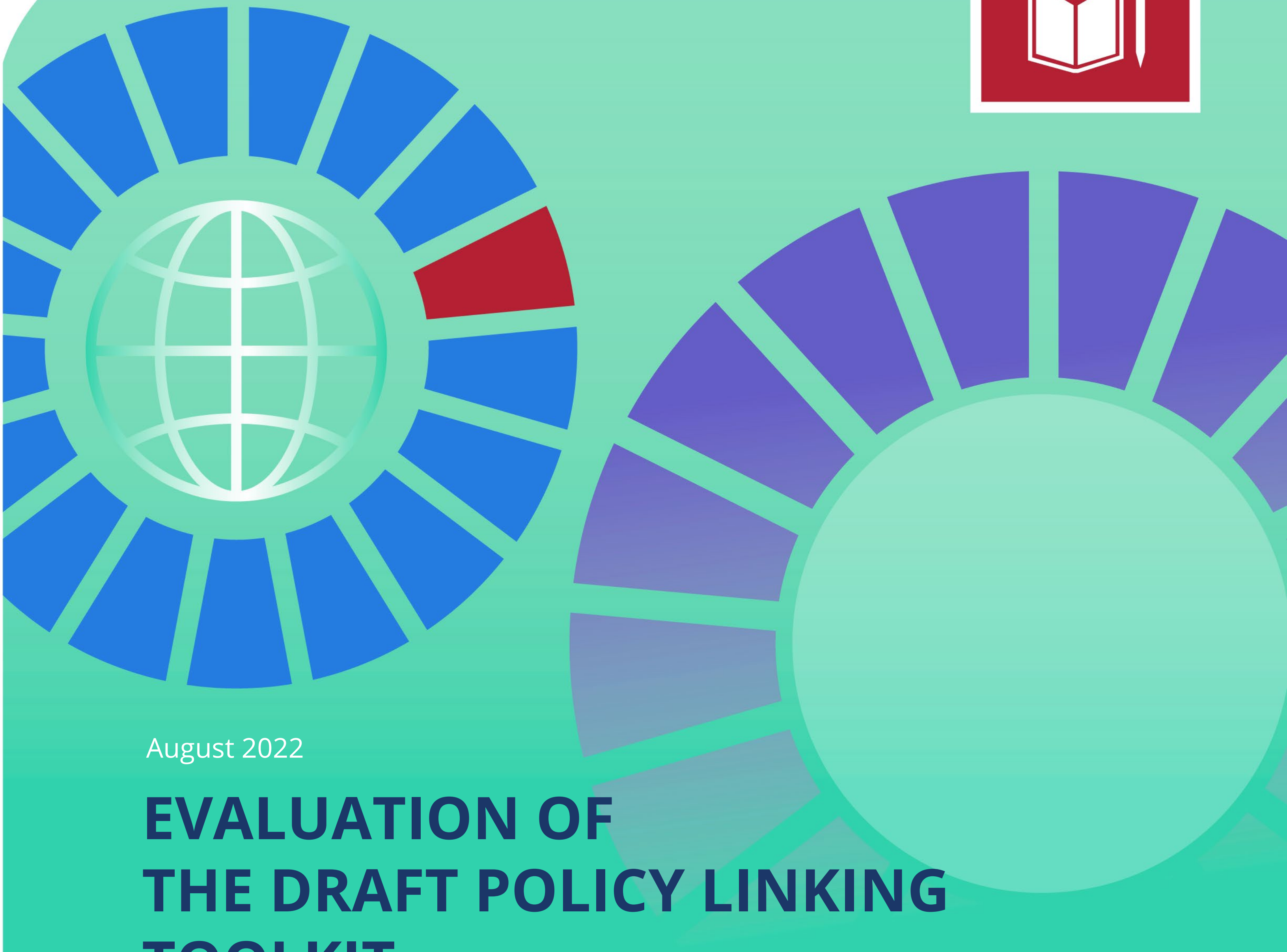
unesco
Institute for Statistics

4 QUALITY EDUCATION



August 2022

EVALUATION OF THE DRAFT POLICY LINKING TOOLKIT

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

The designations employed and the presentation of material throughout this publication do not imply the expression of any opinion whatsoever on the part of UNESCO concerning the legal status of any country, territory, city or area or of its authorities or concerning the delimitation of its frontiers or boundaries. The ideas and opinions expressed in this publication are those of the authors; they are not necessarily those of UNESCO and do not commit the Organization.

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 of people during the pilot of the policy linking workshop.

First, the management support provided by officials from the UNESCO Institute for Statistics (UIS) was instrumental in planning and implementing the workshops. Special thanks to Shailendra Sigdel, Melissa Chiappetta and Georges Boade.

Secondly, the support provided by officials at the National Council of Educational Research and Training (NCERT) in India, the Examinations Council of Lesotho (ECoL), the Education Quality Assurance Department of the Ministry of Education, Youth and Sports of Cambodia (EQAD), the Education Review Office of Nepal (ERO) and the Examinations Council of Zambia (ECZ) was critical for the success of the piloting workshops.

Thirdly, the hands-on support of all the local content facilitators was most important for realizing the goal of these workshops.

Finally, the dedication and effort devoted by the panelists in India, Lesotho, Cambodia, Nepal and Zambia was indispensable in establishing the pilot global benchmarks and drawing lessons learned from the workshops.

Sanneke Schouwstra

Cor Sluijter

Sjoerd Crans

Michel Hop

Margreet van Aken

Anneke de Graaf

Angela Verschoor

Melchior de Vries

Table of Contents

Acronyms and Abbreviations

USAID	U.S. Agency for International Development
ECoL	Examinations Council of Lesotho
EQAD	Education Quality Assurance Department of the Ministry of Education, Youth and Sports of Cambodia
ERO	Education Review Office of Nepal
ECZ	Examinations Council of Zambia
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
NAEP	The Lesotho National Assessment of Educational Progress Survey NCERTNational Council of Educational Research and Training in India,
PLT	Policy Linking Toolkit
SDG	Sustainable Development Goal
SEM	Standard Error of Measurement
UIS	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.

Contents of the draft policy linking toolkit

Draft Policy Linking Toolkit (version dd. Oct. 2020; version dd. Dec. 2020)

Content Facilitator Slides

Timed Assessment Slides

Untimed Assessment Slides

Global Proficiency Framework

Mathematics – pdf, Excel (version dd. Dec. 2020; version dd. Apr. 2021)

Language– pdf, Excel (version dd. Dec. 2020; version dd. Apr. 2021)

1. Executive Summary

This document contains Cito's feedback on the Policy Linking methodology. In 2021 and 2022, Cito piloted the policy linking workshop in five countries: India, Lesotho, Cambodia, Nepal and Zambia. Officials from the UNESCO Institute for Statistics (UIS) supported the five countries in organizing the pilots. The support provided by officials at the National Council of Educational Research and Training (NCERT) in India, the Examinations Council of Lesotho (ECoL), the Education Quality Assurance Department of the Ministry of Education, Youth and Sports of Cambodia, the Education Review Office of Nepal (ERO) and the Examinations Council of Zambia (ECZ) was critical for the success of the piloting workshops.

Due to the pandemic, the workshops were conducted completely remotely or in a hybrid mode with panelists meeting in-person and the international facilitators joining virtually. All five countries used an assessment from a national survey for policy linking. After each policy linking workshop a report was written about the recommended benchmarks.

In this report Cito evaluates the policy linking methodology, the toolkit and underlying Global Proficiency Framework. In this report we report on the strengths, weaknesses, opportunities and challenges noted during the hybrid as well as the completely remote Policy Linking workshops.

In the report we describe our observations and recommendation in each phase of the Policy Linking Workshop: during preparation, implementation, analyses and reporting. An overview of the recommendations and further research are given in Section 7. We recommend that the continued development of the toolkit and the implementation of Policy Linking should focus on increasing the usability and standardization and conducting research. Three out of five countries wanted to have benchmarks on a national assessment that employed a survey design and IRT modelling. The PLT did not contain methods or procedures to apply in such a situation. We therefore recommend to complement the PLT with standardized methodology to be used in such instances.

2. Introduction

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).

In order 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 India, Bangladesh and Nigeria. Moreover, in 2020 a pilot was conducted in Kenya and Nigeria to set benchmarks for International Common Assessment of Numeracy (ICAN). 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 took place in several countries, using remote workshops rather than in-person workshops.

Cito piloted the policy linking toolkit in five countries: India, Lesotho, Cambodia, Nepal and Zambia. Officials from the UNESCO Institute for Statistics (UIS) supported the five countries in organizing the pilots. The support provided by officials at the National Council of Educational Research and Training (NCERT) in India, the Examinations Council of Lesotho (ECoL), the Education Quality Assurance Department of the Ministry of Education, Youth and Sports of Cambodia, the Education Review Office of Nepal (ERO) and the Examinations Council of Zambia (ECZ) was critical for the success of the piloting workshops. After each policy linking workshop a report was written about the recommended benchmarks. In these reports each workshop was also reviewed and recommendations for the implementation of the Policy Linking workshop were specified.

As final task, Cito evaluates in this report the policy linking methodology, the toolkit and underlying Global Proficiency Framework. In this report we report on the strengths, weaknesses, opportunities and challenges noted in each phase of the Policy Linking Workshop: during preparation (Section 3), implementation (Section 4), analyses (Section 5) and reporting (Section 6). An overview of the recommendations and further research are given in Section 7.

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

3. Preparation of the workshop

Objective of the workshops

The objective of the workshops was setting global benchmarks on the national assessments at the end of primary education in language and mathematics. The workshops had a piloting function and should increase the capabilities of the national teams to conduct similar workshops in the future. Because of the COVID-19 pandemic it was agreed that the workshops would be conducted remotely.

The five piloting countries were reluctant about conducting the workshop completely remotely, predominantly for technical reasons. The five countries expected that the stability of the internet connectivity for all panelists could be problematic. Therefore, they preferred to approach as closely as possible an in-person workshop. For this reason, Cito developed the hybrid mode in which the international facilitators and UIS participated virtually and, nationally, the team and panelists met face-to-face in one or several regional locations. In the end, only Cambodia conducted the workshop completely remotely, albeit in a shorter period than the remote workshop mentioned in the PLT.

- *Given the preference of the countries for a hybrid mode, we recommend extending the Policy Linking Toolkit with a description of the hybrid mode.*

Table 1. Overview of the workshops

	India	Lesotho	Cambodia	Nepal	Zambia
Organizer	The National Council for Educational Research and Training (NCERT)	The Examinations Council of Lesotho (ECoL)	The Education Quality Assurance Department of the Ministry of Education, Youth and Sports in Cambodia (EQAD)	Education Review Office of Nepal (ERO)	Examinations Council of Zambia (ECZ)
Assessment	National Achievement Survey 2017	National Assessment of Educational Progress Survey 2016	National Learning Assessment 2016	National Assessment of Student Achievement 2018	National Achievement Survey 2016
Grade	Grade 8	Grade 5	Grade 6	Grade 5	Grade 5
Language	Hindi	English	Khmer	Nepali	English
Number of benchmarks	3	3	3	3	3
Mode	Hybrid with several locations	Hybrid with one location	Remote	Hybrid with one location	Hybrid with one location
Language during workshop	English & Hindi	English & Sesotho	Khmer	Nepali	English
Platform	Teams	Zoom	Zoom	Zoom	Zoom
Date workshop	14 March 2021 until 19 March 2021	May 31, 2021 until June 5, 2021	July 5, 2021 until July 16, 2021	September 26, 2021 until October 1, 2021	May 9, 2021 until May 14, 2022
Agenda	6-day	6-day	11-day	6-day	6-day

First three policy linking stages

After the initial engagement, the country governments or assessment agencies were meant to collate evidence of curriculum and assessment validity and alignment (stage 2 of policy linking). “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. The 4.1.1. Review Panel were to review this collated evidence.

At the start of piloting, the 4.1.1. Review Panel was not yet in place. Prior to the workshops, Cito was never informed whether the assessments met reliability and validity standards required to proceed with policy linking for reporting global outcomes. For this reason, Cito made an initial assessment of whether the assessments met the standards required to proceed with policy linking. All countries provided us with a Technical Report or Survey report. These reports allowed us to evaluate broadly, the Assessment validity and appropriateness for the population (criterion 2) and the reliability of the assessment (criterion 3) as outlined in the draft criteria for policy linking validity (Annex U of the PLT). The first criterium (Alignment between the assessment, the assessment framework, and the curriculum) could only partially be evaluated. The assessment frameworks were usually described in the provided Technical Reports. However, the countries were not asked to provide information on the curriculum and the item development and review process were usually not described, so the alignment between assessment and curriculum could not be assessed. In one country, we did not obtain the assessment itself to evaluate the first criterium.

