Institute for Statistics

Credit: EQAD/MoEYS

Report on the Cambodian National Grade 6 Learning Assessment Policy Linking for Measuring Global Learning Outcomes Workshop (July 2021)

Setting Global Benchmarks for Khmer and Mathematics in Cambodia

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 and people to make the Cambodian policy linking pilot workshop a success.

First, the organizational support provided by Chinna Ung, Director of the Education Quality Assurance Department of the Ministry of Education, Youth and Sports of Cambodia was critical for the success of the workshop.

Second, the support provided by Shailendra Sigdel and Melissa Chiappetta from the UNESCO Institute for of Statistics (UIS) in Montreal and Danny Padilla and Don Tan of the UNESCO regional office in Phnom Penh was instrumental in planning and implementing the workshop.

Third, the hands-on support of our local content facilitators, Hav Khou and Pagna Tol and the local workshop coordinator, Sarin Sar, all of the Education Quality Assurance Department, and the rest of their colleagues involved was of the utmost importance for realizing the goals of this workshop.

Finally, the dedication and effort devoted by the panelists for Khmer and Mathematics during the two weeks in which the workshop took place, were indispensable in establishing the pilot global benchmarks and drawing lessons learned from the workshop.

Cor Sluijter, lead facilitator

Margreet van Aken, content facilitator Khmer

Michel Hop, content facilitator Mathematics

Sanneke Schouwstra lead facilitator/data analyst

Angela Verschoor, data analyst

Table of Contents

List of tables

List of figures

Acronyms and Abbreviations

EQAD	Education Quality Assurance Department, Ministry of Education, Youth and Sports
GPD	Global Proficiency Descriptor
GPF	Global Proficiency Framework
GPL	Global Proficiency Level
IRT	Item Response Theory
JE	Just Exceeds Minimum Proficiency
JM	Just Meets Minimum Proficiency
JP	Just Partially Meets Minimum Proficiency
MoEYS	Ministry of Education, Youth and Sports
NLA	National Learning Assessment
PLT	Policy Linking Toolkit
SDG	Sustainable Development Goal
SEM	Standard Error of Measurement
UIS	UNESCO Institute for Statistics
UNESCO	United Nations Educational, Scientific and Cultural Organization
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.

The photo was taken in 2021 during the Grade 6 National Learning Assessment (G6-NLA-2021) in Preyveng province. The photo is provided by the Education Quality Assurance Department (EQAD) of the Ministry of Education, Youth and Sport (MoEYS), Kingdom of Cambodia.

1. Executive summary

This document contains the report on the Cambodian online policy linking workshop that took place from July 5, 2021 until July 16, 2021. The Education Quality Assurance Department of the Ministry of Education, Youth and Sports in Cambodia (EQAD) and the UNESCO Institute for Statistics (UIS) organized this workshop as a pilot. The objective of the workshop was to set global benchmarks on the 2016 National Learning Assessment (NLA) at grade 6 in Khmer and mathematics through organizing a fully remote policy linking workshop.

This was the first time Cambodia participated in a policy linking workshop. UIS hosted the workshop using the Zoom videoconferencing platform. All the participants worked individually from home. The participants performed their tasks with dedication and engaged in lively discussions during the tasks. To mitigate the risk of an unstable internet connection several important sessions were recorded so that participants who missed parts could review the session afterwards. The content facilitators and the participants performed their tasks with full dedication and with excellent commitment. They were eager to learn, and at the end of the workshop were grateful for what they had learned and for the opportunity to participate. Consequently, all the activities, from the familiarization at the start to the benchmarking at the end, were carried out with full engagement and with lively and relevant discussions. Every step of the process produced important outcomes. The participants gave very positive feedback, both in person and in their evaluation forms. Although some panelists did encounter problems with the internet connectivity during the workshop, this did not affect its quality.

The workshop was formally closed with an inspirational speech by the Cambodian Secretary of State, Ministry of Education Youth and Sport, emphasizing the importance of monitoring the quality of education in Cambodia through activities like policy linking and thanking all the workshop participants for their commitment. The local organizers expressed their hope and belief that the workshop would have a catalyzing effect on the future of Cambodia's education and expressed their eagerness to organize another workshop with the next NLA as the instrument to set benchmarks upon.

The participants' work showed that the NLA for Khmer is strongly aligned to the Global Proficiency Framework (GPF) for grade 6, both in depth and in breadth. After the alignment session, the conclusion was that the NLA for Mathematics was in breadth also strongly aligned to the GPF for grade 6, but in depth additionally aligned. After the matching sessions, where both for Khmer and Mathematics complete consensus was reached, the latter conclusion changed also into strong alignment. The final benchmarks of the panelists show a good consistency, which makes the benchmarks useable for comparing, aggregating, and tracking learning outcomes for the NLA. To sum up: the piloting of the policy linking workshop in a fully remote mode in Cambodia can be considered a success.

2. Background

Policy Linking Overview

In September 2015, Member States of the United Nations formally adopted the 2030 Agenda for Sustainable Development in New York. The agenda contains 17 goals, including a new global education goal (SDG 4). SDG 4 is to ensure inclusive and equitable quality education and promote lifelong learning opportunities for all and has seven targets (UNESCO, 2021). The first target focusses on primary and secondary education (target 4.1): By 2030, ensure that all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes. To monitor progress the indicator 4.1.1 is used: Proportion of children and young people (a) in grades 2/3; (b) at the end of primary; and (c) at the end of lower secondary achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex (United Nations, 2021).

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

Global Proficiency Framework

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

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

The policy linking methodology

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

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

The policy linking methodology is elaborated in the Policy Linking Toolkit (PLT), 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 PLT 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 (GPL) and Global Proficiency Descriptor. Each panelist determines the levels of knowledge and skills required from students to correctly answer each aligned item. The panelists should work in groups to reach consensus
- Task 3 — The panelists set three global benchmarks for each assessment using a standardized method (a modified version of the Angoff methodology) through two rounds of ratings.

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

Overview of the National Learning Assessment (NLA) 2016

After its creation in 2009, EQAD has been conducting National Learning Assessments since 2012. The NLA from 2016 was the fourth national assessment completed by EQAD and was held in grade six. Earlier assessments were held in grade three (2014-15), grade eight (2013-14), and grade six (2012-2013) (EQAD, 2017, p.4). Before that, the General Secondary Education Department of the MoEYS was responsible for the National Learning Assessments. For Grade 6 a first round was held in 2007. Thus, the NLA 2016 was the third round. A fourth round was planned in 2021, but was delayed because of the Covid-19 outbreak.

The ultimate goal of Cambodia's National Learning Assessments is to assure the development of the Cambodian education sector, particularly in primary school level and to monitor achievement of specific objectives:

- To diagnose student achievement compared to intended curriculum, curriculum standards and detailed curriculum
- To identify improvement—or decline—in student achievement
- To figure out strengths, weaknesses, and skills of students

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

- To determine what influences student achievement
- To determine whether MoEYS reached its defined standard, ultimate goal, or indicators compared to inputs (resources)
- To share lessons learned and recommendations to improve the quality of education

Content and design of the NLA in grade 6

The NLA is a low stake system level assessment that summarizes students' achievement for Khmer and Mathematics at national and subnational levels. Not all items were administered to all learners. Items were divided into three overlapping nominally equivalent booklets, making it possible to report the outcomes on one and the same scale by using techniques from Item Response Theory(IRT). Each booklet for Khmer contained 33 items and each booklet for Mathematics 32 items. The technical report provided by EQAD (EQAD, 2017) does not contain information on the specific IRT model used for reporting, but the data that were provided indicate that the two-parameter Birnbaum model (Birnbaum, 1968) must have been used.

The 70 NLA-items for Khmer measured five different content areas: Punctuation, Reading Comprehension, Grammar, Writing and Dictation. The 121 NLA items for Mathematics also measured five different content areas: Statistics, Algebra, Numbers, Measurement and Geometry.

Sample and data analysis

EQAD employed a two-stage random sampling design for the NLA 2016. Table 1 (EQAD, 2017, p. 6) provides an overview of the sample of learners that participated in the NLA against the total Cambodian population in grade 6 The sample consisted of 210 randomly selected public schools and 18 private schools that were chosen manually. Thus, the sample consisted of 228 schools. The proportion of urban and rural schools in the sample reflects the proportion in the total population of schools. In addition to this, all 25 provinces in Cambodia are represented in the sample.

Table 1. Comparison of the NLA 2016 Sample against the population

Stratum	Population		Sample	
	Schools	Percentage	Schools	Percentage
Urban/Plains	221	9.4%	20	9.5%
Urban/Tonle Sap	169	5.2%	11	5.2%
Urban/ Plateau	83	2.0%	5	2.4%
Urban/Coastal	32	1.0%	3	1.4%
Rural/Plains	1860	39.4%	80	38.1%
Rural/Tonle Sap	1490	25.4%	53	25.2%
Rural/Plateau	625	11.4%	24	11.4%
Rural/Coastal	326	6.3%	14	6.7%
Total	4806	100%	210	100%

3. Pilot Workshop Preparation

Objective of the workshop

The objective of the workshop was setting global benchmarks on the 2016 NLA at grade 6 in Khmer and Mathematics using a fully remote policy linking workshop. The workshop had a piloting function and should increase the capabilities of EQAD to conduct similar workshops in the future. EQAD requested to set three benchmarks.

First three policy linking stages

After the engagement of Cambodia, on Wednesday 03-03-2021, a kick-off meeting took place between UNESCO, EQAD and Cito. Cito was contracted to facilitate the policy linking workshop and provided the lead facilitator, two content facilitators and a data analyst. After the initial engagement, the country governments or assessment agencies should collate evidence of curriculum and assessment validity and alignment (stage 2 of policy linking) and the 4.1.1. Review Panel should review this collated evidence. However, after the initial engagement of Cambodia, the 4.1.1. Review Panel was not yet in place. “This stage of the process involves the country government sharing standard-, curriculum-, and assessment-related documents (including the most recent round of data) with the project team and examination of those documents by the project team and the 4.1.1 Review Panel to determine whether the assessment(s) meets reliability and validity standards required for a country to proceed with policy linking for reporting global outcomes.” (PLT, p. 170). The 4.1.1. Review Panel uses three criteria: Alignment between the assessment and the curriculum, Appropriateness of the assessment for the population, Reliability of the assessment.

The 4.1.1 Review Panel was not in place in place. Therefore, the Cito team made an initial assessment of whether the NLA met reliability and validity standards required to proceed with policy linking. The Technical Report of the 2016 NLA (EQAD, 2017) does not contain concrete information on reliability, but the Technical Report states that the design and sampling of the NLA was reviewed by two experts of USAID who concluded that “The development and refinement of the instrument [the NLA]has met international standards for sample-based assessment instrument development, and EQAD has demonstrated the instruments’ reliability and validity in a statistically rigorous approach.” (EQAD, 2017, p.8)

The evidence presented in the Technical Report of the 2016 NLA shows that the NLA also seems appropriate for the population. The items have been reviewed to determine their validity. EQAD piloted the items and tested them also in a field trial (EQAD, 2017, p. 7).The implemented sampling procedure (EQAD, 2017, p. 5) ensures that the learners who carried out the assessment are representative of the population against which results are reported. The Technical Report also contained information on the alignment between the NLA and the curriculum. All in all the 2016 NLA seems to live up to all the requirements of the policy linking procedure.

General preparation of the workshop

UIS, EQAD and Cito planned to facilitate the workshop remotely, due to the COVID-19 pandemic. There were three possible options. Of these options, the most preferred one was having all the panelists gather in one place. The second preferred option was to have the panelists gather in the provincial EQAD headquarters in their province for an in-person workshop. The least preferred option was to have all the panelists working from home. The main reason for this being the least preferred option was the expectation that the stability of the internet connectivity for all panelists could not be guaranteed. From the beginning of May and onwards, weekly meetings were held to organize the workshop and to monitor the COVID-19 situation. Initially UIS, EQAD and Cito decided to wait until the situation allowed the panelists to travel, so the most preferred option could be carried out. However, because the situation did not improve, in the meeting of June 16 the team decided to carry out a fully remote workshop to ensure its continuation.

To mitigate the issue of bad internet connectivity, Cito developed an 11-day agenda to make sure that panelists with internet problems could watch recorded sessions and discussion afterward and would have enough time to receive complete information, complete their activities and turn in their output. The agenda was adapted to the workday in Cambodia and adjusted to allow for data entry. Before finalizing the agenda, it was shared with EQAD and UIS for suggestions and improvements. Note that the detailed agenda has a maximum of 3,5 hours of online activity with several comfort breaks for the participants to warrant participants could maintain focus. Also note that, as stated earlier, the agenda for each day only contained a restricted number of hours of online contact time. Because of the five hour time difference between Cambodia and the Netherlands the online time was in the afternoon. The mornings in Cambodia were reserved for having panelists watch recordings of presentations which they might have missed the day before and for follow up activities based on the activities on the day before under the guidance of the local content facilitators.

After approval from the Ministry of Education, Youth and Sports (MoEYS) on June 23, the workshop took place from Monday 05-07-2021 until Friday 16-07-2021. Because EQAD was not sure all panelists mastered English sufficiently, two interpreters were hired for simultaneous translation during the workshop. In addition to this, all relevant material for the workshop was not only made available in English, but also translated into Khmer.

EQAD sought a group of teachers and subject matter experts (SMEs) as representative for these professional groups as possible. Table 2 gives an overview of the panelists' background information. In total 46 panelists participated 23 for Khmer and 23 for Mathematics. Because the workshop was also seen as a means of capacity building, a larger group of subject matter experts participated then advised in the Policy Linking Toolkit. All teachers participating were certified teachers. Furthermore, international observers were present during some of the sessions.

Table 2. Panelists' background information

	Khmer		Mathematics		Total
	Teachers	SME's	Teachers	SME's	
Total					
Gender					
F	5	3	4	1	13
M	5	10	6	12	33
Level of education					
Completed 4-year College	5	3	7	3	18
Completed Master's Education	3	1	1	2	7
N/A	2	9	2	8	21
Grand Total	10	13	10	13	46

Materials for the workshop and pre-workshop analyses

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

Collecting materials and pre-workshop analyses

Before the workshop, EQAD shared all items with the Cito team, after the team members had signed a non-disclosure agreement. Because of confidentiality, the NLA itself could not be shared with the panelists before the workshop. Therefore, it was not possible for panelists to administer the NLA to nine learners as the Toolkit requires.

In preparation for the workshop the distribution of the sum scores and the p-values of the selected items was calculated (see Annex E).

Because of the large number of items in the NLA, it was decided that is was not practically possible to use the complete item set in the workshop. Instead a selection was made optimally reflecting the complete content of the NLA. The alignment, matching and Angoff rating activities would have taken too much time to fit within the agenda of the workshop. And the amount of effort required of the panelists would have been too high

Because, the NLA data and results of the analyses were also shared before the workshop, including the IRT-parameter values of the items, it was possible to only select a subset of all the items that were originally part of the NLA for the workshop. . Roughly speaking, one of the booklets was selected, because they were nominally equivalent. The selection consisted of 33 items for Khmer and 31 items for Mathematics.

Because an IRT-analysis had taken place, the benchmarks established in the workshop can be used to calculate the corresponding positions on the underlying NLA 2016 ability scale for Khmer or Mathematics. And then, based on these ‘ability scale benchmarks’, the expected item score (expected p-value, given the specific ability scale score) corresponding with these benchmarks can be calculated for all NLA items. Thus, for any subset or the complete set of NL-items or any subset, GPF benchmarks can be calculated. This can be done, by simply adding the expected item scores.

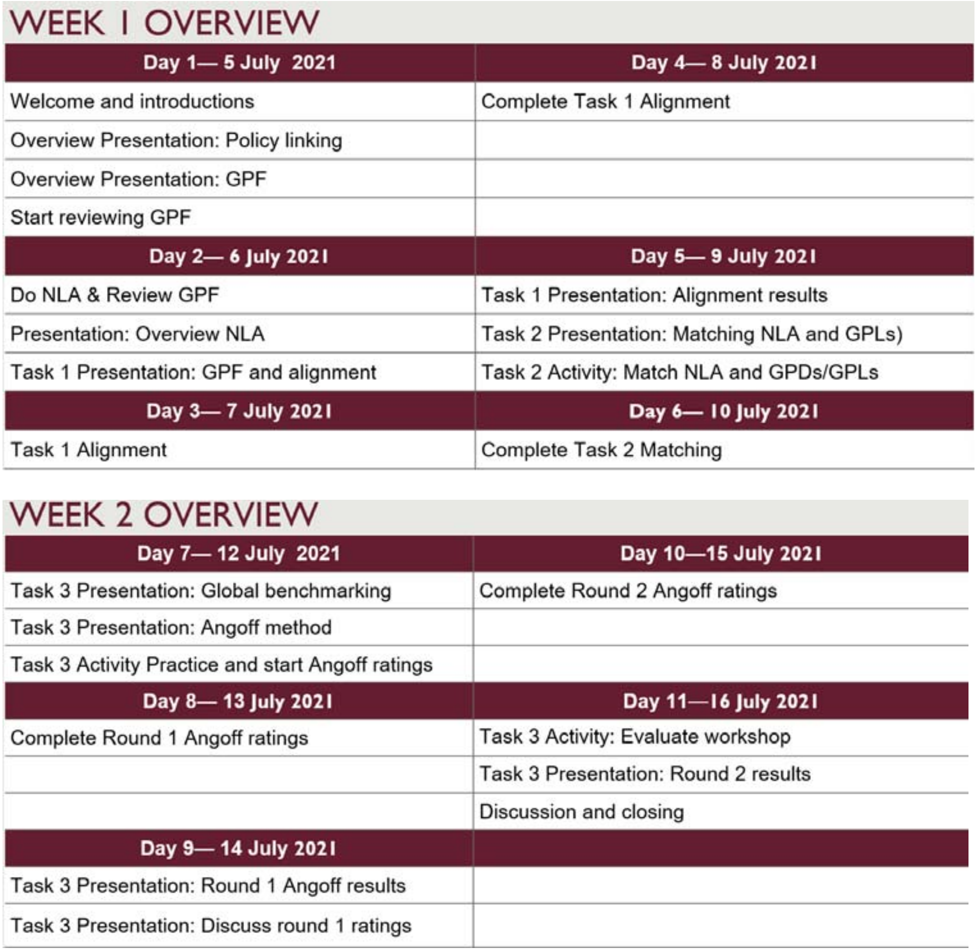
Creating workshop materials

To limit the hours of online activity and to mitigate the risk of instable internet connectivity, an eleven-day workshop was planned (see the overview in Table 3, in Annex A the complete agenda is presented). Because the PLT did not contain digital forms for remote workshops yet, for each of the three tasks Cito developed digital forms, separate for Khmer and Mathematics (see Annex B). Forms were created for the alignment ratings (Annex B, Figure 9), matching ratings, (Annex B, Figure 10) and the item ratings (Annex B, Figure 11). It was decided to translate the evaluation questionnaire into Khmer and to convert the form into a Google Docs document. Next to that, forms were also created for the entry of the alignment ratings (Annex B, Figure 12), item ratings (Annex B, Figure 13) and for the entry of the evaluation forms (Annex B, Figure 14). The digital forms were designed to ease the task of the panelists, to prevent inconsistent ratings and to speed-up the data analyses during the workshop. To increase the efficiency of the data collection, EQAD also recreated the alignment and item rating forms in Google Docs.

Cito prepared a package for panelists containing all workshop materials, to be printed and distributed on location. The package contained the agenda for the workshop, a unique panelist ID, the GPF for Grades 5 to 7, the glossary and acronym list, a handout of the slides of all presentations and the items selected from the NLA. Furthermore, the package contained the Alignment rating form and the Item rating form. Where necessary, the material was translated into Khmer. Panelists received the information both in paper and in digital format. As already stated, all forms were transformed into Google Docs documents to increase the efficiency of data collection and processing. The URL was shared with the panelists on the days they had to provide output. The matching form was only shared with the local content facilitators, because they were supposed to summarize the outcomes of the matching activity during the matching session

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

Table 3. Agenda for the eleven-day fully remote workshop



WEEK 1 OVERVIEW

Day 1— 5 July 2021	Day 4— 8 July 2021
Welcome and introductions	Complete Task 1 Alignment
Overview Presentation: Policy linking	
Overview Presentation: GPF	
Start reviewing GPF	
Day 2— 6 July 2021	**Day 5— 9 July 2021**
Do NLA & Review GPF	Task 1 Presentation: Alignment results
Presentation: Overview NLA	Task 2 Presentation: Matching NLA and GPLs)
Task 1 Presentation: GPF and alignment	Task 2 Activity: Match NLA and GPDs/GPLs
Day 3— 7 July 2021	**Day 6— 10 July 2021**
Task 1 Alignment	Complete Task 2 Matching

WEEK 2 OVERVIEW

Day 7— 12 July 2021	Day 10—15 July 2021
Task 3 Presentation: Global benchmarking	Complete Round 2 Angoff ratings
Task 3 Presentation: Angoff method	
Task 3 Activity Practice and start Angoff ratings	
Day 8— 13 July 2021	**Day 11—16 July 2021**
Complete Round 1 Angoff ratings	Task 3 Activity: Evaluate workshop
	Task 3 Presentation: Round 2 results
	Discussion and closing
Day 9— 14 July 2021	
Task 3 Presentation: Round 1 Angoff results	
Task 3 Presentation: Discuss round 1 ratings	

Training the local content facilitators

The local content facilitators and the local workshop coordinator participated in the weekly meetings between EQAD, UIS and Cito. Thus, they already became globally aware of the purposes and content of the workshop relatively long before the actual start of the workshop. But to ensure that they could perform all their activities correctly, several additional measures were taken.

First of all, a four hour interactive online training was designed. This training consisted of several parts. It started with an introduction of the generics and specifics of policy linking for both local content facilitators, followed by two separate one hour sessions for Khmer and mathematics. These two simultaneous sessions focused on the relevant parts of the GPF for either Khmer or Mathematics and on the specific activities of the local content facilitators during the different parts of the workshop (Alignment, Matching and Benchmarking). The final two hours were spent to follow the workflow of the workshops over the 11 days, again focusing on the specific activities of the local content facilitators. In addition to this, Cito produced a detailed script for the workshop which was shared with and reviewed by the local content facilitators. Furthermore, Cito also produced a document containing specific instructions for them on their tasks during the different parts of the workshop. And last but not least, they received an additional training and instructions on processing the output of the panelists, including the data entry Excel sheets to help them with entering the data from paper rating forms received from panelists.

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

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

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

A successful Technical Test of the Zoom platform was performed on Friday July 2 with the interpreters and most panelists and staff involved. But only on the first day of the workshop it was found out that it was not possible to have two parallel simultaneous translations running in the break-out rooms for Khmer and Mathematics. This was solved on the first day by having separate Zoom meetings for Khmer and Mathematics.

Training for local data entry

Because the data had to be checked and transformed into Excel files before processing, data entry was needed, and a special 2-hour data entry training was given to the local content facilitators at the end of the second day of the workshop. On fourdays (day 4, 8, 10 and 11) data entry had to occur. The local content facilitators and the logistic coordinator collected the panelist document from Google Docs and after all panelists had completed their work, the data had to be entered into the Excel files and sent to Cito. During the training the schedule and times for data entry were shown. For the sake of completeness these were also in a separate document with detailed written instructions. Next, Cito discussed the steps in data entry and gave a demonstration of data entry for each of the different forms.

The global steps in data entry were:

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

5. Send all forms to Cito

4. Implementing the fully remote workshop

Familiarization

Following feedback from other policy linking workshops, the workshop started with several background sessions. After the formal welcome, in the afternoon, the first day focused on familiarizing panelists with policy linking and the GPF. Its key objectives were that panelists understood the purpose of policy linking and get globally acquainted with the GPF.

The first presentation gave background information on policy linking, including a chronology of the development of the method in response to the global indicators. The second one provided information on the structure and content of the GPF. Next, in two simultaneous separate sessions, the content facilitators from EQAD and Cito started the training on the GPF and its role in policy linking. The example of the benchmarks and the proficiency levels is shown in Figure 1.

In the separate meetings for Khmer and Mathematics, the content facilitators introduced -with the help of the local content facilitators- each of the domains, constructs, subconstructs, statements of knowledge and/or skill(s), and GPLs and GPDs. An example from part of the mathematics GPF is shown in Table 4.

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

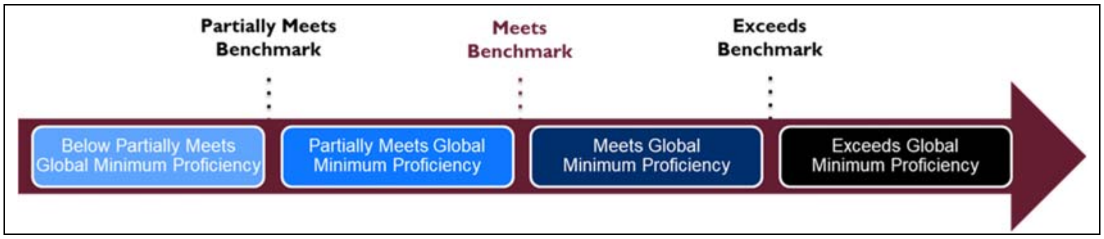

Partially Meets Benchmark
Meets Benchmark
Exceeds Benchmark
Below Partially Meets Global Minimum Proficiency
Partially Meets Global Minimum Proficiency
Meets Global Minimum Proficiency
Exceeds Global Minimum Proficiency

Table 4. Part of the GPF 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

In the morning of day 2, panelists made the selection of items from the NLA themselves in order to get better acquainted with the items and with the skills and knowledge necessary to answer these items correctly. While answering the items of the NLA, the panelists were asked to note stumble blocks and aspects of the items that might make the item easy or difficult for Grade 6 learners. The morning was also used for studying the GPF. The afternoon of the second day of the workshop started with the continued reviewing of the GPF and identification of elements that were still unclear. This was followed by a presentation with an overview of the NLA by EQAD. The familiarization part of the workshop ended with a discussion in the Khmer and Mathematics group on the NLA and the GPF.

