unesco
Institute for Statistics

Credit: GPE/Alexandra Humme

Report of the National Assessment of Student Achievement Policy Linking for Measuring Global Learning Outcomes Workshop (May 2022)

Setting Global Benchmarks for Grade 5 Reading and Mathematics in Zambia

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

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

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

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

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

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

Acknowledgements

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

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

Acknowledgements

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

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

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

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

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

Sjoerd Crans

Gerben Veerbeek

Margreet van Aken

Sanneke Schouwstra

Angela Verschoor

Table of Contents

Acronyms and Abbreviations

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
NAS	The National Assessment Survey
PLT	Policy Linking Toolkit
SDG	Sustainable Development Goal
SEM	Standard Error of Measurement
UIS	UNESCO Institute for Statistics (UIS)
USAID	U.S. Agency for International Development

Glossary of Terms from the Policy Linking Toolkit

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1. Executive Summary

This document contains the report on the hybrid policy linking workshop that took place from Monday May 9, 2021 until Saturday May 14, 2022. The Examinations Council of Zambia (ECZ) and UNESCO Institute for Statistics (UIS) organized this workshop as a pilot. The objective of the workshop was to set global benchmarks on the 2021 National Assessment Survey (NAS) at grade 5 in English language and in Mathematics.

This was the first time Zambia participated in a policy linking workshop. However, Zambia had previous experience with standard setting in the Monitoring Impacts on Learning Outcomes (MILO) project. The local participants met physically in a hotel of Lusaka, Zambia, whereas the international participants joined via a videoconferencing platform (Zoom). The presence of the international participants was enhanced by excellent facilities provided by the local organizers: microphones, cameras, big screens and a close cooperation between local and international content facilitators via mail, chat and telephone contact.

The participants performed their tasks with dedication and engaged in lively discussions during the tasks. Every step of the process produced important outcomes. The participants gave very positive feedback, both in person and in their evaluation forms. In the closing remarks of the workshop, the wider benefit of this workshop for Zambia, namely capacity building in assessment for the participants, was repeatedly mentioned.

The participants' work showed that the NAS for English language is strongly aligned to the GPF for grade 5 and both in depth and in breath. The NAS for Mathematics is additionally aligned in depth to the GPF for grade 5 and strongly aligned in breath. Furthermore, the panelists managed to reach almost complete consensus on the matching, albeit sometimes after long discussion. The final benchmarks of the panelists show an adequate consistency, which makes the "Meets" benchmark useable for comparing, aggregating, and tracking learning outcomes for the NAS in Zambia.

The piloting of the policy linking workshop in this hybrid mode can be considered a success. One point that came up in earlier policy linking workshops that we conducted in other countries in Africa came up again in this workshop, namely the question of the validity and viability of Policy Linking for language and the results of such Policy Linking when the language of the assessment is not the native language of the learners. We recommend a careful consideration of this point. This and further recommendations coming from this piloting workshop for the conducting of future policy linking workshops are given in Section 8.

2. Background

Policy Linking Overview

In September 2015, Member States of the United Nations formally adopted the 2030 Agenda for Sustainable Development in New York. The agenda contains 17 goals, including a separate 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 and three implementation targets (UNESCO, 2021). The first target focusses on primary and secondary education (target 4.1): By 2030, ensure that all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes. To monitor progress the indicator 4.1.1 is used: *Proportion of children and young people (a) in grades 2/3; (b) at the end of primary; and (c) at the end of lower secondary achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex* (United Nations, 2021).

To allow countries to use their existing – sub-national, national, and cross-national – assessments to report against Sustainable Development Goal (SDG) 4.1.1, the policy linking methodology was developed (USAID, 2019). Policy linking makes use of a standard-setting methodology (the Angoff approach) to set benchmarks on learning assessments. While it is an existing standard-setting methodology, UIS and its partners have extended its use to help countries set benchmarks using the Global Proficiency Framework (GPF) for producing and reporting SDG4.1.

Global Proficiency Framework (GPF)

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 countries' national assessments to the GPF, countries and donors/partners can compare learning outcomes across language groups in countries as well as across countries and over time.

The policy linking methodology

There are seven stages of policy linking for measuring global learning outcomes that must be completed to facilitate global reporting for SDG4.1 (USAID at all, 2020c). Countries/assessment agencies and their partners must complete each of these stages for their results to be accepted for reporting against SDG 4.1.1.

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

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

Policy linking workshop

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

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

The policy linking methodology was piloted in several countries in 2019 and 2020, among which in India, Bangladesh and Nigeria. Also, the ICAN pilot was conducted in 2020. Following these piloting workshops, adjustments were made to the methodology, toolkit, and GPF. Due to the COVID-19 pandemic the piloting was delayed. In 2021 and 2022 further piloting of the Policy Linking Toolkit took place in several countries.

Overview of the National Achievement Survey (NAS)

The Examinations Council of Zambia (ECZ) had been conducting the National Achievement Survey (NAS) since 1998 in grade 5 and grade 9. The NAS targets students in public, private, grant-aided and community schools. The NAS was developed in response to the demand for a systematic approach to the accountability for learning outcomes in schools. NAS results function as performance indicators for the educational system and they allow verifying learning achievement targets.

The major objective of the NAS is to provide feedback about the learning achievements and the trends in learning achievement over the time. NAS has the following:

- Measure the extent to which learners have mastered the literacy and numeracy skills appropriate to their level;
- To identify variations in learning achievement by gender and region and identify geographical disparities in the levels of learning achievement;
- To provide information on the impact of educational inputs on learning achievement
- To provide information on learning achievements and how they are changing over time relative to changes in educational inputs and processes;

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

- To provide the baseline as the point of reference in the future.

The NAS serves the following purposes: - designing individualized instructional plans - supporting teachers (training, relevant materials, etc.) - school or educator accountability - sub-national level monitoring of learning outcomes - monitoring education quality levels - planning education policy reforms - measuring impacts of educational inputs.

Content and sample of the NAS in grade 5

The NAS is a low-stake assessment. It is a written assessment, administered face-to-face and delivered through paper-pencil tests. All test-takers are presented with the same cognitive booklets or tests, which are aligned with the national curriculum.

The NAS summarizes pupils' achievement. Each grade 5 pupil receives one paper-pencil test consisting of English language, Life skills and mathematics. Apart from the cognitive tests each pupil also receives a background questionnaire. Their teachers and head teacher also received a questionnaire.

The sampling design used for NAS is a multi-stage sample design. Schools were selected using probability proportionate to Size (PPS). Public, community, grant-aided and private schools were included. In total 400 schools were selected. In the second stage of the sampling, from each selected school 20 pupils were selected using random numbers. The school response rate (including replacement schools) was 99% and the student response rate (including replacement) was 93%. The overall response rate was 91%.

Pilot Workshop Preparation

Objective of the workshop

The objective of the workshop was setting global benchmarks on the 2021 National Assessment Survey (NAS) at grade 5 in English language and Mathematics using a hybrid policy linking workshop. The workshop had a piloting function and should increase the capabilities of ECZ to conduct similar workshops in the future.

First three policy linking stages

After the engagement of Zambia, Cito joined the meeting between UIS and ECZ on Monday March 7, 2022. Cito was contracted to facilitate the policy linking workshop and provided the lead facilitator, two content facilitators and a data analyst. After the initial engagement, the country governments or assessment agencies should collate evidence of curriculum and assessment validity and alignment (stage 2 of policy linking) and the 4.1.1. Review Panel should review this collated evidence. Cito did not receive information about stage 2 of policy linking from the 4.1.1 Review Panel. “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.

Prior to the workshop, Cito was not informed whether the assessment meets reliability and validity standards required for Zambia to proceed with policy linking for reporting global outcomes For this reason, Cito made an initial assessment of whether the assessment(s) meets the standards required to proceed with policy linking. Cito’s content facilitators gave an estimate whether enough items would align. ECZ shared the codebooks, the item parameters and items of the NAS with UIS and Cito for preparing the workshop.

The NAS English language consists of one booklet containing 35 items. About two-thirds of the items assess reading. The remaining third of the items assesses writing, vocabulary or punctuation and has no link to the Global Proficiency Framework. ECZ and UIS decided to implement the suggestion of Cito to exclude these items from the procedure.

The NAS mathematics consists of one booklet containing 45 items. After consultation with UIS and Cito, ECZ followed Cito’s suggestion use all items in the procedure.

The implemented sampling procedure, as described in the MILO 2022 report (UNESCO Institute for Statistics, 2022), ensures that the learners who carried out the assessment are representative of the population against which results are reported. The item development, review process and pretesting as reported seems appropriate. The reliability of the NAS as reported in the 2022 MILO report is good: both English language and Mathematics had a reliability higher than 0.82.

General preparation of the workshop

UNESCO/UIS, ECZ and Cito planned to facilitate the workshop in a hybrid form, due to the COVID-19. The videoconferencing platform used was Zoom. The data analyses were done remotely by the Cito analysts in The Netherlands.

Two rooms were reserved for the workshop: one for the plenary meetings and the break-out session for mathematics and one for the break-out sessions of Reading. The international facilitators participated in the workshop through a big screen, one in each room. During the workshop, the panelists gave their ratings on paper. For this reason, Cito provided Excel-files for data entry and a two-hour data entry training. After each task the data were entered on location in the developed Excel-files and sent to Cito.

A draft agenda for the workshop was shared with the stakeholders (ECZ, UIS) for suggestions and improvements; consequently, the final agenda was made and shared (see the overview in Table 1, in Annex A the complete agenda is presented). Following this, the draft slides for the workshop were shared with the stakeholders (ECZ, UIS) for suggestions and improvements; consequently, the final slides made and shared. The workshop took place from Monday May 9 until Saturday May 14 2022.

Table 1 Agenda for the workshop

Day 1 — May 9	**Day 4 — May 12**
Welcome and introduction Policy Linking	Task 1 Presentation: Alignment results
Familiarization: Global Proficiency Framework (GPF)	Task 2 Activity: Matching NAS items and GPDs/GPLs
Familiarization: National Assessment Survey (NAS)	Task 2 Plenary discussion: Matching NAS items and GPDs/GPLs and results of matching
Day 2 — May 10	**Day 5** — May 13
Familiarization: Review GPF + NAS	Task 3 Presentation: Benchmarking
Task 1 Presentation: GPF and alignment	Task 3 Activity: Angoff Round I
Task 1 Activity: Align the NAS and the GPF	Task 3 Presentation: Angoff Round 1 results
Day 3 — May 11	**Day 6** — May 14
Familiarization: GPLs and GPDs	Task 3 Activity: Angoff Round 2
Task 2 Presentation: Matching NAS and GPDs/GPLs	Task 3 Presentation: Angoff Round 2 results
Task 2 Activity: Matching NAT and GPDs/GPLs	Closing and logistics

ECZ sought teachers and specialists from each province. Both in the reading panel and in the mathematics panel all ten provinces were represented (see Table 2). In the mathematics panel a bit more men than women were present (62% male), whereas in the reading panel the vast majority (81%) was female. In both panels 10 panelists were teacher in primary education, the other panelists were specialists from the Ministry of Education, the Examinations Council of Zambia or retired specialists.

Table 2 Panelist' background information

	Mathematics	Reading	Total
Province			
Central Province	1	1	2
Copperbelt	1	1	2
Eastern	1	1	2
Luapula Province	1	2	3
Lusaka	2	5	7
uchinga Province	1	2	3
Northern	2	1	3
North-Western	1	1	2
Southern	2	1	3
Western	1	1	2
Total	13	16	29
Gender			
Female	5	13	18
Male	8	3	11
Total	13	16	29

Materials for the workshop and pre-workshop analyses

During the preparation of the workshop, ECZ, UIS and the lead facilitator from Cito tried to have weekly meetings. Actually holding these meetings was hindered by busy schedules of most of the people involved and by internet and power problems. A week-by-week timeline for the Policy Linking Workshop as described in the UIS Activity plan (see Annex C) served as a guideline.