- *We recommend establishing the 4.1.1 Review Panel as described in the Policy Linking Toolkit to “determine whether the assessment(s) meets reliability and validity standards required for a country to proceed with policy linking for reporting global outcomes”. This will also ensure all relevant material is provided in a timely manner.*
- *To allow the 4.1.1 Review Panel to review the assessment, countries should also be asked to provide information on the curriculum and the item development process (including the item review process) to the Review Panel 4.1.1*

General preparation of the workshop

UIS supported the countries with the general preparations of the workshop. UIS supported the countries in realizing the legal, governmental and financial conditions for the workshop, which was also the most time-consuming part of the preparations. Setting a date for the workshop proved also quite difficult, in part due to the pandemic. In two of the workshops, it remained uncertain until the very last moment whether the workshop would take place, because not all formal conditions were met. After conducting the workshop in the first country, the UIS consultant developed the UIS Activity plan in which a week-by-week timeline for the Policy Linking Workshop is described (see Annex C). This activity planner shows all partners which actions they and the other partners must take week by week. Sharing this timeline in advance and strictly following this timeline helps realizing the workshops in a timely and organized manner. The timeline could be extended with an indication of the amount of time needed for each activity. The local facilitators should allocate sufficient time for the preparation of the workshop, especially in the two weeks before the workshop.

- *In the activity planner, one could also indicate the agreed upon choices of the participating country with respect to the assessment, the grade, the language assessed, the number of benchmarks and the mode of the workshop. Last minute chances in those agreed upon choices should be strongly discouraged, as well as changes in the agenda and workshop dates. We propose to add the Activity Planner to the Policy Linking Toolkit. A strict adherence to the activity planner and preventing last minute changes in workshop dates and mode, will ensure well-prepared workshops and high quality.*

Selection of team and panel

The selection of the team in all countries was carried out relatively smoothly. Appointing the national workshop coordinator and a national logistician proved to be slightly more difficult than appointing the local content facilitators (the counterparts of the international content facilitators).

Each national council or department organizing the PLT workshop received instructions and information regarding the selection of panelists from UIS (or consultant of UIS). In chapter 3, pp. 24-25 the requirements are described, and an example of the invitation letter is included as well (Annex I). The Policy Linking Toolkit also includes a form to collect the panelists demographic information. Cito noticed that, in most countries, the demographic information was not collected systematically in advance of the workshop. The countries selected panelists that were diverse in terms of geography, gender and experience.

The panelists did not always receive the key documents, such as the Global Proficiency Framework well ahead of the workshop in their own language. The countries often tried to reduce translation costs by using the original material in English.

- *Countries should be strongly encouraged to translate the key information into (one of) their first language(s), especially the Global Proficiency Framework (especially the Tables 3 and 5 and example materials).*
- *A digital form and digital file to obtain all relevant background information would be a useful addition to the toolkit. The Policy Linking Toolkit could be slightly adjusted to emphasize that a diverse panel is required rather than a representative panel.*

Collecting materials and pre-workshop analyses

After their initial engagement, all five countries immediately shared their published technical report and/or final report about their survey. However, the process of acquiring permission to share the assessment and data with the international team was sometimes more time consuming. Four out of five countries shared their assessment well ahead of the workshop (translated into English whenever necessary), so the content facilitators could prepare for the workshop and select adequate examples for practice and the presentations. Sharing the assessment was also necessary to evaluate the assessments validity for policy linking and to evaluate the adequacy of the items to be used in the workshop. As not all five countries shared their assessment prior (or not even during) the workshop with the international facilitators, , the necessity of sharing the assessment and purpose of sharing might be explained more extensively, in the policy linking toolkit or material for the organizers.

In all three language assessments that were shared before the workshop, some items referred to skills and knowledge that had no link to the Global Proficiency Framework (e.g., writing, vocabulary, punctuation and grammar). These items were excluded from the policy linking procedure.

Four out of five countries shared their raw data in advance of the workshop even though the codebooks were not always included. In all five countries, the assessment was sample based rather than a census. To allow for estimating the impact of the benchmark at population level, sampling weights are needed, however this data were seldomly included in the data file provided. Furthermore, whenever a complex survey design with IRT modelling is used the item parameters are needed prior to the workshop (three out of the five countries used IRT modelling). All countries have shared (part) of the item parameters. In their Final Reports or Technical Reports about their surveys some general information about the modelling could be found, such as the type of IRT model used and the type of ability estimate. For replication, more specific information is needed as well as access to the software and code used. In some of the piloting countries, the software license was expired. For this reason, they could not recalculate the item parameters.

The description of the pre-workshop analyses in the toolkit are based on a simple assessment design with one booklet and Classical Test Theory. As this situation did not apply in three out of five cases, we suggest extending the description of the pre-workshop analyses for analyses using IRT. Also checks are needed to evaluate the quality of the item calibration. The calibration needs to be of a high quality to be sure of the position of the benchmarks on the underlying ability scale and a selection of items that is sufficiently aligned to the GPF.

- *In the policy linking toolkit, the requirement of sharing the assessment and purpose of sharing might be explained more extensively. A clear instruction can be added to the Policy Linking Toolkit, describing when to exclude items that assess (language) skills not mentioned in the policy linking toolkit.*
- *In the policy linking toolkit, the materials to be collected are clearly described. We suggest adding the assessment design, sampling weights, item parameters and ability estimates (or plausible values) to the list of materials that need to be obtained (p. 27) in case of a complex survey design using IRT modelling. A description of the reason for providing these materials and format might be useful.*
- *A separate description must be developed describing the analyses to perform when a complex survey design is used and IRT modelling.*
- *The 4.1.1. Review Panel should also look at the quality of the IRT calibration to find out beforehand if the national assessment is suited for policy linking.*

Technical preparation

Because all five workshops were conducted remotely or in a hybrid mode, a technical test of the facilities of the workshop was planned. Preferably, this test takes place well ahead of the workshop. However, the venues in which the workshops took place, and the technical setup were not available ahead of the workshop. For this reason, the test took place shortly before the workshop and usually with a limited set of people. A technical test with all locations and participants will make clear in advance if back-up material or equipment is needed (e.g., the WhatsApp contact) and to troubleshoot any technology issues. The technical facilities required for a completely remote workshop or a hybrid workshop are quite different from each other. Currently, a list of the technical facilities needed is not included in the Policy Linking Toolkit. Also, fall back options should be planned ahead of the workshop in case internet and/or power problems are likely to occur.

In Teams (and Zoom) the options are limited when participating without a license. It is vital to work with a platform for which the organizing national organization has a license. In the completely remote workshop, it is imperative that the assistant handling the platform can easily appoint people to the break-out rooms. Switching between different break-out rooms should be practiced beforehand with the panelists. Also, panelists must be able to ascertain easily in which room they are. Adding different backgrounds for Khmer and Mathematics panelists was an excellent idea of the EQAD team. This made checking if all participants were in the correct session simple and efficient. Finally, the lead facilitator (and some of the international observers) should always be able to switch between rooms at will.

- *A list of the technical facilities needed for the hybrid and completely remote mode should be added to the Policy Linking Toolkit.*
- *The necessity of a technical test ahead of the workshop should be added to the Policy Linking Toolkit and one could consider developing a testing protocol (e.g., testing the audio and switching between break-out rooms). The technical test should include switching between break-out rooms with all people that have to switch rooms during the workshop.*

Content facilitator training

During the last week before the workshops, a training for the local content facilitators was held. The Policy Linking Toolkit also contains content facilitator training slides which describe for each day of the 5-day in-person workshop the objective of that day and the role of the content facilitator. After the first workshop, Cito concluded that the local content facilitators might benefit from a more intensive training or general rehearsal. After the first workshop, 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.

Cito invited the entire local teams to join the introduction (1) and the general rehearsal (3). The interactive sessions were intended for Cito's content facilitators and their local counter parts. It is important that the international content facilitator and their counterparts create 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. When the local content facilitators studied the Global Proficiency Framework thoroughly prior to the training, the training developed seemed to work well. We did notice during the workshops, that sometimes the local content facilitators had the tendency to participate in the discussion as a panelist rather than act as a facilitator only.

- *The Policy Linking Toolkit could be extended with a clear program for the content facilitator training.*
- *Currently, in Annex E the slides of the content facilitator training are missing.*
- *The slides of the content facilitator training in the Policy Linking Toolkit could be developed further showing how to train local assessment experts in benchmarking, the Global Proficiency Framework and Policy Linking. We suggest adding a slide about dealing with group dynamics and about common pitfalls during facilitation.*
- *The role of the international content facilitator on the one hand and the local content facilitator on the other hand should be explained, both in the (local) content facilitator training and in the Policy Linking Toolkit.*

Training for local data entry

In one country the panelists directly sent digital forms, but this yielded many problems due to different versions of hard- and software on each device. In four out of five countries, the panelists worked on paper, therefore data entry was needed. Cito developed special data entry files and a special 2-hour data entry training. Cito gave this data entry instruction on the second day of the workshop (in the 6-day version). On three days (day 3, 5 and 6) data entry had to be carried out. 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 data entry went very smoothly in all three countries.

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

- *We recommend including instructions in the Policy Linking Toolkit for data management. The Policy Linking Toolkit could extend the Team description with data entry personnel, include data entry files and add the data entry training. A preferred option is to develop a digital tool to process and manage all the data of the workshops.*

Materials for the workshop

Manual & appendices

The manual and appendices contain all relevant information and all the steps in the process are clearly described. It is an overall manual describing everything from selecting the panelists, to implementing the workshop and the analyses. It should be clear which information is meant for the local team and which for the international facilitators. Some of the information is needed for the local team, e.g., the instruction for selecting and inviting the workshop panelists and the sample invitation letter. Part of the toolkit are documents that must be adapted and sent to the panelist (see Table 2). The documents have to be adapted to the choices of the country (agenda, workshop mode, the assessment and grade in which the assessment has been administered).

Table 2. Contents of the panelist package

Content	Adaptation required	Format of material
Agenda	Yes	pdf
Glossary and acronym list	No	pdf
GPF Mathematics of applicable grades	Yes	pdf
GPF Language of applicable grades	Yes	pdf
Hand-out workshop slides for language	Yes	Pdf
Hand-out workshop slides for mathematics	Yes	Pdf
Alignment form language	Yes	Pdf and Excel
Alignment form mathematics	Yes	Pdf and Excel
Item rating form	Yes	Pdf and Excel
Evaluation form	Yes	Pdf and Microsoft forms

- *We suggest creating separate manuals for different people (both international and national), containing only the information relevant to them (e.g., data analysis, content facilitation, logistics, selection of panelists).*
- *We suggest creating a separate list of materials to provide to the content facilitators and to the data analysts*

Agenda

The Policy linking Toolkit contains two agendas: one for the in-person workshop and one for a 3-week completely remote workshop. The agenda should be adapted to the needs of the country in terms of starting and closing time, breaks etc. However, as we were asked to conduct the workshop in a hybrid mode, a new agenda had to be developed. To limit the number of travelling days and the stay in a hotel, instead of a 3-week workshop, a six-day workshop was developed (see Appendix A). Also, for the completely remote workshop, we were asked to develop a more intensive workshop in which the number of days were limited as much as possible. We observed some differences between the two agendas that are currently part of the Policy Linking Toolkit. The reference to the presentations and activities is not identical and the time allotted for each presentation and activity is different. Some presentations are also missing in the completely remote agenda, for example “Task 1 Presentation: Alignment results”.

The schedule in the six-day blended workshop is very tight. In a six-day workshop, there is very little room for adapting to unforeseen circumstances or solving technical problems, such as frequently occurred during the first day. Furthermore, the time allotted for data entry and data analyses is extremely short. Performing data entry and analyses in such a short time requires a completely standardized procedure in which also the code for data analyses is prepared beforehand. Another disadvantage of the developed agenda is that it does not allow for the time differences between the international team and national team.

- *We suggest adding the agenda for the hybrid mode and the 11-day completely remote agenda to the policy linking toolkit. The reference to the different presentations and activities should be completely standardized across the different agendas and the time dedicated to each activity should be standardized in the different agendas.*

GPF

The Global Proficiency Framework (GPF or Framework) defines, for both reading and mathematics, the minimum proficiency levels that learners are expected to attain at the end of each of grades one through nine. A few inconsistencies were removed in the latest update. The GPF was developed based on review of national assessment and curriculum frameworks. For reading, the description of domains and skills to be addressed in each grade seems to be built strongly upon the curricula of Indo-European languages and curricula.