Observations

Although a technical test showed no problems, on the first day it was discovered that it was not possible to have two parallel sessions in break-out rooms with simultaneous translation. However, simultaneous translation was necessary to ensure that all panelists fully understood all presentations and instructions. To solve this problem separate Zoom meetings were scheduled for Khmer and Mathematics. The Mathematics Zoom meeting also served as the meeting for all plenary activities. Thus, this specific problem was solved. A check of the data from the evaluation forms shows that this caused no issues with the panelists: there are no significant differences in approval between this part of the workshop and the other parts.

On this day and during the whole workshop, there was good and frequent contact via WhatsApp chat, telephone and e-mail between the local and Cito content facilitators. This helped both sides staying informed. The content facilitators used these communication means to confer about content and organizational issues as well. In addition to this the local content facilitators had set up an extra means of contact with all panelists via Telegram.

EQAD, UIS and Cito had planned to facilitate the workshop remotely, due to the COVID-19 pandemic. As EQAD could not provide each panelist with an individual laptop with headset, panelists had to work from their own devices. Therefore, it was expected that some of the panelists would have to participate with a tablet or a smartphone. In practice, this was not a real issue; only 2 out of 46 panelists had to use a smartphone for the workshop. To make data entry as efficient as possible, Cito developed Excel-files for data entry and a two-hour data entry training for the local content facilitators. The local team collected all data from the alignment, item rating and workshop evaluation sessions through Google Docs, exported the data to the Cito-Excel files and sent them after checking for errors to Cito upon completion. There were no serious issue with this procedure

Having local content facilitators and interpreters available helped a lot in the communication. Because most panelists did not feel sure about communicating in English, a lot of the discussions went on in Khmer. And this could be followed by the Cito facilitators through the simulataneous translation going on. Working with the Zoom platform proved to be an advantage, because there was no interference from the ongoing discussions and facilitators could concentrate on the translation.

However, the presentations about policy linking and about the GPF did not succeed well in engaging the panelists. This is partly understandable, because both policy linking and the GPF were unfamiliar and policy linking is a complex procedure, while the GPF contains a lot of detailed and multifaceted information. A possible other reason for the lack of engagement is the form of the presentation, which is one-directional.

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

It was also discovered that the panelists and the Cito team had not completely identical versions of the GPF, especially table 5. This caused quite some confusion and delay. The issue was resolved by deciding that the panelists would use the version they had, and the local content facilitators would make a shadow document in which the differences, when they occurred, would be noted down. This issue was caused by the time period between completing the material for the workshop and the starting date. In this period some changes to the GPF were made that were absent from the Cambodian version.

The number of items selected from the NLA for Khmer and Mathematics was almost equal: 33 and 31 items respectively[2]. Despite this fact, it proved difficult to keep both groups in synch and have them both ready for the plenary parts at the same time. Moreover, the fact that there were two Zoom meetings simultaneously made it more difficult to follow the progress in both panelist groups by the lead facilitator.

Task 1: Alignment

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

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

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

The alignment method in the PLT 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 NLA items and GPF knowledge and/or skill(s) statement(s) and in the second step the facilitators compile and summarize the ratings to check the alignment between the assessments and the GPF.

The afternoon of the third day started with group discussions in the Khmer and Mathematics meetings on the first five items under the guidance of the local and Cito content facilitators. Next, some sample items were aligned. 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

[2] In Task 2 for Khmer the number of items was reduced to 32 and in Task 3 for Mathematics to 30 for reasons explained elsewhere See footnotes 3 and 4.

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.

Panelists were then asked to work individually and independently on day 4 to rate the alignment between each NLA 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 could be approached by them on the internet via Google Docs (see Annex B).

After all panelists had completed their alignment forms on day four, the EQAD team finalized the second step. All alignment ratings forms were merged into one Excel-file, checked and sent to Cito for analysis.

Alignment Khmer

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

Averaging the panelists' ratings, we see that all 33[3] items (on average) aligned to Reading comprehension. At least 10 items were aligned to Retrieve information; at least 11 items were aligned to Interpret information (on average 11,3) and at least 11 were aligned to Reflect on Information. The NLA Khmer is therefore strongly aligned in depth (see Table 5).

We see that on average all subconstructs of Reading comprehension are covered (see Table 21 in Annex D). The NLA Khmer assessment was therefore strongly aligned in breadth (see the criteria in Table 5).

[3] During Alignment it was discovered that one of the items was not a Reading comprehension item and it was therefore eliminated.

Table 5. Reading Alignment Criteria for Grades 1–9

Level of Alignment	Category	Grade 1–2 Criteria	Grade 3–6 Criteria Grade	Grade 7–9 Criteria
Minimally Aligned	Domain/Construct (depth):	D (minimum five items)	R (minimum five items)	R (minimum five items)
		C (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: R1 (minimum 5 items)
			R: B2 (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 almost 28 of the 31 items, aligned to grade 6 subconstructs. One item was excluded from the ratings, because correct information was missing for the item[4]. In the GPF 24 subconstructs are mentioned for grade 6 and the NLA covered 20 of those subconstructs (an average of >0.5, see Table 22 in Annex D). In breadth the NLA is strongly aligned to the GPF for Grade 6 as the items covered more than 50% of all grade 6 subconstructs.

The NLA Mathematics items covered all five domains and 9 out of 12 constructs for grade 6. According to the new criteria in the Policy Linking Toolkit, for strong alignment in Depth at least 5 items should align to the domain Number and Operations, at least 5 items to Measurement and Geometry and at least 5 items to Statistics and Probability and Algebra (see Table 6). On average 15.6 items covered the domain of Number and Operations, 7.3 items the domains Measurement and Geometry, and 4.7 items the domains Statistics and Probability and Algebra. For this reason, according to the panelists for Mathematics, the NLA is additionally aligned to

[4] The data from NLA Mathematics made clear that this item was in fact a meta item containing three separate items. At this point in time it would have caused a lot of confusion with panelists and a lot of extra work for the EQAD and Cito team if this would have been taken into account. So it was decided to exclude the item from the next steps in policy linking.

the GPF in depth, because the number of items related to the domains Statistics and Probability and Algebra should be larger than 5 to warrant the conclusion that the NLA is strongly aligned in depth.

Table 6. Mathematics Alignment Criteria for Grades 1–9

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

Observations

From the alignment task onwards, the Khmer group and mathematics group remained in their separate Zoom meetings, coming together digitally only for the plenary activities. In the plenary presentation on alignment, examples were presented of the three types of fit, but only for mathematics. It would have helped the Khmer panelists if there would also have been similar examples for language in the presentation.

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

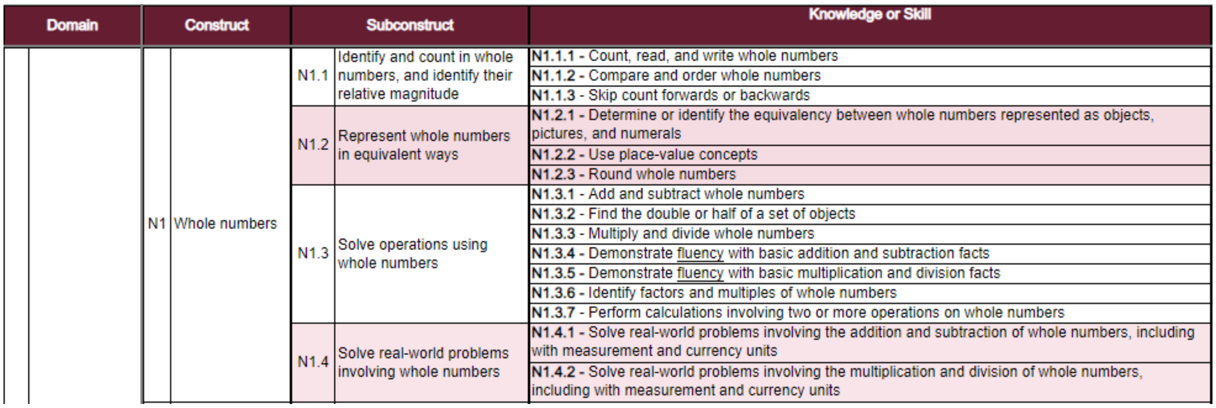
For Mathematics the panelist group concluded that an extra overview showing vertical alignment between grades would increase the efficiency of the alignment activity. Such an overview would help in converting, for instance a Meets descriptor in Grade 5 to a Partially Meets descriptor in Grade 6, or an Exceeds descriptor in Grade 5 to a Meets descriptor in Grade 6.

The filling in of the alignment forms went smoothly, as well as the data entry process by the local content facilitators for Khmer and Mathematics. The resulting data sets were sent in time to Cito to allow for the necessary analyses.

The addition of codes for the knowledge or skill statements is a big improvement compared to earlier versions of the GPF. See Table 7 for an example. However, in the mathematics GPF some inconsistencies were still found.

Only after the alignment session it was discovered that one of the items consisted of several sub items and that it therefore should have been treated as three individual items (see footnote 4).

Table 7. The new knowledge or skill codes for mathematics

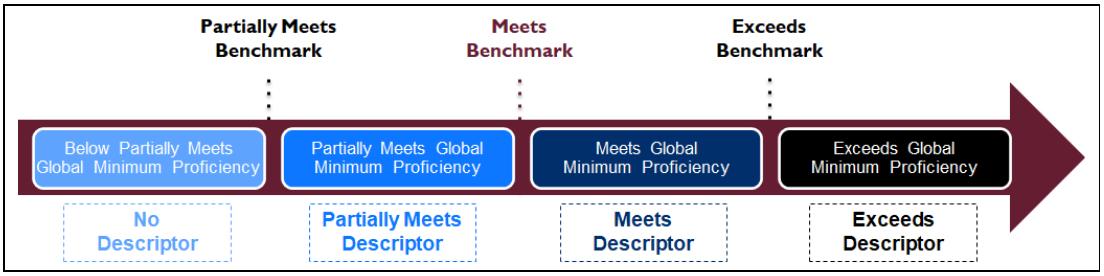


Domain	Construct		Subconstruct		Knowledge or Skill
	N1	Whole numbers	N1.1	Identify and count in whole numbers, and identify their relative magnitude	**N1.1.1** - Count, read, and write whole numbers
					N1.1.2 - Compare and order whole numbers
					N1.1.3 - Skip count forwards or backwards
			N1.2	Represent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

Task 2: Matching

On the afternoon of the fifth day, the panelists received the outcome of their alignment tasks. Subsequently, they received training for the next task: matching the NLA items with the Global proficiency levels and descriptors. Task 2 builds on the panelists' understanding of the items and GPF gained through the alignment activity. The purpose of Task 2 is to further narrow down the expectations of learners measured by each assessment item. The panelists should identify the descriptors (GPDs) of global minimum proficiency that match with the items.

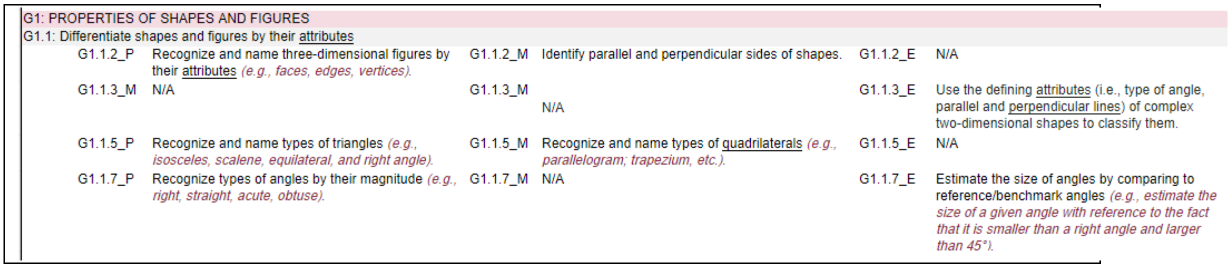
Figure 2. GPLs and GPDs in the Global Proficiency Framework


Partially Meets Benchmark
Meets Benchmark
Exceeds Benchmark
Below Partially Meets Global Minimum Proficiency
Partially Meets Global Minimum Proficiency
Meets Global Minimum Proficiency
Exceeds Global Minimum Proficiency
No Descriptor
Partially Meets Descriptor
Meets Descriptor
Exceeds Descriptor

A Global Proficiency Descriptor (GPD) 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 GPF a learner should be able to demonstrate within a subject at a grade level. The GPDs describe 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 GPDs are organized by domain, construct and subconstruct, with descriptors for each subconstruct. In Table 8 an example is displayed of GPDs for the three GPLs (partially meets, meets and exceed global minimum proficiency).

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



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

They had the morning of the sixth day to work together on the matching task. On the afternoon of day six, they finished Task 2 together for Khmer and Mathematics. For reasons of efficiency, it was decided not to have discussions in several subgroups for both subjects. Both for Khmer and Mathematics full consensus was reached. In both groups there was one item for which consensus was only reached after a long discussion. Both the Khmer and the Mathematics group discussed the outcome of Task 2 at the end of the day.

Observations

The Mathematics group felt that there was a lack of material to practice with; especially the descriptors for matching. Good examples for different issues, like finding the lowest descriptor in the GPF and what the GPL is related to Grade 6 helped more than generically explaining the descriptors.

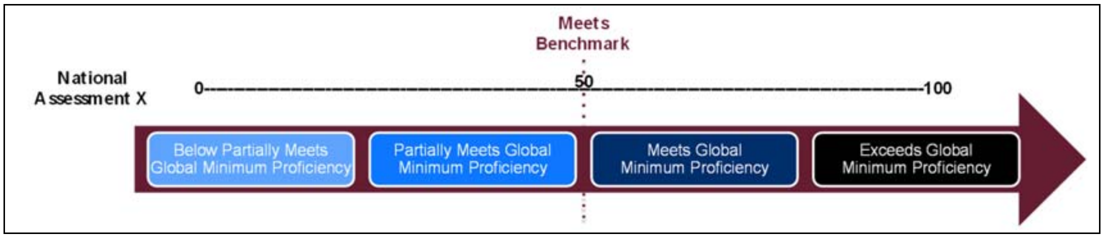
Both groups felt matching was a complex task. Again, the availability of local content facilitators and interpreters was of enormous benefit to the discussion. But the number of items to be discussed and the intricacy of the task took the panelists longer than planned. The agenda of the workshop was robust enough to solve this problem, because it allowed panelists to finalize their discussions on the items on Monday morning.

The matching activity turned out to have consequences for the earlier conclusions on alignment. Not only the number of items considered to be covering subconstructs changed, but also conclusions on alignment. For Mathematics after alignment it was concluded that the NLA was additionally aligned in depth with the GPF, but after matching the conclusion was that it was strongly aligned in depth. Given the fact that after matching there was complete consensus between all panelists for Mathematics, the latter conclusion on alignment might be considered to be more valid.

Task 3: Benchmarking

On the seventh day the panelists received training in setting global benchmarks using the Angoff method. The facilitator first presented a hypothetical example of how the benchmarking method would link a national assessment to the GPF, thus allowing for the calculation of the percentages of students attaining minimum proficiency (see Figure 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


Meets
Benchmark
National
Assessment X
0
50
100
Below Partially Meets Global Minimum Proficiency
Partially Meets Global Minimum Proficiency
Meets Global Minimum Proficiency
Exceeds Global Minimum Proficiency

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)?

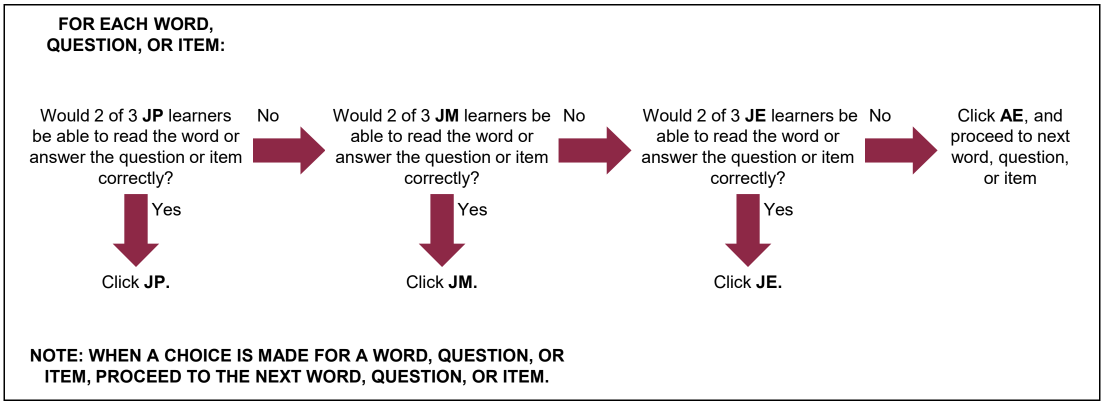
- If "yes," place an "X" under JP and proceed to the next item.
- If "no," ask whether minimally proficient JM learners would be able to answer the item correctly?
    - If "yes," place an "X" under JM and proceed to the next item.
    - If "no," ask whether minimally proficient JE learners would be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


FOR EACH WORD, QUESTION, OR ITEM:
Would 2 of 3 JP learners be able to read the word or answer the question or item correctly?
No
Yes
Click JP.
Would 2 of 3 JM learners be able to read the word or answer the question or item correctly?
No
Yes
Click JM.
Would 2 of 3 JE learners be able to read the word or answer the question or item correctly?
No
Yes
Click JE.
Click AE, and proceed to next word, question, or item
NOTE: WHEN A CHOICE IS MADE FOR A WORD, QUESTION, OR ITEM, PROCEED TO THE NEXT WORD, QUESTION, OR ITEM.

After all panelists finished their first ratings on the eight day, their input was exported from the Google Doc forms and entered in the Excel data entry files for Khmer and Mathematics. The local content facilitators 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
- A table containing a rank order of the items, starting with the item on which disagreement was highest and ending with the item on which disagreement was lowest.
- 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 (not shown in the presentation)

The lead facilitator presented the preliminary results of Round 1 in the afternoon of the ninth day. 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. Subsequently, the lead facilitator described what the panelists had to do in Round 2.

Round 2

On day ten, panelists had to complete their second rating using the same procedure. After the panelists conducted their second ratings, their output was exported from Google Docs to the data entry Excel sheets for Khmer and Mathematics. Like the day before, the local content facilitators tracked the submission of the forms and checked the forms. After the data entry, the file was sent to Cito and the data analyst analyzed the data. On the last day, the results were shared with the panelists after they all had returned the Google Docs workshop evaluation form.

Observations

As expected, the conceptualization of three Just Partially Meets (JP), three Just Meets (JM), and three Just Exceeds (JE) learners based on an understanding of the GPF provided to be the most challenging part of the workshop for the panelists. This is not because something went wrong, but because this is inherently difficult. First, to "switch off" your own intuitions and knowledge based on your own experience in your own country and on the country's curriculum, and instead building a picture of a JP "global learner" based on all the descriptors of the Partially Meets level. Secondly, to decide what it means, based on this picture, that this learner is *just* in the PM level: which tasks at Below Partially Meets level is such a learner able to carry out and which tasks at PM level? And the same for the other two levels. And then to apply this to the actual items on the NLA.

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

The filling in of the forms, by the panelists and by the local content facilitators, went as smoothly as it did with the alignment task.

To help the panelists to have an efficient discussion as possible on differences in rating after the first round, a table was created with a rank order of the items, based on the level of disagreement between all panelists. This table was created by taking the range of ratings into account as well as their dispersion.

A last important observation is that, apparently, in the Policy Linking Toolkit a national assessment is considered to be a linear test which is the same for all members of the specific population, e.g. Grade 6 learners. However, the NLA is not such a test, but consists of several booklets with a certain overlap of items administered to different sets of learners. This makes it an impossibility to have panelists align and match all items to the GPF, because of the large numbers of items. And in addition to this, data analysis working with Item Response Theory instead of Classical Test Theory. There are no guidelines in the Policy Linking Toolkit on the methodology to be used when working with national large scale assessments that are aimed at measuring educational progress in a detailed way. We think there should be, as the same issue will probably be encountered in other countries as well.

Workshop evaluation

At the start of the eleventh day, all panelists were asked to share their opinion about the workshop. Their evaluations are completely anonymous. They were informed that their opinion was important to improve the workshop and to evaluate the validity and reliability of the standard setting process. The panelists had about one hour to answer the questions about:

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

The questions included are presented in the PLT (see also Annex F). To make this activity as simple as possible the questionnaire was translated into Khmer and could be filled in via Google Docs. The evaluation consists of Likert-type scales and open-ended questions on the panelists' satisfaction with the orientation, training, and process.

Observations

One question had to be removed from the questionnaire, because it referred to an activity the panelists did not perform: administering the NLA-items to a group of their own learners

In turned out that another question that had to be in the questionnaire was missing. This was the question asking panelists whether they had had sufficient time to complete the Round 1 ratings. This was discovered before the start of this workshop. But there was a long time between the delivery of all the materials and the actual start of the workshop. And therefore this change in the questionnaire had not taken place.

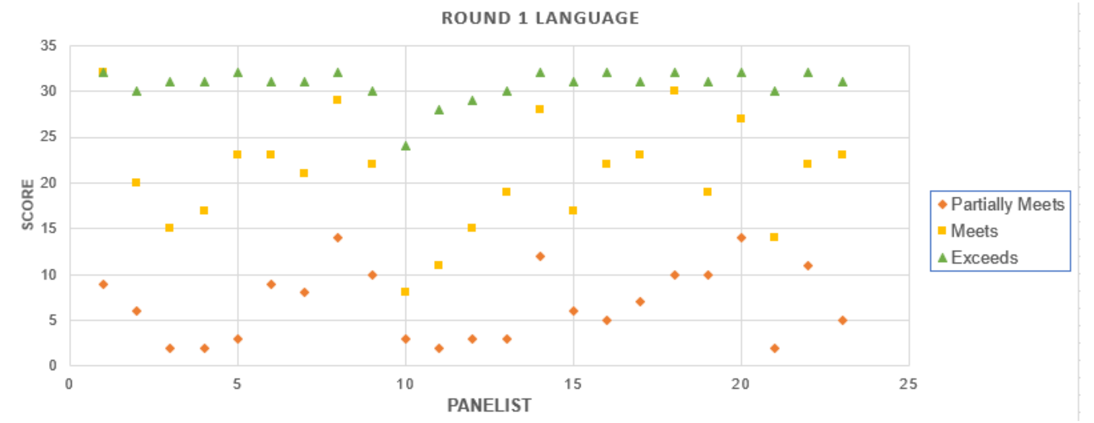
After the data entry it was discovered that there were more respondents than panelists. The group of respondents for Khmer consisted of 24 persons, while there were only 23 panelists. And the number of respondents for Mathematics was 27, while the number of panelists was also 23. The reason for this is that some of the local content facilitators and coordinators also filled in the evaluation form. And because respondents were anonymous, these responses could not be filtered out. However, given the number of panelists and the large similarities in the responses, confounding of the results is negligible.

5. Results of the benchmarking

Round 1

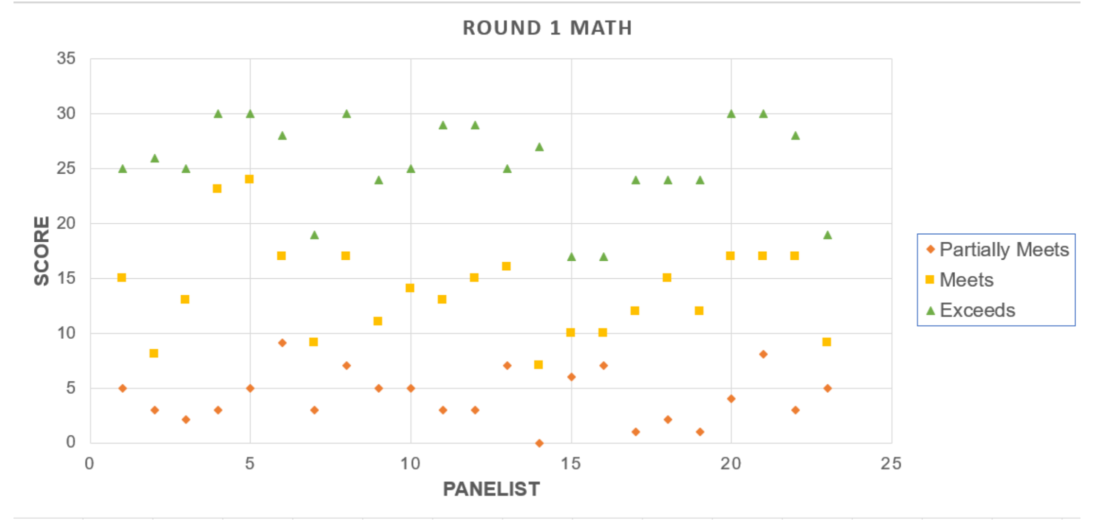
The data analyst and lead facilitator produced summary tables and graphs for the first round, which showed the initial benchmarks, score ranges, and impact data for each Minimum Proficiency Level (see Table 9 and Table 10). In the plenary room the panelists were presented with anonymous normative information on the panelists ratings (see Figure 5 and Figure 6). For Khmer, we saw that the ratings of panelists varied considerably, both for the lowest (Partially meets) and the middle benchmark (Meets). We also see a ceiling effect with the Exceeds benchmark. Exceeds is with a few exceptions almost at the maximum (32).

Figure 5. Anonymous information on the panelists' ratings for Khmer Round 1



For Mathematics, we saw that the ratings of panelists also varied considerably, both for the lowest (Partially meets) and the middle benchmark (Meets). We also see a small ceiling effect with the Exceeds benchmark. Five of the panelists put the Exceeds benchmark at the maximum score of 30.