Collecting materials and pre-workshop analyses

Before the workshop, ECZ shared the assessments. This was not without difficulty, however: for a long time, there was confusion about the relation and the difference between NAS and MILO, and consequently which data set to use. The panelists were not asked to administer the NAS to students before the workshop, but were asked to take the assessment themselves during the workshop. For language, it turned out that a number of the items did not address reading comprehension, but grammar or vocabulary. In the end, it was decided to use only 22 (out of the 35) items for the workshop.

We received the MILO data and the raw data of Zambia, including the keys. The raw data were scored using the keys. In preparation for the workshop the distribution of the sum scores was calculated and the p-values using Classical Test Theory (see Appendix F). After the workshop the impact of the benchmarks was calculated using the sampling weights from the MILO dataset.

Creating workshop materials

Cito prepared a package for panelists containing all workshop materials, to be printed on location. The package contained:

1. Agenda
2. Panelist ID

3. Glossary of Terms
4. Acronym list
5. Grade 4, 5, 6 from the Global Proficiency Framework
6. National Assessment Survey (NAS) – grade 5
7. Slides (printed in notes format)
8. Alignment rating form
9. Item rating form
10. Evaluation form

During the workshop, further materials were shared digitally and printed where necessary, for example the Matching form for the content facilitators.

Examples of the forms can be found in Annex B.

Training the local content facilitators

During the last 2 weeks before the workshop, the content facilitator training was held. Cito planned a 6-hour training consisting of 3 different parts for both the local content facilitators for Language and Mathematics:

1. A two-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 two-hour general rehearsal of the workshop for both Language and Math.

The whole team was invited for the introduction (1) and the general rehearsal (3). The interactive sessions were intended for Cito's content facilitators and their local counterparts (The Gambia's content facilitators). 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. During all these sessions, Cito's content facilitators and their counterparts created a good working relationship and understanding of their respective roles during the workshop.

Again, scheduling was difficult due to busy schedules and internet and power problems. The final session also doubled up as a technical test of all the facilities in the Zambia venue (big screen, audio, microphones, internet and wide-angle cameras).

Training for local data entry

As the panelists worked on paper, data entry was needed, and a special 2-hour data entry training was given on the second day of the workshop (a data entry training is not yet part of the policy linking toolkit). On three days (day 3, 5 and 6) data entry had to occur. The panelists handed over their forms at the end of the morning and during lunch time the data had to be entered. As the data had to be analyzed and the results presented that same afternoon, the window for data entry was narrow. During the training the schedule and times for data entry were shown. Next, Cito discussed the steps in data entry and gave a demonstration of data entry for each of the different forms.

The global steps in data entry were:

1. Received 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

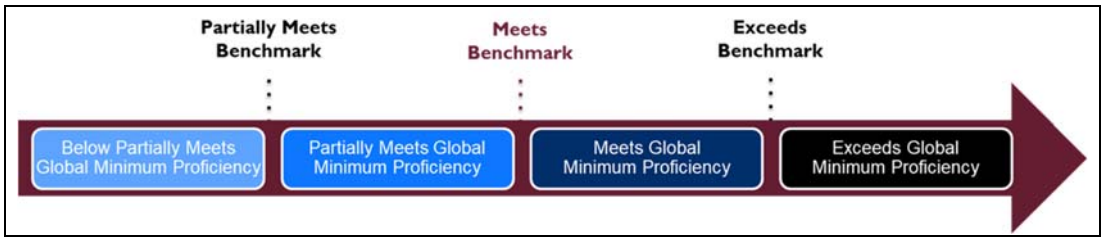
3. Implementing the blended workshop

Familiarization

The workshop started with a welcoming and preparation session. After the formal welcome, the first day focused on familiarizing panelists with policy linking, the Global Proficiency Framework and the National Assessment of Student Achievement. The panelists received the printed workshop materials in the venue (such as the Global Proficiency Framework).

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 adviser of UIS presented the panelists with an overview of the Global Proficiency Framework and its role in policy linking. The example of the benchmarks and the proficiency levels is shown in Figure 1.

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



In the separate sessions for the subjects, the content facilitators introduced each of the domains, constructs, subconstructs, and statements of knowledge and/or skill(s). The GPLs and GPDs were mentioned, but not yet introduced in depth. This was done because this knowledge was not yet needed for the panelists for the Alignment task, and because experience from Cito with previous workshops had shown that this extra information was confusing for the panelists. An example from part of the mathematics GPF is shown in Table 3.

Table 3. Part of the Global Proficiency Framework of Mathematics describing the domain, constructs and subconstructs

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

The day closed with an introduction to the National Assessment Survey and the panelists doing the NAS themselves. In the morning of the second day the panelists reviewed the NAS items and discussed any elements of the GPF (up to and including statements of Knowledge and skills) that were still unclear.

Observations

In the Language group we looked at the GPF, in particular the structure and the way in which it is composed. The slides do not allow for much attention for table 2, but it has proven to be time well spent, when the GPF, which is a very hefty document, is broken down into more bite-sized chunks.

There are a number of mismatchess in the GPF. The tables 3 and 5 were adapted at some point during 2020-2021, whereby the coding was changed. However, the example texts and items have not been adapted alongside this. This means that when you practice, by giving the panelists sample texts and items, the coding in the appendixes does not match that in the tables.

In the agenda, there is no time scheduled for the local and international content facilitators to meet at the end of the day, to look back on and learn from the day and to look forward to and prepare for the next day. Especially in remote or hybrid settings, this typically means this is not done.

In this workshop, there were two local content facilitators per subject scheduled. They all participated in the content facilitator training. However, during the workshop, as far as we could see only one of the local content facilitators for language was present.

Task 1: Alignment

The panelists had to execute three tasks during the workshop:

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

Still on the morning of the second day of the workshop, the panelists received an introduction to their first task: aligning the National Assessment of Student Achievement to the Global Proficiency Framework (GPF). Alignment is important, because it ensures there are enough items in the assessment that measure the knowledge and/or skill(s) depicted in the GPF for policy linking to work. The purpose of the alignment task was to ensure panelists have fully understood the GPF and to allow them to identify which statements of knowledge and/or skill(s) describe the knowledge and/or skill(s) required of students 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 NAS 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 subject groups, 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.

After this practice, panelists were asked to work individually and independently to rate the alignment between each NAS 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 ECZ entered all ratings in an Excel sheet developed for this purpose and sent the completed file to Cito.

After ECZ 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 (grade 5). The data analyst took the average of the number of items that the panelists aligned to each grade 5 subconstruct, construct and domain. Each item was counted only once (even if it was a partial fit), non-fitting items were not counted towards alignment.

Alignment NAS

Alignment NAS reading English

All results were summarized at the subconstruct level. Only the subconstructs were considered with knowledge and/or skill(s) expected at the grade level for which alignment was being conducted (grade 5). The data analyst took the average of the number of items that the panelists aligned to each grade 5 subconstruct, construct and domain. Each item was counted only once (even if it was a partial fit), non-fitting items were not counted towards alignment.

Averaging the panelists' ratings, we see that on average 21 items of the 22 reading items aligned to Reading comprehension in the GPF. On average fourteen items were aligned to Retrieve information and seven items to Interpret information. However, on average only one item was aligned to Reflect on information. Because at least five items were aligned to Retrieve information and also at least five items to Interpret information, the reading part of the Language assessment is strongly aligned in depth (see Table 4).

We see that on average six subconstructs of Reading comprehension are covered (see in Table 18 in Appendix C) out of eight grade 5 Reading comprehension subconstructs. The reading part of the Language assessment was therefore also strongly aligned in breadth (see the criteria in Table 4).

Table 4. Reading Alignment Criteria for Grades 1–9

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

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

Alignment Mathematics

"When summarizing results to the subconstruct level, facilitators and/or data analysts should only consider the subconstructs with knowledge and/or skill(s) expected at the grade level for which alignment is being conducted" (PLT, p. 15). Averaging the panelists' ratings, on average 41 of the 45 items, aligned to grade 5 subconstructs. For thirteen items, one to four panelists' rated that the item did not fit and with one item more than half of the panelists' indicated that the item did not fit. No fit 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 (PLT, p. 13).

In the Global Proficiency Framework 21 subconstructs are mentioned for grade 5 and the assessment covered twelve of those subconstructs (an average of >0.5, see Table 19 in Appendix C). In breadth the NAS is strongly aligned to the Global Proficiency Framework for Grade 5 as the items covered more than 50% of all grade 5 subconstructs.

The Mathematics items covered four of the five domains (only Statistics and probability was not covered). The assessment covered and 7 out of 12 constructs for grade 5. According to the new criteria in the Policy Linking Toolkit, for strong alignment in Depth at least 5 items should align to the domain Number and Operations, at least 5 items to Measurement and Geometry and at least 5 items to Statistics and Probability and Algebra (see Table 5). On average 27 items covered the domain of Number and Operations, 12 items to Measurement and Geometry, but

only 2 items to Statistics and Probability and Algebra. Therefore, the NAS mathematics is therefore additionally aligned to the GPF in depth.

Table 5. Mathematics Alignment Criteria for Grades 1–9

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

Observations

Not yet going into GPLs and GPDs before the Alignment task worked well: there was more alignment between the information presented and the task to be performed.

Also, it worked well that the facilitators clearly distinguished which information was “nice to know”, i.e., how the results of the work of the panelists will be aggregated and how conclusions will be drawn from this, and information that the panelists really “need to know”, i.e., what is directly necessary in order to carry out the task properly.

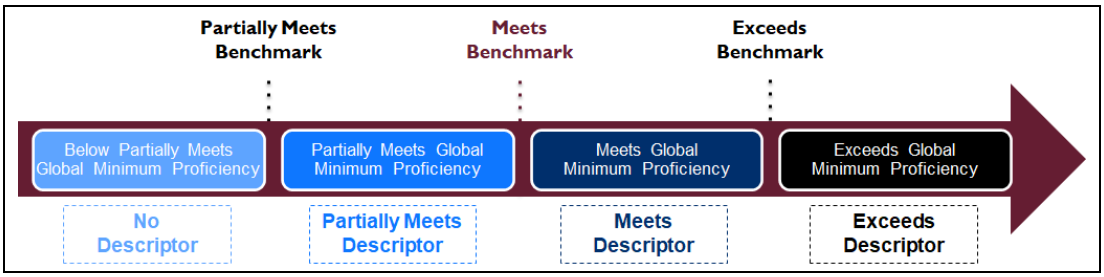
In both groups, the alignment task was finished by the end of the second day. For mathematics, this was despite the fairly large number of items to consider (45). One factor we think certainly contributed to this was the firm guidance on time by the content facilitators. Hereby they carefully tried to avoid putting pressure on the panelists to do things quickly.

Task 2: Matching

On the third day, before the Matching task proper, the GPLs and GPDs of the GPF were introduced in depth, and then discussed further in the subgroups. Only after this was completed to satisfaction, the panelists received training for the next task: Matching the NAT items with the Global proficiency levels and descriptors. After this training, the panelists spent most of days 3 and 4 on this task.

Task 2 builds on the panelists’ understanding of the items and GPF gained through the alignment activity. The purpose of Task 2 is to further narrow down the expectations of learners measured by each assessment item. The panelists should identify the descriptors (GPDs) of global minimum proficiency that match with the items.