For the workshop, the GPF documents had to be adapted to the grade level for the assessment under consideration and one grade level below and one above the grade level of the assessment. Also, after this adaptation, the GPF is a lengthy document and thus expensive to translate. In three countries, the GPF was not translated completely, because the panelists mastered English at a high level. However, we feel that it would be much easier for the panelists to familiarize themselves with the Global Proficiency Framework and to execute the tasks, if they receive key documentation in the form of a hand-out translated into their own language, especially the Global Proficiency Framework.

In two countries, the GPF was used to benchmark an English assessment rather than an assessment of the 1st language. The learning of a second language is very different from learning a first language and needs (a) different framework(s).

- *It would be interesting to study how well the Global Proficiency Framework fits the curricula of different language families.*
- *In the Global Proficiency Framework, it should be emphasized that the GPF Reading is designed for the 1st language, not a 2nd language. Translation of the GPF in (one of) the 1st language(s) of a country should be a requirement.*
- *A version of the GPF should be made that allows easy adaptation to grade (at least a Word-document rather than a pdf).*

Slides

The slides seem to serve several purposes: as a presentation for the panelists, as a presentation for local content facilitators and as a manual for facilitators. The PowerPoint contains all the slides for the entire workshop and therefore the PowerPoint is very lengthy (171 slides for an untimed assessment).

For each workshop, the slides need to be adapted to:

- The assessment
- The grade
- The organizing team (names of facilitators and logo on the first slide every day)
- The breaks
- The agenda
- The examples

Adapting the slides to the 6-day and 11-day agenda required adding and changing the slides with the daily agenda, changing the position of the breaks and adding and changing the slides with the daily objectives and review of the previous day. Less adaptation would be needed if the slides referring to a particular day and moment of the day (the breaks) were removed. Removing those slides from the presentation, would make the PLT also more flexible to last minute changes in the mode and/or agenda of the workshop.

To make the slides more practical for use during the workshop, Cito added sections. Each presentation or activity was a separate section. The name of the section also showed whether and in which break-out room (language or mathematics) the slides were presented.

The slides contain a lot of text, for which reason translation of the slides is important. However, this has not been done. In general, the text should be reduced ("Less is more when it is about the text on slides."). The slides also contain many acronyms and technical words, with which the panelists seemed to experience some difficulties. The wording could be simpler. The content of the slides would be more engaging if more illustrations, and videos or animations were used instead of text.

Apart from adapting to the new agenda, adapting to the grade involved quite some work. Throughout the slides, several grade-specific examples need to be used. Cito's content facilitators had to create these grade-specific examples (for grade 8 and 6). In the slides more information could be provided about the level of the text and references to the examples, now the focus (specifically in the slides) is on the items.

- *We recommend that for each grade appropriate examples are developed and sample items to practice. These examples and sample items could be placed in a small database.*
- *For language, provide more information about the level of the text and references to the examples, now the focus (specifically in the slides) is on the items.*
- *The slides could be improved and easier to adapt by:*
    - *Adding sections*
    - *Removing agenda, breaks and daily review and objectives*
    - *Reducing text*
    - *Avoiding technical terms and acronyms*
    - *Adding visuals*

Forms

Because the policy linking toolkit does not contain digital forms for remote workshops yet, Cito developed digital alignment rating forms, item rating form and workshop evaluation form (see Appendix B) based on the examples in the toolkit (Annex D and F). The digital forms were designed to ease the task of the panelists, to prevent inconsistent ratings and to speed-up the data analyses during the workshop. The digital forms could also be printed and used during the workshop for paper-based rating. For matching a form was developed for the local content facilitators, to keep track of the matching and the consensus.

- *We recommend adding digital forms to the Policy Linking Toolkit that are easy for adaptation, printing (paper-based rating), data management and processing.*

4. Implementing the Policy Linking Workshop

Homework

In all five countries there were security concerns related to releasing the assessment, because the assessments contained anchoring items that are to be used in future surveys. This prevented the panelists from familiarizing themselves with the assessment and similarly from administering the assessment to 9 leaners prior to the workshop, as described in the Policy Linking Toolkit. According to the Policy Liking Toolkit each panelist should select “three leaners who just barely meet the requirements of the GPF’s Partially Meets Global Minimum Proficiency level for the grade level of the assessment, three who just barely meet the requirements of the Meets Global Minimum Proficiency level, and three who just barely meet the requirements of the Exceeds Global Minimum Proficiency level”. Obviously, selecting these learners requires the panelists to already be very familiar with the Global Proficiency Framework and to be able to assess which learners will meet the requirements. Furthermore, the requirement to administer the assessment to nine learners as described in the Policy Linking Toolkit might interfere with the panelist getting the correct cognitive representation of JP-, JM- and JE-learners. This requirement might lead panelists to focus on weak, mediocre and strong learners in their own population of learners.

Instead of administering the assessment prior to the workshop, the panelists made the assessment themselves during the workshop. As for studying the GPF prior to the workshop, an excellent description and translation is needed. Looking at the GPF without instruction and explanation is extremely difficult. One might consider sending some exercises, and having the familiarization day ahead of the actual workshop.

- *The requirement to administer the assessment to nine learners as described in the Policy Linking Toolkit might be reconsidered given the frequent security concerns about the assessment and given that a full understanding of the GPF ahead of the workshop is unlikely.*
- *The panelists should receive key information well ahead of the workshop, so they can familiarize themselves with the contents. Countries should be strongly encouraged to translate the key information into their own first language, especially the Global Proficiency Framework.*
- *We suggest developing exercises for the familiarization with the GPF ahead of the workshop.*

Familiarization

Following feedback from other policy linking workshops a year earlier, 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. 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. The regional adviser of UIS presented the panelists with an overview of Policy Linking and the Global Proficiency Framework. 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.

The first day closed with an introduction to the national assessment. The national team presented the national assessment, and the content facilitators discussed the first five items of the national assessment 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 take the national assessment themselves. While answering the items of the national assessment, the panelists were asked to make a note of stumbling blocks and aspects of the items that might make the item easy or difficult for the grade specific students.

The familiarization phase is new in the policy linking toolkit. We feel the familiarization is an important addition. However, the vast quantity of information is quite overwhelming for the panelists. In general, the panelists need more time to get acquainted with the Global Proficiency Framework and to get a good understanding of the framework, specifically the GPDs and GPLs. We feel it is important to note that the facilitators are best suited to give the presentations.

- *The familiarization with the Global Proficiency Framework should be more pedagogically informed, with suitable involvement of the panelists. The familiarization should be focused on practicing rather than on listening to presentations. This to enhance engagement of the panelists and to avoid them feeling overwhelmed. Let people with expertise in training teachers develop additional activities for the panelists.*
- *We suggest partitioning the information in smaller, logical pieces in a logical order. 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.*
- *The local content facilitators (or someone else from the agency or governmental organization that has created the assessment) are best suited to give a presentation about the assessment. The lead facilitator should be an experienced trainer with in-depth knowledge of policy linking and of the GPF and suited to familiarizing the panelists with policy linking and the Global Proficiency Framework.*

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 were asked to carry out three tasks during the workshop:

- Task 1 — Rate the alignment between the NASA and the GPF
- Task 2 — Match the NASA items to the appropriate Global Proficiency Level and Global Proficiency Descriptor.
- Task 3 — Set three global benchmarks for the NASA

The panelists received an introduction to their first task: aligning the national assessment 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) as outlined in the GPF to successfully perform policy linking. 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 NASA items and GPF knowledge and/or skill(s) statement(s) and in the second step the data analyst compiles and summarizes 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 assessment 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 in print (see Annex B). After they completed the alignment rating, they had to hand in their rating form. An employee of the national team entered all ratings in an Excel sheet developed for this purpose and sent the completed file to Cito.

After the national team sent the completed Excel file with the alignment ratings, Cito's data analyst completed the second step. 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. The data analyst took the average of the number of items that the panelists aligned to each subconstruct, construct and domain of the appropriate grade. Each item was counted only once (even if it was a partial fit), non-fitting items were not counted towards alignment. A second analysis was done as well, in which the average was taken of the number of items that the panelists aligned to each subconstruct, construct and domain for the appropriate grade and lower.

At the start of the alignment, the panelists' understanding of the Alignment task was often hindered by the overabundance of information given during the familiarization, and particularly information that they would not need until at a later stage. As suggested previously, it would be better to give panelists only the information at the moment when they need it. The panelists should focus on knowledge or skill statements, not whether it is the appropriate grade.

The examples, of the alignment scale to rate the level of alignment of the item, also caused some confusion. The "No fit" example in the slides is rather confusing, because a reference is made to the grade. According to the toolkit, "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." However, the slide states: "the item can be rated as "no fit" since it requires knowledge or skill that is not expected at (or before) the grade level."

In the analysis, non-fitting items were not counted towards alignment as indicated in the Policy Linking Toolkit. However, it is not clear how to deal with these non-fitting items in the subsequent tasks. It also remains unclear what to do in the subsequent tasks if an assessment is not aligned to the Global Proficiency Framework in depth and/or breadth. Furthermore, presenting the results to the panelists when the assessment is not aligned can be extremely demotivating. The panelists do not need to know the level of alignment in order to perform the subsequent task. One might consider presenting the alignment results only at the end of the workshop.

- *We suggest removing this "No fit" example that refers to grade. Use a "No fit" example, showing an item that requires knowledge or skills not described in the GPF.*
- *We also suggest to include examples for the level of alignment for languages in the plenary presentation on alignment.*
- *The Policy Linking Toolkit should include clear instructions on how to handle non-fitting items in the subsequent tasks and how to handle an assessment that does not align to the GPF.*
- *One could consider presenting the results of the alignment only at the end of the workshop.*

Matching

After the panelists completed task 1, they received instructions for the next task: Matching the assessment 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 zoom in on what is expected of learners as measured by each item in the assessment. The panelists were asked to identify the descriptors (GPDs) of global minimum proficiency that match with the items.

A Global Proficiency Descriptor is a detailed definition drawn up by subject matter experts which clarifies how much of the content described in 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). The Global Proficiency Descriptors are organized by domain, construct and subconstruct, with descriptors for each subconstruct.

The matching is a group activity (in contrast to alignment and benchmarking). The panelists are asked to work in groups to reach consensus on three aspects:

- What knowledge and/or skill(s) are required to answer the items correctly?
- What makes the item easy or difficult?
- What is the lowest GPL that is most appropriate for the item?

The various groups of panelists went about reaching consensus in two different ways. Some groups worked 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. In other groups, the whole group proceeded together item by item, the discussion being facilitated by the content facilitators.

We also noticed that sometimes the matching outcome was not in line with the alignment outcome. For example, while in the alignment the majority might consider an item to (completely or partially) fit the GPF, in the matching procedure the outcome might be that that item does not fit. Similarly, the consensus might be reached that answering an item requires, for example, primarily the skills to “Solve equations and inequalities” whereas during the alignment most panelists indicated that the item required the skills to “Describe the position and direction of objects in space”.

The matching task often took up more time than scheduled. Also, language and mathematics could be out of sync when both subjects had a different number of items. Reaching consensus is time-consuming; the more items the more time is needed. Sometimes the panelists did not manage to reach complete consensus on all items. We also noticed that the consensus requirement and the limited time put pressure on the local content facilitators. Sometimes the local content facilitator felt pressured to explain the “correct” understanding – correct according to the content facilitator that is – instead of facilitating the panelist’s thinking and allowing the panelists to reach their own conclusion. Within a six-day workshop, we estimate that about 30 to 35 items can be aligned, matched and rated.

- *We suggest evaluating the necessity to reach consensus for the quality of the benchmarks. If consensus is a requirement, more time should be scheduled for this task.*
- *We suggest evaluating the different ways the matching activity was carried out – first find consensus in subgroups or work immediately with the whole group – and choose one of them (or a different one) for future workshops.*
- *We recommend evaluating the consistency between the results of the three tasks and to describe how to resolve inconsistencies.*
- *When conducting a workshop for more subjects and/or grades, it would be helpful if the assessments for the different groups were of similar length.*

Benchmarking

After the matching task, 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. 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.

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 reasonably expected to be made.

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 1): 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)?