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



After round 1 the benchmark was calculated as the average of the panelists' benchmarks. The average benchmark was truncated, as stipulated in the policy linking toolkit. For Khmer, the impact information shows that only 3.4% of the learners would fall at the Below Partially Meets Minimum Global Proficiency level; that 39.1% would fall at the Partially Meets Global Minimum Proficiency Level; 49.2% at the Meets Minimum GPL and 8.3% at the Exceeds Global Minimum Proficiency level using Round 1 benchmarks (see Table 9).

Table 9. Round 1 benchmarks, score range and impact for Khmer with 32 items

Minimum Proficiency Level	Round 1 Benchmark	Score Range			Percentage of Learners
			Female	Male	Total
Below Partially Meets	N/A	0 - 5	2.8%	4.3%	3.4%
Partially Meets	6.7	6 - 19	34.5%	43.8%	39.1%
Meets	20.9	20 - 29	52.7%	45.4%	49.2%
Exceeds	30.7	30 - 32	10.0%	6.6%	8.3%

For Mathematics, the impact information shows that only 1.1% would fall in the Below Partially Meets Minimum Global Proficiency level; that 37.3% would fall at the Partially Meets Global Minimum Proficiency Level; 52.2% at the Meets Minimum GPL and 9.3% at the Exceeds Global Minimum Proficiency level using Round 1 benchmarks (see Table 10).

Table 10. Round 1 benchmarks, score range and impact for Mathematics with 30 items

Minimum Proficiency Levels	Round 1 Benchmark	Score Range			Percentage of Learners
			Female	Male	Total
Below Partially Meets	N/A	0 - 3	1.2%	1.5%	1.1%
Partially Meets	4.4	4 - 12	36.6%	37.7%	37.3%
Meets	13.9	13 - 24	53.1%	50.7%	52.2%
Exceeds	25.4	25 - 30	9.1%	10.2%	9.3%

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, based on the ordering of items presented after round 1. After the discussion the panelists individually conducted the Round 2 ratings and submitted their forms. The data analyst produced a parallel set of summary tables and graphs with final benchmarks.

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

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

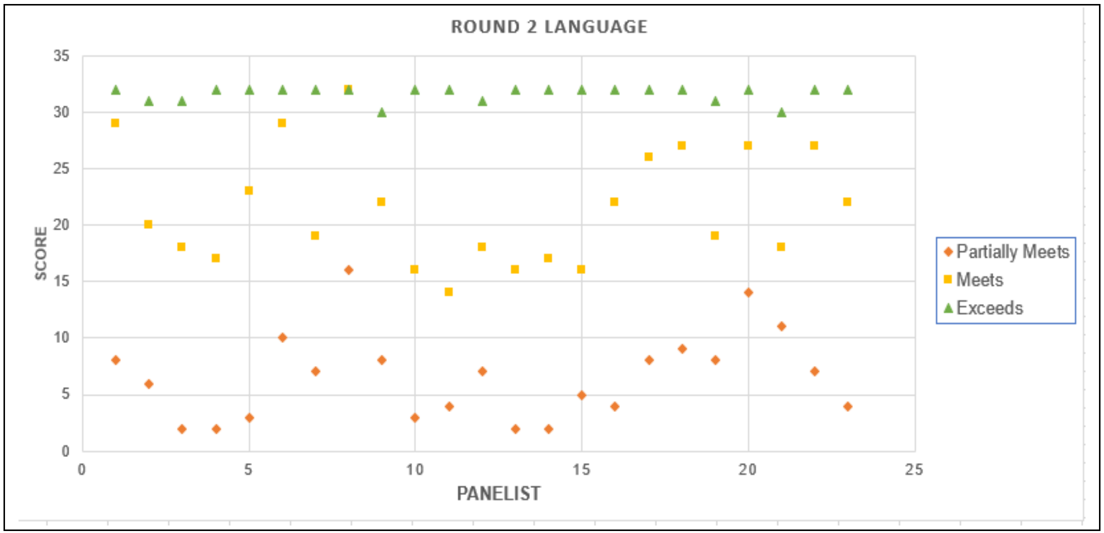
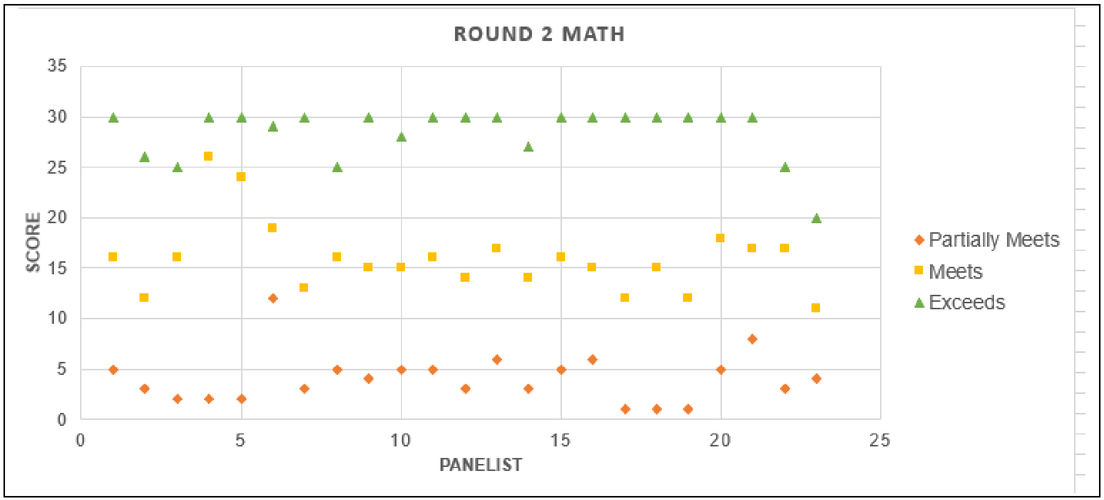


Figure 8. Anonymous information on the panelist's ratings of Mathematics Round 2



For Khmer, the results show that in Round 2 only 3.4% fall in the Below Partially Meets level and 43.3 % fall in the Partially Meets Level (see Table 11). Furthermore, 48.7% fall in the Meets level and only 4.6% in the Exceeds level. The benchmarks were set slightly higher in round 2 than in round 1. The Below Partially Meets benchmark remains stable between rounds 1 and 2. Both the Meets and Exceeds benchmarks increase by one score point (see Table 12). The Exceeds benchmark is set at almost at the top of the scale, which is an indication of a ceiling effect.

Table 11. Round 2 benchmarks, score range and impact for Khmer with 32 items

Minimum Proficiency Level	Round 2 Benchmark	Score Range			Percentage of Learners
			Female	Male	Total
Below Partially Meets	N/A	0 - 5	2.8%	4.3%	3.4%
Partially Meets	6.5	6 - 20	38.8%	47.8%	43.3%
Meets	21.5	21 - 30	52.7%	44.2%	48.7%
Exceeds	31.7	31 - 32	5.6%	3.8%	4.6%

Table 12. Comparison of Round 1 benchmarks and Round 2 benchmarks for Khmer with 32 items

Minimum Proficiency Level	Round 1 Benchmark	Round 1 Percentage of Learners	Round 2 Benchmark	Round 2 Percentage of Learners
Below Partially Meets	N/A	3.4%	N/A	3.4%
Partially Meets	6.7	39.1%	6.5	43.3%
Meets	20.9	49.2%	21.5	48.7%
Exceeds	30.7	8.3%	31.7	4.6%

For Mathematics, the results show that in Round 2 only 1.1% of learners fall in the Below Partially Meets level and 54.1% fall in the Partially Meets Level (see Table 13). Furthermore, 41.7% fall in the Meets level and only 3.1% in the Exceeds level. Comparison of Rounds 1 and 2 shows that Partially Meets benchmark remains stable. Both the Meets and Exceeds benchmarks go upwards with three score points (see Table 14). After round 2 a higher percentage of learners falls in the Partially Meets proficiency level and a lower percentage in the Meets proficiency level. Only 3.1% of the learners fall in the Exceeds level.

Table 13. Round 2 benchmarks, score range and impact for Mathematics with 30 items

Minimum Proficiency Level	Round 2 Benchmark	Score Range			Percentage of Learners
			Female	Male	Total
Below Partially Meets	N/A	0 - 3	1.2%	1.5%	1.1%
Partially Meets	4.0	4 - 15	53.4%	53.8%	54.1%
Meets	16.2	16 - 27	42.8%	41.1%	41.7%
Exceeds	28.7	28 - 30	2.6%	3.6%	3.1%

Table 14. Comparison of Round 1 benchmarks and Round 2 benchmarks for Mathematics with 30 items

Minimum Proficiency Level	Round 1 Benchmark	Percentage of Learners	Round 2 Benchmark	Percentage of Learners
Below Partially Meets	N/A	1.1%	N/A	1.1%
Partially Meets	4.4	37.3%	4.0	54.1%
Meets	13.9	52.2%	16.2	41.7%
Exceeds	25.4	9.3%	28.7	3.1%

6. Evaluation of the Standard Setting Process

Internal Evaluation SEM, Panelist Consistency and Panelists' Agreement

In addition to calculating benchmarks and impact data, the PLT also requires calculating measures of consistency and presenting evaluation feedback results. These measures of consistency are reported in Table 15 and Table 16.

As shown in Table 15, 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 Mathematics with 30 items, and not much higher for Khmer with 32 items The results show that the SEM is relatively small for Khmer for the Exceeds benchmarks. This is a consequence of a ceiling effect for this benchmark.

Table 15. Standard Error of Measurement by Round

	SEM by Benchmark					
	Round 1			Round 2		
Subjects	Partially Meets	Meets	Exceeds	Partially Meets	Meets	Exceeds
Khmer	0.80	1.07	0.13	0.80	1.21	0.36
Mathematics	0.52	0.70	0.42	0.48	0.91	0.87

As panelist consistency and panelists' agreement are concerned, the results show that the inter-rater consistency for both Khmer 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 PLT values of 0.80 or greater are desirable, as they indicate substantial agreement between the panelists. Both for English and Mathematics the inter-rater consistency was above the 0.80 (see Table 16).

The intra-rater consistency index evaluates the panelists' overall consistency in estimating item difficulty. Intra-rater consistency is calculated for each panelist across all items on the assessment. The value ranges between 0 and 1. A lower value indicates high consistency and a higher value indicates low consistency. We see that the intra-rater consistency is quite high (given the scale of 0 to 1): with the exception of Round 1 for Mathematics the values are above .7.

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

	Round 1		Round 2	
Subject	Inter-Rater Consistency	Intra-Rater Consistency	Inter-Rater Consistency	Intra-Rater Consistency
Khmer	0.81	0.72	0.84	0.72
Mathematics	0.81	0.48	0.87	0.85

Procedural Evaluation

All panelists shared their opinion about the workshop through a questionnaire (see Annex F). This questionnaire was translated into Khmer to cater for panelists with a lower mastery of English. 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. A distinction was made between the two groups to be able to notice relevant differences in appraisal of the workshop between Khmer and Mathematics panelists.

Note that the number of respondents differs from the number of panelists. For the Khmer group there are 24 respondents and 23 panelists, while there are 27 respondents from the Mathematics group where there were also 23 respondents. The reason for this is that some of the local content facilitators and coordinators also filled in the evaluation form. And because respondents were anonymous, these responses could not be filtered out. However, given the number of panelists and the large similarities in the responses, confounding of the outcomes is negligible.

On average, we see that the respondents were quite positive about the workshop, both for Khmer and Mathematics (See Table 17 and Table 18). For the Khmer group all six aspects received an average score above 4 (on a scale of 1 to 5). The overall evaluation shows that the respondents for Khmer are overall very positive: 4.60 on a scale of 1 to 5 (the neutral category has been added to the scale, which was missing in the example in the Policy Linking Toolkit).

Table 17. Workshop evaluation results for Khmer

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	4.50	0.36	24
The training on the NLA Survey[5]	1-5	5	4.35	0.38	24
The training on the alignment methodology	1-5	5	4.34	0.41	24
The training on the matching methodology	1-5	5	4.29	0.36	24
The training on the benchmark-setting (Angoff) methodology[6]	1-5	10	4.35	0.32	24
Benchmark Round 2 evaluation	1-5	8	4.28	0.31	24
Overall evaluation	1-5	3	4.60	0.48	24

For the Mathematics group the results are comparable: here also all six aspects received an average score above 4 (on a scale of 1 to 5). The overall evaluation shows that the respondents for Mathematics are overall very positive: 4.52 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).

[5] 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”).

[6] One question was missing on the form “I was able to follow the instructions and complete the Round 1 form accurately”.

Table 18. Workshop evaluation results for Khmer

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	4.50	0.38	27
The training on the NLA[7]	1-5	5	4.38	0.38	27
The training on the alignment methodology	1-5	5	4.30	0.37	27
The training on the matching methodology	1-5	5	4.31	0.41	27
The training on the benchmark-setting (Angoff) methodology[8]	1-5	10	4.33	0.37	27
Benchmark Round 2 evaluation	1-5	8	4.24	0.30	27
Overall evaluation	1-5	3	4.52	0.44	27

[7] 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”).

[8] One question was missing on the form “I was able to follow the instructions and complete the Round 1 form accurately”.

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

The results of the policy linking workshop in Cambodia are summarized in Table 19 and Table 20. In the PLT (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 Khmer (Table 19), the intra-rater and inter-rater reliability meet the requirements. The standard error of measurement is low. However, the third benchmark ("Exceeds") might not be valid. There is not much variation for the Exceeds benchmark and a number of panelists set the benchmark at the maximum score, so there is a ceiling effect (even though this is not mentioned as a criterium). The panel has good gender representation and a good geographical representation. There is a good ratio of teachers to subject matter experts (see Table 2). All teacher panelists are experienced and certified teachers. The panelists rated their understanding of the GPF, assessment, and policy linking methodology above 4 and they felt on average comfortable with their Round 2 evaluations and final benchmarks. The adequacy of the policy linking workshop for Khmer in Cambodia can be considered to be good.

For Mathematics (Table 20), the intra-rater and inter-rater reliability meet the requirements. The standard error of measurement is low. The panel has good gender representation and a good geographical representation. There is a good ratio of teachers to subject matter experts (See Table 2).). All teacher panelists are experienced and certified teachers. The panelists rated their understanding of the GPF, assessment, and policy linking methodology above 4 and they felt on average comfortable with their Round 2 evaluations and final benchmarks. The adequacy of the policy linking workshop for mathematics in Cambodia can be considered to be good.

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

Question		Criteria	Response
a)	What was the intra-rater reliability for the second round of ratings?	The intra-rater reliability will vary depending on the number of items on the assessment. The panel will provide guidance on how they determined acceptability.	0.72
b)	What was the inter-rater reliability for the second round of ratings?	The inter-rater reliability should be at least .80.	0.84
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: 32 0.48 (Partially Meets) 0.91 (Meets) 0.87 (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.	• Teachers: 50% female; 50% male SME's: 23% female, 77% male • N/A • N/A • NA
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.	• Teacher mean > 15 years SME mean > 7 years • 23 of 23 • 23 of 23 • Yes • Yes • Yes

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 – **4.46** • I understand the relationship between domains, constructs, subconstructs, knowledge and skills, and GPDs - **4.46** • The GPDs were clear and easy to understand - **4.33** **NLA** • I understand the purpose of the assessment - **4.42** • I understand the constructs assessed in the assessment - **4.38** • I understand how the assessment is administered - **4.33** **Alignment** • I understand the purpose of alignment - **4.38** • I understand the alignment methodology - **4.29** • I understand the difference between no fit, partial fit, and complete fit - **4.29** **Matching** • I understand the purpose of matching - **4.21** • I understand the matching methodology - **4.38** • I understand how the alignment activity links to the matching activity - **4.29** **Benchmarking methodology** • I understand the process I need to follow to complete the benchmarking exercise - **4.38** • I understand how the benchmarking methodology links to the steps on alignment and matching - **4.33** • I understand the difficulty level of the assessment items - **4.29** **Benchmark round 2** • I understand the data on others' ratings - **4.25** • I understand the item difficulty data and how it relates to this process - **4.42** • I understand the impact data and how it relates to this process - **4.25** **Comfortable with Round 2** • How comfortable are you with your final performance predictions? - **4.79**

Table 20. Summary of Results for Criteria for Policy Linking Validity Mathematics Grade 6

Question		Criteria	Response
g)	What was the intra-rater reliability for the second round of ratings?	The intra-rater reliability will vary depending on the number of items on the assessment. The panel will provide guidance on how they determined acceptability.	0.85
h)	What was the inter-rater reliability for the second round of ratings?	The inter-rater reliability should be at least .80.	0.87
i)	What was the Standard Error of Measurement (SEM) at each global proficiency level?	SEM should be appropriate for each global proficiency level reported. There is no maximum SEM provided in this document, since it will depend on the number of items in the assessment.	Number of items: 30 0.48 (Partially Meets) 0.91 (Meets) 0.87 (Exceeds)
j)	To what extent were the panelists representative of the target population of schools being reported on?	Panelists should be selected to ensure: • Gender representation – The panelists must be selected to ensure gender balance, both for the teachers and non-teachers. • Geographical representation – The teachers (and non-teachers, if possible) must be selected to ensure representation from regions, provinces, and/or states. • Ethnic and/or linguistic representation (where applicable) • Representation of crisis-and-conflict-affected areas.	• Teachers: 40% female; 60% male SME's: 8% female, 92% male • N/A • N/A • NA
k)	To what extent did the panelists meet the other selection criteria described in the Policy Linking Toolkit?	Panelists should all have: • Several years of teaching experience in the grade level for which they are providing ratings (classroom teachers) • Skills in the subject area (all panelists) • Skills in the different languages of instruction and assessment (all panelists) • Knowledge of learners of different proficiency levels, including at least some who would meet the requirements of the meets minimum proficiency level and some who would meet the requirements of the exceeds minimum proficiency level (all panelists) • Knowledge of the instructional environment (all panelists) • Experience administering the assessment(s) being used for the policy linking workshop.	• Teacher mean > 12 years SME mean > 13 years • 23 of 23 • 23 of 23 • Yes • Yes • Yes

l)	To what extent did panelists report understanding the GPF, assessment, and policy linking methodology? And, to what extent did they feel comfortable with their Round 2 evaluations and final benchmarks?	On a five-point Likert scale, with 1 being strongly disagree, very uncomfortable, etc. and 5 being strongly agree, very comfortable, etc., the average rating for each of these criteria should be 4 or above.	**GPF** • I understand the purpose of the GPF - **4.44** • I understand the relationship between domains, constructs, subconstructs, knowledge and skills, and GPDs - **4.52** • The GPDs were clear and easy to understand - **4.41** • **NLA** • I understand the purpose of the assessment - **4.44** • I understand the constructs assessed in the assessment - **4.41** I understand how the assessment is administered - **4.30** **Alignment** • I understand the purpose of alignment - **4.37** • I understand the alignment methodology - **4.30** • I understand the difference between no fit, partial fit, and complete fit - **4.30** **Matching** • I understand the purpose of matching - **4.37** • I understand the matching methodology - **4.37** • I understand how the alignment activity links to the matching activity - **4.30** **Benchmarking methodology** • I understand the process I need to follow to complete the benchmarking exercise - **4.30** • I understand how the benchmarking methodology links to the steps on alignment and matching - **4.22** • I understand the difficulty level of the assessment items - **4.26** **Benchmark round 2** • I understand the data on others' ratings - **4.30** • I understand the item difficulty data and how it relates to this process - **4.33** • I understand the impact data and how it relates to this process - **4.26** **Comfortable with Round 2** • How comfortable are you with your final performance predictions? - **4.74**

8. Conclusions and Recommendations

Due to the travel restrictions of COVID-19, UIS hosted the workshop using a videoconferencing platform (Zoom). All participants worked from home. For most if not all participants, this was the first time they participated in an international workshop and the first time using a videoconferencing platform. The Cito facilitators had earlier experience with running standard setting workshops from a distance, either from an earlier workshop in this UIS series or elsewhere. But for all this was the first fully remote workshop.

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

According to the panelists for Khmer at the end of the alignment activity, the NLA is both in breadth and in depth strongly aligned to the GPF. In the eyes of the panelists for Mathematics, at the end of the alignment exercise the NLA is strongly aligned in breadth and additionally aligned in depth. However, after the matching activity agreement increased and based on these results, the conclusion is that the NLA is also strongly aligned in depth as Mathematics is concerned. Mathematics is both in depth and breadth strongly aligned to the GPF for grade 6. Furthermore, the panelists managed to reach complete consensus on the matching both for English and for mathematics. The final benchmarks of the panelists show a good consistency, which makes the benchmarks useable for comparing, aggregating, and tracking learning outcomes for the NLA in Cambodia.

Recommendations

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

Workshop Preparation

Collecting workshop materials and pre-workshop analyses

- In the policy linking toolkit, the materials to be collected, such as the assessment instrument and the data file, are clearly described. The UIS activity plan ensured the workshop materials were exchanged in a timely manner.
- It is important that the Review Panel 4.1.1 is in place. To ensure the reliability of the results of the workshop, an independent panel needs to evaluate before the workshop whether an assessment meets the standards required to proceed with policy linking.

Creating workshop materials

- A technical test should be held well in advance of the workshop. A technical test with all locations and participants will also make clear in advance if back-up material or equipment is needed (e.g. the WhatsApp contact) and to troubleshoot any technology issues.
- The fact that a lot of the key documentation was translated into Khmer, made it easier for the panelists to familiarize themselves with the GPF and to execute the tasks. This compensated for the fact that the workshop had to be organized online and not all panelists probably had a sufficient mastery of English to understand the relatively complex topics involved with policy linking,

- Working with two virtual separate Zoom meeting rooms (and a digital plenary room) worked well. It prevented a lot of confusion, which often occurs when people participate for the first time in Zoom and work with digital break-out rooms. Adding different backgrounds for Khmer and Mathematics panelists was an excellent idea of the EQA team. This made checking if all participants were in the correct session simple and efficient.

Training the local content facilitators

- The local content facilitators and the coordinator proved to have more than enough expertise to perform all their tasks. The knowledge and skills already present, helped the efficiency of the training and the understanding of all the different parts of the workshop. However, if the local content facilitators are less well equipped, the training provided, might prove to be not effective enough.

Implementing the fully remote workshop

- To facilitate the sessions and discussions, the presence of translation from English to the local language and vice versa is a necessity. Two interpreters should assist lead and content facilitators with their communication. Simultaneous translation should be planned for all sessions.
- A two-week workshop as implemented in this instance is possible for a fully remote format. The schedule has enough room to mitigate the issue of an unstable internet connection. The recording of sessions makes it possible for panelists who missed parts of a session to review everything in time to be ready for the next session againin the six-day blended workshop is very tight and forms a risk for the quality of the results. In a six-day workshop, there is very little room for adapting to unforeseen circumstances or solving technical problems, such as occurred during the first day. With this schedule there is also enough room for adapting to unforeseen circumstances or solving technical problems, although we were lucky enough to encounter only one small moment of about a minute where the internet connection of the lead facilitator was completely lost. Panelists did have moments of losing connection, but none of these had an impact of the quality of their output.
- The process of collecting and checking the forms and doing data-entry locally, made the process much smoother. The fact that the EQAD team decided to convert the different rating forms to Google Docs was a brilliant addition to the procedure. This made the data entry for all the panelists easy and prevented the use of paper forms.
- When conducting a fully remote workshop with all panelists joining from home, there should be enough room in the agenda to account for unforeseen circumstances. So a schedule tighter than the one used is not recommended. Also, a good and frequent contact between local and international content facilitators, for example via WhatsApp and/or telephone, and between local content facilitators and panelists, via Telegram in this instance, is a necessity.

Familiarization

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

- The agency or governmental organization that has created the assessment, is best suited to give a presentation about the assessment, instead of the lead facilitator.
- The presentations, both plenary and in the subgroups, should be more pedagogically informed, with suitable involvement of the panelists: more practicing than presenting. This to enhance engagement of the panelists and to avoid them feeling overwhelmed.

- The presentations should take the starting point of the panelists more into account. The panelists seem to have difficulty with the many acronyms and technical words. A didactical approach can help in making the slides clearer and less word-based aiming at more language independent information. A translation of the slides helps as well.
- The two plenary starting presentations/activities on the first day: Overview of the policy linking and Overview of the GPF should be given by an experienced trainer with in-depth knowledge of policy linking and of the GPF.
- Perform the familiarization of the GPF in two steps: up to and including the knowledge or skill statements before the Alignment task, and the GPD and GPL between the Alignment task and the Matching task. This avoids possible confusion by the panelists and a possible overload of information on the first day.
- In conducting a workshop for more subjects and/or grades, it would be helpful if the assessments for the different groups were of similar length.

Task 1: Alignment

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

Task 2: Matching

- Give clearer instructions in the PLT on how to deal with items that match with a descriptor from a grade other than the one under consideration.
- Perform an extra check by letting both the local and the international content facilitator administer the conclusions and comparing afterwards.
- Schedule more time for the matching task, especially for the consensus discussions.

Task 3: Benchmarking

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

9. References

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

Birnbaum (1968). Some latent trait models and their use in inferring an examinee's ability. In: F.M. Lord & M.R. Novick (Eds.) *Statistical theories of mental test scores.* (pp 397-424). Reading: Addison-Wesley

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

The reason for this is that some of the local content facilitators and coordinators also filled in the evaluation form. And because repondents were anonymous, these responses could not be filtered out. However, given the number of panelists and the large similarities in the responses, confounding of the outcomes is negligible.

EQAD (2017). *Results of Grade Six Student Achievement from the National Assessment in 2016.* Ministry of Education, Youth and Sports Cambodia.