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



A Global Proficiency Descriptor is a detailed definition crafted by subject matter experts that clarifies how much of the content described under the statements of knowledge and/or skill(s) in the Global Proficiency Framework a learner should be able to demonstrate within a subject at a grade level. The Global Proficiency Descriptors (GPD) describes the minimum proficiency for the Global Proficiency Levels (GPLs), i.e., the minimum knowledge or skill(s) necessary for classification into each GPL (by grade and subject), see Figure 2.

The Global Proficiency Descriptors are organized by domain, construct and subconstruct, with descriptors for each subconstruct. In Table 6 an example is displayed of Global Proficiency Descriptors for the three GPLs (partially meets, meets and exceed global minimum proficiency).

Table 6. Example of the Global Proficiency Descriptors for three Proficiency Levels.

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

In both groups, consensus was reached on all items.

Observations

The mathematics panelists found several errors in the numbering in the GPF.

- Page 71: N3.3 should be: N3.3.1 (4x)
- Page 72: M1.2.2 should be: M1.2.3 (3x)
- Page 73: M2.2.1 must be: M2.2.2 (3x); M2.2.2 should be: M2.2.3 (3x)
- Page 68: N1.3.3_E talks about “with remainder”. The equivalent at page 79 (N1.3.3-M) talks about “with and without remainder”.

They also indicated that the crosses in table 3 are not always correct: for example, there is no cross, but there is a descriptor at Exceeds level. Is that right? Or does a cross mean there is a descriptor at the Meets level?

In the mathematics group, the idea of the 'lowest GPD' turned out to be very difficult. It was confused with: ‘Grade 5 Meets is Grade 6 Partially meets, the lowest is Partially meets, so you write Grade 6 partially meets’. No, it's about the content of the descriptor, it indicates a lesser amount of knowledge and/or skills. And only with Partial Fit do you take the *highest*. level because that is the *lowest* you *must* have in order to be able to answer the item correctly.

In the language group, the panelists first worked in small groups of 3 or 4 panelists. When all groups had completed the task, the local content facilitator led the group through the discussions. Each group in turn was asked to give their judgment, upon which the other groups responded. By rotating turns, the local content facilitator ensured that every group was given ample opportunity to voice their opinions.

In the mathematics group, they proceeded by groups of 5 items: form your own opinion, then discuss them one by one to come to consensus.

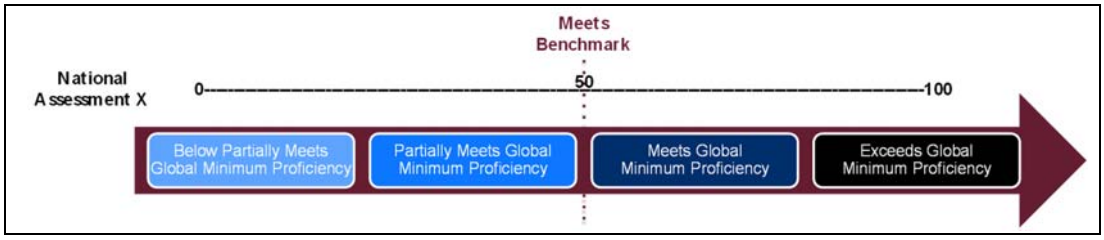
They needed the whole of day 4 to finish the matching task, but they did finish it on day 4, thanks to the perseverance of all.

The practice items were of grade 6 whereas the NAS was on grade 5. This necessitated an extra translation step for the panelists, in that what they did in practice was not 1-on-1 applicable in the task.

Task 3: Benchmarking

On the fifth day the panelists were given an introduction into global benchmarking and they received training in setting global benchmarks using the Angoff method. The facilitator first presented a hypothetical example of how the benchmarking method would link a national assessment to the GPF, thus allowing for the calculation of the percentages of students attaining minimum proficiency (see Figure 3). This example was extended to three national assessments of different difficulties, and how this would lead to a different benchmark for each assessment. The facilitators discussed how the benchmarking results – when applied to the assessment data sets – could be used for comparing and aggregating assessment results, as well as tracking those results over time.

Figure 3. Example of an assessment and a benchmark



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

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

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

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

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

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

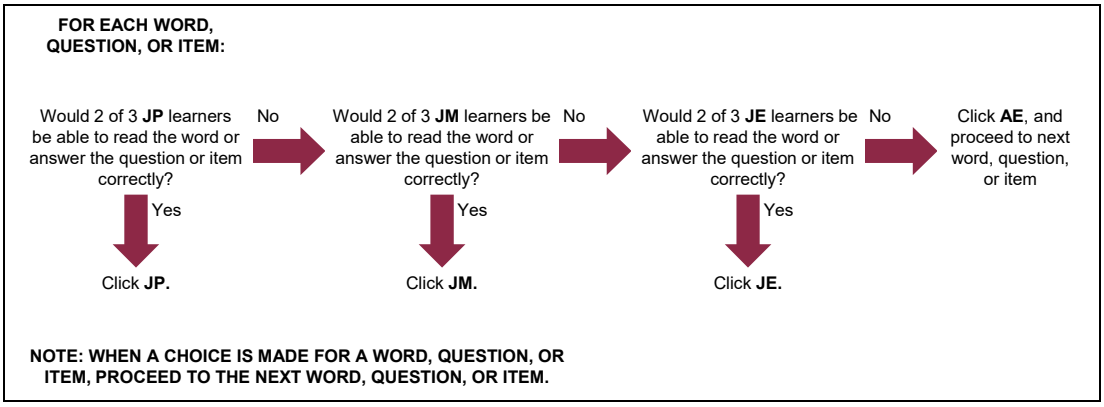
- If "yes," place an "X" under JP and proceed to the next item.
- If "no," ask whether minimally proficient JM learners would be able to answer the item correctly?
    - If "yes," place an "X" under JM and proceed to the next item.
    - If "no," ask whether minimally proficient JE learners would be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

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

Round 1

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

Figure 4. Steps for Rating Items



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

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

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

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

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

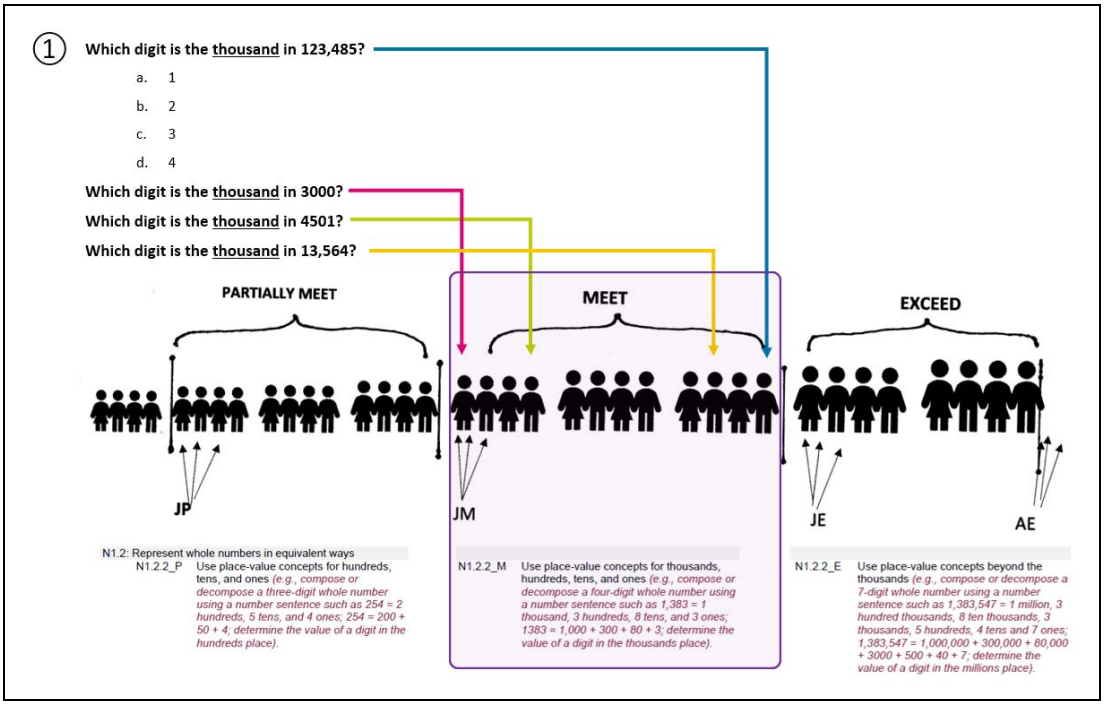
Round 2

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

Observations

It remains quite a task for the participants to get a clear picture of what exactly they have to do. Also, how they can (should?) use the results of the Matching task for this. For the mathematics group, the content facilitators used several pictorial representations of learners in an attempt to clarify this further. An example is contained in Figure 5. In the reading group, similar presentations were used.

Figure 5. Pictorial representation of differences within a level with example items



We observed that primary school teachers of English cannot always fully grasp the terminology used in the GPF, particularly in table 5 where the differences in the levels are extremely subtle and much more open to interpretation. (e.g. the difference between P/M/E could depend on: P=….in consecutive sentences when there is no competing information…./ M=……..from a paragraph, but not in consecutive sentences, when there is limited competing information…. / E= …from one or more paragraphs when there is more distance between the pieces of information and/or a lot of competing information.)

The results of round 1 for language turned out to be rather disappointing (see below). Possible explanations were discussed between the facilitators and the data analysts. The most important idea was the fact that in Zambia English is a national language, but not the mother tongue of the learners, and learners in grade 5 have only had a maximum of 2 years of English at that point.

The results of round 2 for language were considerably “better”, i.e., fewer learners in the below partially meets level, and also seemed to be in line with the results from the MILO.

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.

4. Results of the benchmarking

Round 1

Cito received an Excel file with the item ratings of 16 workshop participants for reading English and of 13 workshop participants for Mathematics. The data analysts produced summary tables and graphs from the first round, which showed the initial benchmarks, score ranges, and impact data for each Minimum Proficiency Level (see Table 7 and Table 8). In the plenary room the panelists were presented with anonymous normative information on the panelists ratings (see Figure 6 and Figure 7). We saw that the ratings of panelists varied considerably for mathematics. In contrast, for reading we see a clear ceiling effect with reading. Exceeds is at the maximum (22) for all panelists. The majority of the panelists (12) also put the Meets benchmark at the maximum. Only the Partially meets benchmark shows more variation.