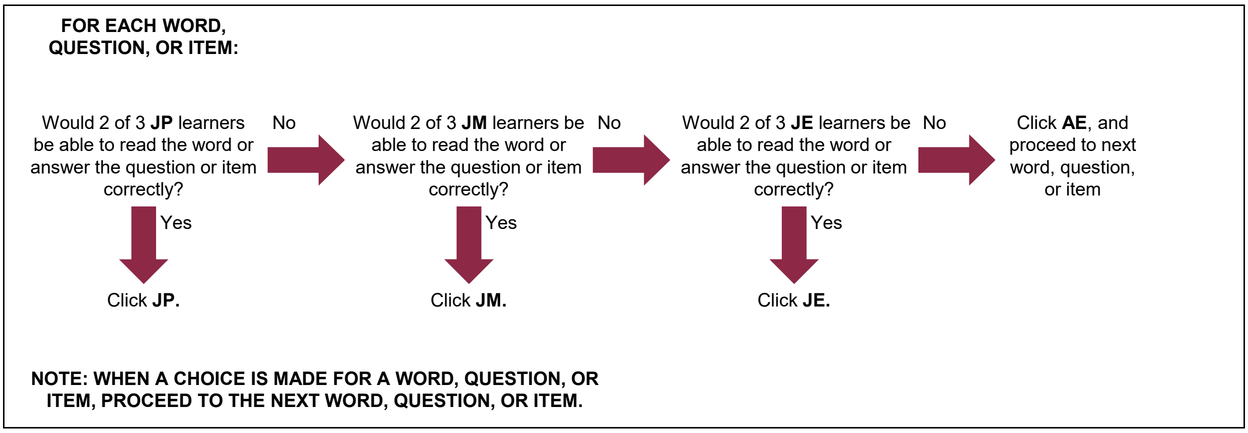
- If "yes," place an "X" under JP and proceed to the next item.
- If "no," ask whether minimally proficient JM learners would be able to answer the item correctly?
  - If "yes," place an "X" under JM and proceed to the next item.
  - If "no," ask whether minimally proficient JE learners would be able 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. 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 consideration the difficulty of the item. To obtain realistic ratings, the panelists was instructed to consider what a learner *would* answer at the respective GPL, rather than what a learner *should* answer.

Figure 1. Steps for Rating Items



FOR EACH WORD, QUESTION, OR ITEM:

Would 2 of 3 **JP** learners be able to read the word or answer the question or item correctly?

Click **JP.**

Would 2 of 3 **JM** learners be able to read the word or answer the question or item correctly?

Click **JM.**

Would 2 of 3 **JE** learners be able to read the word or answer the question or item correctly?

Click **JE.**

Click **AE**, and proceed to next word, question, or item

NOTE: WHEN A CHOICE IS MADE FOR A WORD, QUESTION, OR ITEM, PROCEED TO THE NEXT WORD, QUESTION, OR ITEM.

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. These members of staff 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. The report contained the following :

- 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. During the discussion the content facilitators focused 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 in 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.

In the general presentation for Benchmarking, and in earlier presentations, the language panelists expressed their concern about the lack of examples with language items. They did not feel that the information from the mathematics examples was fully relevant, and neither were the practice items (too abstract).

Working remotely, it proved to be difficult to ascertain whether the ratings were made individually and independently. Working on location and the presence of experts might have had an influence on the alignment ratings and item ratings. It is crucial that the experts only observe and do not join the discussions.

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. During the discussion, it became apparent that the panelists had the tendency to visualize their own students instead of the learners as described in the GPF. The facilitators helped the panelists to refocus on the Global Proficiency Framework.

In one country, the assessment contained polytomous items. Even though the Angoff method can be extended for polytomous items, the forms, data entry and analyses need to be adapted to polytomous items. Also, a clear description must be added to the Policy Linking Toolkit of the instruction to give to panelists.

In one country, several items were considered non-fitting during the matching. However, the panelists rated the items anyway during the benchmarking. The Policy Linking Toolkit does not specify how to approach items that are considered non-fitting. We should point out that a last-minute change in the forms is not possible. In the agenda, not enough time is foreseen to redo the preparatory analyses (new frequency distribution without the non-fitting items).

Finally, in one group the panelists asked for more explanation of the impact information between the rounds. We should be careful with impact information between the rounds and explaining how it is calculated, because panelists could steer the ratings in the desired direction in round 2 (e.g., more students that meet the minimum proficiency).

- *We suggest also providing examples for languages in all plenary presentations.*
- *Give a clear instruction in the Policy Linking Toolkit on the role and expected behavior of participants other than the facilitator and panelists.*
- *Provide a clear description in the Policy Linking Toolkit or GPF of the Just Partially Meets (JP), Just Meets (JM), and Just Exceeds (JE) learners and how this relates to their own leaners and the GPF. Some exercises could be developed to help panelists to keep focusing on the Global Proficiency Framework while rating the items.*
- *Extend the Policy Linking Toolkit with a description of how to deal with polytomous items in the presentation, activity, forms and analyses.*
- *Add to the Policy Linking Toolkit a clear description of how to handle non-fitting items during the benchmarking task and in the analyses.*
- *Consider not giving impact information after round 1 to prevent manipulated outcomes.*

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 Student Achievement
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.

The evaluation went well. A minor correction was implemented in the evaluation form. The response scale of the last three questions (about the overall evaluation) was changed from a 4-point scale to a 5-point scale, like the response scales of all the other questions. Rather than ask panelist to evaluate daily (as in the completely remote workshop), we asked the panelists to evaluate the workshop at the end (as in the in-person workshop). Without a daily evaluation, the program was already quite full and overwhelming for the panelists, and several times lack of time was experienced.

- *We suggest evaluating the workshop only once, at the end of the workshop for all modes.*

5. Data management and analyses

During the workshop analyses are needed of the ratings of the panelists at three specific moments: after the alignment task, after the first round of the benchmarking and after the second round of benchmarking. Cito developed special data entry files for data management, as these are not included in the Policy Linking Toolkit.

According to the Policy Linking Toolkit, the facilitators should summarize the alignment results by taking an average of the number of items that the panelists aligned to each domain, construct, and subconstruct. We let a data analyst create these summaries.

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. The data analyst took the average of the number of items (rounded to the nearest whole number) that the panelists aligned to each subconstruct, construct and domain for the appropriate grade. Each item was counted only once (even if it was a partial fit), non-fitting items were not counted towards alignment. We evaluated also other criteria for alignment. Cito's data analyst performed a second analysis considering the subconstructs with knowledge and/or skill(s) expected at the grade level and lower grades for which alignment was being conducted. In one case, this distinction made a difference in the outcome.

For the analysis of the alignment, it is important that the panelists only consider which knowledge and skills are required to answer each item correctly, not whether the grade was appropriate. If the panelists would have also considered grade, we would not have been able to evaluate other alignment criteria (such as the appropriate grade and lower grades).

For the analysis of task 3 (the benchmarking), the Policy Linking Toolkit contains several appendices in which these analyses are clearly described. Prior to the workshop, using the assessment data, the item difficulty, the conditional item difficulty and the sum score distribution must be calculated (Annex L – Pre-workshop statistics). During the workshop the panelists receive feedback, for which their ratings are plotted, and impact data is shown (described in Annex M). Furthermore, the benchmarks are calculated (Annex R) and the intra-rater and inter-rater consistency indices and the SE (Annex G).

As discussed previously, only in two countries the assessment was a linear test that was the same for all learners participating and the data were analyzed with Classical Test Theory. In the other countries, the assessments consisted of several booklets with a certain overlap of items administered to different sets of learners. In most cases, this made it an impossibility to have panelists align and match all items to the GPF, because of the large numbers of items. We were faced with four different scenarios:

(1) One entire booklet was selected for the benchmarking (and one group of students made this booklet).
(2) Some items from one booklet were selected for the benchmarking and one group of students made all these selected items.
(3) Items from several booklet were selected for the benchmarking and none of the students made all these selected items
(4) All items administered were included in the benchmarking procedure, but none of the students made all items because the students made only one booklet containing only part of the items.

Using IRT you can deal with all four situations, even though in some of these scenarios more steps in the analysis must be taken. We should point out that the Policy Linking Toolkit does not anticipate on using IRT and assessment designs with several booklets. We should also emphasize that several IRT models exist in which the sum score is not a sufficient statistic. Basically, in those models it matters which items learners answer correctly rather than how many. For the benchmarking procedure and the analyses, this also has consequences. For the benchmarking it means that it matters *which* items two out of three Just Meets minimum proficiency learners can answer correctly according to the panelist and not how *many* items two out of three Just Meets minimum proficiency learners can answer correctly according to the panelist.

In the results, we encountered two situations on which the Policy Linking Toolkit does not provide guidance. Several times we found ceiling effects (or a bottom effect) in the benchmark. Basically, this means that almost all panelists place the Exceeds Minimum Proficiency cut score at the maximum score. In case of a ceiling effect (or bottom effect), the benchmark is not valid. In the validation of the benchmarks clear rules should be provided to determine whether ceiling or bottom effects occur, and the associated benchmark is not valid.

The second situation in which the Policy Linking Toolkit does not provide guidance are outliers. Sometimes it occurs that a panelist completely disagrees with all other panelists or adapts his or her rating to counteract the ratings of the other panelists. The Policy Linking Toolkit could be expanded to include outlier analyses and a decision rule on how to deal with such outliers. Another addition would be calculating the confidence interval of the benchmarks or displaying the ability distribution and the position of the benchmarks and items on the underlying ability scale.

In several countries, it proved difficult, or impossible, to obtain the sampling weights. The use of sampling weights is crucial for reporting on Sustainable Development Goal (SDG) 4.1.1:

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.

The percentage of learners in each category can be quite different when looking only at the learners who took the assessment and ignoring the sampling design and weights. Currently the policy linking toolkit only provides a suggestion regarding the sampling weights: “The officials should be encouraged to talk about next steps with the benchmarks, i.e., using percentages by category for global reporting. There may need to be additional work on using sampling weights to generalize to the population if the assessment was a sample-based assessment rather than a census.” (p. 44).

- *In a remote workshop, more time is needed for collecting, checking, merging, analyzing and reporting the results of the alignment and two Rounds of Item rating. To ease the process, we suggest that the collecting and checking of the forms is done locally.*
- *For the purpose of capacity building, one could consider developing a data analysis training.*
- *We suggest creating dedicated tools for data entry and analyses to standardize the data management and analyses process. Using a dedicated tool should help prevent errors, inconsistencies or missing ratings. Also, a separate data analysis manual will help.*
- *The analyses and measures described in the Policy Linking Toolkit should be extended with IRT analyses and cover the three situations mentioned. Also, guidance should be given for booklet and item selection in case the assessment contains too many items for the policy linking workshop. Also the minimum number of items for a benchmark should be specified.*
- *The analyses should be expanded to include the analysis of outliers and calculation of the confidence intervals of the benchmarks. We suggest also to visualize the ability distribution and the position of the benchmarks and items on the underlying ability scale.*
- *The Policy Linking Toolkit should clearly outline that the use of sampling weights is necessary for validly reporting on SDG 4.1.1 and provide guidance for the analyses with sampling weights.*
- *The 4.1.1 criteria for policy linking workshop validity should contain clear rules regarding the validity of the benchmarks in case of ceiling or bottom effects.*

6. Reporting the results

The Policy Linking Toolkit contains an outline for a policy linking technical report. This outline is an adaptation from the technical report on setting benchmarks for the National Assessment of Educational Progress (NAEP) in the United States. The outline is not targeted towards the actual set-up of the policy linking workshop with the familiarization and three tasks.

Cito adapted the outline based on the pilot workshop and previous examples of draft reports (see Appendix D). The technical report starts with describing the policy linking methodology and the national assessment used in the policy linking. After this background information, the report continues by describing the workshop preparation and the implementation of the workshop. After that the results of the benchmarking are presented, and the standard setting process is evaluated. The results are summarized for the 4.1.1 review panel and the report finishes with conclusions and recommendations.

- *The outline of the technical report could be aligned more with the policy linking workshop as described in the toolkit.*
- *One might consider creating a complete sample report*

7. Overview of recommendations and further research

The Policy Linking Toolkit is a well-developed and extensively piloted toolkit. It gives suitable guidance for executing policy linking, to those acquainted with standard setting methods. In some areas some further clarification is needed. We recommend that the continued development of the toolkit and the implementation of Policy Linking should focus on increasing the usability and standardization and conducting research.

Additional clarification

Especially in the familiarization phase, we see that additional clarification is needed. The familiarization has been added to the workshop after the first round of piloting. We feel that the familiarization is a very important addition, but it still did not work optimally in the hybrid workshops. Prior to the workshop panelists should receive all key information translated in their 1st language and exercises to familiarize themselves with the GPF. We advise not to ask the panelists to administer the assessment to leaners for two reasons. First, often some or all items must remain secret for future test administrations. Secondly, this homework assignment might lead panelist to focus on weak, average and strong learners in their own country instead of on the Just Partially Meets (JP), Just Meets (JM), and Just Exceeds (JE) learners.

Familiarization during the first day of the workshop could be improved by partitioning information in smaller steps and using more exercises. We suggest that experienced trainers with in-depth knowledge of policy linking and of the GPF give the presentations about policy linking methodology and the Global Proficiency Framework and help with developing additional exercises.