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

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

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

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

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

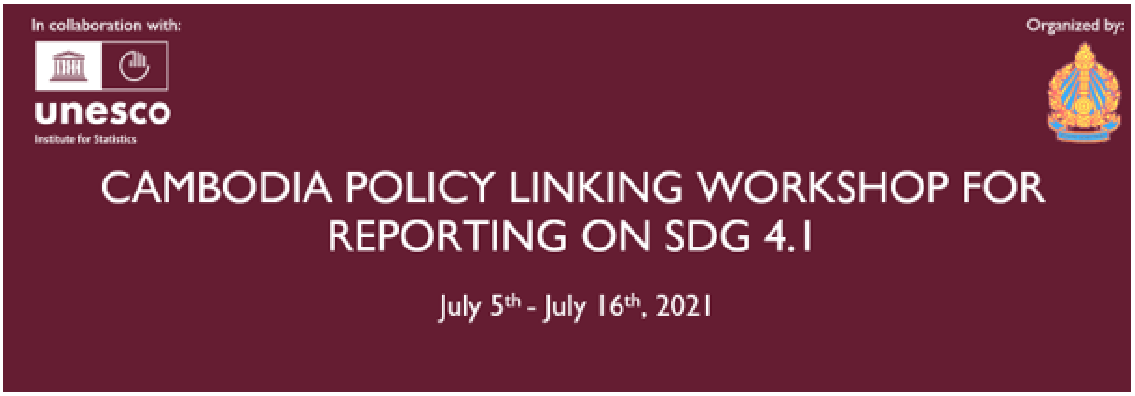
USAID, UNESCO Institute for Statistics (UIS), UK's Foreign, Commonwealth, and Development Office (FCDO), The Bill and Melinda Gates Foundation, Australian Council for Education Research (ACER), World Bank (2020b). *Global Proficiency Framework for Reading Grades 1 to 9*. Downloaded from https://www.edu-links.org/sites/default/files/media/file/GPF_Reading_Final_D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

USAID, World Bank, UNESCO Institute for Statistics (UIS), UK's Foreign, Commonwealth, and Development Office (FCDO), Australian Council for Education Research (ACER), MSI (2020c). Policy Linking for Measuring Global Learning Outcomes Toolkit: Linking Assessments to the

Global Proficiency Framework. Downloaded from https://www.edu-links.org/sites/default/files/media/file/Policy_Linking_for_Measuring_Global_Learning_Outcomes_Final.pdf.

10. Annexes

Annex A: Agenda for the fully remote 11-day workshop



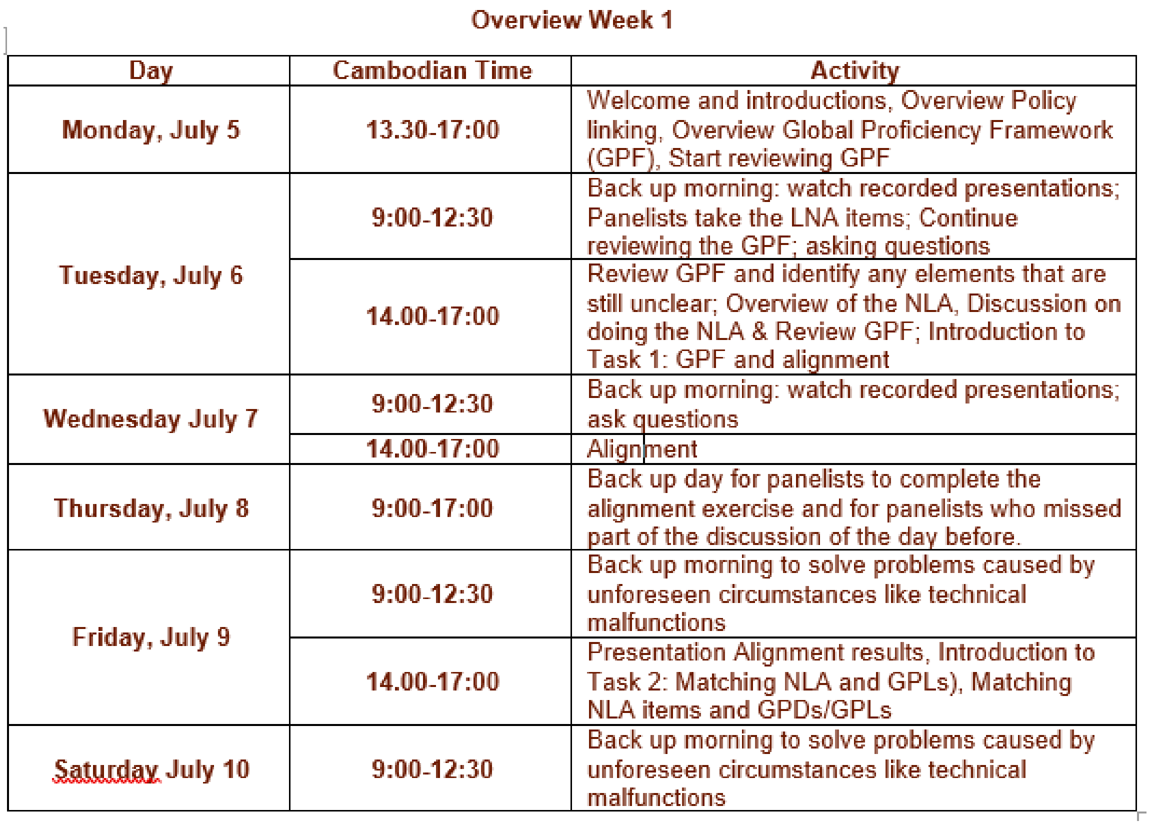
In collaboration with:

unesco
Institute for Statistics

Organized by:

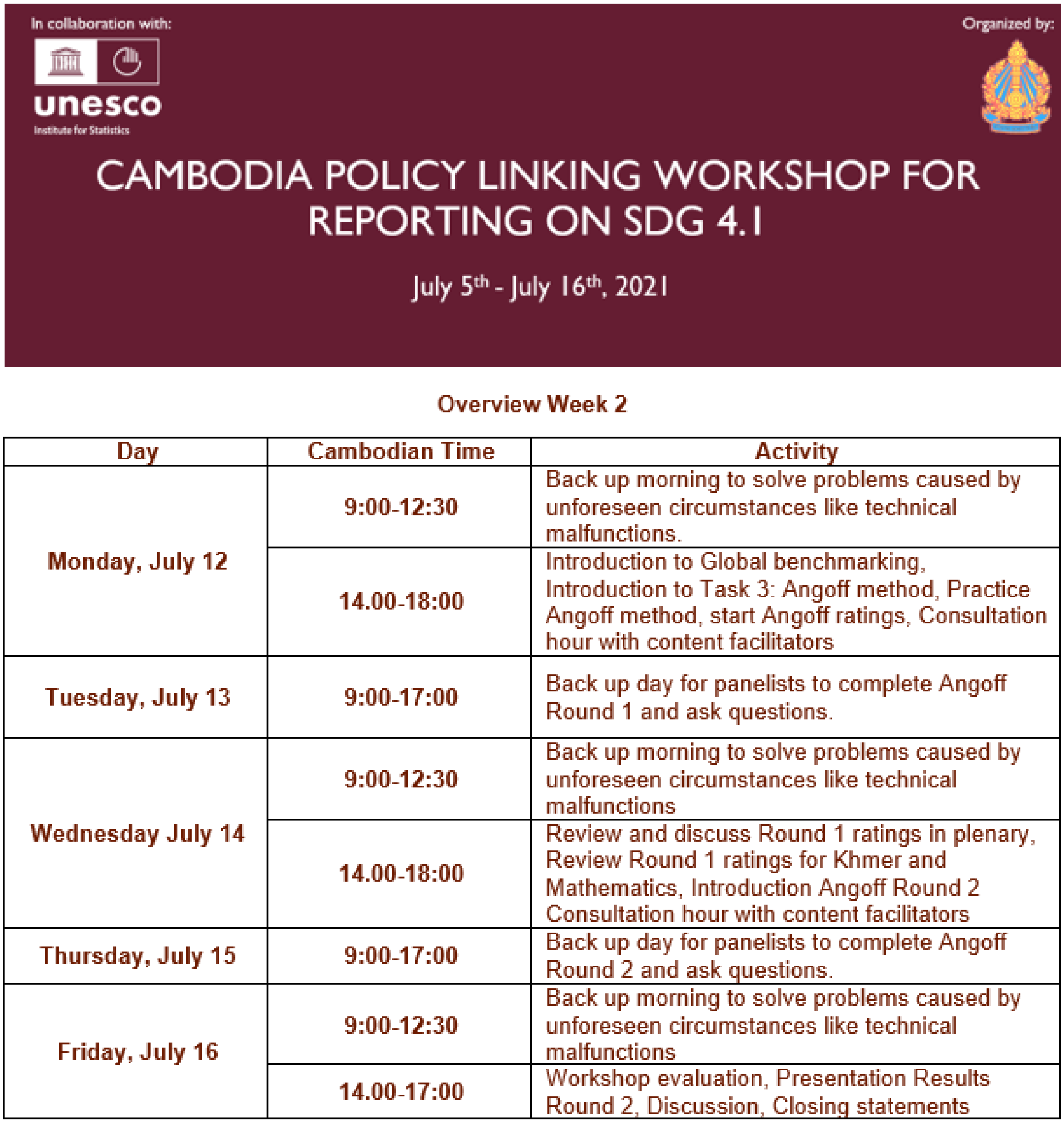
CAMBODIA POLICY LINKING WORKSHOP FOR REPORTING ON SDG 4.1

July 5th - July 16th, 2021



Overview Week 1

Day	Cambodian Time	Activity
Monday, July 5	13.30-17:00	Welcome and introductions, Overview Policy linking, Overview Global Proficiency Framework (GPF), Start reviewing GPF
Tuesday, July 6	9:00-12:30	Back up morning: watch recorded presentations; Panelists take the LNA items; Continue reviewing the GPF; asking questions
	14.00-17:00	Review GPF and identify any elements that are still unclear; Overview of the NLA, Discussion on doing the NLA & Review GPF; Introduction to Task 1: GPF and alignment
Wednesday July 7	9:00-12:30	Back up morning: watch recorded presentations; ask questions
	14.00-17:00	Alignment
Thursday, July 8	9:00-17:00	Back up day for panelists to complete the alignment exercise and for panelists who missed part of the discussion of the day before.
Friday, July 9	9:00-12:30	Back up morning to solve problems caused by unforeseen circumstances like technical malfunctions
	14.00-17:00	Presentation Alignment results, Introduction to Task 2: Matching NLA and GPLs), Matching NLA items and GPDs/GPLs
Saturday July 10	9:00-12:30	Back up morning to solve problems caused by unforeseen circumstances like technical malfunctions



In collaboration with:

unesco

Institute for Statistics

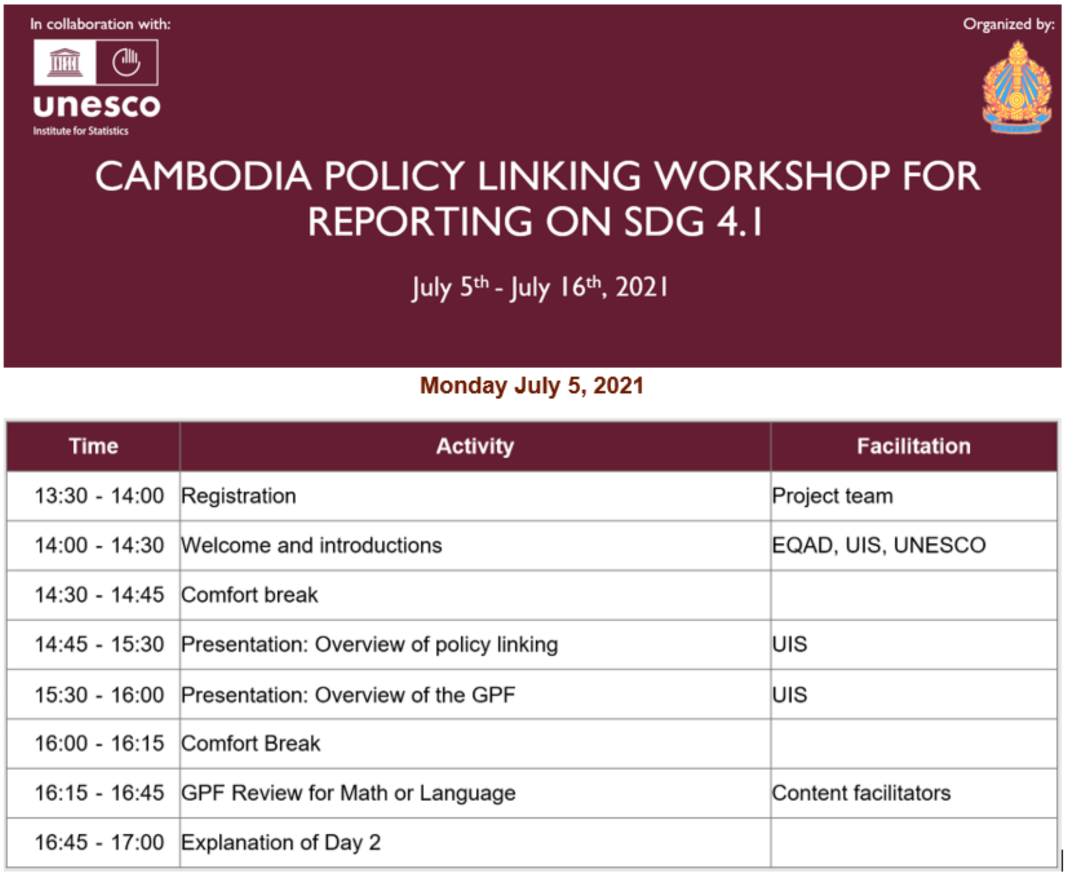
Organized by:

CAMBODIA POLICY LINKING WORKSHOP FOR REPORTING ON SDG 4.1

July 5th - July 16th, 2021

Overview Week 2

Day	Cambodian Time	Activity
Monday, July 12	**9:00-12:30**	Back up morning to solve problems caused by unforeseen circumstances like technical malfunctions.
	14.00-18:00	Introduction to Global benchmarking, Introduction to Task 3: Angoff method, Practice Angoff method, start Angoff ratings, Consultation hour with content facilitators
Tuesday, July 13	**9:00-17:00**	Back up day for panelists to complete Angoff Round 1 and ask questions.
Wednesday July 14	**9:00-12:30**	Back up morning to solve problems caused by unforeseen circumstances like technical malfunctions
	14.00-18:00	Review and discuss Round 1 ratings in plenary, Review Round 1 ratings for Khmer and Mathematics, Introduction Angoff Round 2 Consultation hour with content facilitators
Thursday, July 15	**9:00-17:00**	Back up day for panelists to complete Angoff Round 2 and ask questions.
Friday, July 16	**9:00-12:30**	Back up morning to solve problems caused by unforeseen circumstances like technical malfunctions
	14.00-17:00	Workshop evaluation, Presentation Results Round 2, Discussion, Closing statements



In collaboration with:

unesco

Institute for Statistics

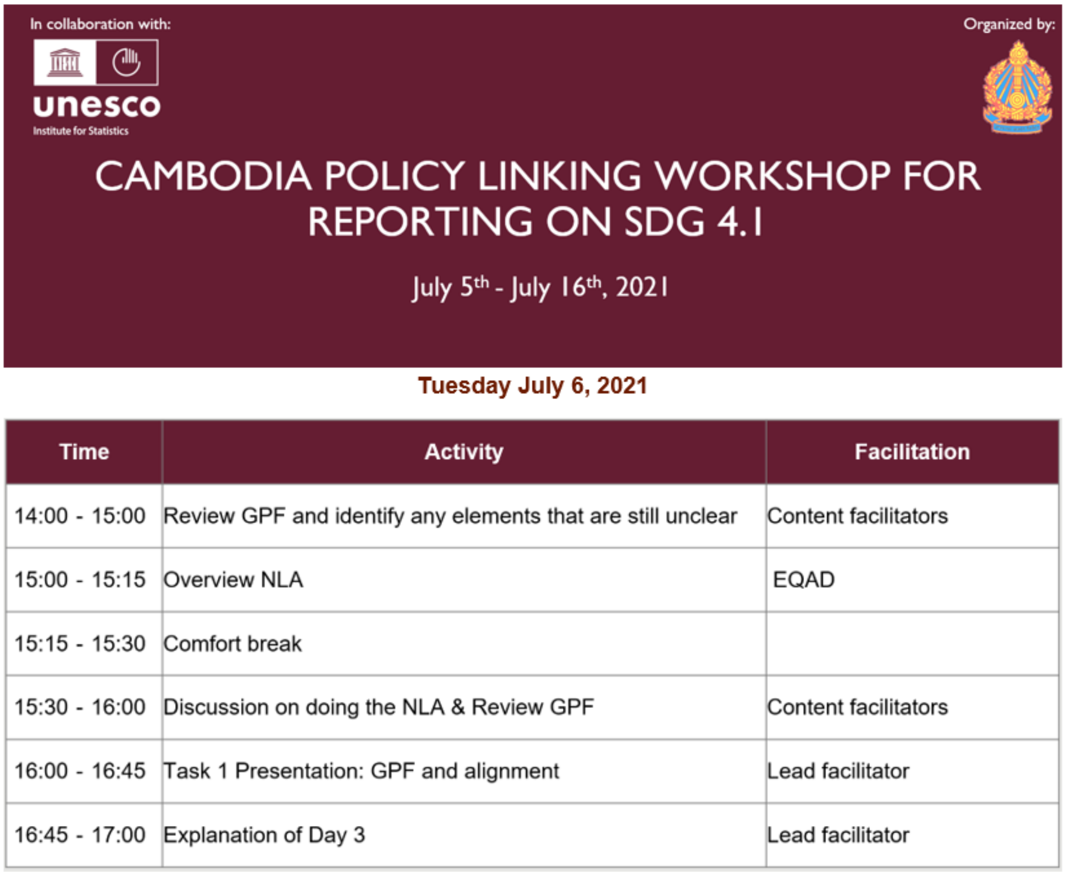
Organized by:

CAMBODIA POLICY LINKING WORKSHOP FOR REPORTING ON SDG 4.1

July 5th - July 16th, 2021

Monday July 5, 2021

Time	Activity	Facilitation
13:30 - 14:00	Registration	Project team
14:00 - 14:30	Welcome and introductions	EQAD, UIS, UNESCO
14:30 - 14:45	Comfort break	
14:45 - 15:30	Presentation: Overview of policy linking	UIS
15:30 - 16:00	Presentation: Overview of the GPF	UIS
16:00 - 16:15	Comfort Break	
16:15 - 16:45	GPF Review for Math or Language	Content facilitators
16:45 - 17:00	Explanation of Day 2	



In collaboration with:

unesco

Institute for Statistics

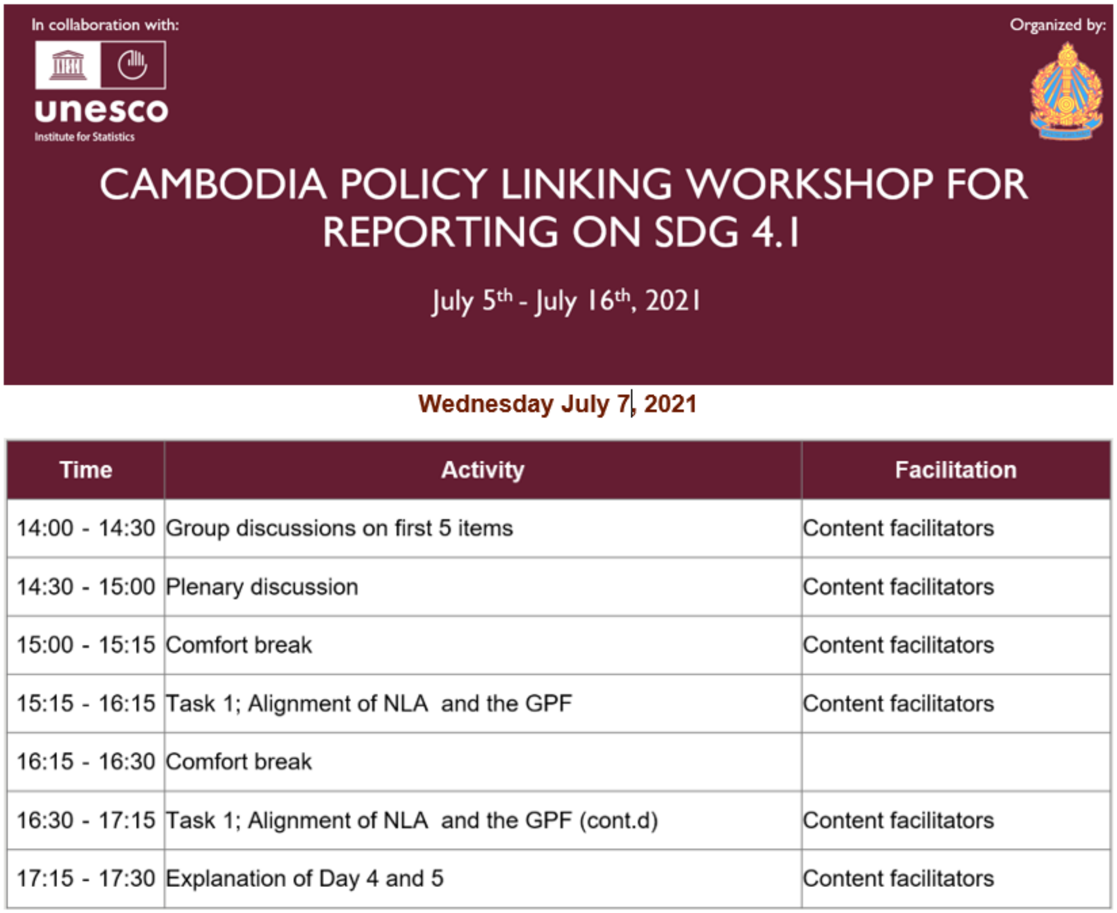
Organized by:

CAMBODIA POLICY LINKING WORKSHOP FOR REPORTING ON SDG 4.1

July 5th - July 16th, 2021

Tuesday July 6, 2021

Time	Activity	Facilitation
14:00 - 15:00	Review GPF and identify any elements that are still unclear	Content facilitators
15:00 - 15:15	Overview NLA	EQAD
15:15 - 15:30	Comfort break	
15:30 - 16:00	Discussion on doing the NLA & Review GPF	Content facilitators
16:00 - 16:45	Task 1 Presentation: GPF and alignment	Lead facilitator
16:45 - 17:00	Explanation of Day 3	Lead facilitator



In collaboration with:

unesco
Institute for Statistics

Organized by:

CAMBODIA POLICY LINKING WORKSHOP FOR REPORTING ON SDG 4.1

July 5th - July 16th, 2021

Wednesday July 7, 2021

Time	Activity	Facilitation
14:00 - 14:30	Group discussions on first 5 items	Content facilitators
14:30 - 15:00	Plenary discussion	Content facilitators
15:00 - 15:15	Comfort break	Content facilitators
15:15 - 16:15	Task 1; Alignment of NLA and the GPF	Content facilitators
16:15 - 16:30	Comfort break	
16:30 - 17:15	Task 1; Alignment of NLA and the GPF (cont.d)	Content facilitators
17:15 - 17:30	Explanation of Day 4 and 5	Content facilitators

In collaboration with:

unesco
Institute for Statistics

Organized by:

CAMBODIA POLICY LINKING WORKSHOP FOR REPORTING ON SDG 4.1

July 5th - July 16th, 2021

Friday July 9, 2021

Time	Activity	Facilitation
14:00 - 14:30	Task 1 Presentation: Alignment results	Lead facilitator
14:30 - 15:00	Task 2 Presentation: Matching assessments and GPDs/GPLs	Content facilitators
15:00 - 15:15	Comfort break	
15:15 - 16:15	Task 2 Presentation: Matching assessments and GPDs/GPLs (continued)	Content facilitators
16:15 - 16:30	Comfort break	
16:30 - 17:15	Task 2 Activity: Matching assessment items and GPDs/GPLs	Content facilitators
17:15 - 17:30	Explanation Day 6	Content facilitators

In collaboration with:

unesco
Institute for Statistics

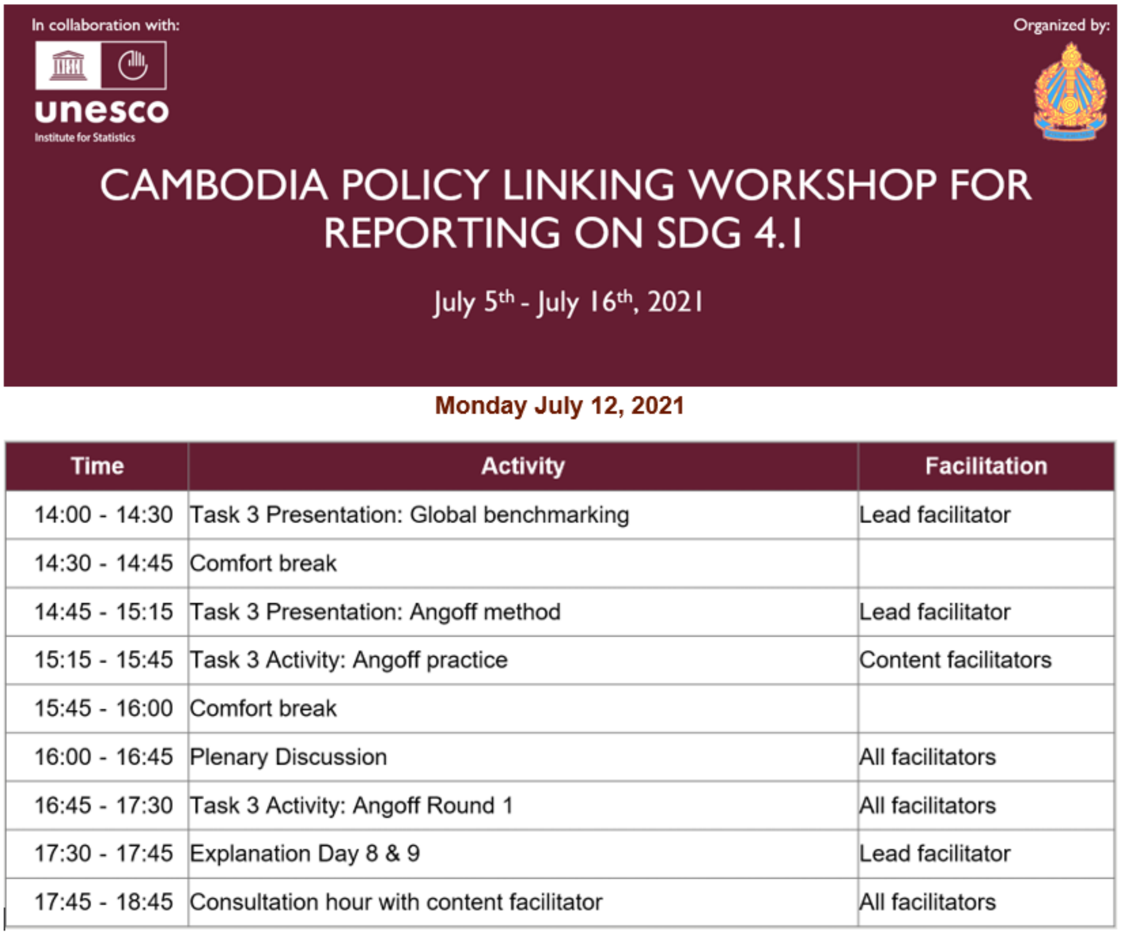
Organized by:

CAMBODIA POLICY LINKING WORKSHOP FOR REPORTING ON SDG 4.1

July 5th - July 16th, 2021

Saturday July 10, 2021

Time	Activity	Facilitation
14:00 - 15:15	Small groups complete Task 2 together	Content facilitators
15:15 - 15:30	Comfort break	
15:30 - 16:45	Plenary discussion: Matching assessment items and GPDs/GPLs and results of matching	Content facilitators
16:45 - 17:00	Explanation Day 7	Content facilitators



In collaboration with:

unesco
Institute for Statistics

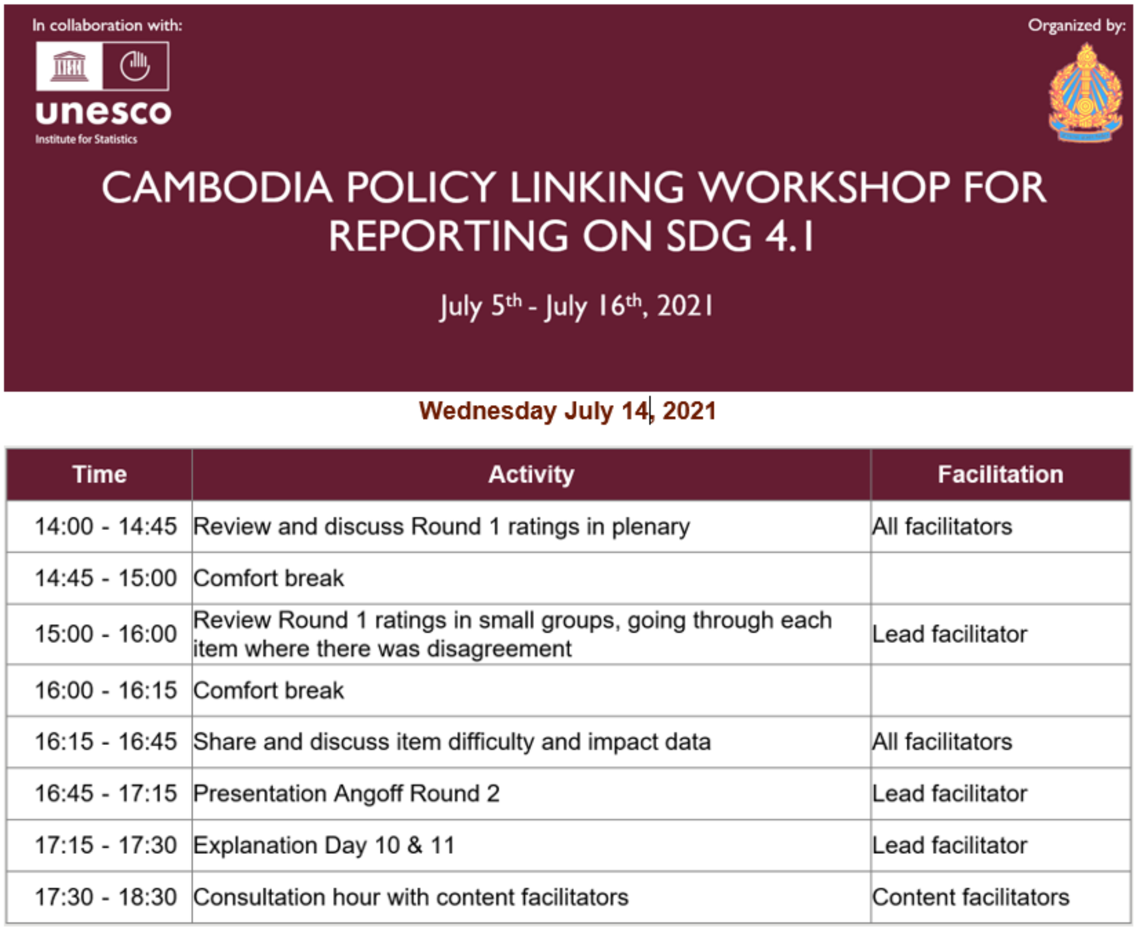
Organized by:

CAMBODIA POLICY LINKING WORKSHOP FOR REPORTING ON SDG 4.1

July 5th - July 16th, 2021

Monday July 12, 2021

Time	Activity	Facilitation
14:00 - 14:30	Task 3 Presentation: Global benchmarking	Lead facilitator
14:30 - 14:45	Comfort break	
14:45 - 15:15	Task 3 Presentation: Angoff method	Lead facilitator
15:15 - 15:45	Task 3 Activity: Angoff practice	Content facilitators
15:45 - 16:00	Comfort break	
16:00 - 16:45	Plenary Discussion	All facilitators
16:45 - 17:30	Task 3 Activity: Angoff Round 1	All facilitators
17:30 - 17:45	Explanation Day 8 & 9	Lead facilitator
17:45 - 18:45	Consultation hour with content facilitator	All facilitators



In collaboration with:

unesco
Institute for Statistics

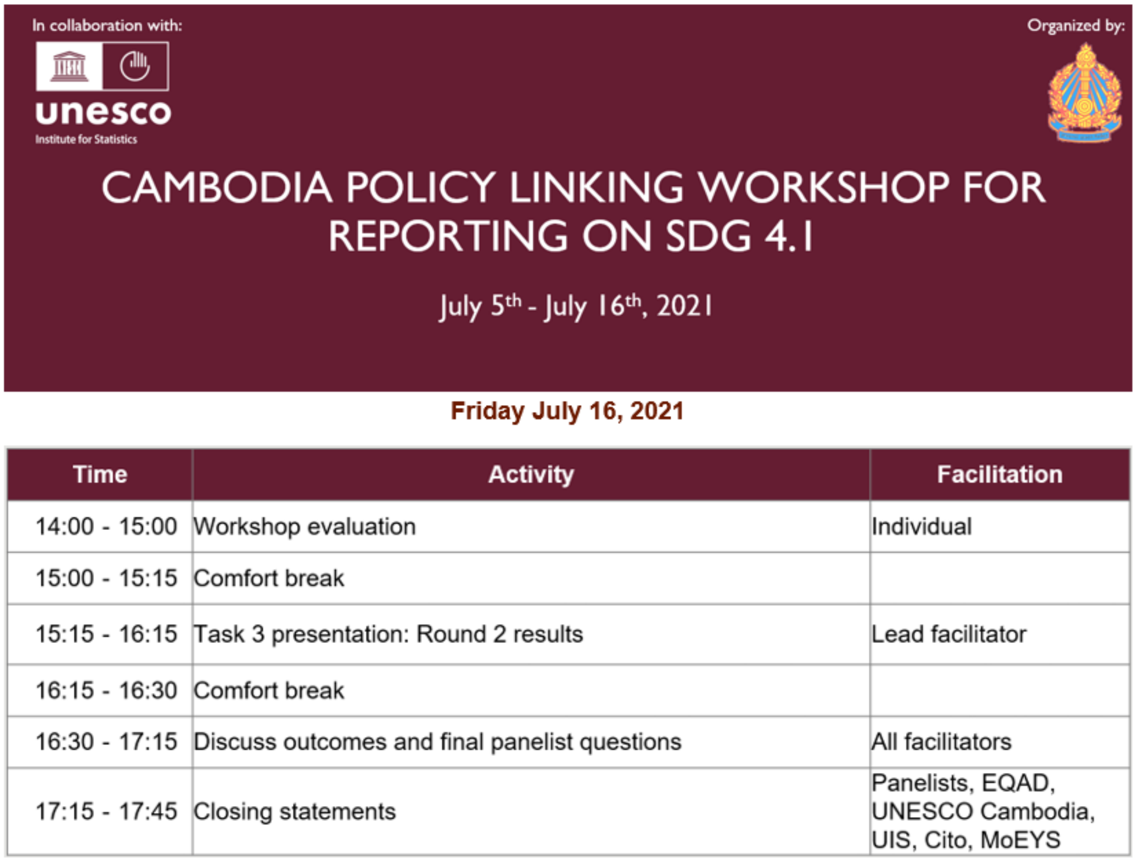
Organized by:

CAMBODIA POLICY LINKING WORKSHOP FOR REPORTING ON SDG 4.1

July 5th - July 16th, 2021

Wednesday July 14, 2021

Time	Activity	Facilitation
14:00 - 14:45	Review and discuss Round 1 ratings in plenary	All facilitators
14:45 - 15:00	Comfort break	
15:00 - 16:00	Review Round 1 ratings in small groups, going through each item where there was disagreement	Lead facilitator
16:00 - 16:15	Comfort break	
16:15 - 16:45	Share and discuss item difficulty and impact data	All facilitators
16:45 - 17:15	Presentation Angoff Round 2	Lead facilitator
17:15 - 17:30	Explanation Day 10 & 11	Lead facilitator
17:30 - 18:30	Consultation hour with content facilitators	Content facilitators



In collaboration with:

unesco
Institute for Statistics

Organized by:

CAMBODIA POLICY LINKING WORKSHOP FOR REPORTING ON SDG 4.1

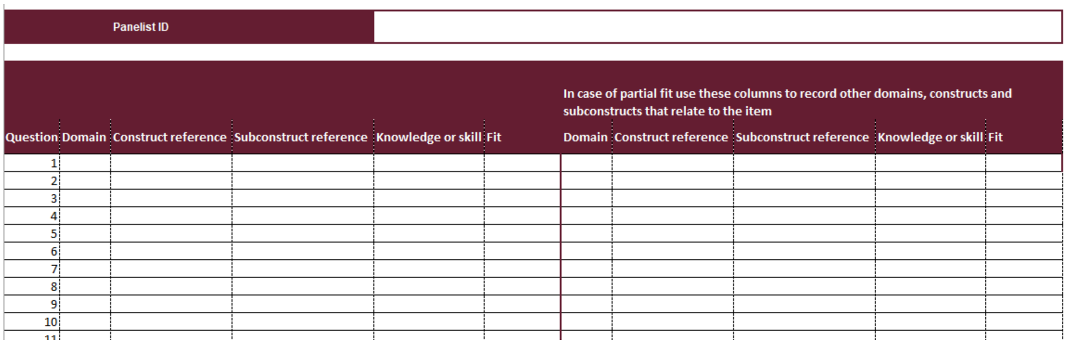
July 5th - July 16th, 2021

Friday July 16, 2021

Time	Activity	Facilitation
14:00 - 15:00	Workshop evaluation	Individual
15:00 - 15:15	Comfort break	
15:15 - 16:15	Task 3 presentation: Round 2 results	Lead facilitator
16:15 - 16:30	Comfort break	
16:30 - 17:15	Discuss outcomes and final panelist questions	All facilitators
17:15 - 17:45	Closing statements	Panelists, EQAD, UNESCO Cambodia, UIS, Cito, MoEYS

Annex B: Example of the forms

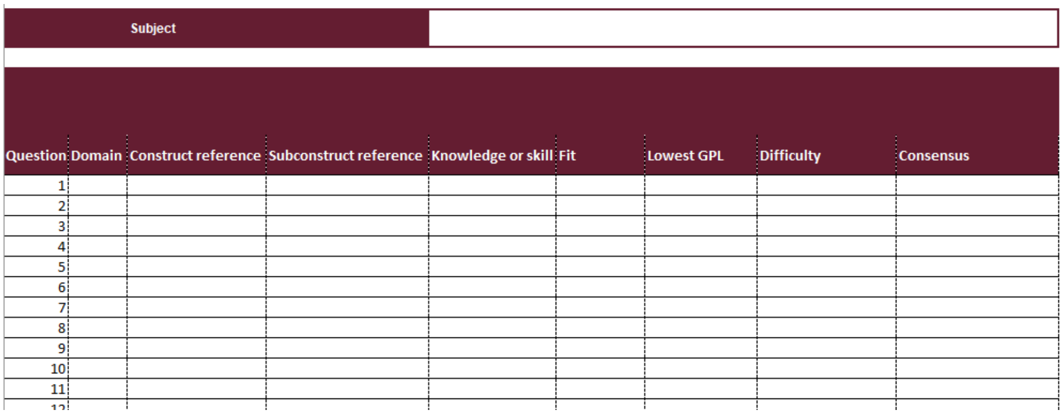
Figure 9. Alignment rating form Khmer and Mathematics (English version)



Panelist ID	

						In case of partial fit use these columns to 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										

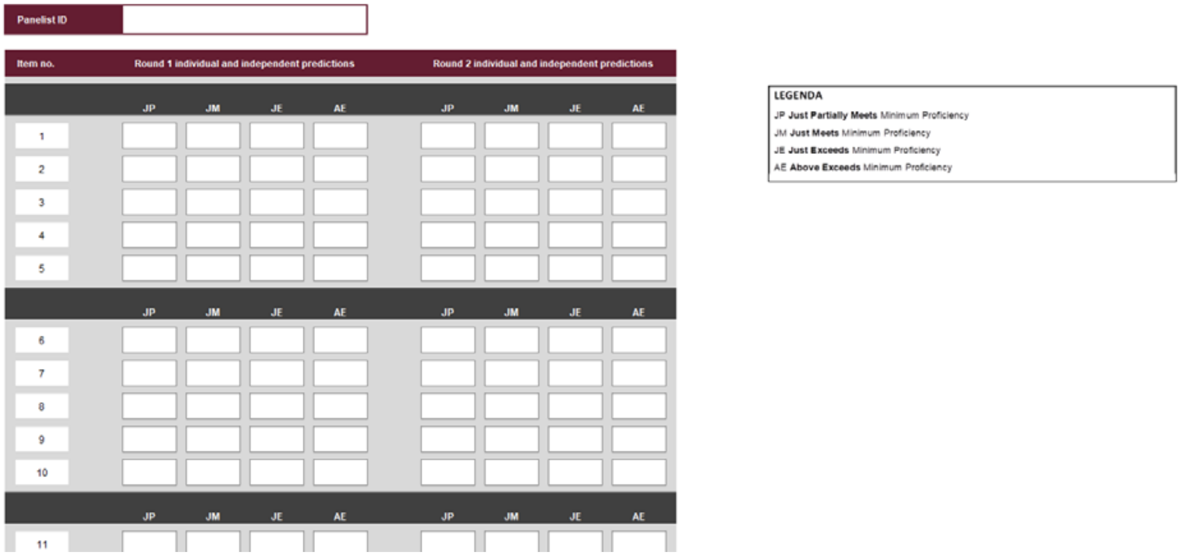
Figure 10. Matching form for the local content facilitator (English version)



Subject	

Question	Domain	Construct reference	Subconstruct reference	Knowledge or skill	Fit	Lowest GPL	Difficulty	Consensus
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								

Figure 11. Item rating form (English version)

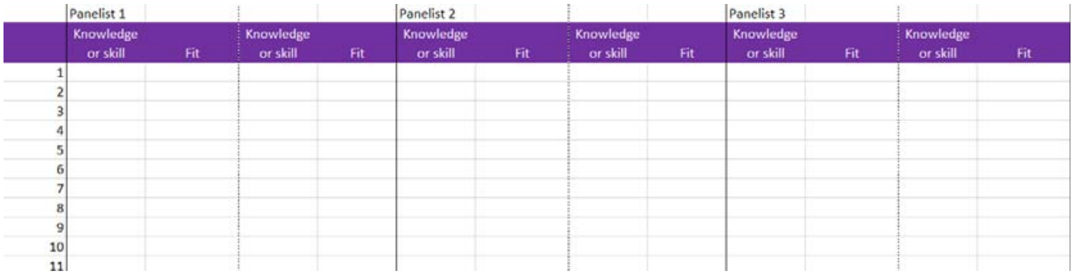


Panelist ID	

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								
	JP	JM	JE	AE	JP	JM	JE	AE
11								

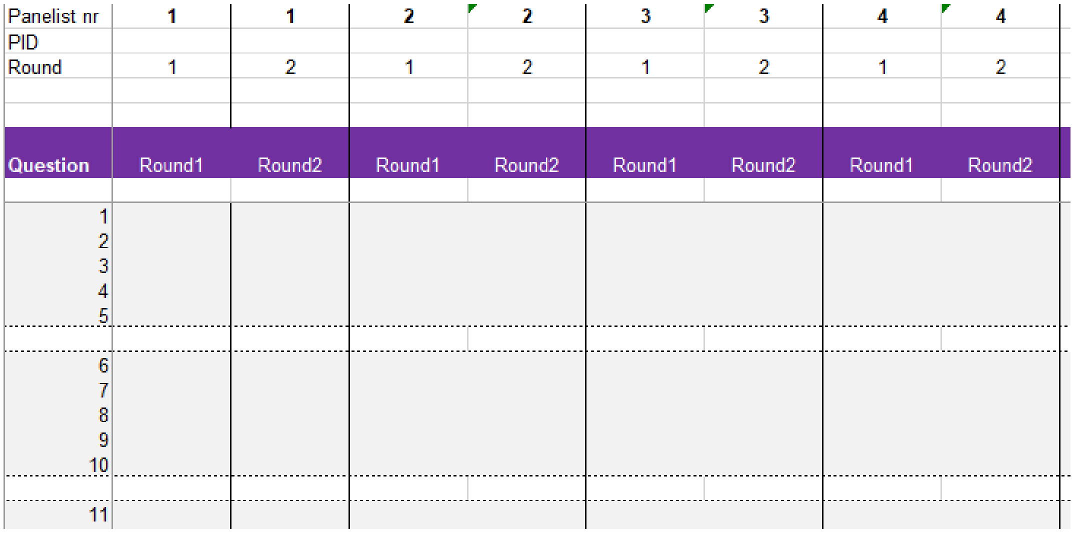
LEGENDA
JP **Just Partially Meets** Minimum Proficiency
JM **Just Meets** Minimum Proficiency
JE **Just Exceeds** Minimum Proficiency
AE **Above Exceeds** Minimum Proficiency

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



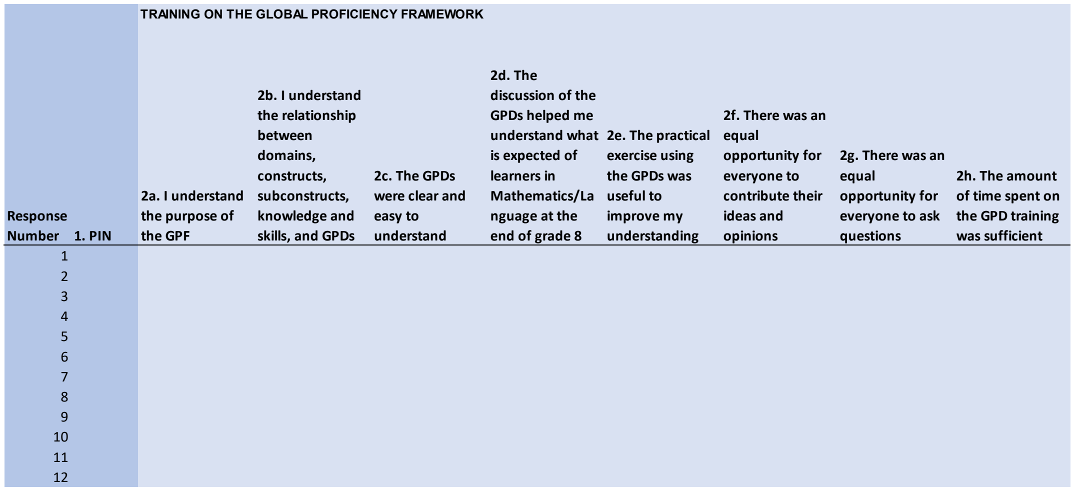
	Panelist 1				Panelist 2				Panelist 3			
	Knowledge or skill	Fit	Knowledge or skill	Fit	Knowledge or skill	Fit	Knowledge or skill	Fit	Knowledge or skill	Fit	Knowledge or skill	Fit
1												
2												
3												
4												
5												
6												
7												
8												
9												
10												
11												

Figure 13. Data entry file for Item rating results



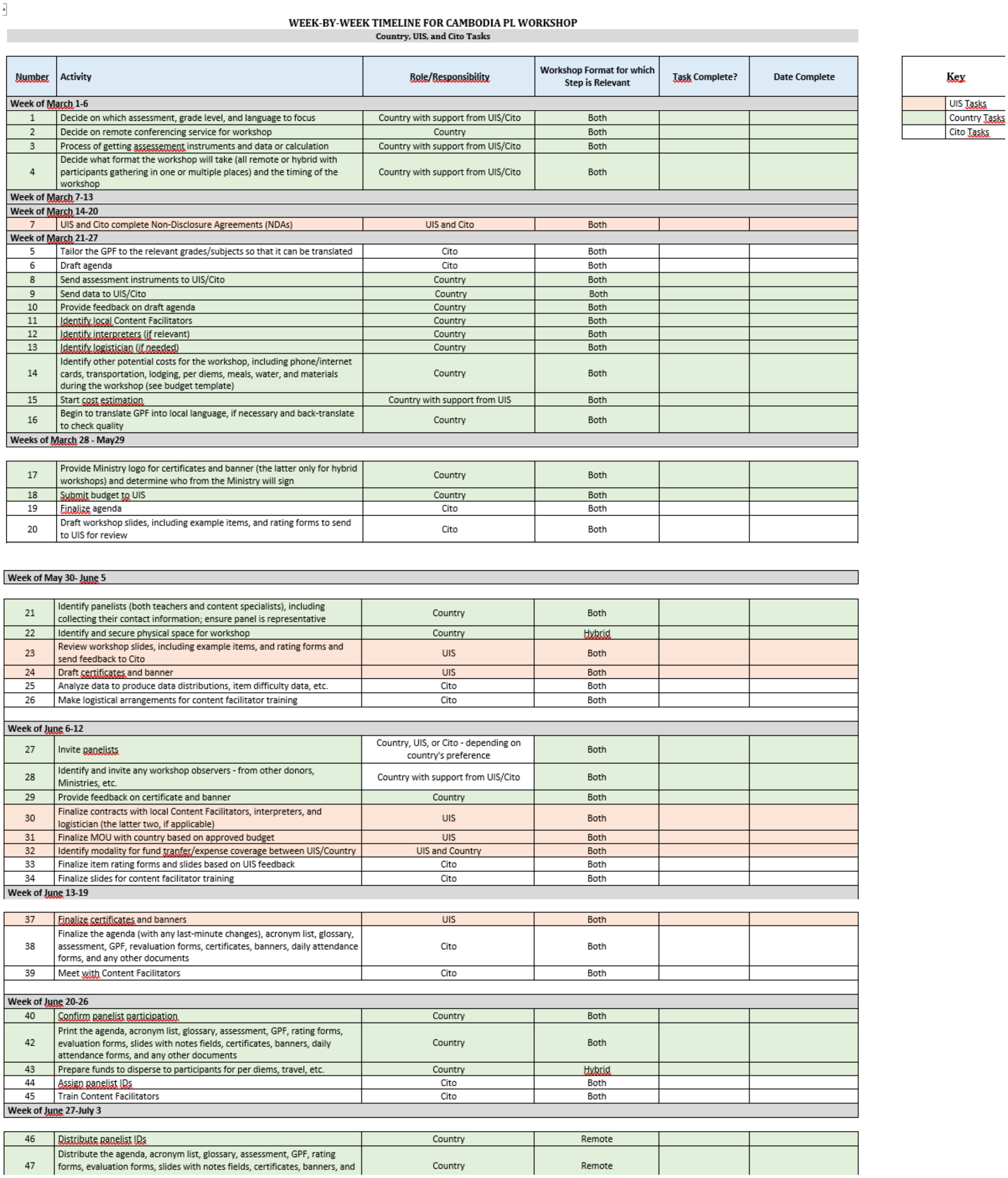
Panelist nr	1	1	2	2	3	3	4	4
PID								
Round	1	2	1	2	1	2	1	2
Question	Round1	Round2	Round1	Round2	Round1	Round2	Round1	Round2
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								

Figure 14. Data entry file for the Evaluation form (English version)