Figure 6 Anonymous information on the panelists' ratings of reading English Round 1

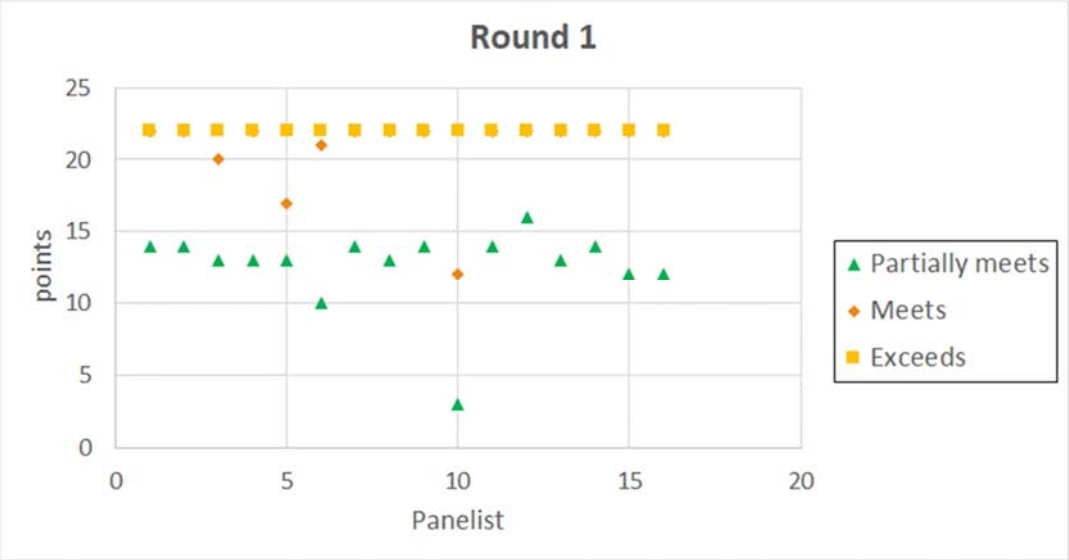
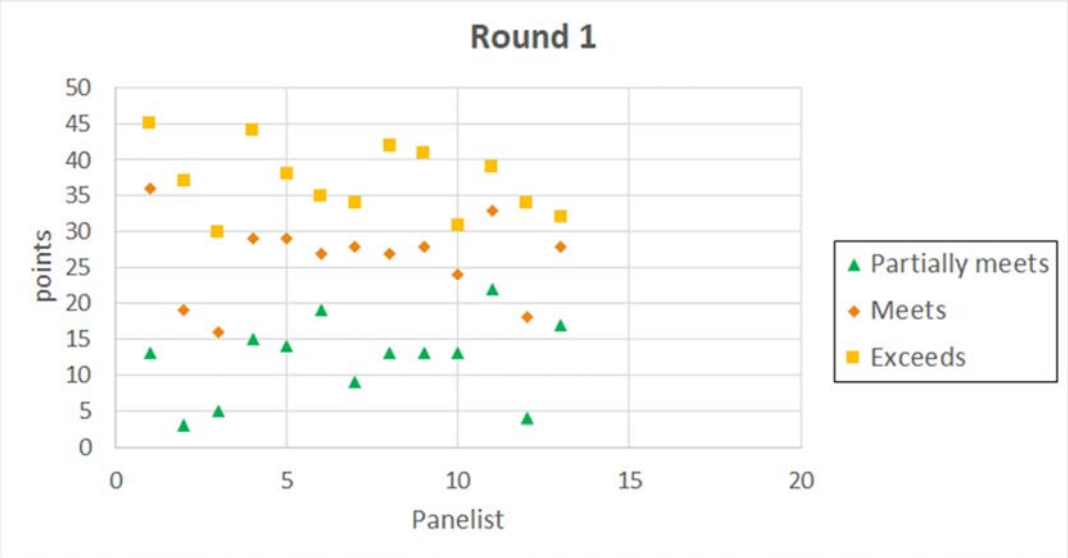


Figure 7 Anonymous information on the panelists' ratings of Mathematics Round 1



After round 1 the benchmark was calculated as the average of the panelists' benchmarks. The average benchmark was rounded down, as stipulated in the policy linking toolkit. For reading, the impact information shows 84% of the learners would fall in the Below Partially Meets Proficiency level and 14% in the Partially Meets level (see Table 7).

Table 7 Reading English round 1 benchmarks, score range and impact for pupils taking the NAS

Minimum Proficiency Levels	Round 1 Benchmark	Score Range	Percentage of pupils		
			Female	Male	Total
Below Partially Meets	N/A	0 - 11	82.7%	84.4%	83.6%
Partially Meets	12.63	12 - 19	15.3%	13.5%	14.4%
Meets	20.88	20 - 21	1.6%	1.5%	1.6%
Exceeds	22	22 - 22	0.4%	0.6%	0.5%

For Mathematics, the impact information shows that only 28 percent would fall in the Below Partially Meets Proficiency level and more than half (66%) would fall in the Partially Meets level (see Table 8) using round 1 benchmarks.

Table 8 Mathematics round 1 benchmarks, score range for only dichotomous items and impact for pupils taking the NAS

Minimum Proficiency Levels	Round 1 Benchmark	Score Range	Percentage of pupils		
			Female	Male	Total
Below Partially Meets	N/A	0 - 11	25.0%	25.0%	24.9%
Partially Meets	12.31	12 - 25	68.0%	68.8%	68.4%
Meets	26.31	26 - 36	6.3%	5.7%	6.0%
Exceeds	37.08	37 - 45	0.7%	0.6%	0.7%

Round 2

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

In round 2 with reading (Figure 8), we see the ratings went down, but still a ceiling effect remains for the Exceeds benchmark with almost all panelists putting the benchmark at the maximum of 22. For mathematics, we see that in Round 2 the ratings of panelists varied less than in Round 1 (Figure 9).

Figure 8 Anonymous information on the panelists' ratings of reading Round 2

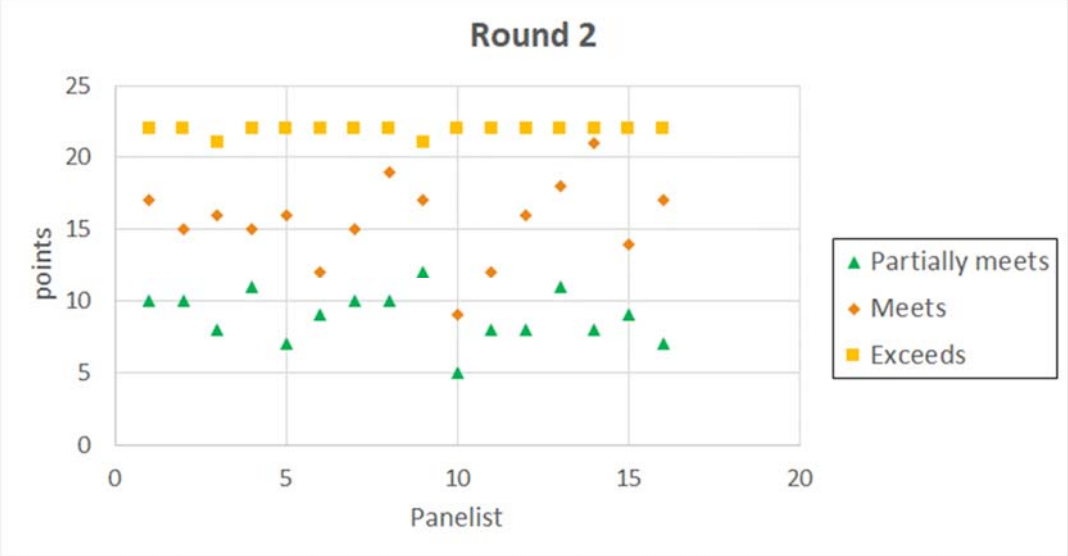
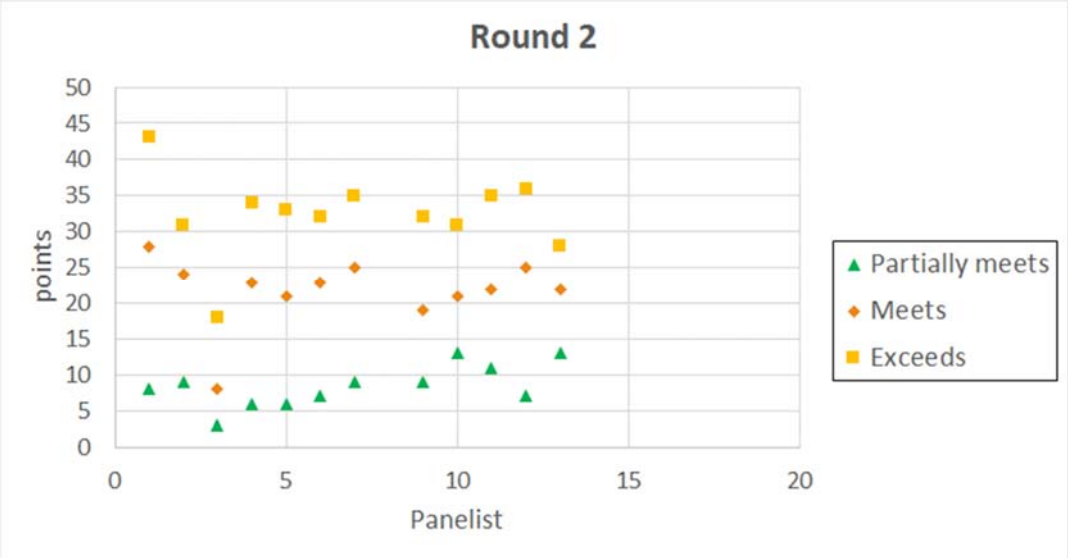


Figure 9 Anonymous information on the panelist's ratings of Mathematics Round 2



For reading English, the results show that in Round 2 more than half of the students (58%) fall in the Below Partially Meets level (see Table 9). More than half of the learners who took the NAS (58%) fall in the Below Partially Meets level. One third (33%) falls in the Partially Meets level and 9% in the Meets or Exceeds level. The benchmarks were set lower in round 2 than in round 1. Consequently, after round 2 a lower percentage of learners falls in the Below Partially meets proficiency level. The Exceeds benchmark is set at the top of the scale, which is a clear ceiling effect.

Table 9 English reading round 2 benchmarks, score range and impact for pupils taking the NAS

Minimum Proficiency Levels	Round 1 Benchmark	Score Range	Percentage of pupils		
			Female	Male	Total
Below Partially Meets	N/A	0 - 7	55.9%	58.7%	57.2%
Partially Meets	8.94	8-14	34.5%	33.1%	33.9%
Meets	15.56	15 - 20	8.6%	6.9%	7.8%
Exceeds	21.88	21 - 22	1.0%	1.2%	1.1%

Table 10. Comparison of Round 1 benchmarks and Round 2 benchmarks for English reading

Minimum Proficiency Levels	Round 1 Benchmark	Round 1 Percentage of pupils	Round 2 Benchmark	Round 2 Percentage of pupils
Below Partially Meets	N/A	83.6%	N/A	57.2%
Partially Meets	12.63	14.4%	8.94	33.9%
Meets	20.88	1.6%	15.56	7.8%
Exceeds	22	0.5%	21.88	1.1%

For Mathematics, we see that in Round 2 the benchmarks were set at a lower score (see Table 12). Consequently, after round 2 a higher percentage of learners falls in the Partially Meets and Meets proficiency level. A lower percentage of learners than was the case in Round 1 fall in the Below Partially Meets level. Five percent of the learners fall in the Below Partially Meets level (Table 12). More than three quarters of the students (80%) fall in the Partially Meets level, 14% in the Meets level and 2% in the Exceeds level (which might be a ceiling effect).

Table 11. Mathematics round 2 benchmarks, score range for only dichotomous items and impact for pupils taking the NAS

Minimum Proficiency Levels	Round 1 Benchmark	Score Range	Percentage of pupils		
			Female	Male	Total
Below Partially Meets	N/A	0 - 7	4.0%	3.3%	3.9%
Partially Meets	8.42	8-20	80.2%	79.0%	79.6%
Meets	21.75	21 - 31	13.9%	16.0%	14.7%
Exceeds	32.33	32 - 45	1.9%	1.7%	1.8%

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

Minimum Proficiency Levels	Round 1 Benchmark	Round 1 Percentage of pupils	Round 2 Benchmark	Round 2 Percentage of pupils
Below Partially Meets	N/A	24.9%	N/A	3.9%
Partially Meets	12.31	68.4%	8.42	79.6%
Meets	26.31	6.0%	21.75	14.7%
Exceeds	37.08	0.7%	32.33	1.8%

5. Evaluation of the Standard Setting Process

Internal Evaluation SEM, Panelist Consistency and Panelists' Agreement

In addition to calculating benchmarks and impact data, the Policy Linking Toolkit also requires calculating measures of consistency and presenting evaluation feedback results. These measures of consistency are reported in Table 13 and Table 14.