For the other three tasks, the description in the Policy Linking Toolkit is adequate for a good implementation. We can see that the actual workshop has been piloted before and that the toolkit has received several updates. Some minor clarifications have been suggested, especially on how to deal with less common situations, such as non-fitting items, polytomous items and an assessment that does not align with the GPF. It is also crucial for valid and comparable benchmarks that the Policy Linking Toolkit contains a clearer description of the Just Partially Meets (JP), Just Meets (JM), and Just Exceeds (JE) learners and how this relates to their own leaners and the GPF. Some exercises could be developed to help panelists to keep focusing on the Global Proficiency Framework while rating the items.

Usability

The usability of the policy linking can be expanded by increasing the user friendliness and by extending the situations in which policy linking can be applied. The implementation of the workshop is thoroughly described in the manual and in the slides, but both the manual and the slides have become lengthy documents. We suggest creating separate manuals for different target groups, containing only the information relevant to them (e.g., data analysis, content facilitation, logistics, selection of panelists). The material and especially the slides could be designed in such a way, that easier adaptation to grade is possible. We suggest making a library containing relevant examples, sample items and exercises for each grade, both for language and for mathematics. A communication expert could review all the materials for panelists.

During the pilots, we extended the situation in which policy linking was applied in two ways. First, the pandemic necessitated working remotely. The countries clearly preferred to meet in-person when possible and therefore the hybrid mode was created. Furthermore, an 11-day completely remote workshop was developed as the remote workshop currently described in the toolkit was deemed too long (3 weeks) for the countries. We propose to add the agenda for the hybrid mode (and the 11-day completely remote) to the toolkit, as well as the technical requirements, digital tools and adapted slides for the hybrid mode. These modes are also useful for implementing policy linking in countries and areas when travelling is unsafe, to include teachers and experts from remote areas and to reduce travelling and accommodation costs.

Secondly, we used the policy linking for educational surveys using IRT modelling and stratified samples. We strongly recommend adapting the Policy Linking Toolkit for use with complex educational surveys using stratified samples, (complex) booklet designs and IRT. The material provided, the measures and analyses described should be updated for such situations. Also, guidance should be given for booklet and item selection in case the assessment contains too many items for the policy linking workshop and for the minimum number of items needed for a benchmark. In case of national assessments with a complex survey designs using IRT, one could also consider other standard setting methods like the 3DC (Keuning, Straat & Feskens, 2017) in which more items can be used. Finally, we recommend that in calculating the benchmarks he sampling weights are always used. The analyses after the workshop should also include outliers' analyses, several measures of reliability and graphical visualization of the confidence interval of the benchmark.

Standardization

After increasing the usability, the second general recommendation is to standardize the preparation and implementation further. Standardization will ease preparing and implementing the workshop but will also increase the quality and comparability of the results. At the start of this round of piloting, the preparation was not standardized to a high extent. For example, the 4.1.1 Review Panel was not in place. The 4.1.1 Review Panels can determine before the actual preparation of the workshops starts whether the assessment(s) meets reliability and validity standards required for a country to proceed with policy linking for reporting global outcomes. We suggest that the panel also looks at the quality of the IRT calibration and the distribution of the item parameters compared to the ability distribution in the population beforehand to determine whether a national survey is suited for policy linking. We really need a high-quality calibration to be sure of the position of the benchmarks on the underlying ability scale and a selection of items that is sufficiently aligned to the Global Proficiency Framework. Selecting items with suitable item parameters could mitigate the risk of floor or ceiling effects with benchmarks.

The work of the 4.1.1 Review Panel will also help to ensure that all relevant material is provided in a timely manner. The list of materials to be collected and prepared should be extended with the assessment design, sampling weights, item parameters and ability estimates (or plausible values) in case of a complex survey design using IRT modelling. In the policy linking toolkit, the requirement of sharing this information, the assessment and raw data and the purpose of sharing might be explained more extensively. In our experience, the necessity of sharing this information was not clear to the countries.

Another clear step towards a more standardized preparation is the Activity Planner. The UIS consultant developed an Activity Planner in which week-by-week the preparation activities are described. We propose to add this Activity Planner to the Policy Linking Toolkit, as a strict adherence to the planner will ensure well-prepared workshop of a higher quality.

Two aspects of the workshop itself could be more standardized. First, the roles of participants other than the actual panelists should be explained carefully in the policy linking toolkit. The presence of experts and their expressed opinions can have an influence on the alignment ratings and item ratings of the panelists. The Policy Linking Toolkit should provide a clear instruction on the role and expected behavior of those experts and the roles of the international content facilitator on the one hand and the local content facilitator on the other hand could be explained in more detail. Furthermore, the training of the local content facilitator could focus more on dealing with group dynamics and about common pitfalls during facilitation.

The most important step towards further standardization would be to include data management in the Policy Linking Toolkit. Digital forms should be developed and included, as well as data entry files. Data entry personnel should be included in the Team description and a data entry training. Preferably dedicated digital tools should be developed to process and manage all the data of the workshop. Also, for the data analysis, dedicated stand-alone tools and a separate data analysis manual could be developed to ease the analysis and increase standardization.

Research

During the piloting some questions were raised that cannot be answered with piloting. These questions could be studied in experimental research, simulation research or with secondary data analysis. After the first year of piloting, the requirement of consensus was added to the matching task. We suggest conducting experiments to evaluate the effect of consensus on the quality of the benchmarks. Also, an experiment could be conducted to test which way of reaching consensus works best: first find consensus in subgroups or work immediately with the whole group.

During the pilot workshops, we noticed some inconsistencies between the results of the three tasks. Using the data obtained during the workshops, the consistency between the results should be charted. The results of this secondary analysis could help creating instructions for resolving inconsistencies.

As for the benchmarks themselves, the robustness, generalizability and comparability should be studied. Both in experiments and with simulation the robustness of the outcomes can be studied. An important question is whether the same benchmarks and results would have been obtained with another group of raters. Also, the impact of the sampling weights and the relation between the ability distribution and distribution of item parameters should be studied. The suitability of the assessment for policy linking is not only a question of its content, but also depends on the ability distribution and the distribution of the item parameters.

Experiments are also needed to test the generalizability and comparability of the policy linking outcomes. A group of international experts could replicate two or more policy linking workshops. The benchmarks of the original policy linking workshop could then be compared with the benchmarks of the international experts. Cross validation is also possible when different national assessments are equated. The benchmarks established through policy linking can be projected on the next national assessment and a new policy linking workshop can be performed (preferably with the same raters). Next, the equated benchmarks and the new policy linking outcomes can be compared.

8. 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.

Keuning, J., Straat, J.H. Straat & Feskens, R. (2017). The Data-Driven Direct Consensus (3DC) Procedure: A New Approach to Standard Setting. In: Blömeke, S.K & Gustafsson *J.E. Standard Setting in Education: The Nordic Countries in an International Perspective.* 10.1007/978-3-319-50856-6.

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

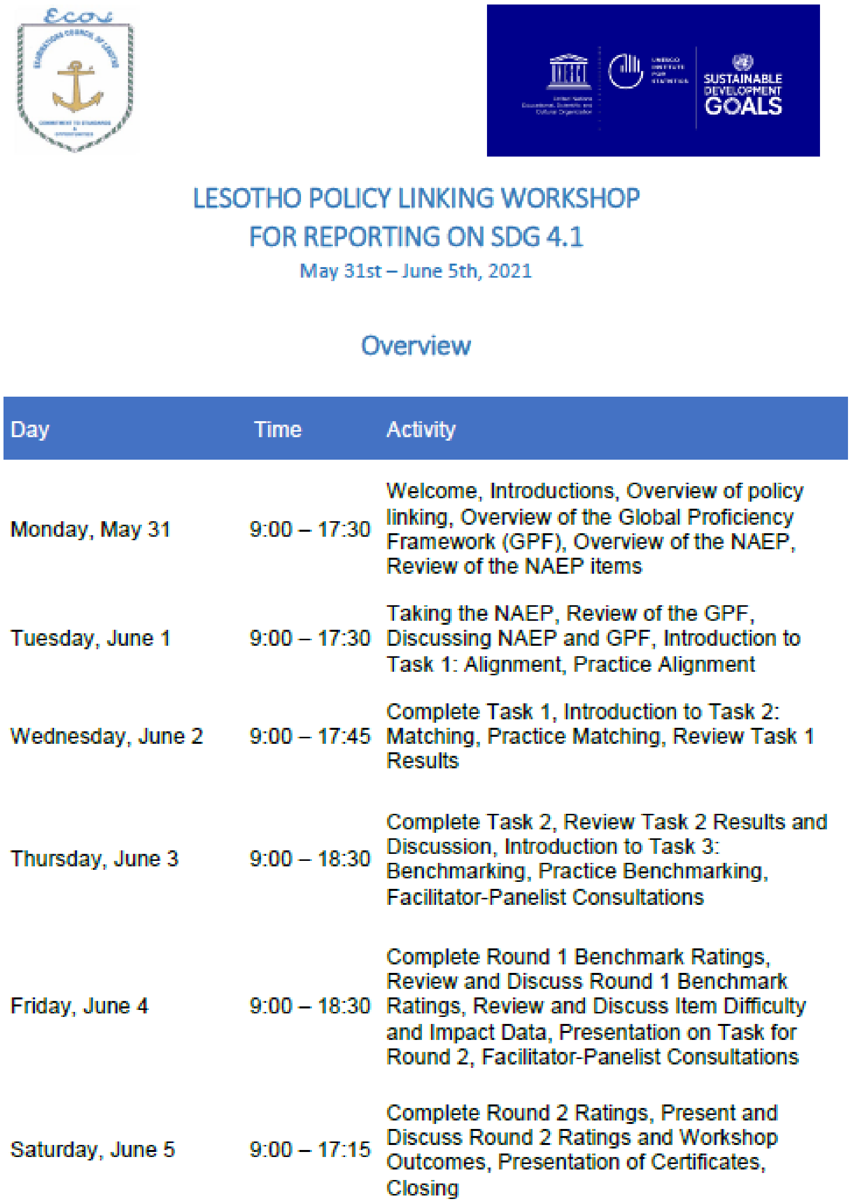
USAID, UNESCO Institute for Statistics (UIS), UK's Foreign, Commonwealth, and Development Office (FCDO), The Bill and Melinda Gates Foundation, Australian Council for Education Research (ACER), World Bank (2020b). *Global Proficiency Framework for Reading Grades 1 to 9*. Downloaded from https://www.edu-links.org/sites/default/files/media/file/GPF_Reading_Final_D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