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

Annex C: UIS Activity plan



WEEK-BY-WEEK TIMELINE FOR CAMBODIA PL WORKSHOP

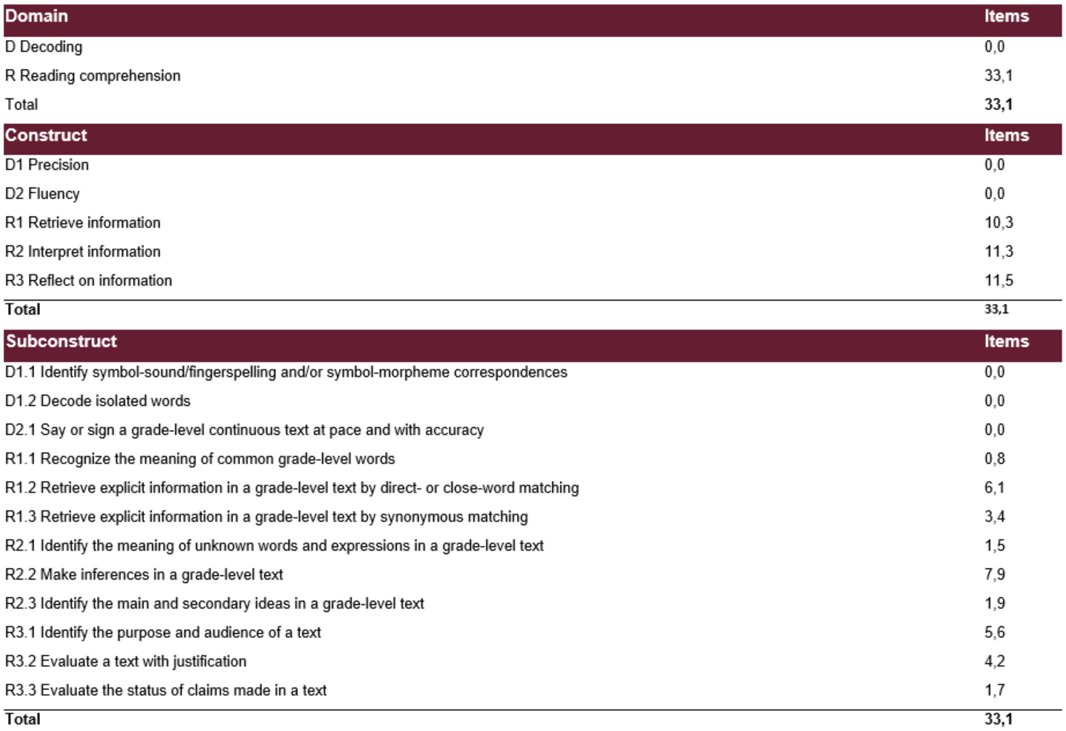
Country, UIS, and Cito Tasks

Number	Activity	Role/Responsibility	Workshop Format for which Step is Relevant	Task Complete?	Date Complete
Week of March 1-6					
1	Decide on which assessment, grade level, and language to focus	Country with support from UIS/Cito	Both		
2	Decide on remote conferencing service for workshop	Country	Both		
3	Process of getting assessement instruments and data or calculation	Country with support from UIS/Cito	Both		
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		
Week of March 7-13					
Week of March 14-20					
7	UIS and Cito complete Non-Disclosure Agreements (NDAs)	UIS and Cito	Both		
Week of March 21-27					
5	Tailor the GPF to the relevant grades/subjects so that it can be translated	Cito	Both		
6	Draft agenda	Cito	Both		
8	Send assessment instruments to UIS/Cito	Country	Both		
9	Send data to UIS/Cito	Country	Both		
10	Provide feedback on draft agenda	Country	Both		
11	Identify local Content Facilitators	Country	Both		
12	Identify interpreters (if relevant)	Country	Both		
13	Identify logistician (if needed)	Country	Both		
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		
Weeks of March 28 - May29					
17	Provide Ministry logo for certificates and banner (the latter only for hybrid workshops) and determine who from the Ministry will sign	Country	Both		
18	Submit budget to UIS	Country	Both		
19	Finalize agenda	Cito	Both		
20	Draft workshop slides, including example items, and rating forms to send to UIS for review	Cito	Both		
Week of May 30- June 5					
21	Identify panelists (both teachers and content specialists), including collecting their contact information; ensure panel is representative	Country	Both		
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		
26	Make logistical arrangements for content facilitator training	Cito	Both		
Week of June 6-12					
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		
Week of June 13-19					
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		
Week of June 20-26					
40	Confirm panelist participation	Country	Both		
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		
Week of June 27-July 3					
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	Country	Remote		

Key	
🟧	UIS Tasks
🟩	Country Tasks
⬜	Cito Tasks

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

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

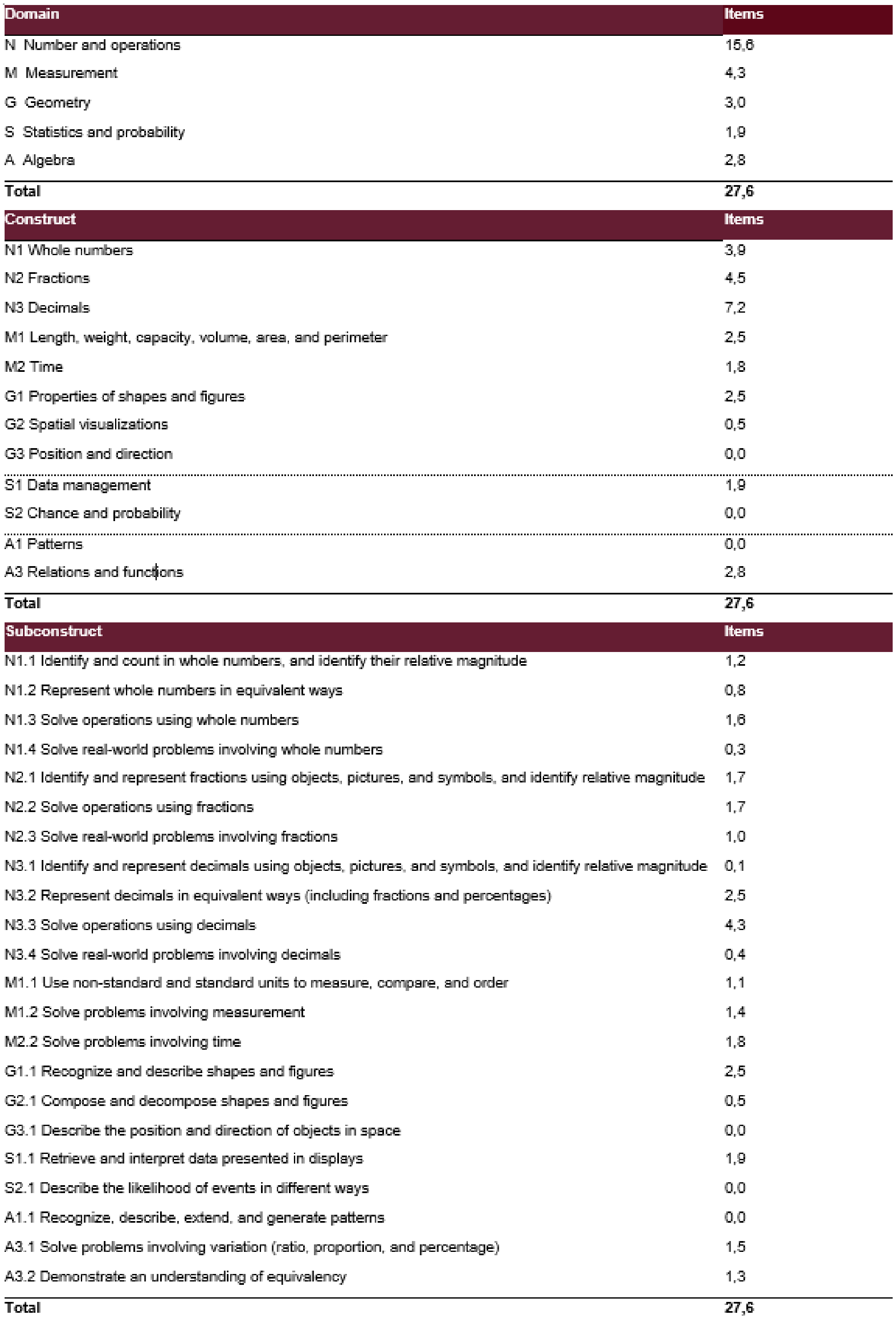


Domain	Items
D Decoding	0,0
R Reading comprehension	33,1
Total	**33,1**

Construct	Items
D1 Precision	0,0
D2 Fluency	0,0
R1 Retrieve information	10,3
R2 Interpret information	11,3
R3 Reflect on information	11,5
Total	**33,1**

Subconstruct	Items
D1.1 Identify symbol-sound/fingerspelling and/or symbol-morpheme correspondences	0,0
D1.2 Decode isolated words	0,0
D2.1 Say or sign a grade-level continuous text at pace and with accuracy	0,0
R1.1 Recognize the meaning of common grade-level words	0,8
R1.2 Retrieve explicit information in a grade-level text by direct- or close-word matching	6,1
R1.3 Retrieve explicit information in a grade-level text by synonymous matching	3,4
R2.1 Identify the meaning of unknown words and expressions in a grade-level text	1,5
R2.2 Make inferences in a grade-level text	7,9
R2.3 Identify the main and secondary ideas in a grade-level text	1,9
R3.1 Identify the purpose and audience of a text	5,6
R3.2 Evaluate a text with justification	4,2
R3.3 Evaluate the status of claims made in a text	1,7
Total	**33,1**

Table 22. Mathematics: Number of items aligned to each grade 6 domain, construct and subconstructs



Domain	Items
N Number and operations	15,6
M Measurement	4,3
G Geometry	3,0
S Statistics and probability	1,9
A Algebra	2,8
Total	**27,6**

Construct	Items
N1 Whole numbers	3,9
N2 Fractions	4,5
N3 Decimals	7,2
M1 Length, weight, capacity, volume, area, and perimeter	2,5
M2 Time	1,8
G1 Properties of shapes and figures	2,5
G2 Spatial visualizations	0,5
G3 Position and direction	0,0
S1 Data management	1,9
S2 Chance and probability	0,0
A1 Patterns	0,0
A3 Relations and functions	2,8
Total	**27,6**

Subconstruct	Items
N1.1 Identify and count in whole numbers, and identify their relative magnitude	1,2
N1.2 Represent whole numbers in equivalent ways	0,8
N1.3 Solve operations using whole numbers	1,6
N1.4 Solve real-world problems involving whole numbers	0,3
N2.1 Identify and represent fractions using objects, pictures, and symbols, and identify relative magnitude	1,7
N2.2 Solve operations using fractions	1,7
N2.3 Solve real-world problems involving fractions	1,0
N3.1 Identify and represent decimals using objects, pictures, and symbols, and identify relative magnitude	0,1
N3.2 Represent decimals in equivalent ways (including fractions and percentages)	2,5
N3.3 Solve operations using decimals	4,3
N3.4 Solve real-world problems involving decimals	0,4
M1.1 Use non-standard and standard units to measure, compare, and order	1,1
M1.2 Solve problems involving measurement	1,4
M2.2 Solve problems involving time	1,8
G1.1 Recognize and describe shapes and figures	2,5
G2.1 Compose and decompose shapes and figures	0,5
G3.1 Describe the position and direction of objects in space	0,0
S1.1 Retrieve and interpret data presented in displays	1,9
S2.1 Describe the likelihood of events in different ways	0,0
A1.1 Recognize, describe, extend, and generate patterns	0,0
A3.1 Solve problems involving variation (ratio, proportion, and percentage)	1,5
A3.2 Demonstrate an understanding of equivalency	1,3
Total	**27,6**

Annex E. Difficulty Level of the Items

Table 23. P-value and Item-Total correlation of the NLA Khmer items

Item	Itemcode	N	P-value	P0-25	P26-50	P51-75	P76-100	Rit
1	KA17_3S5	2135	0.89	0.24	0.62	0.85	0.96	0.44
2	KA6_1S5	2121	0.57	0.21	0.39	0.55	0.78	0.38
3	KA23_1S6	2122	0.40	0.25	0.32	0.39	0.52	0.27
4	KA25_1S6	2124	0.67	0.23	0.44	0.66	0.85	0.36
5	KA1_8O6	2143	0.66	0.12	0.34	0.68	0.93	0.56
6	KA1_9O6	2136	0.66	0.16	0.38	0.66	0.90	0.51
7	KA14_8O6	2127	0.65	0.11	0.37	0.66	0.91	0.48
8	KA17_2S5	2144	0.81	0.29	0.55	0.78	0.94	0.44
9	KA8_1O6	2128	0.40	0.15	0.30	0.45	0.71	0.33
10	KA6_2S5	2132	0.57	0.11	0.32	0.60	0.87	0.48
11	KA1_31O6ox	6351	0.69	0.24	0.37	0.59	0.86	0.41
12	KA1_34O6	2135	0.57	0.10	0.32	0.59	0.87	0.49
13	KA1_35O6ox	6318	0.71	0.21	0.37	0.63	0.88	0.42
14	KA1_29O6	2122	0.68	0.13	0.41	0.69	0.91	0.46
15	KA18_3S5	2115	0.77	0.22	0.51	0.75	0.94	0.43
16	KA18_2S5ox	6287	0.61	0.09	0.24	0.47	0.83	0.49
17	KA11_1S5	2127	0.70	0.19	0.44	0.68	0.91	0.52
18	KA8_3O6	2121	0.67	0.27	0.47	0.67	0.86	0.38
19	KA3_11O6	2135	0.50	0.23	0.38	0.52	0.69	0.27
20	KA24_2S6x	6315	0.50	0.17	0.28	0.39	0.65	0.32
21	KA1_44O6	2136	0.68	0.21	0.43	0.68	0.90	0.43
22	KA1_30O6	2135	0.70	0.14	0.42	0.71	0.93	0.51
23	KA18_1S5	2131	0.77	0.15	0.42	0.78	0.96	0.60
24	KA23_2S6	2120	0.57	0.13	0.34	0.60	0.85	0.40
25	KA1_21O6	2128	0.50	0.21	0.36	0.52	0.71	0.46
26	KA14_1O6	2124	0.58	0.23	0.42	0.59	0.80	0.46
27	KA9_1O6x	6325	0.54	0.11	0.25	0.45	0.71	0.39
28	KA3_5O6ox	6319	0.81	0.11	0.42	0.79	0.96	0.55
29	KA3_6O6	2121	0.73	0.29	0.48	0.69	0.87	0.54
30	KA3_2O6	2133	0.79	0.25	0.56	0.78	0.93	0.36
31	KA24_6S6	2133	0.70	0.16	0.44	0.70	0.90	0.42
32	KA11_6S5	2122	0.57	0.17	0.34	0.57	0.82	0.39

Table 24. P-value and Item-Total correlation of the NLA mathematics items

Item	Itemcode	N	P-value	P0-25	P26-50	P51-75	P76-100	Rit
1	MAS8_1	2116	0.41	0.17	0.31	0.53	0.81	0.44
2	MAO1_1x	6264	0.46	0.19	0.30	0.51	0.77	0.40
3	MAO2_5	2097	0.63	0.12	0.40	0.77	0.97	0.57
4	MAO2_6x	6241	0.46	0.09	0.23	0.54	0.88	0.56
5	MAO3_1	2133	0.72	0.28	0.54	0.78	0.92	0.46
6	MAO3_3x	6279	0.55	0.24	0.39	0.60	0.86	0.40
7	MAO5_6	2099	0.50	0.19	0.37	0.59	0.84	0.39
8	MAO7_5	2111	0.58	0.18	0.40	0.70	0.91	0.52
9	MAS1_2ox	6286	0.49	0.24	0.33	0.53	0.81	0.40
10	MAS5_1	2132	0.36	0.13	0.27	0.48	0.79	0.45
11	MAS5_4ox	6281	0.73	0.39	0.60	0.82	0.94	0.39
12	MAS6_5	2113	0.73	0.26	0.53	0.79	0.94	0.49
13	MAS8_3	2099	0.30	0.23	0.31	0.40	0.59	0.30
14	MAS8_4ox	6161	0.50	0.14	0.29	0.57	0.89	0.50
15	MAO6_4x	6186	0.39	0.17	0.26	0.39	0.74	0.39
16	MAS6_3	2105	0.38	0.10	0.25	0.51	0.85	0.51
17	MBO2_2	2107	0.48	0.31	0.43	0.53	0.74	0.26
18	MBS1_3x	6274	0.72	0.29	0.56	0.83	0.97	0.44
19	MBO3_1x	6193	0.39	0.22	0.28	0.37	0.70	0.32
20	MBS4_1	4092	0.32	0.11	0.21	0.39	0.68	0.37
21	MCS1_4ox	6233	0.55	0.14	0.35	0.62	0.91	0.47
22	MCS1_5ox	6279	0.59	0.22	0.41	0.68	0.90	0.44
23	MCS4_5	2099	0.27	0.21	0.29	0.36	0.56	0.20
24	MCS5_1x	6235	0.62	0.29	0.47	0.69	0.90	0.38
25	MCS5_2	2093	0.51	0.16	0.36	0.62	0.87	0.47
26	MDO1_1	2103	0.79	0.48	0.66	0.77	0.88	0.21
27	MES2_5ox	6179	0.68	0.27	0.50	0.78	0.97	0.46
28	MEO1_3	2030	0.52	0.16	0.36	0.61	0.88	0.45
29	MDS1_4	2021	0.47	0.14	0.32	0.60	0.87	0.49
30	MES2_1	2011	0.74	0.34	0.58	0.76	0.91	0.31

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

EVALUATION OF THE WORKSHOP (English version)

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

1. PIN

TRAINING ON THE GLOBAL PROFICIENCY FRAMEWORK

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

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

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

TRAINING ON THE NLA

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 fifth and sixth 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 seventh 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[9]					

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					
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					

[9] This is an additional question on request of observers. This question was not include in the questionnaire for Cambodia, so responses are not available..

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[10]
    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[11]
    d) Unsuccessful
    e) Totally Unsuccessful
25. How would you rate the organization of the workshop?
    a) Totally Successful
    b) Successful
    c) Neutral[12]
    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.

[10] Added the Neutral on request of UIS project leader
[11] Added the Neutral on request of UIS project leader
[12] Added the Neutral on request of UIS project leader

11. Addendum to the Report on the Cambodian National Grade 6 Learning Assessment Policy Linking for Measuring Global Learning Outcomes Workshop: Setting Global Benchmarks for Khmer and Mathematics

Summary Addendum

This document contains an additional report on the Cambodian online policy linking workshop that took place from July 5, 2021 until July 16, 2021. The Education Quality Assurance Department of the Ministry of Education, Youth and Sports in Cambodia (EQAD) and the UNESCO Institute for Statistics (UIS) organized this workshop as a pilot. The objective of the workshop was to set global benchmarks on the 2016 National Learning Assessment (NLA) at grade 6 in Khmer and Mathematics by organizing a fully remote policy linking workshop. For details on the preparation, logistics and outcomes of the workshop we refer to the original report.

The reason for this additional report is that the results of the standard setting part of the workshop were surprisingly positive for EQAD, with an unexpectedly high proportion of pupils in the Partially Meets, Meets and Exceeds categories. Therefore, EQAD decided to have a third round of alignment, matching and benchmarking, without the involvement of Cito staff. This third round took place from 23 until 27 July and featured the set of items from the NLA that were *not* selected for the original workshop. The complementary items were aligned and matched with the Global Proficiency Framework and these items were also used in a single-round modified Angoff rating procedure that was identical to the one used in the original workshop itself. For Khmer there were 28[13] complementary items and mathematics there were 40[14] complementary items.

For this additional report, Cito carried out several additional activities:

1. We checked the procedures employed and the data to find out if all processes ran as they should have run and if any mistakes were made in data processing and analyses.
2. We analyzed the additional new data produced in the third round after the Cambodian workshop.
3. We put forward plausible explanations for the unexpectedly positive results for the Cambodian workshop, taking into account the data collected in the third round.
4. We described a procedure to further ensure the validity of the outcomes in situations comparable to the Cambodian one.

Check on procedures and data analyses
Procedures used for alignment, matching and standard setting in the workshop were checked and all data and performed analyses were inspected to find out if there were any errors. Everything seemed to be carried out correctly. This is supported by the fact that the workshop lives up to the 4.1.1. Review Panel criteria and the evaluations of the workshop are relatively positive. So there are no reasons to mistrust the outcomes of the workshop. Therefore, we have to look for other explanations for the unexpectedly positive results.

Alignment

[13] There was one polytomous item that had to be discarded for the benchmarking analysis. So the number of items used there was 27

[14] There was one polytomous item that had to be discarded for the benchmarking analysis. And there were 11 items that were not IRT calibrated. So the number of items used in the benchmarking procedure was 28.

The analyses of the additional new data show that, as far as alignment is concerned, the 28 complementary Khmer items used in Round 3 are only *minimally aligned* in depth to the GPF, whereas the selected items in the original workshop were strongly aligned. Both in the original workshop and in this additional round the selected items were strongly aligned in breadth. If we take the complementary set of items and the original set together, then this total set of 61 items is both strongly aligned in depth and breadth to the GPF. However, this can be expected, because the chances of being strongly aligned in depth and breadth increase on a par with the absolute number of items used in the alignment exercise.

The 39 complementary Mathematics items used in Round 3 are strongly aligned in depth and in breadth. This is comparable to the results of the original workshop, where the selected items were strongly aligned in breadth and additionally aligned in depth, but where the results of the matching showed that this conclusion could be changed to strongly aligned, because the matching increased the level of agreement between the raters. If we take the complementary set of items and the original set together, then this total set of 70 items is both strongly aligned in depth and breadth to the GPF. However, as with Khmer, this can be expected, because the chances of being strongly aligned in depth and breadth increase on a par with the absolute number of items used in the alignment exercise.

Matching

The analyses of the additional new data show that, as far as matching is concerned, both for the 28 complementary Khmer items and the 39 complementary Mathematics items, full consensus was reached. Whereas for the set of 33 Khmer items and 31 Mathematics items, full consensus was also reached, but in both groups this was only achieved after a long discussion for one of these items.

Benchmarking

To give direct insight into the robustness of the benchmarks found, 95% confidence intervals for the percentages of learners in the Below Partially Meeting, Partially Meeting, Meeting and Exceeding categories were calculated both for the results of Round 1 and Round 2 as well as for Round 3. This exercise makes clear that percentages of learners corresponding to the benchmarks vary widely, especially in the Partially Meeting and Meeting categories, both for Khmer and Mathematics.

For instance, the benchmarks for Khmer in Round 1 are respectively 6.7, with a 95% confidence interval of 5.1 – 8.3; 21.0, with a confidence interval of 18.6 – 23.3 and 30.7, with a confidence interval of 30.0 – 31.4. The corresponding percentages of learners are respectively 1.7% for Below Partially Meeting with a confidence interval of 1.1% – 3.0%; 28.3% for Partially Meeting with a confidence interval of 19.8% – 42.5%; 58.9% for Meeting, with a confidence interval of 38.8% –71.5%, and 11.1% for Exceeding with a confidence interval of 5.8% – 17.6%. The results for other rounds and for Mathematics are comparable, making clear that the benchmark results, particularly those for Below Partially Meeting and Partially Meeting, lack robustness.

Plausible explanations

There are several plausible explanations for the unexpected positive results. The most plausible one is the lack of robustness of the outcomes of the standard setting process. This lack of robustness is caused by the ability distribution in the Cambodian population in relation to the standard error of measurement (SEM) in the benchmark scores. The lack of robustness can be demonstrated by adding a 95% reliability interval to the benchmark scores based on the SEMs calculated. The lower and upper boundaries of the benchmark scores correspond with lower and upper percentages of learners in the JPM-, JM and JE categories of the GPF. This exercise shows that small changes in scores on the benchmarks do result in large differences in percentages in said categories.

A second possible explanation might be found by the choice of the raters. However, their representativeness seem to be in order. And an inspection of the values for inter- and intra-rater consistency makes clear that there are no real outliers that clearly affect the outcome of the standard setting process.

Another factor that could influence the results is sampling weights and plausible values from the NLA were not provided. Tthe benchmarks calculated only hold for the sample of learners that made the NLA and cannot be generalized to the Cambodian population as a whole. In addition to this, the results could partly be caused by the quality of the IRT calibration of the items in the NLA. The NLA data received contained IRT parameters of the items, but there was no information on the accuracy of the parameters, so the estimation error of the IRT parameters could not be taken into account.

A fourth explanation could be that the outcomes of the policy linking workshop are completely valid, but that they are not in accordance with the original benchmarks (Below Basic; Basic; Proficient and Advanced) from the NLA, because these were established through a different procedure. The validity of this last explanation could be examined by a detailed comparison between the Grade Six Performance Standard Skills from the NLA with the Global Proficiency Framework.

All in all, there are several plausible explanations for the outcomes found in the workshop. But there are no clear-cut criteria to make clear what the real reasons are for the unexpectedly positive outcomes of the workshop. Our best guess is that, if the results are really unrealistically high, this is caused by a lack of robustness in the standard setting combined with the ability distribution in the population. The results make clear that small differences in benchmarks lead to relatively large differences in percentages and that the confidence intervals around these percentages are also large. However, there are two assumptions underlying this best guess. The first one is that the sample of learners that took the NLA is really representative for the Cambodian population. And the second one is that the IRTanalyses performed led to valid estimates of item and learner parameters. Additional explanations might be found in the different standard setting procedures used for establishing the original benchmarks on the NLA and the ones produces in the workshops and Round 3.

If we look at the different benchmarks produced, both in the workshop and in Round 3, probably the best estimates of the 'real' benchmarks are those for the total set of items, because here the benchmarks have the smallest standard error of measurement and the percentages therefore have the smallest confidence intervals.

Towards a procedure with survey designs and IRT modelling

The procedure that has been developed to ensure the validity of the outcomes in situations comparable to the Cambodian one is described in chapter 5. The policy linking toolkit will have to be expanded with standardized procedures that can be used in situations where the assessment, on which benchmarks have to be set has used for system evaluation and employs an assessment design with several booklets and a set of anchoring items and IRT-analyses. Using a standard setting more suited in cases like these, for instance the Bookmark or 3DC method might also be considered. The number of items to be rated can be comparable to the linear assessments that were the subject of the PLT in earlier policy linking procedures, provided the IRT-parameters are estimated with enough precision.