As shown in Table 13, the Standard Error of Measurement (SEM), which measures how much panelists' benchmarks are spread around a "true" benchmark, was in both rounds under 1.0 for reading English with 22 items and under 2.00 for Mathematics with 45 items. The results show that the Standard Error of Measurement is smaller for the Exceeds benchmarks of reading English (even zero in round 1 for reading). This is a consequence of a ceiling effect for this benchmark. For reading, almost all panelists have put the Exceeds benchmark at the maximum sum score (see the previous section).

Table 13. Standard Error of Measurement by Round

	SEM by Benchmark					
	Round 1			Round 2		
Subjects	Partially Meets	Meets	Exceeds	Partially Meets	Meets	Exceeds
English reading	0.72	0.68	0	0.45	0.72	0.09
Mathematics	1.59	1.59	1.37	0.85	1.42	1.68

The results show that the inter-consistency for both English reading and Mathematics was higher in Round 2 than in Round 1. The inter-rater consistency index evaluates the panelists' overall agreement or consensus across all possible pairs of panelists. Inter-rater consistency is calculated at the item level and for the entire assessment. The value ranges between 0 and 1. According to the Policy Linking Toolkit values of 0.80 or greater are desirable, as they indicate substantial agreement between the panelists. For English reading the interrater consistency was equal or above the 0.80 in round 2 (see Table 14), for Mathematics it was a little bit lower (0.76).

The intra-rater consistency index evaluates the panelists' overall consistency in estimating item difficulty. Intra-rater consistency is calculated for each panelist across all items on the assessment. The value ranges between 0 and 1. A lower value indicates low consistency and a higher value indicates high consistency. We see that the intra-rater consistency for English reading in round 2 is quite high (higher than .80) and for mathematics somewhat lower (given the scale of 0 to 1).

Table 14. Inter-rater consistency and intra-rater consistency by subject and round

	Round 1		Round 2	
Subjects	Inter-Rater Consistency	Intra-Rater Consistency	Inter-Rater Consistency	Intra-Rater Consistency
English reading	0.90	0.88	0.87	0.86
Mathematics	0.74	0.81	0.76	0.76

Procedural Evaluation

All panelists shared their opinion about the workshop through a questionnaire (see Annex F). The panelists indicated on a five-point scale (Strongly Disagree-Disagree-Neutral-Agree-Strongly Agree) how strongly they agreed with several statements about six aspects of the

workshop. On average, we see that the respondents were positive about the workshop. All six aspects received an average score of 4 or higher (on a scale of 1 to 5). The overall evaluation shows that the panelists are overall positive: 4.1 on a scale of 1 to 5 (the neutral category has been added to the scale, which was missing in the example in the Policy Linking Toolkit).

Table 15. Workshop evaluation results

Part of the workshop	Scale	Number of statements	Average scale score	Standard deviation of scale score	N
The training on the Global Proficiency Framework	1-5	8	4,5	0,8	29
The training on the NAS[2]	1-5	5	4,3	0,9	28
The training on the alignment methodology	1-5	5	4,5	0,4	29
The training on the matching methodology	1-5	5	4,3	0,7	29
The training on the benchmark-setting (Angoff) methodology[3]	1-5	11	4,4	0,5	29
Benchmark Round 2 evaluation	1-5	8	4,0	0,5	29
Overall evaluation	1-5	3	4,1	0,7	29

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

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

The results of the policy linking workshop in Zambia are summarized in Table 16 and Table 17. In the policy linking toolkit (Annex U, p. 164) six criteria are mentioned for the validity of policy linking workshop. The evaluation of the validity is based on the intra-rater and inter-rater reliability, the standard error of measurement, the representativeness of the panel and panelists' understanding of the procedures.

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

For reading English (Table 16), the inter-rater reliability and the intra-rater reliability meet the requirements. However, the third benchmark ("Exceeds") might not be valid. Almost all panelists put the Exceeds benchmark at the maximum, so there is little variation and a clear ceiling effect (even though this is not mentioned as a criterium). The panel represents all ten provinces. More panelists are from the Lusaka province (5) than from the other provinces (each 1 or 2 panelists). All but three panelists are female. Ten of the panelists are teachers, but their experience is unknown. The panelists rated their understanding of the GPF, assessment, and policy linking methodology for almost all aspects above 4 and they felt on average comfortable with their Round 2 evaluations and final benchmarks. The adequacy of the policy linking workshop for English in Zambia can be considered to be good.

For mathematics (Table 17), the intra-rater and inter-rater reliability are slightly below 0.80. The standard error of measurement is moderate given the number of items. The panel all ten provinces. From each province one or two panelists participated. There are slightly more male panelists than female. Ten of the panelists are teachers, but their experience is unknown. The panelist rated their understanding of the GPF, assessment, alignment and matching for almost all aspects above 4. They felt on average fairly comfortable with their Round 2 evaluations and final benchmarks (3.2). The adequacy of the policy linking workshop for mathematics in Zambia can be considered to be satisfactory.

Table 16. Summary of Results for Criteria for Policy Linking Validity reading English Grade 5

Question	Criteria	Response
a) What was the intra-rater reliability for the second round of ratings?	The intra-rater reliability will vary depending on the number of items on the assessment. The panel will provide guidance on how they determined acceptability.	0.86
b) What was the inter-rater reliability for the second round of ratings?	The inter-rater reliability should be at least .80.	0.87
c) What was the Standard Error of Measurement (SEM) at each global proficiency level?	SEM should be appropriate for each global proficiency level reported. There is no maximum SEM provided in this document, since it will depend on the number of items in the assessment.	Number of items: 22 0.45 (Partially Meets) 0.72 (Meets) 0.09 (Exceeds)
d) To what extent were the panelists representative of the target population of schools being reported on?	Panelists should be selected to ensure: • Gender representation – The panelists must be selected to ensure gender balance, both for the teachers and non-teachers. • Geographical representation – The teachers (and non-teachers, if possible) must be selected to ensure representation from regions, provinces, and/or states. • Ethnic and/or linguistic representation (where applicable) • Representation of crisis-and-conflict-affected areas.	• 81% female, 19% male • From all ten provinces one to five teachers (6% to 31%)
e) To what extent did the panelists meet the other selection criteria described in the Policy Linking Toolkit?	Panelists should all have: • Several years of teaching experience in the grade level for which they are providing ratings (classroom teachers) • Skills in the subject area (all panelists) • Skills in the different languages of instruction and assessment (all panelists) • Knowledge of learners of different proficiency levels, including at least some who would meet the requirements of the meets minimum proficiency level and some who would meet the requirements of the exceeds minimum proficiency level (all panelists) • Knowledge of the instructional environment (all panelists) • Experience administering the assessment(s) being used for the policy linking workshop.	• 63% currently employed as teacher

f)	To what extent did panelists report understanding the GPF, assessment, and policy linking methodology? And, to what extent did they feel comfortable with their Round 2 evaluations and final benchmarks?	On a five-point Likert scale, with 1 being strongly disagree, very uncomfortable, etc. and 5 being strongly agree, very comfortable, etc., the average rating for each of these criteria should be 4 or above.	**GPF** • I understand the purpose of the GPF – 4.3 • I understand the relationship between domains, constructs, subconstructs, knowledge and skills, and GPDs - 4.4 • The GPDs were clear and easy to understand – 3.8 **NAS** • I understand the purpose of the assessment - 4.5 • I understand the constructs assessed in the assessment - 4.5 • I understand how the assessment is administered - 4.3 **Alignment** • I understand the purpose of alignment - 4.7 • I understand the alignment methodology - 4.4 • I understand the difference between no fit, partial fit, and complete fit - 4.8 **Matching** • I understand the purpose of matching - 4.5 • I understand the matching methodology - 4.3 • I understand how the alignment activity links to the matching activity - 4.3 **Benchmarking methodology** • I understand the process I need to follow to complete the benchmarking exercise - 4.4 • I understand how the benchmarking methodology links to the steps on alignment and matching - 4.4 • I understand the difficulty level of the assessment items - 4.6 **Benchmark round 2** • I understand the data on others' ratings - 4.5 • I understand the item difficulty data and how it relates to this process - 4.3 • I understand the impact data and how it relates to this process - 3.9 **Comfortable with Round 2** • How comfortable are you with your final performance predictions? - 3.9

Table 17. Summary of Results for Criteria for Policy Linking Validity Mathematics Grade 5

Question	Criteria	Response
a) What was the intra-rater reliability for the second round of ratings?	The intra-rater reliability will vary depending on the number of items on the assessment. The panel will provide guidance on how they determined acceptability.	0.76
b) What was the inter-rater reliability for the second round of ratings?	The inter-rater reliability should be at least .80.	0.76
c) What was the Standard Error of Measurement (SEM) at each global proficiency level?	SEM should be appropriate for each global proficiency level reported. There is no maximum SEM provided in this document, since it will depend on the number of items in the assessment.	Number of items: 45 0.85 (Partially Meets) 1.42 (Meets) 1.68 (Exceeds)
d) To what extent were the panelists representative of the target population of schools being reported on?	Panelists should be selected to ensure: • Gender representation – The panelists must be selected to ensure gender balance, both for the teachers and non-teachers. • Geographical representation – The teachers (and non-teachers, if possible) must be selected to ensure representation from regions, provinces, and/or states. • Ethnic and/or linguistic representation (where applicable) • Representation of crisis-and-conflict-affected areas.	• 38% female, 62% male • From all ten provinces one to two teachers (6% to 13%)
e) To what extent did the panelists meet the other selection criteria described in the Policy Linking Toolkit?	Panelists should all have: • Several years of teaching experience in the grade level for which they are providing ratings (classroom teachers) • Skills in the subject area (all panelists) • Skills in the different languages of instruction and assessment (all panelists) • Knowledge of learners of different proficiency levels, including at least some who would meet the requirements of the meets minimum proficiency level and some who would meet the requirements of the exceeds minimum proficiency level (all panelists) • Knowledge of the instructional environment (all panelists) • Experience administering the assessment(s) being used for the policy linking workshop.	• 77% currently employed as teacher

f)	To what extent did panelists report understanding the GPF, assessment, and policy linking methodology? And, to what extent did they feel comfortable with their Round 2 evaluations and final benchmarks?	On a five-point Likert scale, with 1 being strongly disagree, very uncomfortable, etc. and 5 being strongly agree, very comfortable, etc., the average rating for each of these criteria should be 4 or above.	**GPF** • I understand the purpose of the GPF - 4.6 • I understand the relationship between domains, constructs, subconstructs, knowledge and skills, and GPDs - 4.8 • The GPDs were clear and easy to understand - 4.2 **NAS** • I understand the purpose of the assessment - 4.1 • I understand the constructs assessed in the assessment - 4.1 • I understand how the assessment is administered - 4 **Alignment** • I understand the purpose of alignment - 4.7 • I understand the alignment methodology - 4.3 • I understand the difference between no fit, partial fit, and complete fit - 4.7 **Matching** • I understand the purpose of matching - 4.3 • I understand the matching methodology - 4.2 • I understand how the alignment activity links to the matching activity - 4 **Benchmarking methodology** • I understand the process I need to follow to complete the benchmarking exercise - 4.2 • I understand how the benchmarking methodology links to the steps on alignment and matching - 4.3 • I understand the difficulty level of the assessment items - 4.3 **Benchmark round 2** • I understand the data on others' ratings - 3.8 • I understand the item difficulty data and how it relates to this process - 4.4 • I understand the impact data and how it relates to this process - 4 **Comfortable with Round 2** • 22. How comfortable are you with your final performance predictions? - 3.2

7. Conclusions and Recommendations

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

After getting used to this mode the first day, the participants engaged in discussions regarding the alignment of the NAS items with the Global Proficiency Framework, the matching and the Item ratings. The participants performed their tasks with critical dedication. Every step of the process produced important outcomes. The participants gave positive feedback in their evaluation forms. In this respect the piloting of the policy linking workshop in this hybrid mode can be considered a success. In fact, this hybrid mode, which arose out of Covid limitations, could also be an option for other situations where travel by international facilitators to the country is deemed not to be feasible or desirable.