9. Annexes

Annex A: Agenda for the blended 6-day workshop

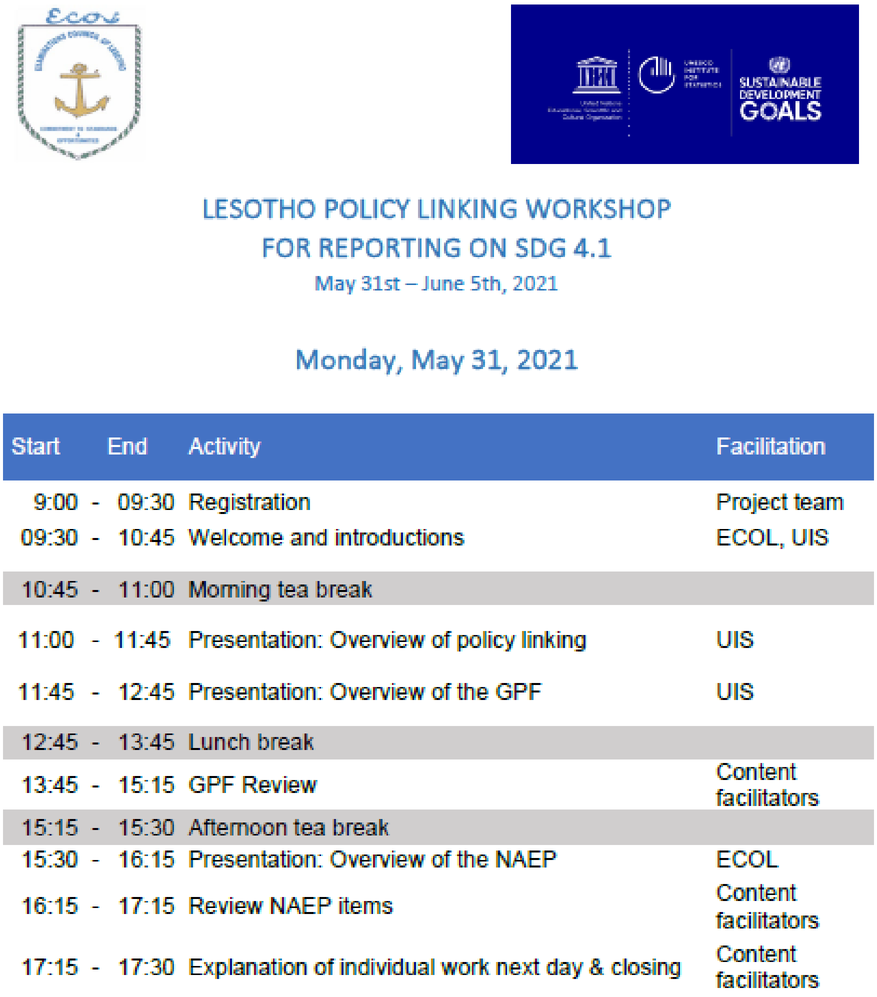


LESOTHO POLICY LINKING WORKSHOP
FOR REPORTING ON SDG 4.1

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

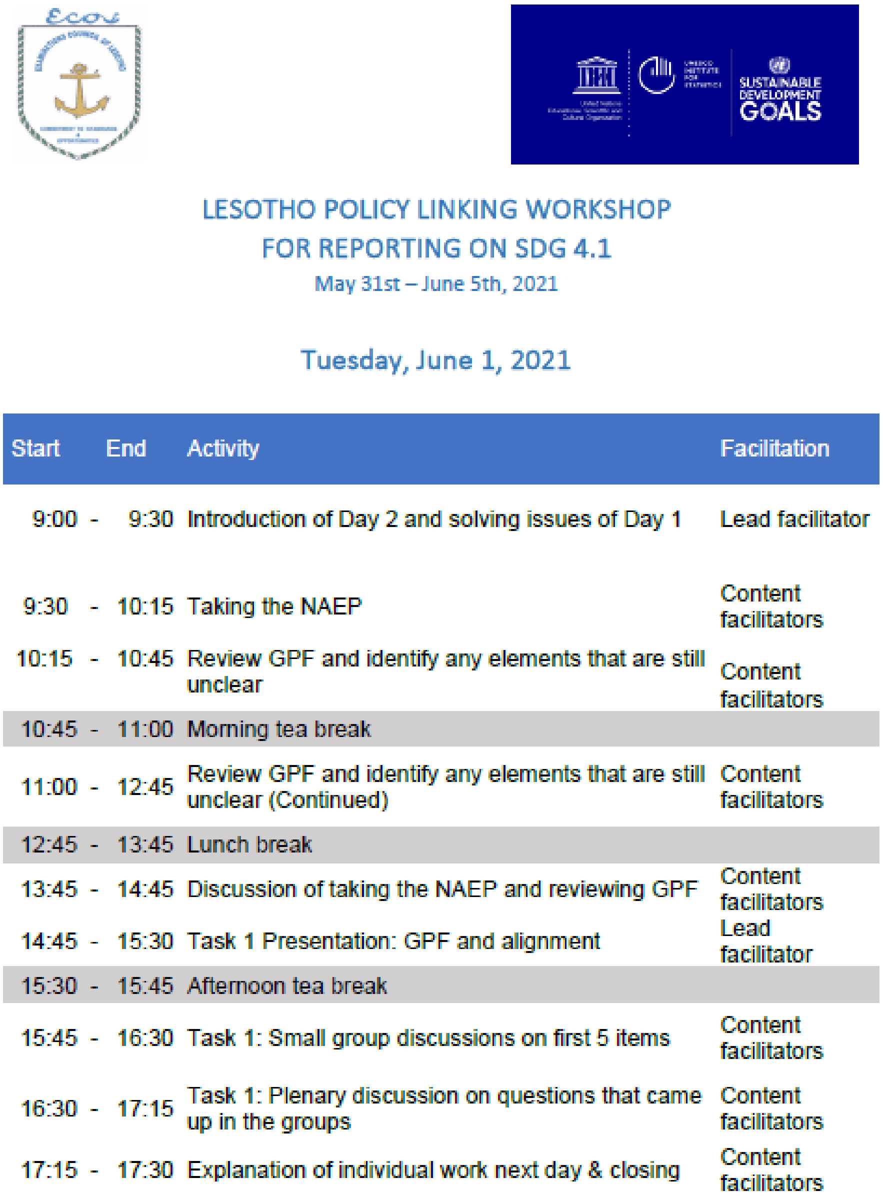


LESOTHO POLICY LINKING WORKSHOP FOR REPORTING ON SDG 4.1

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

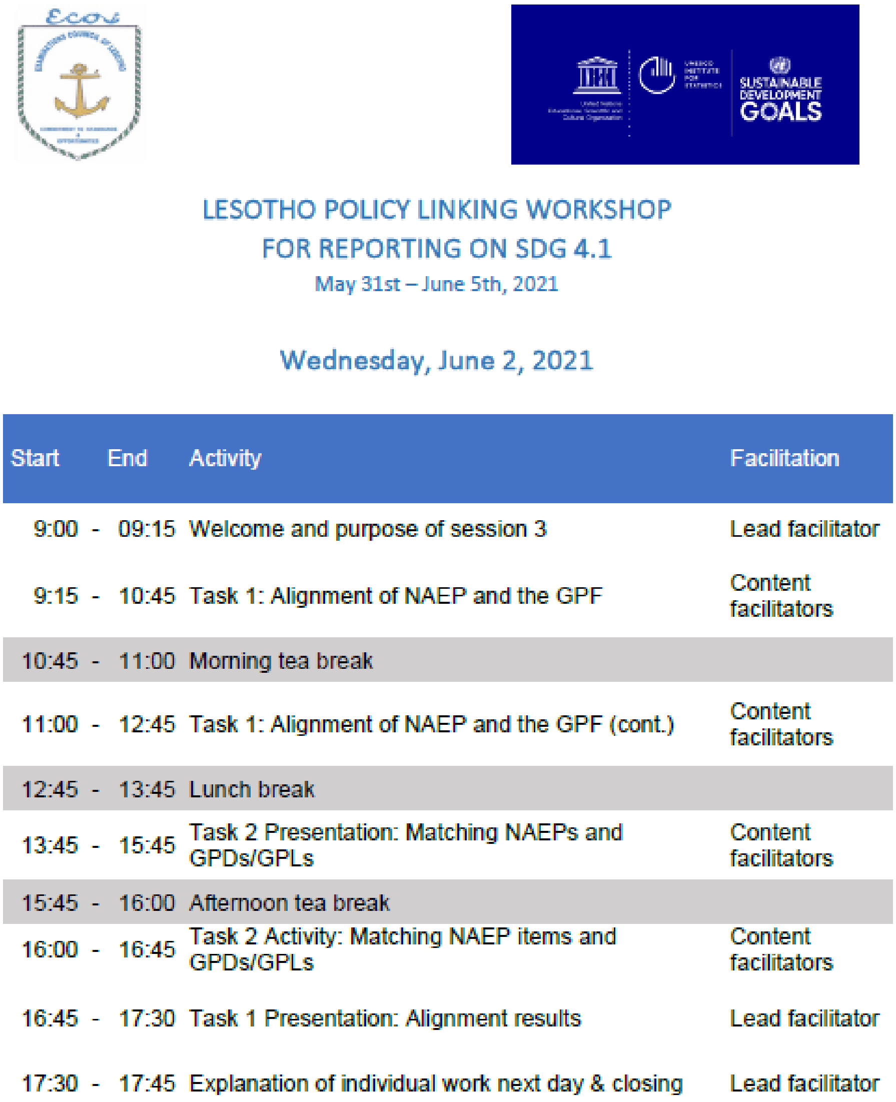


LESOTHO POLICY LINKING WORKSHOP FOR REPORTING ON SDG 4.1

May 31st – June 5th, 2021

Tuesday, June 1, 2021

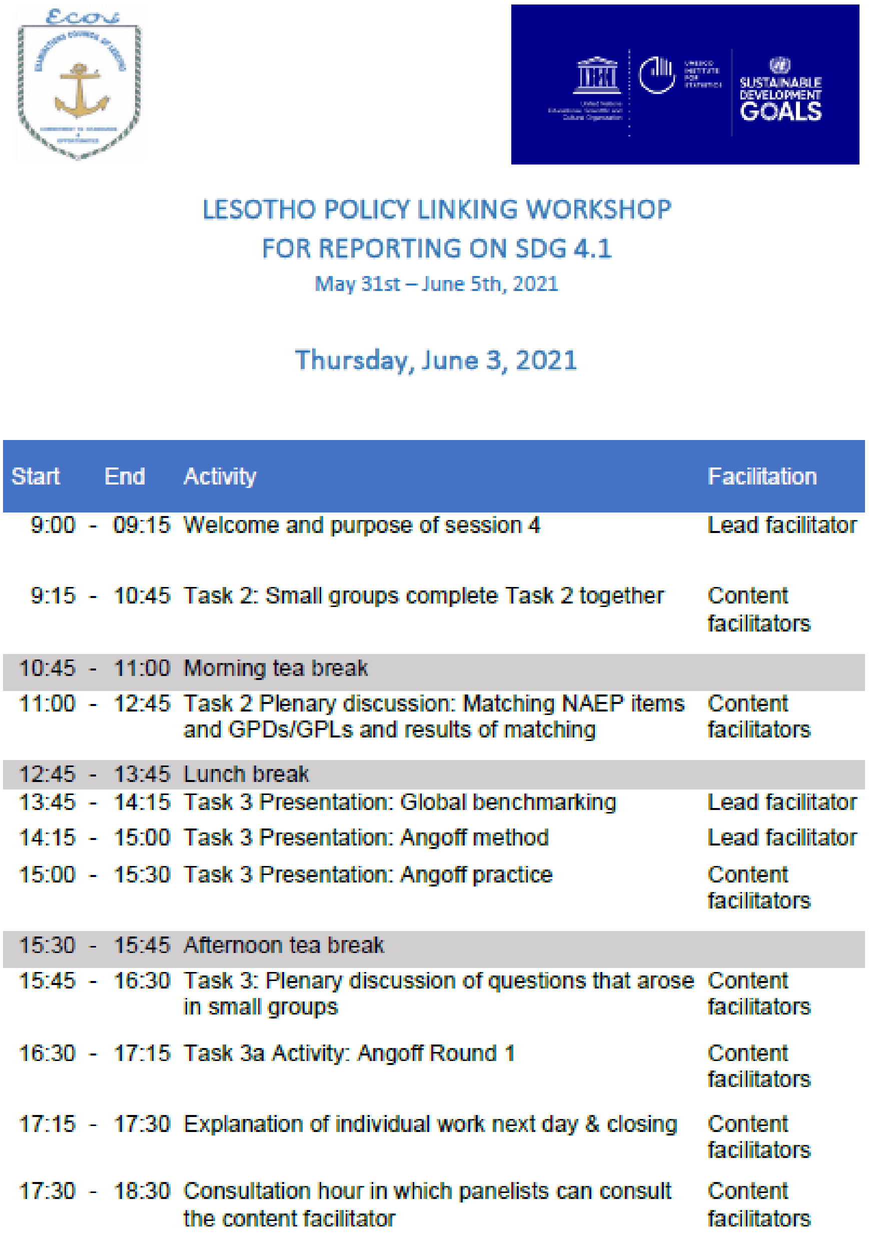
Start		End	Activity	Facilitation
9:00	-	9:30	Introduction of Day 2 and solving issues of Day 1	Lead facilitator
9:30	-	10:15	Taking the NAEP	Content facilitators
10:15	-	10:45	Review GPF and identify any elements that are still unclear	Content facilitators
10:45	-	11:00	Morning tea break	
11:00	-	12:45	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