The results from Round 3

Introduction

We found no inconsistencies or errors in the procedures used for alignment, matching and standard setting in the workshop. All data and performed analyses were inspected to check for errors or mistakes and were found to be in order. In addition to this, the workshop lives up to the 4.1.1. Review Panel criteria. In the PLT (Annex U, p. 164) six criteria are mentioned for the validity of a 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, the extent which the panelists meet a set of selection criteria and panelists' understanding of the procedures. Furthermore, the evaluations of the workshop were are relatively positive. Evaluations for training on the GPF, on the NLA itself, on the alignment, matching and benchmark-setting methodology, and the benchmark round 2 and overall evaluation scored, well above 4 on a 5-point scale for both Khmer and Mathematics. Standard deviations were small and there were no outliers. An explanation for the unexpectedly positive results therefore will have to be sought elsewhere. The main objective of this additional report is to explore plausible explanations for these unexpected results. But first we describe the procedures employed in Round 3 and their results.

Task 1: Alignment

As in the workshop, the panelists had to execute three tasks in the third round:

- Task 1 — Rate the alignment between the *complementary* NLA-items and the GPF
- Task 2 — Match the *complementary* NLA items to the appropriate GPL and Global Proficiency Descriptor.
- Task 3 — Perform a modified Angoff rating procedure for the complementary items from the NLA. Note that there was only a single round instead of two rounds as described in the PLT.

The alignment method was identical to the method used in the workshop. In the first step, panelists independently rated the alignment between the NLA items and GPF knowledge and/or skill(s) statement(s) and in the second step the facilitators compiled and summarized the ratings to check the alignment between the assessments and the GPF.

Again, panelists rated each item using the scale of Complete Fit, Partial Fit, and No Fit:

- 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.

Alignment for Khmer language

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

Complementary items
Averaging the panelists' ratings, we see that in the third round all 28 complementary items (on average 27,5) aligned to Reading comprehension. At least 17 (on average 17,3) items were aligned to Retrieve information; only 2 items were aligned to Interpret information and at least 8 (on average 8,2) were aligned to Reflect on Information. This means that this subset of the NLA is only *minimally aligned* in depth (see the criteria in Table 25). This in contrast to the results of the original workshop where the selected items were strongly aligned in depth.

We see that on average almost all subconstructs of Reading comprehension are covered (see Table 43 in the annex). This means that this subset of NLA-items is strongly aligned in breadth (see the criteria in Table 25). This was also the case in the original workshop.

Total set of items
If we also take into account the set of *original* items that were used in the workshop, we have a total set of 61 (28+33) items. Of these 61 items all (on average) aligned to Reading comprehension. At least 27 items were aligned to Retrieve information; at least 13 items were aligned to Interpret information and at least 19 were aligned to Reflect on Information. The total set of NLA Khmer items is therefore strongly aligned in depth (see Table 25).

We see that on average all subconstructs of Reading comprehension are covered. The total set of NLA Khmer items is therefore strongly aligned in breadth (see the criteria in Table 25). However, both conclusions should not come as a surprise because the chances of being strongly aligned in depth and breadth increase with the absolute number of items.

Table 25. 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)	R (minimum five items)	R (minimum five items)
		C (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: R1 (minimum 5 items)
			R: B2 (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 for Mathematics

Complementary items

Averaging the panelists' ratings, more than 35 (on average 35,3) out of the 39 [15]complementary items aligned to grade 6 subconstructs. In the GPF 22 subconstructs are mentioned for grade 6 and the complementary items covered 18 of those subconstructs (an average of >0.5 (see Table 44 in the annex). In breadth the complementary items are strongly aligned to the GPF for Grade 6 as the items covered more than 50% of all grade 6 subconstructs. The same result was found in the original workshop

The complementary NLA Mathematics items covered all five domains and 10 out of 12 constructs for grade 6. According to the new criteria in the Policy Linking Toolkit, for strong alignment in depth at least 5 items should align to the domain Number and Operations, at least 5 items to Measurement and Geometry and at least 5 items to Statistics and Probability and Algebra (see Table 26). On average 21.1 items covered the domain of Number and Operations, 9.1 items the domains Measurement and Geometry, and 5.0 items the domains Statistics and Probability and Algebra. For this reason, the complementary NLA items are also strongly aligned to the GPF in depth. This is a marginally better result than in the original workshop, where additional alignment was found. However after matching, it could also be concluded that there was strong alignment then.

Total set of items

If we also take into account the set of *original* items that were used in the workshop, we have a total set of 70 (39+31) items. Averaging the panelists' ratings, more than 63 of the 70 items aligned to grade 6 subconstructs. One item from the original items was excluded from the ratings, because correct information was missing for the item[16]. In the GPF 22 subconstructs are mentioned for grade 6 and the total set of NLA items covered 20 of those subconstructs. In breadth this total set of NLA-items is strongly aligned to the GPF for Grade 6 as the items covered more than 50% of all grade 6 subconstructs.

The NLA Mathematics items covered all five domains and 10 out of 12 constructs for grade 6. According to the new criteria in the Policy Linking Toolkit, for strong alignment in Depth at least 5 items should align to the domain Number and Operations, at least 5 items to Measurement and Geometry and at least 5 items to Statistics and Probability and Algebra (see Table 26).

On average over 36 items covered the domain of Number and Operations, over 16 items the domains Measurement and Geometry, and almost 10 items the domains Statistics and Probability and Algebra. For this reason, according to the panelists for Mathematics, the total set of NLA-items is strongly aligned to the GPF in depth. However, as with the Khmer NLA items, both conclusions should not come as a surprise because the chances of being strongly aligned in depth and breadth increase with the absolute number of items.

[15] One item had to be excluded, because it was a polytomous item.

[16] The data from NLA Mathematics made clear that this item was in fact a meta item containing three separate items. At the specific point in time in the original workshop it would have caused a lot of confusion with panelists and a lot of extra work for the EQAD and Cito team if this would have been taken into account. So it was decided to exclude the item from the next steps in policy linking.

Table 26. 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):	Number (minimum 5 items) and Measurement and Geometry (minimum 5 items)
	Subconstructs (breadth):	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

Task 2: Matching

After the panelists received the outcome of their alignment tasks, they continued by matching the complementary NLA items with the Global proficiency levels and descriptors. 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.

Both for the 28 complementary Khmer-items and for the 39 complementary Mathematics items full consensus was reached by all raters without long discussions. In the original workshop the outcome was almost alike. There, also full consensus was reached, but both for Khmer and Mathematics for one item a long discussion preceded reaching consensus.

Task 3: Benchmarking

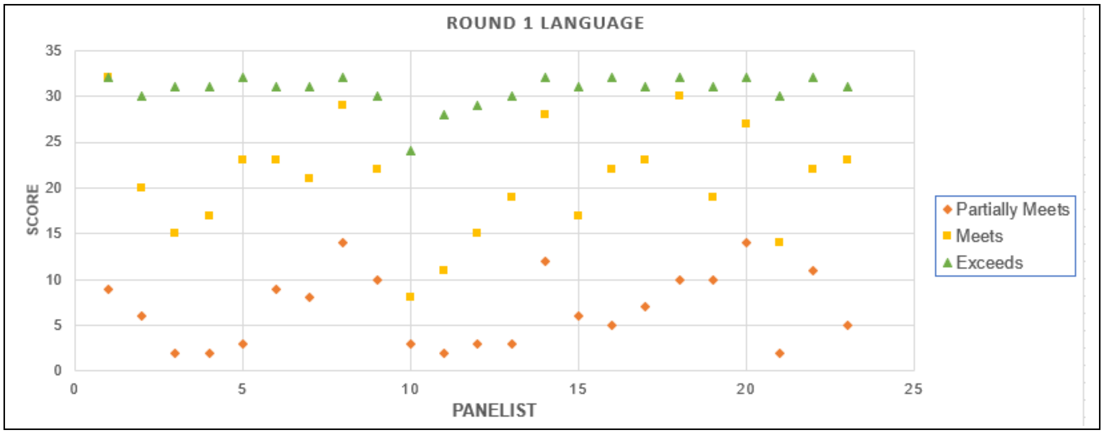
To facilitate a comparison between the workshop benchmarking results and the results of Round 3 , the benchmarking results of the two rounds of the workshop are repeated here. In this report, however, we take into account the standard error of measurement (SEM) of the calculated benchmarks to show the impact of the SEMs for each benchmark on the percentage of learners in the different categories:

- Below Partially Meets Minimum Global Proficiency Level (MGPL)
- Partially Meets MGPL
- Meets MGPL
- Exceeds MGPL

Round 1

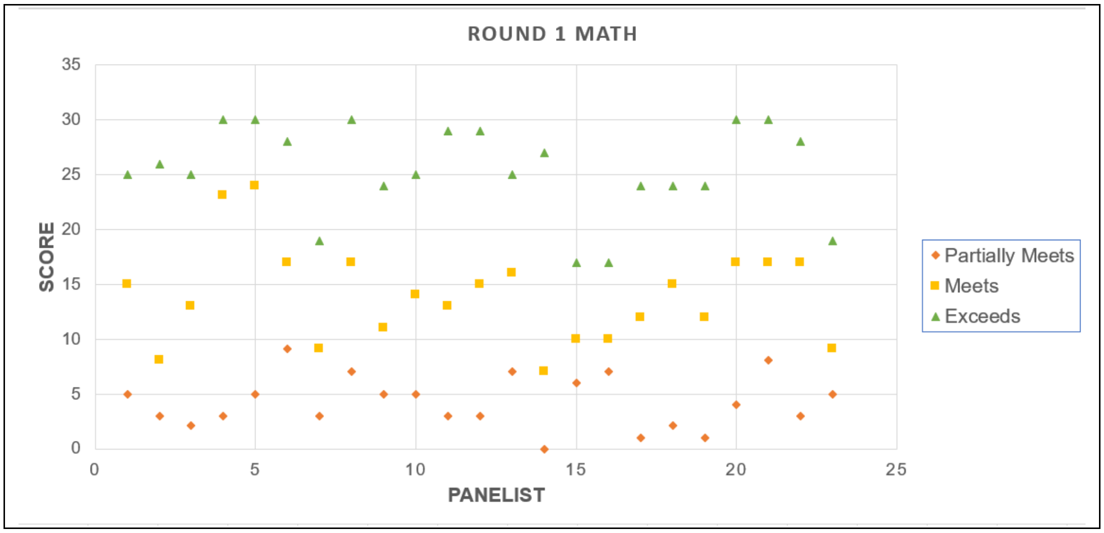
For Khmer, in Round 1 we saw that the ratings of panelists varied considerably (Figure 15), both for the lowest (Partially Meets MGPL) and the middle benchmark (Meets MGPL). We also saw a ceiling effect with the Exceeds MGPL benchmark. Exceeds is with a few exceptions almost at the maximum (32).

Figure 15. Anonymous information on the panelists' ratings for Khmer Round 1



For Mathematics, in Round 1, we saw that the ratings of panelists also varied considerably (Figure 16), both for the lowest (Partially Meet MGPL) and the middle benchmark (Meets MGPL). We also see a small ceiling effect with the Exceeds MGPL benchmark. Five of the panelists put the Exceeds MGPL benchmark at the maximum score of 30.

Figure 16. 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 truncated, as stipulated in the policy linking toolkit. For Khmer, the impact information showed that only 3.4% of the learners would fall at the Below Partially Meets MGPL; that 39.1% would fall at the Partially Meets MGPL; 49.2% at the Meets MGPL and 8.3% at the Exceeds MGPL using Round 1 benchmarks (see Table 27). However, taking into account the 95% confidence intervals around the benchmarks based upon the SEMs, it becomes clear that the percentages of learners, especially in the Partially Meets and Meets GMPL categories, vary considerably. In the Partially Meets MGPL the lower and upper boundaries of the confidence interval are 31.94% and 53.39% and in the Meets MGPL these boundaries are 31.58% and 57.36 percent.

Table 27. Round 1 benchmarks, score range and impact for Khmer with 32 items (95% confidence intervals within parentheses)

Minimum Proficiency Level	Round 1 Benchmark	Score Range			Percentage of Learners
			Female	Male	Total
Below Partially Meets	N/A	0–5 (0)(4-7)	2.8% (1.90-4.92%)	4.3% (3.07%-7.67%)	3.4% (2.42%-6.08%)
Partially Meets	6.7 (5.1–8.3)	6–19 (5-8)(17-22)	34.5% (24.75%-48.2%)	43.8% (32.26%-57.62%)	39.1% (31.94%-53.39%)
Meets	20.9 (18.5–23.3)	20–29 (18-23)(29-30)	52.7% (39.9%-64.7%)	45.4% (32.75%-56.27%)	49.2% (31.58%-57.36%)
Exceeds	30.6 (29.9–31.3)	30–32 (30-31)(32)	10.0% (5.63%-10.0%)	6.6% (3.8%-6.56%)	8.3% (4.63%-12.61%)

For Mathematics, the impact information showed that only 1.1% would fall in the Below Partially Meets MGPL; that 37.3% would fall at the Partially Meets MGPL; 52.2% at the Meets MGPL and 9.3% at the Exceeds MGPL using Round 1 benchmarks (see Table 28). However, taking into account the 95% confidence intervals around the benchmarks based upon the SEMs, it becomes clear that the percentages of learners, especially in the Partially Meets and Meets GMPL categories, vary considerably, although to a lesser extent than with Khmer. In the Partially Meets MGPL the lower and upper boundaries of the confidence interval are 29.43% and 49.17% and in the Meets MGPL these boundaries are 34.65% and 62.98 percent.

Table 28. Round 1 benchmarks, score range and impact for Mathematics with 30 items (95% confidence intervals in parentheses)

Minimum Proficiency Levels	Round 1 Benchmark	Score Range			Percentage of Learners
			Female	Male	Total
Below Partially Meets	N/A	0–3 (0) (2-4)	1.2% (0.35%-2.52%)	1.5% (0.73%-2.85%)	1.1% (0.54%-2.7%)
Partially Meets	4.4 (3.4–5.3)	4–12 (3-5)(11-14)	36.6% (29.71%-48.55%)	37.7% (30.7%-49.32%)	37.3% (29.43%-49.17%)
Meets	13.9 (12.1–15.7)	13–24 (12-15)(22-26)	53.1% (35.99%-63.2%)	50.7% (33.69%-61.23%)	52.2% (34.65%-62.98%)
Exceeds	25.4 (23.7–27.1)	25–30 (23-27)(30)	9.1% (4.57%-15.11%)	10.2% (5.22%-16.26%)	9.3% (4.88%-15.65%)

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, based on the ordering of items presented after round 1. After the discussion the panelists individually conducted the Round 2 ratings.

We see that in Round 2 the ratings of panelists varied less than in Round 1, especially for Mathematics (Figure 17 and Figure 18).

Figure 17. Anonymous information on the panelists' ratings of Khmer Round 2

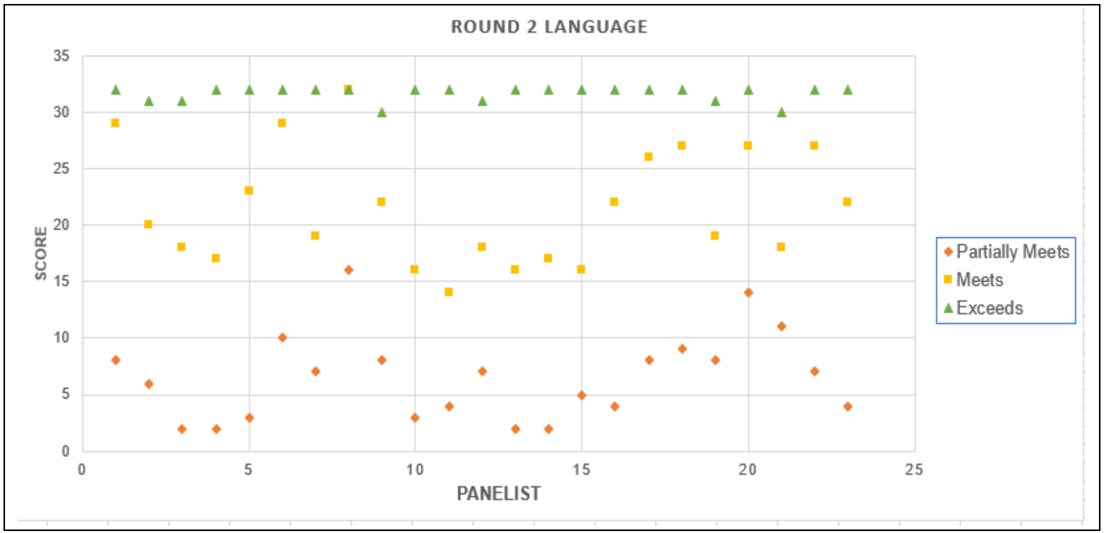
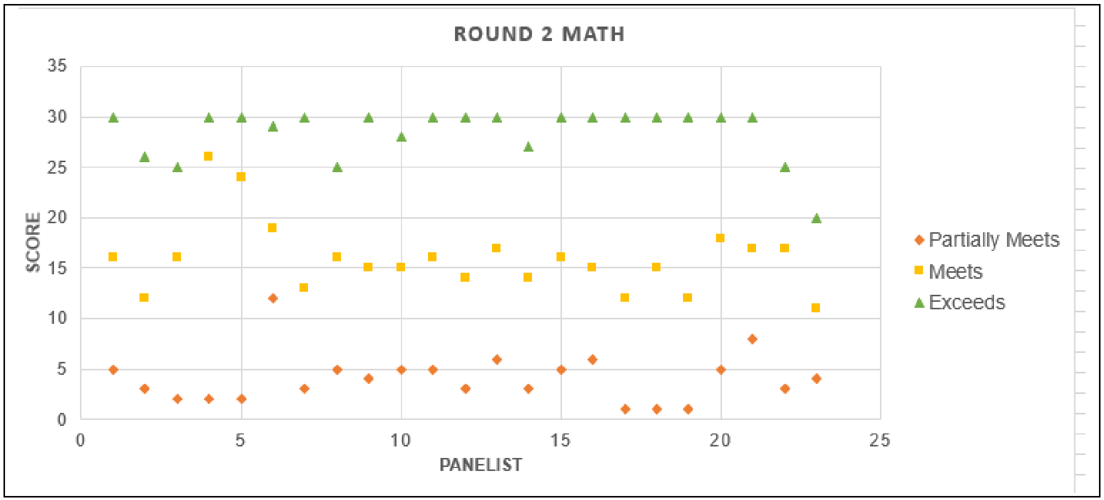


Figure 18. Anonymous information on the panelist's ratings of Mathematics Round 2



For Khmer, the results showed that in Round 2 only 3.4% fall in the Below Partially Meets level and 43.3% fall in the Partially Meets Level (see Table 29). Furthermore, 48.7% fall in the Meets level and only 4.6% in the Exceeds level. However, taking into account the 95% confidence intervals around the benchmarks based upon the SEMs, it becomes clear that the percentages of learners, especially in the Partially Meets and Meets GMPL categories, vary considerably. In the Partially Meets MGPL the lower and upper boundaries of the confidence interval are 31.94% (first round 31.94%) and 54.25% (first round 53.39%) and in the Meets MGPL these boundaries are 39.57% (first round 31.58%) and 57.36 (first round 57.36%) percent.

Table 29. Round 2 benchmarks, score range and impact for Khmer with 32 items (95% confidence intervals in parentheses)

Minimum Proficiency Level	Round 2 Benchmark	Score Range			Percentage of Learners
			Female	Male	Total
Below Partially Meets	N/A	0–5 (0) (4-7)	2.8% (1.9%-4.92%)	4.3% (3.07%-7.67%)	3.4% (1.55%-6.08%)
Partially Meets	6.5 (4.9–8.0)	6–20 (5-8)(18-22)	38.8% (28.43%-48.2%)	47.8% (36.41%-57.62%)	43.3% (31.94%-54.25%)
Meets	21.4 (19.3–23.5)	21–30 (19-23)(30-30)	52.7% (44.27%-61.02%)	44.2% (35.51%-52.12%)	48.7% (39.57%-57.36%)
Exceeds	31.6 (31.3–31.9)	31–32 (31-31)(32)	5.6% (5.63%-5.63%)	3.8% (3.8%-3.8%)	4.6% (4.63%)

For Mathematics, the results show that in Round 2 only 1.1% fall in the Below Partially Meets level and 54.1% fall in the Partially Meets Level. Furthermore, 41.7% fall in the Meets level and only 3.1% in the Exceeds level (see Table 30). However, taking into account the 95% confidence intervals around the benchmarks based upon the SEMs, it becomes clear that the percentages of learners, especially in the Partially Meets and Meets GMPL categories, vary considerably, although to a lesser extent than in Round 1. In the Partially Meets MGPL the lower and upper boundaries of the confidence interval are 40.92 (first round 29.43%) and 59.82 (first round 49.17%) percent and in the Meets MGPL these boundaries are 34.76 (first round 34.65%) and 54.88 (first round 62.98%) percent.

Table 30. Round 2 benchmarks, score range and impact for Mathematics with 30 items (95% confidence intervals in parentheses)

Minimum Proficiency Level	Round 2 Benchmark	Score Range			Percentage of Learners
			Female	Male	Total
Below Partially Meets	N/A	0–3 (0) (2-4)	1.2% (0.35%-2.52%)	1.5% (0.73%-2.85%)	1.1% (0.54%-2.7%)
Partially Meets	4.0 (3.0–5.0)	4–15 (3-5)(13-16)	53.4% (40.42%-59.73%)	53.8% (41.63%-59.83%)	54.1% (40.92%-59.82%)
Meets	16.1 (14.7–17.5)	16–27 (14-17)(26-28)	42.8% (35.35%-55.67%)	41.1% (34.22%-53.99%)	41.7% (34.76%-54.88%)
Exceeds	28.7 (27.8–29.5)	28–30 (27-29)(30)	2.6% (1.39%-4.57%)	3.6% (1.53%-5.22%)	3.1% (1.5%-15.65%)

Round 3

For Khmer, in Round 3 a different complementary sets of 28 items from the NLA were rated using the same procedure that was employed in Round 1 and 2. Of these items only 27 could be used for the analyses, because the 28th was a polytomous writing item which had to be discarded. If we look at the results of Round 3 separately, we see as we did in Round 1 and 2 that the ratings of panelists vary considerably (Figure 19), both for the lowest (Partially meets) and the middle benchmark (Meets). We also see a strong ceiling effect with the Exceeds benchmark. Exceeds is with only one exception at the maximum (27).

For Mathematics in Round 3 a different complementary sets of 40 items from the NLA were rated using the same procedure that was employed in Round 1 and 2. Of these items only 28 could be used for the analyses, because one of the items was a polytomous item which had to be discarded. And in addition to this, 11 items were open ended items which were not IRT calibrated. Since IRT-parameters are necessary to calculate the corresponding positions for the benchmarks on the underlying NLA 2016 ability scale for Mathematics, these items also had to be discarded. If we look at the results of Round 3 separately, We see as we did in Round 1 and 2 that the ratings of panelists vary considerably (see Figure 20), both for the lowest (Partially meets) and the middle benchmark (Meets). We also see a ceiling effect with the Exceeds benchmark. For 16 of the 23 raters Exceeds is at the maximum score of 28.

Figure 19. Anonymous information on the panelists' ratings for Khmer Round 3 (27 complementary items)

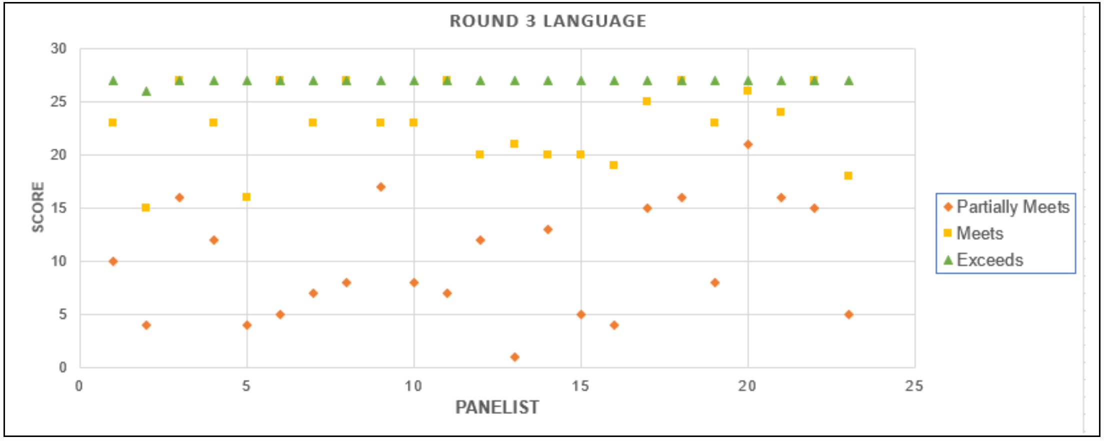
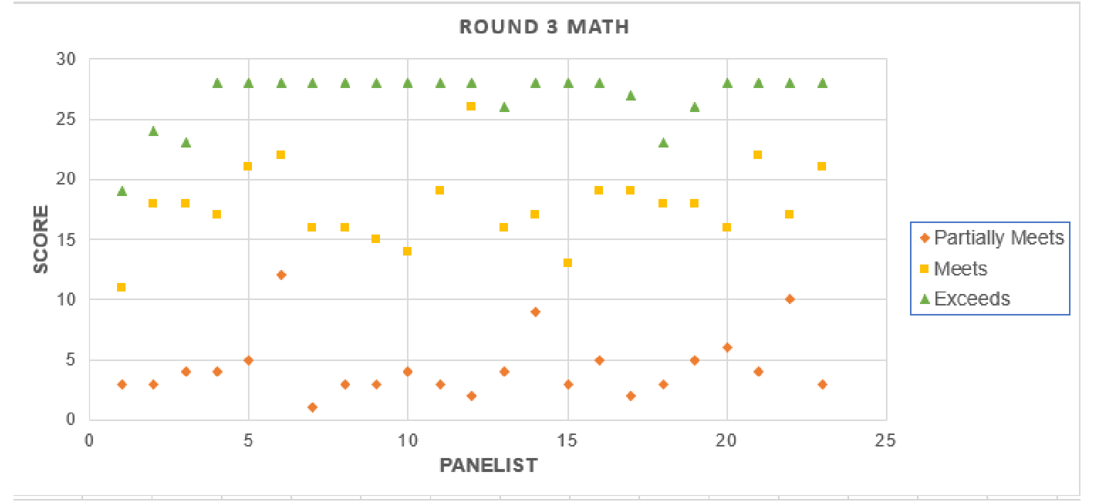


Figure 20. Anonymous information on the panelists' ratings for Mathematics Round 3 (28 complementary items)



In Round 3, the ratings on the 27 complementary items for Khmer result in 13.7% of the learners falling in the Below Partially Meets level and62.9% falling in the Partially Meets Level (see Table 31). Furthermore, 18.7% of the learners fall in the Meets level and only 4.6% in the Exceeds level. However, taking into account the 95% confidence intervals around the benchmarks based upon the SEMsit becomes clear that the percentages of learners, especially in the Partially Meets and Meets GMPL categories, vary considerably. In the Partially Meets MGPL the lower and upper boundaries of the confidence interval are 50.06 and 73.78 percent and in the Meets MGPL these boundaries are 13.35 and 27.14 percent.