The participants' work showed that the NAS for Reading is strongly aligned to the Global Proficiency Framework both in depth and in breath. The participants' work showed that the NAS for Mathematics is in additionally aligned in depth to the Global Proficiency Framework for grade 5 and strongly aligned in breadth. Furthermore, the panelists managed to reach complete consensus on the matching both for reading and for mathematics. The final benchmarks of the panelists show a satisfactory to good consistency, which makes the benchmarks useable for comparing, aggregating, and tracking learning outcomes for the NAS in Zambia. We did find for both reading and mathematics a ceiling effect, for mathematics we also found a bottom effect. Irrespective of these effects, the benchmarks for the "Meets" category seem valid. The benchmarks can be used to estimate with IRT models the impact at a population level using the data of all students and sampling weights.

For reading, based on the benchmarks obtained, 57 percent of the learners who took the NAS fall in the Below Partially Meets level, 34 percent in the Partially Meets level, 8 percent in the Meets level and 1 percent in the Exceeds level (see Table 9). For mathematics, based on the benchmarks obtained, 4 percent of the learners who took the NAS fall in the Below Partially Meets level, 80 percent in the Partially Meets level, 15 percent in the Meets level and 2 percent in the Exceeds level (see Table 11).

We cannot avoid to question the suitability of the GPF for English where it is not the first language. In this respect the situation in Zambia is similar to a number of other African countries. One of the consequences at least is that one must be careful in interpreting the results, especially in comparing them over countries.

Recommendations

Based on Cito's observations during the workshop, several lessons can be drawn that are useful for coming workshops.

Workshop Preparation

Collecting workshop materials and pre-workshop analyses

- It would be good if organizers and facilitators would be relieved from other duties as much as possible in the preparation of the workshop, or at least that they allocate and block sufficient time in their agendas for this. This holds particularly (but not exclusively) for the lead facilitator.

- To support this, we recommend that at the start of the preparation period – about 7 weeks before the workshop – a full preparation program is (developed,) agreed to and scheduled.
- If internet and power problems are likely expected to occur, there should be fall back options in the preparation program to alleviate the impact of these on the preparation.
- We recommend that the country unambiguously selects the assessment to be used for Policy linking right at the moment when committing to holding a Policy Linking Workshop. This notwithstanding that in the process it might turn out that the selected assessment is not so suitable or that another assessment is more suitable.

Creating workshop materials

- In the PLT, "workshop materials" is typically interpreted as materials for the panelists. It is good that the focus here is on the panelists, in the form of the Panelist package. However, there are other important materials, for the content facilitators or for the data analysis, for example the Matching form, which get less attention in the preparation. We recommend to make the overview of workshop materials more complete and make Facilitator packages and an Analysis package too.
- We recommend to incorporate as a standard that the GPF be translated into the panelists' L1 if this is not English. This even if one of the official languages of the country is English. Primary school teachers whose L1 is not English cannot always fully grasp the English terminology used in the GPF, particularly in table 5 where the differences in the levels are extremely subtle and open to interpretation.

Training the local content facilitators

- Similar to the general preparation, it would be good if the content facilitators would be relieved from other duties as much as possible for the training, or at least that they allocate and block sufficient time in their agendas for this.
- Again similar to the general preparation, if internet and power problems are likely expected to occur, there should be fall back options in the training program to alleviate the impact of these on the training.
- The content facilitator training tries to achieve two goals: to inform the content facilitators about the policy linking, and to prepare them for their role in the workshop. It is not much of an exaggeration to say that they are more or less supposed to achieve in 3 online sessions what the panelists take the whole workshop for. To alleviate this and to make things more manageable for the content facilitators, we recommend to limit the general information and to focus the training on the tasks of the content facilitators during the workshop. The rest of the information will come to them during the workshop anyway.
- It was very helpful that some of the local content facilitators had previous experience with standard setting. We recommend to recruit content facilitators with such experience where possible.

Implementing the blended workshop

Familiarization

We recommend that in the presentation on the GPF, more attention is given to table 2, and that the information to the panelists, both in the presentation and in the subject group meetings, is broken down into more bite-sized chunks.

We recommend to spend more time on the familiarization with the GPF. The better understanding of the GPF of the panelists increases the quality of the results of the tasks, might even save some time later in the workshop, and increases the ability of the panelists to use

what they have learned in their own situation. We think these benefits even warrant extending the workshop to 7 days, or at least to make the standard length of a workshop more than 5 days.

We recommend a check on the consistency of the GPF, particularly in the numbering and in the cross referencing.

We recommend to incorporate in the agenda of the workshop a daily meeting at the end of the day of all facilitators, to look back on and learn from the day and to look forward to and prepare for the next day.

Task 1: Alignment

We recommend that the GPLs and GPDs are mentioned before the Alignment task, but only introduced in depth just before the Matching task.

We recommend that the information that is “nice to know” is clearly distinguished from the information that the panelists “need to know”, both in the presentations and in the emphasis that the facilitators put on each.

We were really happy that we were able to finish each task at the end of a day. We recommend to keep this as a guideline in preparing the detailed agenda for a workshop. We also recommend to limit the number of items to be used for a workshop to make this comfortably possible, to something like 30 items per subject. This can be achieved by designing the assessment to this length, or, if the assessment must be longer for other reasons, to translate the benchmarks on the 30 items to benchmarks on the whole assessment using psychometric techniques.

Task 2: Matching

We recommend in the GPF for mathematics a check on the cross-grade structure, especially in the consistency of this. It is not necessary that this is 100% consistent, but it should be understandable to the panelists why the deviations occur and how they should be interpreted. It would also be good if the interpretation of the “crosses” in table 3 and the “N/A” in table 5 could be clarified.

We recommend in the presentation to spend more time clarifying what is the lowest GPL and GPD that are most appropriate for the item. This is, admittedly, mentioned in the facilitator notes with the slide, but perhaps it would be good to highlight this case in the slide itself.

We recommend that the content facilitators go out of their way to involve the panelists in the workshop, and to involve *all* panelists in the workshop. This by being given and feeling the freedom to organize interaction and exchange of arguments in creative ways.

We recommend that example items in the presentations be as relevant as possible to the panelist. Hence: mathematics examples for the mathematics group, language examples for the language group, examples from the same grade as the grade of the assessment, also examples from neighboring grades to illustrate how to deal with those, and preferably “local” items, for example from another booklet of the same assessment, or items from a practice test for the assessment.

Task 3: Benchmarking

We recommend to evaluate and/or reconsider in which way the results of the Matching task are to be used in the Benchmarking task. In this workshop, we have done this by using for each

item the result of the matching as a focal point for the benchmarking, see Figure 5 above. However, this way is not spelled out in the PLT.

The presentation on setting global benchmarks – and the Benchmarking task in general – presumes prior knowledge about the meaning of what a benchmark is or does. We recommend not taking for granted that panelists are familiar with the word “benchmark” and/or the use of this word in assessment, and to introduce the panelists (more) gently into standard setting and the terminology used.

We recommend to (re)consider carefully the validity and viability of Policy Linking for language and the results of such Policy Linking when the language of the assessment is not the native language of the learners.

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.

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

UNESCO Institute for Statistics (2022). COVID-19 in Sub-Saharan Africa: Monitoring impacts on learning outcomes. Montreal, Canada. Downloaded from https://research.acer.edu.au/cgi/viewcontent.cgi?article=1054&context=monitoring_learning

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.