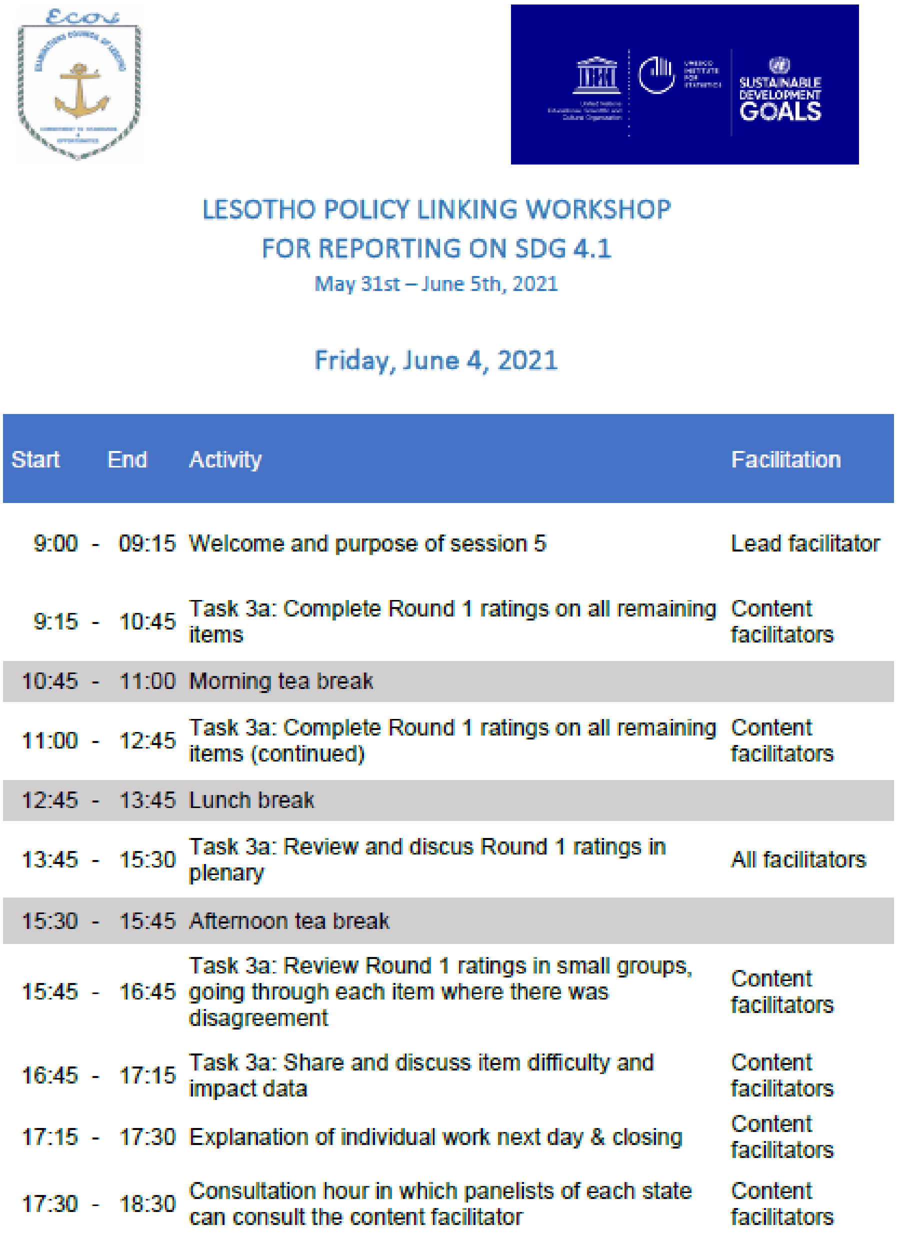


LESOTHO POLICY LINKING WORKSHOP FOR REPORTING ON SDG 4.1

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

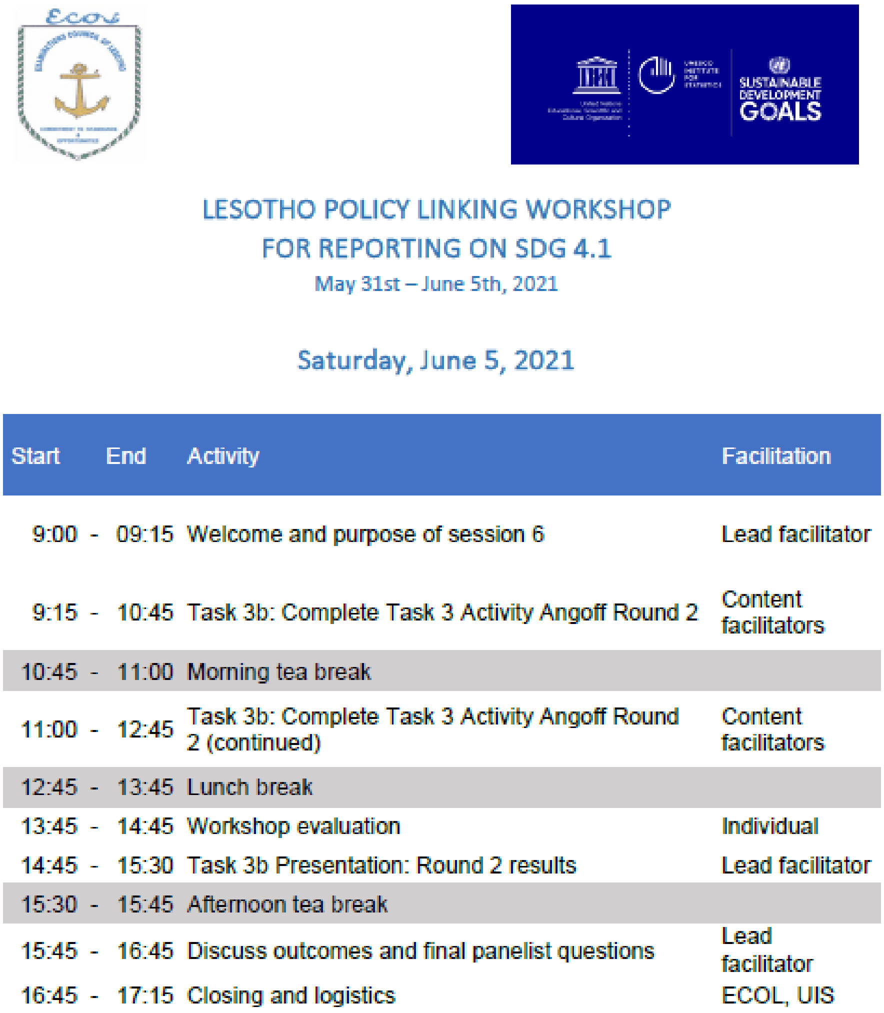


LESOTHO POLICY LINKING WORKSHOP
FOR REPORTING ON SDG 4.1

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



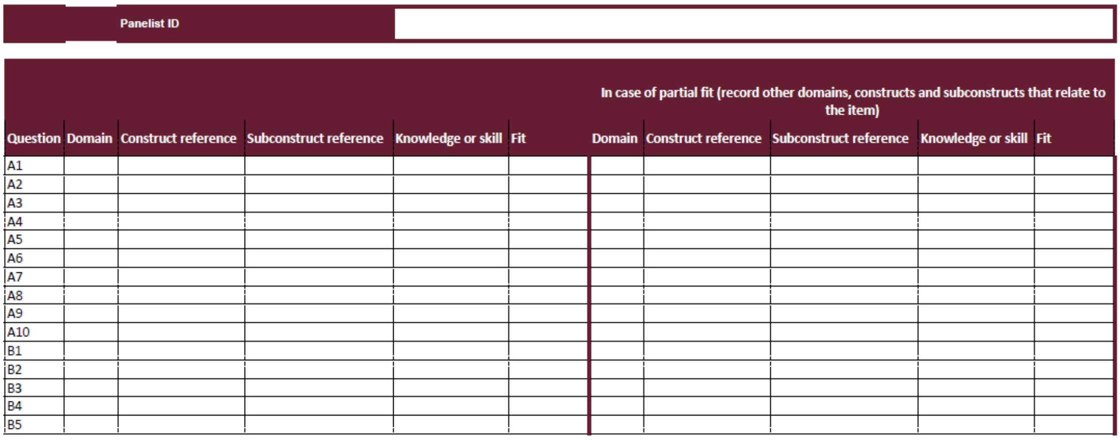
LESOTHO POLICY LINKING WORKSHOP
FOR REPORTING ON SDG 4.1
May 31st – June 5th, 2021

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

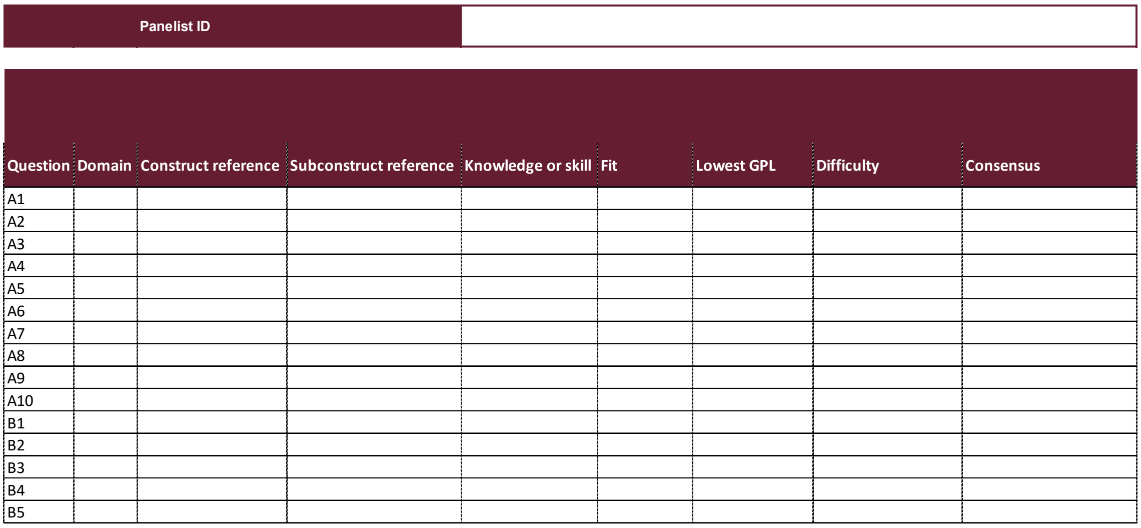
Figure 2. Alignment rating form (English) for paper-based rating



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										

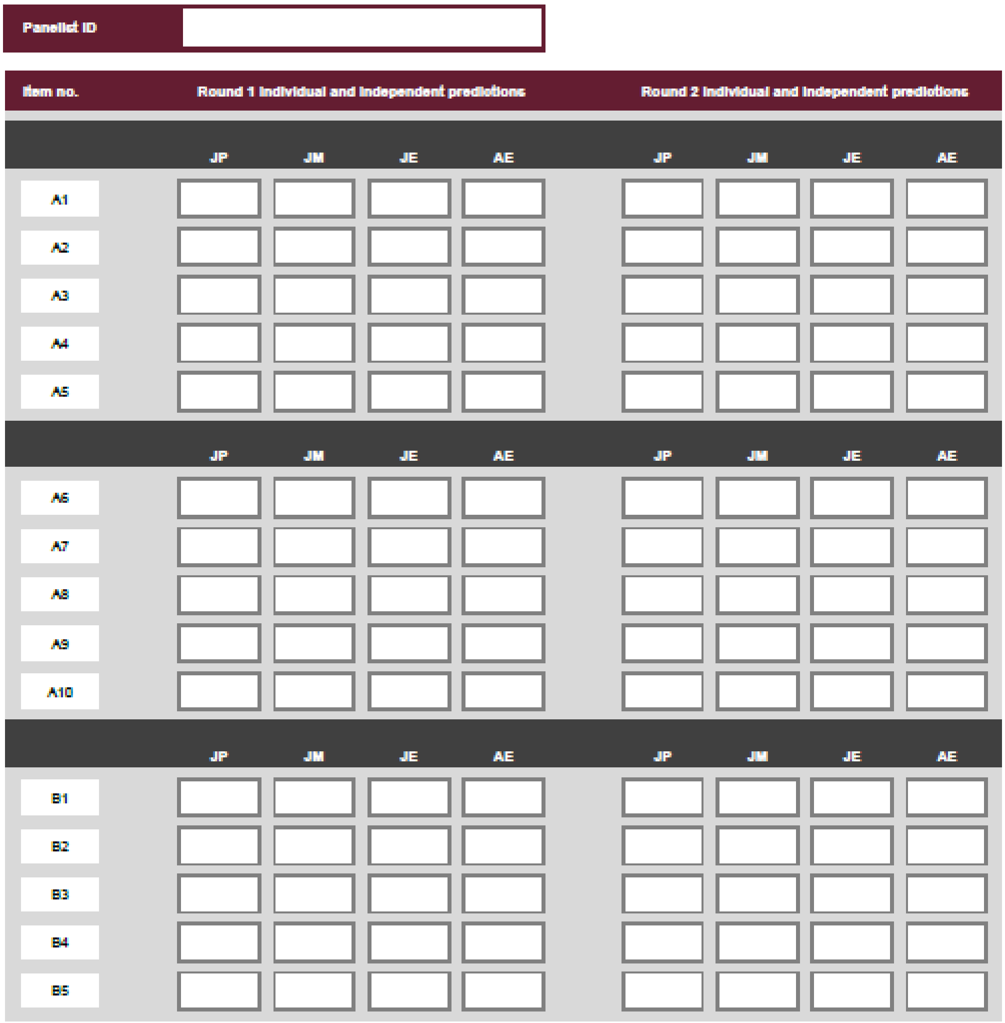
Figure 3. Matching form for the local content facilitator (English)



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								

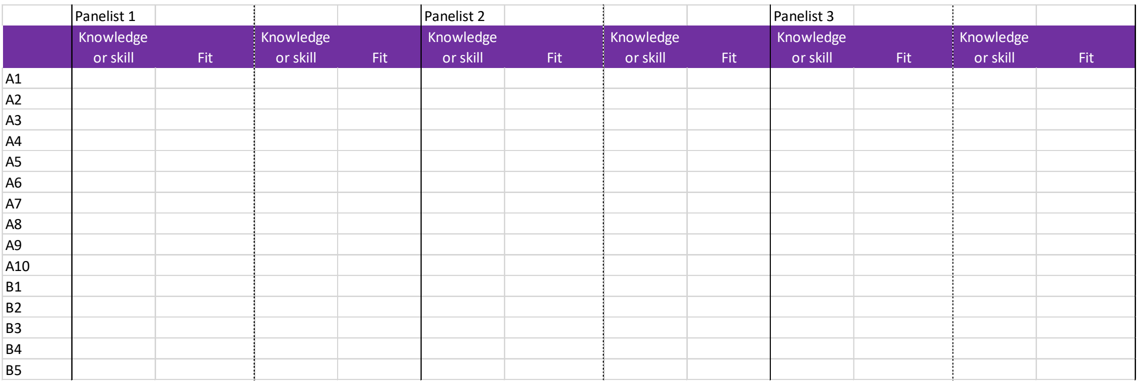
Figure 4. Item rating form (English) for paper-based rating



Panelist ID	

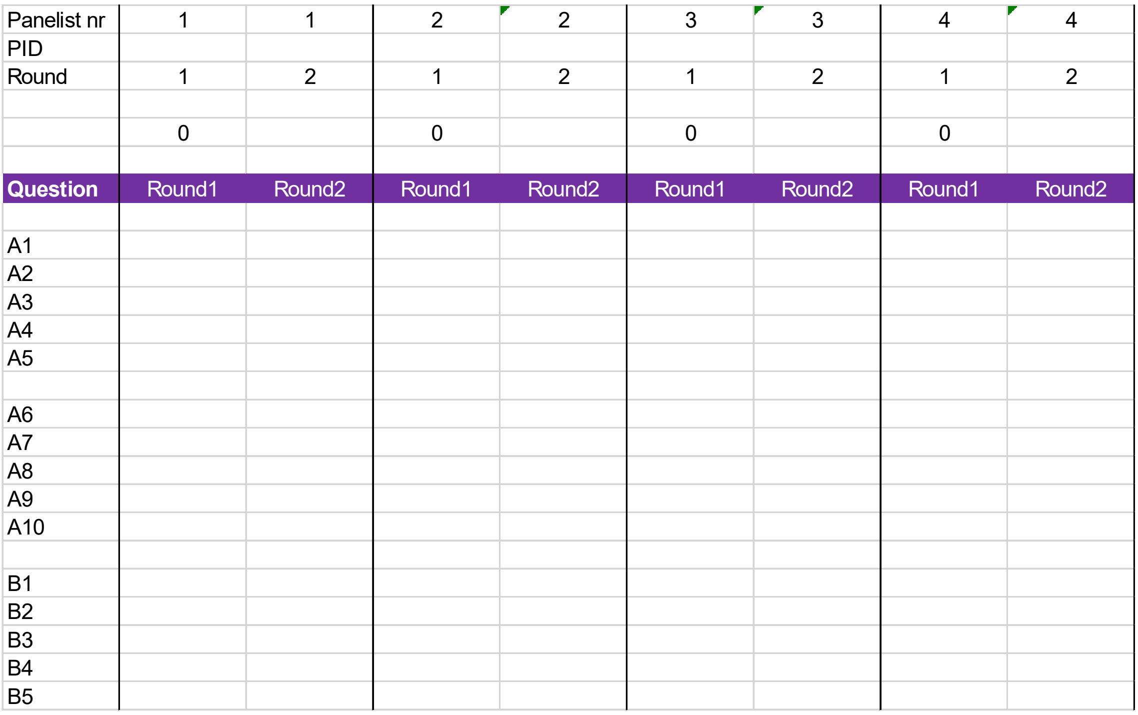
Item no.	Round 1 Individual and independent predictions				Round 2 Individual and independent predictions			
	JP	JM	JE	AE	JP	JM	JE	AE
A1								
A2								
A3								
A4								
A5								
	JP	JM	JE	AE	JP	JM	JE	AE
A6								
A7								
A8								
A9								
A10								
	JP	JM	JE	AE	JP	JM	JE	AE
B1			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

Figure 5. Data entry file for Alignment rating results (English)



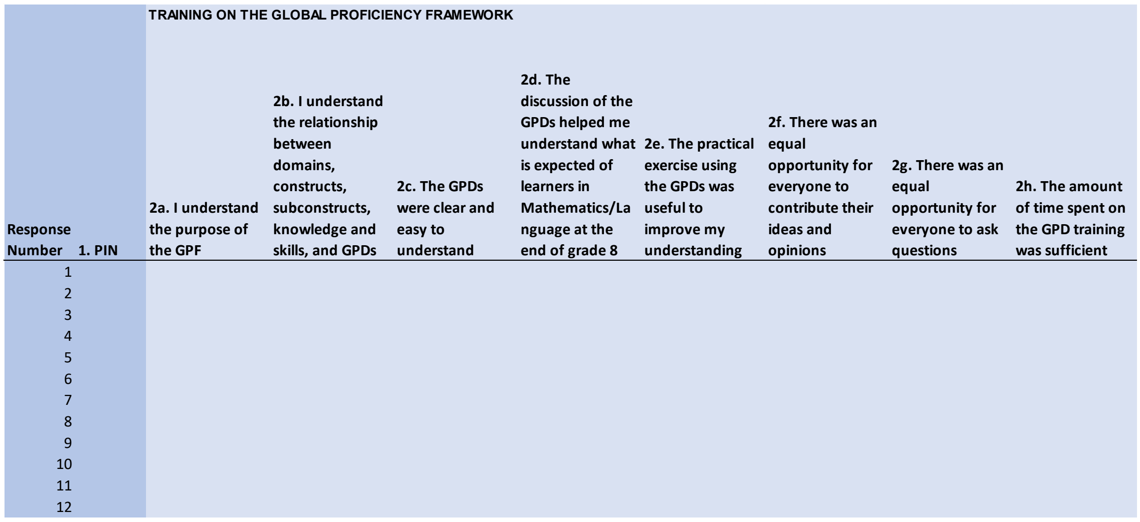
	Panelist 1				Panelist 2				Panelist 3			
	Knowledge or skill	Fit	Knowledge or skill	Fit	Knowledge or skill	Fit	Knowledge or skill	Fit	Knowledge or skill	Fit	Knowledge or skill	Fit
A1												
A2												
A3												
A4												
A5												
A6												
A7												
A8												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

Figure 6. Data entry file for Item rating results



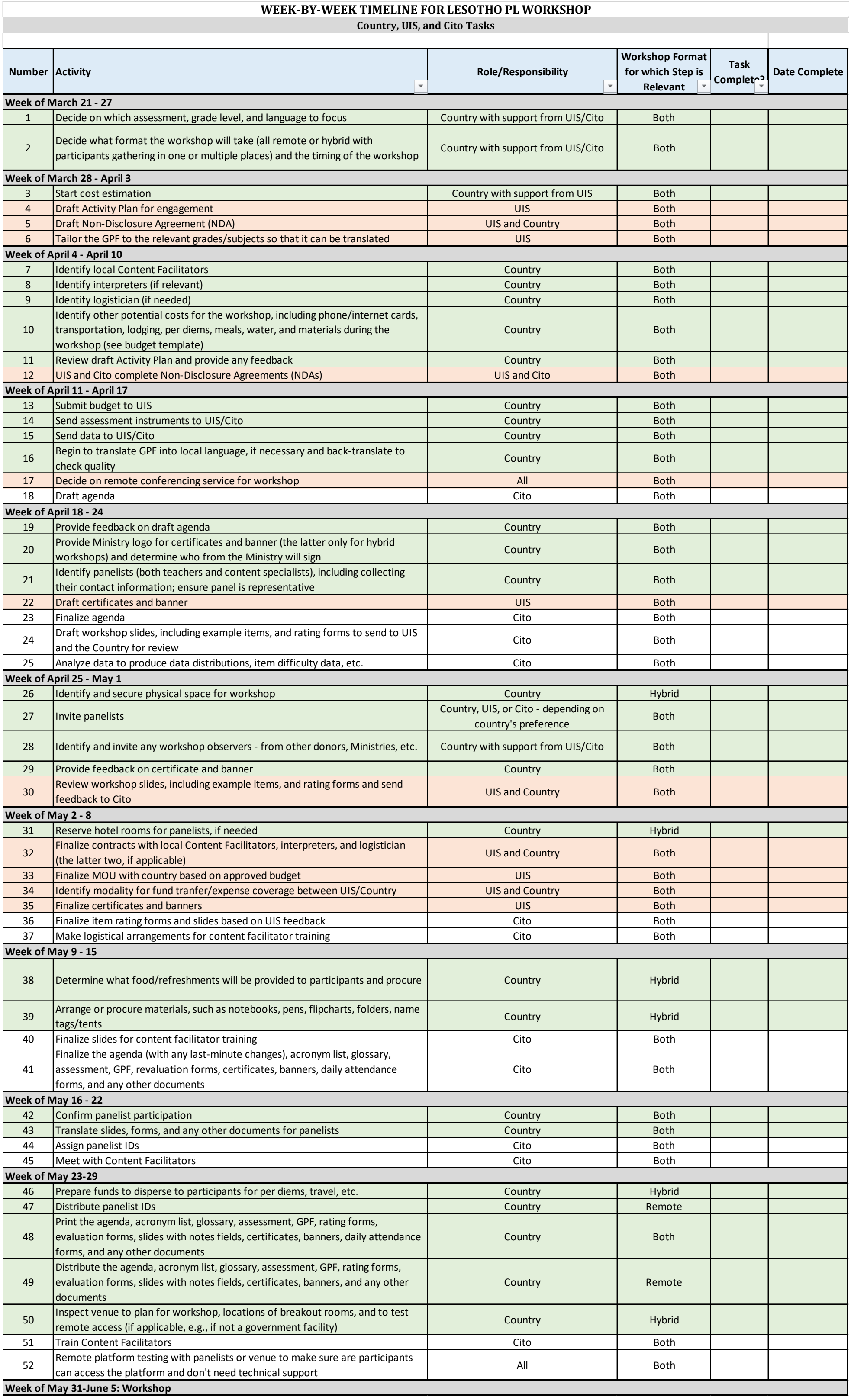
Panelist nr	1	1	2	2	3	3	4	4
PID								
Round	1	2	1	2	1	2	1	2
	0		0		0		0	
Question	Round1	Round2	Round1	Round2	Round1	Round2	Round1	Round2
A1								
A2								
A3								
A4								
A5								
A6								
A7								
A8								
A9								
A10								
B1								
B2								
B3								
B4								
B5								

Figure 7. 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: Outline of policy linking reports

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

ACRONYMS AND ABBREVIATIONS

GLOSSARY OF TERMS FROM THE POLICY LINKING TOOLKIT

1. EXECUTIVE SUMMARY

2. BACKGROUND

POLICY LINKING OVERVIEW

Global Proficiency Framework

The policy linking methodology

OVERVIEW OF THE NATIONAL ASSESSMENT

Content and design of the assessment in grade X

Sample and data analysis

3. PILOT WORKSHOP PREPARATION

OBJECTIVE OF THE WORKSHOP

FIRST THREE POLICY LINKING STAGES

GENERAL PREPARATION OF THE WORKSHOP

MATERIALS FOR THE WORKSHOP AND PRE-WORKSHOP ANALYSES

Collecting materials and pre-workshop analyses

Creating workshop materials

TRAINING THE LOCAL CONTENT FACILITATORS

TRAINING FOR LOCAL DATA ENTRY

4. IMPLEMENTING THE BLENDED WORKSHOP

FAMILIARIZATION

TASK 1: ALIGNMENT

Alignment Language

Alignment Mathematics

TASK 2: MATCHING

TASK 3: BENCHMARKING

Round 1

Round 2

WORKSHOP EVALUATION

5. RESULTS OF THE BENCHMARKING

ROUND 1

ROUND 2

6. EVALUATION OF THE STANDARD SETTING PROCESS

INTERNAL EVALUATION SEM, PANELIST CONSISTENCY AND PANELISTS’ AGREEMENT

PROCEDURAL EVALUATION

7. SUMMARY OF RESULTS OF CRITERION 4 FOR THE 4.1.1 REVIEW PANEL

8. CONCLUSIONS AND RECOMMENDATIONS

RECOMMENDATIONS

Workshop Preparation