9. Annexes

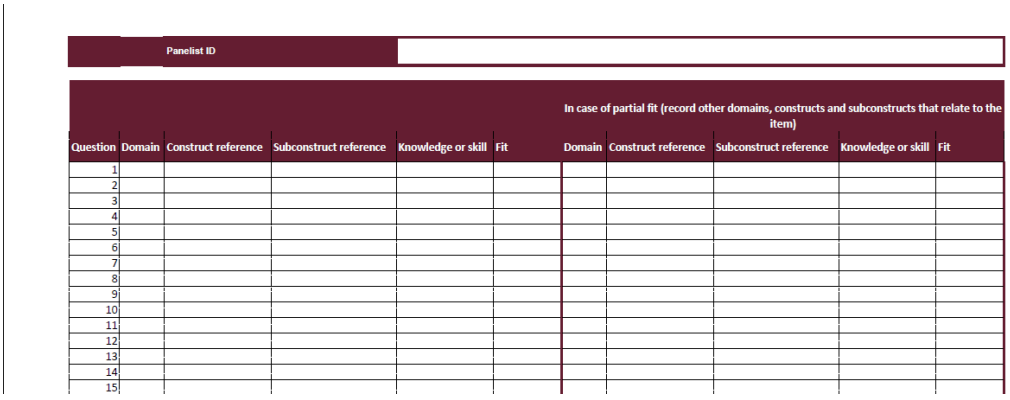
Annex A: Agenda for the workshop

HYBRID WORKSHOP ZAMBIA				
Day	Time	Durat	Activity	Facilitation
			DAY 1: Monday, May 9	
1 Day 1	9:00 - 09:30	**0:30**	Registration	Project team
2 Day 1	9:30 - 10:00	**0:30**	Welcome and introductions	UIS, Cito, ECZ
3 Day 1	10:00 - 10:15	**0:15**	Morning tea break	
4 Day 1	10:15 - 11:00	**0:45**	Presentation: Overview of policy linking	Lead facilitator
5 Day 1	11:00 - 12:00	**1:00**	Presentation: Overview of the GPF	Lead facilitator
6 Day 1	12:00 - 13:00	**1:00**	Activity: GPF Review Knowledge and Skills	Content facilitators
7 Day 1	13:00 - 14:00	**1:00**	Lunch break	
8 Day 1	14:00 - 15:00	**1:00**	Presentation: Overview of the NAS	ECZ
9 Day 1	15:00 - 16:00	**1:00**	Activity: Taking the NAS	Content facilitators
10 Day 1	16:00 - 16:15	**0:15**	Afternoon tea break	
11 Day 1	16:15 - 16:45	**0:30**	Activity: Review NAS items	Content facilitators
12 Day 1	16:45 - 17:00	**0:15**	Looking forward to next day & closing	Lead facilitator
			DAY 2: Tuesday, May 10	
13 Day 2	9:00 - 09:15	**0:15**	Introduction of Day 2 and solving issues of Day 1	Lead facilitator
14 Day 2	9:15 - 10:00	**0:45**	Activity: Review NAS items (cont.)	Content facilitators
15 Day 2	10:00 - 10:15	**0:15**	Morning tea break	
16 Day 2	10:15 - 12:00	**1:45**	Activity: Review GPF and identify any elements that are still unclear	Content facilitators
17 Day 2	12:00 - 13:00	**1:00**	Discussion of taking the NAS and reviewing GPF	Content facilitators
18 Day 2	13:00 - 14:00	**1:00**	Lunch break	
19 Day 2	14:00 - 15:00	**1:00**	Task 1 Presentation: GPF and alignment	Lead facilitator
20 Day 2	15:00 - 16:00	**1:00**	Task 1 Activity: Small group discussions on first 5 items	Content facilitators
21 Day 2	16:00 - 16:15	**0:15**	Afternoon tea break	
22 Day 2	16:15 - 16:45	**0:30**	Task 1 Activity: Alignment of NAS and the GPF	Content facilitators
23 Day 2	16:45 - 17:00	**0:15**	Looking forward to next day & closing	Lead facilitator
			DAY 3: Wednesday, May 11	
24 Day 3	9:00 - 09:15	**0:15**	Introduction of Day 3 and solving issues of Day 2	Lead facilitator
25 Day 3	9:15 - 10:00	**0:45**	Task 1 Activity: Alignment of NAS and the GPF (cont.)	Content facilitators
26 Day 3	10:00 - 10:15	**0:15**	Morning tea break	
27 Day 3	10:15 - 11:15	**1:00**	Presentation: GPF GPLs and GPDs	Lead facilitator
28 Day 3	11:15 - 13:00	**1:45**	Activity: GPF Review GPLs and GPDs	Content facilitators
29 Day 3	13:00 - 14:00	**1:00**	Lunch break	
30 Day 3	14:00 - 15:00	**1:00**	Task 2 Presentation: Matching NAS and GPDs/GPLs	Content facilitators
31 Day 3	15:00 - 16:00	**1:00**	Task 2 Activity: Matching NAS items and GPDs/GPLs	Content facilitators
32 Day 3	16:00 - 16:15	**0:15**	Afternoon tea break	
33 Day 3	16:15 - 16:45	**0:30**	Task 2 Activity: Matching NAS items and GPDs/GPLs	Content facilitators
34 Day 3	16:45 - 17:00	**0:15**	Looking forward to next day & closing	Lead facilitator
			DAY 4: Thursday, May 12	
35 Day 4	9:00 - 09:15	**0:15**	Introduction of Day 4 and solving issues of Day 3	Lead facilitator
36 Day 4	9:15 - 09:30	**0:15**	Task 1 Presentation: Alignment results	Lead facilitator
37 Day 4	9:30 - 10:00	**0:30**	Task 2 Activity: Matching NAS items and GPDs/GPLs (cont.)	Content facilitators
38 Day 4	10:00 - 10:15	**0:15**	Morning tea break	
39 Day 4	10:15 - 12:00	**1:45**	Task 2 Activity: Matching NAS items and GPDs/GPLs (cont.)	Content facilitators
40 Day 4	12:00 - 13:00	**1:00**	Task 2 Plenary discussion: Matching NAS items and GPDs/GPLs and results of matching	Content facilitators
41 Day 4	13:00 - 14:00	**1:00**	Lunch break	
42 Day 4	14:00 - 16:00	**2:00**	Task 2 Plenary discussion: Matching NAS items and GPDs/GPLs and results of matching (cont.)	Content facilitators
43 Day 4	16:00 - 16:15	**0:15**	Afternoon tea break	
44 Day 4	16:15 - 16:45	**0:30**	Task 3 Presentation: Global benchmarking	Lead facilitator
45 Day 4	16:45 - 17:00	**0:15**	Looking forward to next day & closing	Lead facilitator
46 Day 4	17:00 - 18:00	**1:00**	Consultation hour in which panelists can consult the content facilitator	Content facilitators

				DAY 5: Friday, May 13	
47	Day 5	9:00 - 09:15	**0:15**	Introduction of Day 5 and solving issues of Day 4	Lead facilitator
48	Day 5	9:15 - 10:00	**0:45**	Task 3 Presentation: Angoff method	Lead facilitator
49	Day 5	10:00 - 10:15	**0:15**	Morning tea break	
50	Day 5	10:15 - 11:00	**0:45**	Task 3 Activity: Angoff practice	Content facilitators
51	Day 5	11:00 - 13:00	**2:00**	Task 3 Activity: Angoff Round 1	Content facilitators
52	Day 5	13:00 - 14:00	**1:00**	Lunch break	
53	Day 5	14:00 - 16:00	**2:00**	Task 3 Activity: Angoff Round 1 (cont.)	Content facilitators
54	Day 5	16:00 - 16:15	**0:15**	Afternoon tea break	
55	Day 5	16:15 - 16:45	**0:30**	Task 3 Presentation: Angoff Round 1 results	Lead facilitator
56	Day 5	16:45 - 17:00	**0:15**	Looking forward to next day & closing	Lead facilitator
57	Day 5	17:00 - 18:00	**1:00**	Consultation hour in which panelists of each state can consult the content facilitator	Content facilitators
				DAY 6: Saturday, May 14	
58	Day 6	9:00 - 09:15	**0:15**	Introduction of Day 6 and solving issues of Day 5	Lead facilitator
59	Day 6	9:15 - 10:00	**0:45**	Task 3 Plenary discussion: Review Round 1 results and arguments	Content facilitators
60	Day 6	10:00 - 10:15	**0:15**	Morning tea break	
61	Day 6	10:15 - 11:30	**1:15**	Task 3 Plenary discussion: Review Round 1 results and arguments (cont.)	Content facilitators
62	Day 6	11:30 - 13:00	**1:30**	Task 3 Activity: Angoff Round 2	Content facilitators
63	Day 6	13:00 - 14:00	**1:00**	Lunch break	
64	Day 6	14:00 - 15:30	**1:30**	Workshop evaluation	Individual
65	Day 6	15:30 - 16:00	**0:30**	Task 3 Presentation: Angoff Round 2 results	Lead facilitator
66	Day 6	16:00 - 16:15	**0:15**	Afternoon tea break	
67	Day 6	16:15 - 17:00	**0:45**	Closing and logistics	ECZ, UIS, Cito

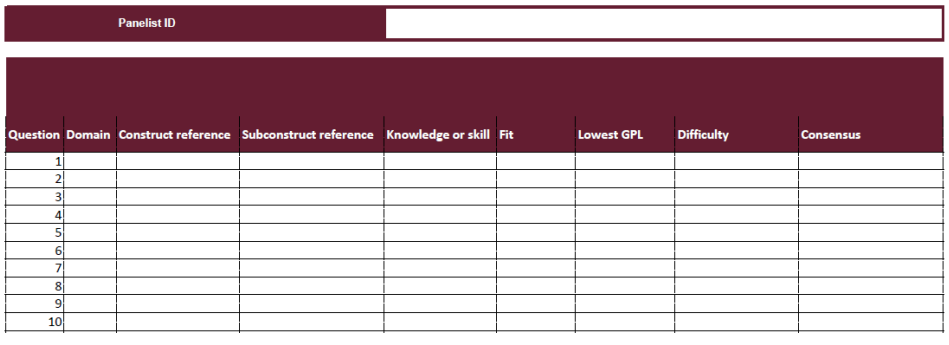
Annex B: Example of the forms

Figure 10. Alignment rating form for paper-based rating of alignment



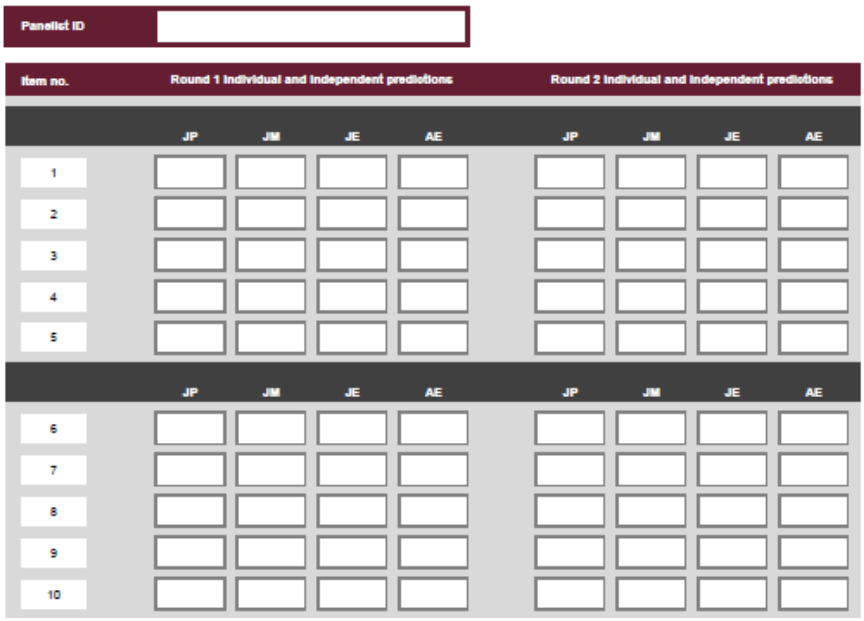
Panelist ID										
						In case of partial fit (record other domains, constructs and subconstructs that relate to the item)				
Question	Domain	Construct reference	Subconstruct reference	Knowledge or skill	Fit	Domain	Construct reference	Subconstruct reference	Knowledge or skill	Fit
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										

Figure 11. Matching form for the local content facilitator



Panelist ID								
Question	Domain	Construct reference	Subconstruct reference	Knowledge or skill	Fit	Lowest GPL	Difficulty	Consensus
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								

Figure 12. Item rating form for paper-based Angoff 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	
1									
2									
3									
4									
5									
	JP	JM	JE	AE	JP	JM	JE	AE	
6									
7									
8									
9									
10									

Figure 13. Data entry file for Alignment rating results

	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
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										

Figure 14. 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	Response 1	Response 2	Response 1	Response 2	Response 1	Response 2	Response 1	Response 2
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								

Figure 15. 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 [x]	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								

Annex C: UIS Activity plan

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

Number	Activity	Role/Responsibility	Workshop Format for which Step is Relevant	Task Complete ?	Date Complete	Comment
Week of March 7-11						
1	Decide on which assessment, grade level, and language to focus	Country with support from UIS/Cito	Both	Yes	21-3-2022	NAS, grade 5, English
2	Decide on remote conferencing service for workshop	Country	Both		29-3-2022	Zoom
3	Process of getting assessement instruments and data or calculation	Country with support from UIS/Cito	Both		1-4-2022	
4	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	Yes		Hybrid: participants in one place, Cito online
Week of March 14-18						
7	UIS and Cito complete Non-Disclosure Agreements (NDAs)	UIS and Cito	Both			Cito completed
Week of March 21-25						
5	Tailor the GPF to the relevant grades/subjects so that it can be translated	Cito	Both			No translation necessary
6	Draft agenda	Cito	Both		4-apr	hopefully 1/4
8	Send assessment instruments to UIS/Cito	Country	Both		april	
9	Send data to UIS/Cito	Country	Both		april	
10	Provide feedback on draft agenda	Country	Both			no feedback
11	Identify local Content Facilitators	Country	Both		april	will be mailed to UIS &Cito
12	Identify interpreters (if relevant)	Country	Both			n/a
13	Identify logistician (if needed)	Country	Both		29-3-2022	
14	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			
15	Start cost estimation	Country with support from UIS	Both			
16	Begin to translate GPF into local language, if necessary and back-translate to check quality	Country	Both			n/a
Week of March 28-April 1						
17	Provide Ministry logo for certificates and banner (the latter only for hybrid workshops) and determine who from the Ministry will sign	Country	Both			in progress
18	Submit budget to UIS	Country	Both			
19	Finalize agenda	Cito	Both		april	
20	Draft workshop slides, including example items, and rating forms to send to UIS for review	Cito	Both		2-mei	
Week of April 4-8						
21	Identify panelists (both teachers and content specialists), including collecting their contact information; ensure panel is representative	Country	Both			80% finished
22	Identify and secure physical space for workshop	Country	Hybrid			
23	Review workshop slides, including example items, and rating forms and send feedback to Cito	UIS	Both			
24	Draft certificates and banner	UIS	Both			
25	Analyze data to produce data distributions, item difficulty data, etc.	Cito	Both		april	
26	Make logistical arrangements for content facilitator training	Cito	Both		april	Sjoerd will send mail
Week of April 11-15						**NB This week Sjoerd absent**
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	Finalize contracts with local Content Facilitators, interpreters, and logistician (the latter two, if applicable)	UIS	Both			
31	Finalize MOU with country based on approved budget	UIS	Both			
32	Identify modality for fund tranfer/expense coverage between UIS/Country	UIS and Country	Both			
33	Finalize item rating forms and slides based on UIS feedback	Cito	Both			
34	Finalize slides for content facilitator training	Cito	Both		april	
Week of April 18-22						
35	Determine what food/refreshments will be provided to participants and procure	Country	Hybrid			
36	Arrange or procure materials, such as notebooks, pens, flipcharts, folders, name tags/tents	Country	Hybrid			
37	Finalize certificates and banners	UIS	Both			
38	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			
39	Meet with Content Facilitators	Cito	Both		april	
Week of April 25-29						
40	Confirm panelist participation	Country	Both			
41	Reserve hotel rooms for panelists, if needed	Country	Hybrid			
42	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			
43	Prepare funds to disperse to participants for per diems, travel, etc.	Country	Hybrid			
44	Assign panelist IDs	Cito	Both			
45	Train Content Facilitators	Cito	Both			ongoing
Week of May 2-6						**NB This week Gerben absent**
46	Distribute panelist IDs	Country	Remote			
47	Distribute the agenda, acronym list, glossary, assessment, GPF, rating forms, evaluation forms, slides with notes fields, certificates, banners, and any other documents	Country	Remote			
48	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			
49	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 9-14: Workshop Begins						

Key	
	UIS Tasks
	Country Tasks
	Cito Tasks

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

Table 18 Reading English: Number of items aligned to each grade 5 domain, construct and subconstructs

Domain	Items
D Decoding	0
R Reading comprehension	21
Total	21
Construct	**Items**
D1 Precision	0
D2 Fluency	0
R1 Retrieve information	14
R2 Interpret information	7
R3 Reflect on information	1
Total	21
Subconstruct	**Items**
D1.1 Identify symbol-sound/fingerspelling and/or symbol-morpheme correspondences	0
D1.2 Decode isolated words	0
D2.1 Say or sign a grade-level continuous text at pace and with accuracy	0
R1.1 Recognize the meaning of common grade-level words	1
R1.2 Retrieve explicit information in a grade-level text by direct- or close-word matching	11
R1.3 Retrieve explicit information in a grade-level text by synonymous matching	2
R2.1 Identify the meaning of unknown words and expressions in a grade-level text	2
R2.2 Make inferences in a grade-level text	5
R2.3 Identify the main and secondary ideas in a grade-level text	0
R3.1 Identify the purpose and audience of a text	0
R3.2 Evaluate a text with justification	1
Total	21

Table 19 Mathematics: Number of items aligned to each grade 5 domain, construct and subconstructs

Domain	Items
N Number and operations	27
M Measurement	9
G Geometry	3
S Statistics and probability	0
A Algebra	2
Construct	**Items**
N1 Whole numbers	21
N2 Fractions	4
N3 Decimals	2
M1 Length, weight, capacity, volume, area, and perimeter	6
M2 Time	3
G1 Properties of shapes and figures	2
G2 Spatial visualizations	0
G3 Position and direction	0
S1 Data management	0
A1 Patterns	2
A3 Relations and functions	0
Subconstruct	**Items**
N1.1 Identify and count in whole numbers, and identify their relative magnitude	4
N1.2 Represent whole numbers in equivalent ways	2
N1.3 Solve operations using whole numbers	11
N1.4 Solve real-world problems involving whole numbers	3
N2.1 Identify and represent fractions using objects, pictures, and symbols, and identify relative magnitude	0
N2.2 Solve operations using fractions	3
N2.3 Solve real-world problems involving fractions	1
N3.1 Identify and represent decimals using objects, pictures, and symbols, and identify relative magnitude	0
N3.2 Represent decimals in equivalent ways (including fractions and percentages)	0
N3.3 Solve operations using decimals	2
M1.1 Use non-standard and standard units to measure, compare, and order	5
M1.2 Solve problems involving measurement	1
M2.1 Tell time	1
M2.2 Solve problems involving time	2
G1.1 Recognize and describe shapes and figures	2
G2.1 Compose and decompose shapes and figures	0
G3.1 Describe the position and direction of objects in space	0
S1.1 Retrieve and interpret data presented in displays	0
S2.1 Describe the likelihood of events in different ways	0
A1.1 Recognize, describe, extend, and generate patterns	2
A3.2 Demonstrate an understanding of equivalency	0
Total	41

Annex E. Difficulty Level of the Items

Table 20. P-value of the NAS reading English items

Question	N	P-value	P0-25	P26-50	P51-75	P76-100
Item 6	0,41	0,19	0,38	0,77	0,96	0,41
Item 7	0,47	0,19	0,48	0,89	0,99	0,47
Item 8	0,33	0,18	0,33	0,48	0,80	0,33
Item 10	0,42	0,19	0,38	0,83	0,99	0,42
Item 11	0,22	0,12	0,19	0,29	0,78	0,22
Item 14	0,48	0,23	0,49	0,80	0,98	0,48
Item 15	0,35	0,20	0,34	0,51	0,87	0,35
Item 16	0,31	0,18	0,29	0,45	0,90	0,31
Item 17	0,27	0,14	0,25	0,37	0,81	0,27
Item 18	0,37	0,17	0,39	0,62	0,74	0,37
Item 19	0,40	0,20	0,40	0,66	0,89	0,40
Item 20	0,34	0,20	0,34	0,49	0,81	0,34
Item 24	0,45	0,21	0,45	0,80	0,92	0,45
Item 25	0,32	0,17	0,30	0,52	0,88	0,32
Item 26	0,30	0,12	0,27	0,55	0,91	0,30
Item 27	0,34	0,13	0,29	0,72	0,95	0,34
Item 28	0,30	0,15	0,29	0,54	0,65	0,30
Item 29	0,28	0,20	0,30	0,30	0,52	0,28
Item 30	0,24	0,13	0,22	0,37	0,72	0,24
Item 32	0,39	0,24	0,40	0,55	0,81	0,39
Item 33	0,30	0,15	0,26	0,57	0,91	0,30
Item 34	0,47	0,25	0,47	0,78	0,95	0,47

Table 21. P-value and Item-Total correlation of the NAS mathematics items

Question	N	P-value	P0-25	P26-50	P51-75	P76-100
Item 1	4469	0,65	0,37	0,73	0,95	0,93
Item 2	4469	0,66	0,40	0,73	0,95	0,96
Item 3	4469	0,54	0,32	0,60	0,72	0,89
Item 4	4469	0,42	0,15	0,45	0,90	0,96
Item 5	4469	0,11	0,11	0,11	0,08	0,41
Item 6	4469	0,39	0,34	0,36	0,70	1,00
Item 7	4469	0,44	0,17	0,48	0,88	0,98
Item 8	4469	0,29	0,13	0,28	0,70	0,96
Item 9	4469	0,33	0,22	0,31	0,68	0,98
Item 10	4469	0,35	0,16	0,35	0,80	0,93
Item 11	4469	0,32	0,14	0,32	0,71	0,96
Item 12	4469	0,41	0,21	0,42	0,80	1,00
Item 13	4469	0,58	0,32	0,63	0,92	0,98
Item 14	4469	0,35	0,16	0,37	0,66	0,89
Item 15	4469	0,44	0,20	0,48	0,89	0,96
Item 16	4469	0,44	0,27	0,45	0,80	0,96
Item 17	4469	0,39	0,15	0,41	0,86	0,96
Item 18	4469	0,26	0,15	0,24	0,64	0,85
Item 19	4469	0,36	0,20	0,37	0,71	0,93
Item 20	4469	0,21	0,12	0,22	0,39	0,65
Item 21	4469	0,46	0,26	0,51	0,71	0,85
Item 22	4469	0,61	0,49	0,63	0,85	0,98
Item 23	4469	0,31	0,18	0,32	0,55	0,91
Item 24	4469	0,27	0,16	0,27	0,57	0,87
Item 25	4469	0,28	0,21	0,29	0,35	0,69
Item 26	4469	0,49	0,25	0,54	0,81	0,93
Item 27	4469	0,30	0,15	0,30	0,61	0,87
Item 28	4469	0,31	0,20	0,33	0,48	0,76
Item 29	4469	0,35	0,23	0,36	0,60	0,85
Item 30	4469	0,42	0,28	0,43	0,71	0,87
Item 31	4469	0,22	0,13	0,21	0,46	0,89
Item 32	4469	0,22	0,18	0,21	0,32	0,72
Item 33	4469	0,19	0,09	0,19	0,42	0,87
Item 34	4469	0,18	0,15	0,17	0,31	0,69
Item 35	4469	0,33	0,24	0,34	0,51	0,89
Item 36	4469	0,24	0,16	0,24	0,41	0,81
Item 37	4469	0,21	0,15	0,24	0,14	0,24
Item 38	4469	0,19	0,12	0,20	0,31	0,76
Item 39	4469	0,30	0,22	0,31	0,44	0,69
Item 40	4469	0,33	0,30	0,34	0,33	0,31
Item 41	4469	0,15	0,14	0,13	0,24	0,78
Item 42	4469	0,17	0,10	0,18	0,32	0,65
Item 43	4469	0,27	0,18	0,27	0,45	0,65
Item 44	4469	0,08	0,07	0,08	0,09	0,31
Item 45	4469	0,26	0,17	0,29	0,36	0,52

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

EVALUATION OF THE WORKSHOP

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

1. PIN

TRAINING ON THE GLOBAL PROFICIENCY FRAMEWORK

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

2. GPD training	**Strongly disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly agree**
I understand the purpose of the GPF					
I understand the relationship between domains, constructs, subconstructs, knowledge and skills, and GPDs					
The GPDs were clear and easy to understand					
The discussion of the GPDs helped me understand what is expected of learners in Mathematics/Language at the end of grade 5					
The practical exercise using the GPDs was useful to improve my understanding					
There was an equal opportunity for everyone to contribute their ideas and opinions					
There was an equal opportunity for everyone to ask questions					
The amount of time spent on the GPD training was sufficient					

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

TRAINING ON THE NAS

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

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

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

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

TRAINING ON ALIGNMENT METHODOLOGY

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

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

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

TRAINING ON MATCHING METHODOLOGY

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

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

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

TRAINING ON THE BENCHMARK-SETTING (ANGOFF) METHODOLOGY

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

16. Policy linking training	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I understand the process I need to follow to complete the benchmarking exercise					
I understand how the benchmarking methodology links to the steps on alignment and matching					
I understand the difficulty level of the assessment items					
The discussion of the procedure was sufficient to allow me to feel confident in the methodology					
I understand how my ratings will result in a final benchmark					
There was an equal opportunity for everyone to contribute their ideas and opinions					
There was an equal opportunity for everyone to ask questions					
The amount of time spent on the policy linking method training was sufficient					
I feel confident in my Round 1 ratings					
I was given sufficient time to complete the Round 1 performance predictions[4]					

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

BENCHMARK ROUND 2 EVALUATION

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

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

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

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

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

OVERALL EVALUATION

22. How comfortable are you with your final performance predictions?
    a) Very uncomfortable
    b) Somewhat uncomfortable
    c) Neutral[5]
    d) Fairly comfortable
    e) Very comfortable
23. If you marked either of the uncomfortable options, please explain why.
24. Overall, how would you rate the success of the policy linking workshop?
    a) Totally Successful
    b) Successful
    c) Neutral[6]
    d) Unsuccessful
    e) Totally Unsuccessful
25. How would you rate the organization of the workshop?
    a) Totally Successful
    b) Successful
    c) Neutral[7]
    d) Unsuccessful
    e) Totally Unsuccessful
26. Please provide any comments you feel would be helpful to us in planning future policy linking workshops.

Thank you for your participation in the workshop.

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