Table 31. Round 3 benchmarks, score range and impact for Khmer with 27 items (95% confidence intervals in parentheses)

Minimum Proficiency Level	Round 3 Benchmark	Score Range			Percentage of Learners
			Female	Male	Total
Below Partially Meets	N/A	0–8 (0) (6-10)	12% (6.64%-18.51%)	16.5% (10.16%-24.59%)	13.7% (8.27%-21.05%)
Partially Meets	9.9 (7.7–12.1)	9–21 (7-11)(20-22)	61.7% (49.21%-79.07%)	62.9% (49.88%-79.48%)	62.9% (50.06%-73.78%)
Meets	22.7 (21.2–24.2)	22–25 (21-23)(25-26)	21.3% (9.28%-30.93%)	17% (6.8%-24.53%)	18.7% (13.35%-27.14%)
Exceeds	26.9 (26.8–27.0)	26–27 (26-27)(27)	5% (1.35%-5.0%)	3.6% (1.0%-3.56%)	4.6% (1.74%-4.6%)

In Round 3, the ratings on the 27 complementary items for Mathematics result in 9.5% of the learners falling in the Below Partially Meets level and 53.8 % falling in the Partially Meets Level (see Table 32). Furthermore, 30.4% of the learners fall in the Meets level and only 6.3% in the Exceeds level. However, taking into account the 95% confidence intervals around the benchmarks based upon the SEMsit becomes clear that the percentages of learners, especially in the Partially Meets and Meets GMPL categories, vary considerably. In the Partially Meets MGPL the lower and upper boundaries of the confidence interval are 46.7 and 64.03 percent and in the Meets MGPL these boundaries are 20.63 and 37.05 percent.

Table 32. Round 3 benchmarks, score range and impact for Mathematics with 28 complementary items (95% confidence intervals in parentheses)

Minimum Proficiency Level	Round 3 Benchmark	Score Range			Percentage of Learners
			Female	Male	Total
Below Partially Meets	N/A	0–3 (0) (2-4)	9% (6.4%-12.02%)	9.9% (6.58%-13.44%)	9.5% (6.51%-12.68%)
Partially Meets	4.3 (3.3–5.4)	4–16 (3-5)(15-18)	53.7% (46.24%-64.19%)	54.1% (46.08%-64.71%)	53.8% (46.7%-64.03%)
Meets	17.7 (16.4–19.1)	17–25 (16-19)(24-26)	30.9% (20.18%-37.84%)	30.1% (19.29%-36.94%)	30.4% (20.63%-37.05%)
Exceeds	26.7 (25.8–27.7)	26–28 (25-27)(28)	6.3% (3.91%-9.24%)	6% (3.55%-9.43%)	6.3% (3.57%-8.84%)

Comparison between Rounds 1, 2 and 3

For Khmer a comparison between the results of Round 1 and Round 2 shows that the feedback given to panelists after Round 1 did not have much effect. There are slight changes in benchmarks and the percentage of learners in the Partially Meets MGPL and the Meets MGPL categories has increased somewhat at the expense of the percentage of learners in the Exceeds MGPL category (see Table 33) . When we compare Round 2 with Round 3 we see a much larger change. The percentage of learners in the Below Partially Meets MGPL category has increased from 3.4 percent to 13.7 percent. The changes in percentages in the Partially Meets and the Meets MGPL category are even more significant: respectively from 43.3 percent to 62.9 percent and from 48.7 down to 18.7%. Using another set of items evidently leads to different outcomes in this case. The most plausible explanation for this difference is that panelists have been triggered by the unexpectedly positive outcomes of the workshop into becoming more severe in their judgments. This, however, does not mean that the outcomes of Round 3 can be seen as more valid.

Table 33. Comparison of Round 1, Round 2 and Round 3 benchmarks for Khmer with 32 items in round 1 and 2 and 27 different items in round 3

Minimum Proficiency Level	Round 1 Benchmark	Round 1 Percentage of Learners	Round 2 Benchmark	Round 2 Percentage of Learners	Round 3 Benchmark	Round 3 Percentage of Learners
Below Partially Meets	N/A	3.4%	N/A	3.4%	N/A	13.7%
Partially Meets	6.7	39.1%	6.5	43.3%	9.9	62.9%
Meets	20.9	49.2%	21.4	48.7%	22.7	18.7%
Exceeds	30.6	8.3%	31.6	4.6%	26.9	4.6%

For Mathematics a comparison between the results of Round 1 and Round 2 shows that the feedback given to panelists after Round 1 did have some effect. There are changes in benchmarks and the percentage of learners in the Partially Meets MGPL category has increased at the expense of the percentage of learners in the Meets and the Exceeds MGPL categories (see Table 34). When we compare Round 2 with Round 3 we see bigger differences. The percentage of learners in the Below Partially Meets MGPL category has increased from 1.1 percent to 9.5 percent. The percentage of learners in the Partially Meets MGPL category has further increased and there is a decrease in the percentage of learners in the Meets MGPL category. The percentage of learners in the Exceeds MGPL has gone up again. As was the case with Khmer, using another set of items evidently leads to different outcomes in this case. The most plausible explanation for this difference is the same as with Khmer: panelists have been triggered by the unexpectedly positive outcomes of the workshop into becoming more severe in their judgments. This, again, does not mean that the outcomes of Round 3 can be seen as more valid.

Table 34. Comparison of Round 1, Round 2 and Round 3 benchmarks for Mathematics with 30 items in round 1 and 2 and 28 different items in round 3

Minimum Proficiency Level	Round 1 Benchmark	Round 1 Percentage of Learners	Round 2 Benchmark	Round 2 Percentage of Learners	Round 3 Benchmark	Round 3 Percentage of Learners
Below Partially Meets	N/A	1.1%	N/A	1.1%	N/A	9.5%
Partially Meets	4.4	37.3%	4.0	54.1%	4.3	53.8%
Meets	13.9	52.2%	16.1	41.7%	17.7	30.4%
Exceeds	25.4	9.3%	28.7	3.1%	26.7	6.3%

Results for the complete sets of items

Of course, it is also possible to look at the results of the standard setting, taking all rated items into account. The results of this exercise are shown in Table 35 and Table 36. For Khmer the ratings on the total set of 59 items result in 6.4% of the learners falling in the Below Partially Meets level and 57.9% falling in the Partially Meets Level (see Table 35). Furthermore, 35.1% of the learners fall in the Meets level and only 0.6% in the Exceeds level. However, taking into account the 95% confidence intervals around the benchmarks based upon the SEMs, it becomes clear that the percentages of learners, especially in the Partially Meets and Meets GMPL categories, vary considerably. In the Partially Meets MGPL the lower and upper boundaries of the confidence interval are 41.81 and 70.99 percent and in the Meets MGPL these boundaries are 25.21 and 47.55 percent. Although the SEMs for the benchmarks are relatively small, because in this analysis the total set of items is used, this does not result in much smaller bounds as far as the percentages are concerned. The ability distribution in the population is such that relatively large differences in the percentages persist.

Table 35. Benchmarks, score range and impact for Khmer with 59 items (95% confidence intervals in parentheses)

Minimum Proficiency Level	Benchmark	Score Range			Percentage of Learners
			Female	Male	Total
Below Partially Meets	N/A	0–15 (0) (12-18)	4.5% (2.56%-7.93%)	8.3% (4.39%-13.51%)	6.4% (3.22%-10.06%)
Partially Meets	16.4 (13.3–19.6)	16–43 (13-19)(39-46)	56.5% (42.31%-68.76%)	60.9% (46.87%-73.75%)	57.9% (41.81%-70.99%)
Meets	44.2 (41.3–47.2)	44–57 (40-47)(57)	38.1% (27.8%-48.86%)	30.6% (21.66%-39.41%)	35.1% (25.21%-47.55%)
Exceeds	58.6 (58.3–58.9)	58–59 (58)(59)	0.9% (0.89%-0.89%)	0.2% (0.21%-0.21%)	0.6% (0.58%)

For Mathematics the ratings on the total set of 58 items result in 1.1% of the learners falling in the Below Partially Meets level and 60.9% falling in the Partially Meets Level (see Table 36). Furthermore, 36.7% of the learners fall in the Meets level and only 1.2% in the Exceeds level. However, taking into account the 95% confidence intervals around the benchmarks based upon the SEMs, it becomes clear that the percentages of learners, especially in the Partially Meets and Meets GMPL categories, vary considerably, although in a lesser extent than with Khmer. In the Partially Meets MGPL the lower and upper boundaries of the confidence interval are 54.37 and 66.46 percent and in the Meets MGPL these boundaries are 31.07 and 42.15 percent. For

Mathematics the smaller SEMs for the benchmarks result in somewhat smaller bounds as far as the percentages are concerned in comparison to Khmer.

Table 36. Benchmarks, score range and impact for Mathematics with 58 items (95% confidence intervals in parentheses)

Minimum Proficiency Level	Benchmark	Score Range			Percentage of Learners
			Female	Male	Total
Below Partially Meets	N/A	0–7 (0) (5-9)	0.9% (0.35%-1.85%)	0.9% (0.38%-2.37%)	1.1% (0.36%-2.61%)
Partially Meets	8.4 (6.6–10.2)	8–32 (6-10)(30-34)	61.7% (55.12%-67.26%)	61.6% (54.89%-67.05%)	60.9% (54.37%-66.46%)
Meets	33.9 (31.9–35.9)	33–54 (31-35)(53-55)	36.5% (30.92%-42.39%)	36.5% (31.07%-42.15%)	36.7% (31.31%-42.45%)
Exceeds	55.4 (54.1–56.7)	55-58 (54-56)(58)	1% (0.64%-1.48%)	1.1% (0.58%-1.51%)	1.2% (0.57%-1.87%)

Conclusions

As usual with standard setting, the outcomes of Round 2 can be considered to be more valid than those of Round 1, because panelists have been provided with several types of concrete feedback on their ratings and their differences. How the outcomes of Round 2 compare with those of Round 3 is hard to say, because a different set of items was used. Of course, raters would have had more experience with standard setting in this round, but no feedback on their ratings was given. The fact that Round 3 led to different outcomes than those in Round 1 and 2 illustrates the lack of robustness of the policy linking procedure in the Cambodian situation: selecting a different set of items for the standard setting procedures leads to divergent percentages in the different GPF categories. Considering the items selected for respectively Round 1 and 2 and Round 3 together leads to smaller SEMs for the benchmarks, but these are only partially reflected in a reduction in the length of the confidence intervals for the percentages, especially for Khmer. If we see the size of SEMs as the most important criterion, then the percentages calculated for the total set of items should be seen as the best estimates of the position of the Cambodian population of learners on the GPF.

Evaluation of the Standard Setting Process

Internal Evaluation SEM, Panelist Consistency and Panelists' Agreement for Round 3

In addition to calculating benchmarks and impact data, the PLT also requires calculating measures of consistency and presenting evaluation feedback results. For Round 3 no evaluation was done. But the measures of consistency are reported in Table 37 and Table 38, together with the corresponding information from Round 1 and 2.

As shown in Table 37, the SEM which measures how much panelists' benchmarks are spread around a "true" benchmark, was in all three rounds under 1.0 for Mathematics and not much higher for Khmer. The results show that the SEM is relatively small for Khmer for the Exceeds benchmarks, especially in Round 1 and 3. This is a consequence of the ceiling effect for this benchmark. To a lesser extent this also holds for the Exceeds benchmark for Mathematics in all rounds.

Table 37. Standard Error of Measurement by Round

	SEM by Benchmark								
	Round 1			Round 2			Round 3		
Subject	Partially Meets	Meets	Exceeds	Partially Meets	Meets	Exceeds	Partially Meets	Meets	Exceeds
Khmer	0.80	1.07	0.13	0.80	1.21	0.36	1.12	0.76	0.04
Mathematics	0.52	0.70	0.42	0.48	0.91	0.87	0.55	0.68	0.49

As far as panelist consistency and panelists' agreement are concerned, the results show that the inter-rater consistency for both Khmer and Mathematics was higher in Round 2 than in Round 1, as should be expected. In Round 3 they were at the level of Round 1, which could also be expected. According to the PLT values of 0.80 or greater are desirable, as they indicate substantial agreement between the panelists. Both for Khmer and Mathematics the inter-rater consistency was above 0.80 in all instances (see Table 38).

The intra-rater consistency index evaluates the panelists' overall consistency in estimating item difficulty. A lower value indicates high consistency and a higher value indicates low consistency. We see that the intra-rater consistency is quite high (given the scale of 0 to 1): the lowest value is 0.48 for Mathematics in Round 1 and the highest 0.85 for Mathematics in Round 2.

Table 38. Inter-rater consistency and intra-rater consistency by subject and round

	Round 1		Round 2		Round 3	
Subject	Inter-Rater Consistency	Intra-Rater Consistency	Inter-Rater Consistency	Intra-Rater Consistency	Inter-Rater Consistency	Intra-Rater Consistency
Khmer	0.81	0.72	0.84	0.72	0.81	0.66
Mathematics	0.81	0.48	0.87	0.85	0.82	0.57

Internal Evaluation SEM, Panelist Consistency and Panelists' Agreement for the total item sets

The measures of consistency for the total set of 59 Khmer and 58 Mathematics items are reported below in Table 39 and Table 40. Both for Khmer and for Mathematics the SEMs are low enough and the inter-rater consistency is above 0.80. The intra-rater consistency is relatively high.

Table 39. Standard Error of Measurement for the total set of 59 Khmer and 58 Mathematics items

	Standard Error of Measurement		
Subject	Partially Meets	Meets	Exceeds
Khmer	1.60	1.51	0.15
Mathematics	0.91	1.02	0.66

Table 40. Inter-rater consistency and intra-rater consistency by subject for the total set of 59 Khmer and 58 Mathematics items

Subject	Inter-Rater Consistency	Intra-Rater Consistency
Khmer	0.83	0.61
Mathematics	0.84	0.54

Summary of results of criterion 4 for the 4.1.1 Review Panel

Round 3

In the PLT (Annex U, p. 164) six criteria are mentioned for the validity of a 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, the extent unto which the panelists meet a set of selection criteria and panelists' understanding of the procedures. These measures for Round 3 are summarized in Table 19 and Table 20.

In this report we review the outcomes of Round 3 as far as intra-rater and inter-rater reliability and standard error of measurement of the benchmarks are concerned. Because the panelists in Round 3 were the same as in the two previous rounds, the information with respect to the other criteria is repeated in the two tables below for the sake of completeness.

For Khmer (Table 19), the intra-rater and inter-rater reliability in Round 3 meet the requirements. In addition to this, the standard error of measurement is low enough. However, as was the case in Round 1 and Round 2, the third benchmark ("Exceeds") might not be valid. There is almost no variation for the Exceeds benchmark as all panelists except one set the benchmark at the maximum score. In other words, there is a clear ceiling effect (even though this is not mentioned as a criterium). The adequacy of the policy linking procedure used in Round 3 can therefore be considered to be good.

For Mathematics (Table 20), the intra-rater and inter-rater reliability in Round 3 meet the requirements. The standard error of measurement is low enough. The adequacy of the policy linking procedure used in Round 3 can therefore be considered to be good.

Total set of items

Table 39 and Table 40 make clear that both for Khmer and Mathematics the intra-rater-, the inter-rater reliability and the SEMS for the total set of items meet the requirements.

Table 41. Summary of Results for Criteria for Policy Linking Validity for Khmer Grade 6 for Round 3

Question	Criteria	Response
m) What was the intra-rater reliability for the third 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.	0.66
n) What was the inter-rater reliability for the second round of ratings?	The inter-rater reliability should be at least .80.	0.81
o) 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: 27 1.12 (Partially Meets) 0.76 (Meets) 0.04 (Exceeds)
p) 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.	• Teachers: 50% female; 50% male SME's: 23% female, 77% male • N/A • N/A • NA
q) 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.	• Teacher mean > 15 years SME mean > 7 years • 23 of 23 • 23 of 23 • Yes • Yes • Yes

r)	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 – **4.46** • I understand the relationship between domains, constructs, subconstructs, knowledge and skills, and GPDs - **4.46** • The GPDs were clear and easy to understand - **4.33** **NLA** • I understand the purpose of the assessment - **4.42** • I understand the constructs assessed in the assessment - **4.38** • I understand how the assessment is administered - **4.33** **Alignment** • I understand the purpose of alignment - **4.38** • I understand the alignment methodology - **4.29** • I understand the difference between no fit, partial fit, and complete fit - **4.29** **Matching** • I understand the purpose of matching - **4.21** • I understand the matching methodology - **4.38** • I understand how the alignment activity links to the matching activity - **4.29** **Benchmarking methodology** • I understand the process I need to follow to complete the benchmarking exercise - **4.38** • I understand how the benchmarking methodology links to the steps on alignment and matching - **4.33** • I understand the difficulty level of the assessment items - **4.29** **Benchmark round 2** • I understand the data on others' ratings - **4.25** • I understand the item difficulty data and how it relates to this process - **4.42** • I understand the impact data and how it relates to this process - **4.25** **Comfortable with Round 2** • How comfortable are you with your final performance predictions? - **4.79**

Table 42. Summary of Results for Criteria for Policy Linking Validity for Mathematics Grade 6 for Round 3

Question		Criteria	Response
a)	What was the intra-rater reliability for the third 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.	0.57
b)	What was the inter-rater reliability for the second round of ratings?	The inter-rater reliability should be at least .80.	0.82
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: 28 0.55 (Partially Meets) 0.68 (Meets) 0.49 (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.	• Teachers: 40% female; 60% male SME's: 8% female, 92% male • N/A • N/A • NA
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.	• Teacher mean > 12 years SME mean > 13 years • 23 of 23 • 23 of 23 • Yes • Yes • Yes

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 - **4.44** • I understand the relationship between domains, constructs, subconstructs, knowledge and skills, and GPDs - **4.52** • The GPDs were clear and easy to understand - **4.41** • **NLA** • I understand the purpose of the assessment - **4.44** • I understand the constructs assessed in the assessment - **4.41** I understand how the assessment is administered - **4.30** **Alignment** • I understand the purpose of alignment - **4.37** • I understand the alignment methodology - **4.30** • I understand the difference between no fit, partial fit, and complete fit - **4.30** **Matching** • I understand the purpose of matching - **4.37** • I understand the matching methodology - **4.37** • I understand how the alignment activity links to the matching activity - **4.30** **Benchmarking methodology** • I understand the process I need to follow to complete the benchmarking exercise - **4.30** • I understand how the benchmarking methodology links to the steps on alignment and matching - **4.22** • I understand the difficulty level of the assessment items - **4.26** **Benchmark round 2** • I understand the data on others' ratings - **4.30** • I understand the item difficulty data and how it relates to this process - **4.33** • I understand the impact data and how it relates to this process - **4.26** **Comfortable with Round 2** • How comfortable are you with your final performance predictions? - **4.74**

Plausible Explanations and Recommendations

Plausible explanations

Lack of robustness of the benchmarks

There are several plausible explanations for the unexpected positive results. If we assume that the results really are too positive, meaning too high percentages of learners in the higher GP categories, then the most plausible one is the lack of "robustness" of the outcomes of the standard setting process: a change of one score point with the benchmarks results in a large change in the percentages of students within the different categories of the GPF. This lack of robustness is caused by the ability distribution in the Cambodian population in relation to the standard error of measurement (SEM) in the benchmark scores. The lack of robustness can be demonstrated by adding a 95% reliability interval to the benchmark scores based on the SEMs calculated. The lower and upper boundaries of the benchmark scores correspond with lower and upper percentages of learners in the JPM-, JM and JE categories of the GPF.

The results of this procedure were already presented in chapter 2, where we presented a number of tables showing these percentages with their confidence interval for Round 1 and 2 in the workshop and the additional Round 3 performed by EQAD. The tables all show that there is indeed a lack of robustness with the benchmarks estimates and that the confidence intervals around the percentages of learners in the JPM-, JM and JE categories of the GPF are large. Results are somewhat better in Round 2 compared to Round 1, because of the feedback that was given to the raters after Round 1. And we see that the results in Round 3 are less positive than in the first two rounds, but that is something that might be expected, because the raters were aware of the fact that the results of the workshop itself were unexpectedly positive for EQAD. The lack of robustness is quite persistent. Even if the ratings on the different item sets of Round 1 and 2 and Round 3 are taken together, confidence intervals around the percentages remain relatively large.

Representativeness or outliers with raters

A second explanation has to do with the raters. However, their representativeness was already checked for the original report on the workshop and seems to be in order. Furthermore, we already reported in chapter 3 of this report (in Tables 13 and 14) on the SEM and the inter- and intra-rater consistency for all three rounds. The SEM was, in all three rounds, under 1.0 for Mathematics and not much higher for Khmer. The inter-rater consistency for both Khmer and Mathematics was higher in Round 2 than in Round 1, as should be expected. In Round 3 they were at the level of Round 1, as also could be expected. Both for Khmer and Mathematics the inter-rater consistency was above 0.80 in all instances. Finally, the intra-rater consistency is somewhat better in Round 3. For Khmer there is a value of 0,66, compared to a value of 0,72 both in Round 1 an 2. And for Mathematics there is a value of 0,57 compared to a value of 0,48 in Round 1 and a value of 0,85 in Round 2. Inspection of all the individual inter-rater consistencies for the items and all the individual intra-rater consistencies for all the raters, showed that there were no outliers. This means that if the unexpectedly positive results are caused by a rater effect, this holds for the group of raters as a whole.

Sampling weights and plausible values

Another factor that could influence the results is the fact that sampling weights from the NLA were not provided. This means that the benchmarks calculated only hold for the sample of learners that took the NLA and cannot be generalized to the Cambodian population as a whole. Furthermore, the ability estimates could be biased, because no use was being made of plausible values. In addition to this, the results could partly be caused by the quality of the IRT calibration of the items in the NLA. The NLA data received contained IRT parameters of the items, but there was no information on the accuracy of the parameters, so the estimation error of the IRT parameters could not be taken into account. However, because we were not provided with information on sampling weights and plausible values were not calculated, it is not possible to check if this had an effect on the results of the workshop.

Different benchmarking procedure

A fourth explanation could be that the outcomes of the policy linking workshop are completely valid, but that they are not in accordance with the original benchmarks (Below Basic; Basic; Proficient and Advanced) from the NLA, because these were established through a different procedure (EQAD, 2017). The validity of this last explanation could be examined by a detailed comparison between the Grade Six Performance Standard Skills from the NLA with the Global Proficiency Framework. However, we are not in a position to compare the Grade Six Performance Standard Skills from the NLA with the GPF. It is therefore not possible for us to acknowledge or deny the plausibility of this explanation.

All in all, there are several plausible explanations for the outcomes found in the workshop. But there are no clear-cut criteria to decide what the real reasons are for the unexpectedly positive outcomes of the workshop. Our best guess is that the outcomes could have been caused by a lack of robustness in the standard setting combined with the ability distribution in the population. Small differences in benchmarks lead to relatively large differences in percentages and the confidence intervals around these percentages are also large. However, there are two assumptions underlying this best guess. The first one is that that the sample of learners that took the NLA is really representative for the Cambodian population. And the second one is that the IRT-analyses performed led to valid estimates of item- and learner parameters. Additional explanations might be found in the different standard setting procedures used for establishing the original benchmarks on the NLA and the ones produces in the workshops and Round 3.

If we look at the different benchmarks produced, both in the workshop and in Round 3, probably the best estimates of the 'real' benchmarks are those for the total set of items, because here the benchmarks have the smallest standard error of measurement and the percentages therefore have the smallest confidence intervals. However, these remain relatively large.

Towards a procedure with survey designs and IRT modelling

The challenge we had to address with the workshop in Cambodia was that the assessment on which benchmarks had to be set, the NLA, was used for national assessment and employed a survey design and IRT modelling. The NLA is a low-stakes system level assessment that summarizes students' achievement for Khmer and Mathematics at national and subnational levels. Not all items were administered to all learners. Items were divided into three partly overlapping nominally equivalent booklets. Each booklet for Khmer contained 33 items and each booklet for Mathematics 32 items. The technical report provided by EQAD (EQAD, 2017) did not contain information on the specific IRT model used for reporting, but the data that were provided indicate that the two-parameter Birnbaum model (Birnbaum, 1968) must have been used.

However, the PLT did not contain guidelines, methods or procedures to apply in such a situation. For reasons of efficiency, it was decided to use only a subset of all items. Roughly speaking both for Khmer and Mathematics, one of the booklets was selected, because they were all nominally equivalent. The selection consisted of 33 items for Khmer and 31 items for Mathematics. The IRT-parameter values of the items were shared before the workshop, including the NLA data.

Although the outcome of the workshop was unexpectedly positive, the approach used is justified in a situation where the assessment is administered through a survey design. Because IRT item parameters are known, the benchmarks needed can be calculated with enough precision for a limited number of items. By using the values of the item parameters of the selected items, the outcome of the standard setting procedures can simply be used to calculate the corresponding values on the underlying NLA 2016 ability scale for Khmer or Mathematics. And because the ability distribution is known, of course, the percentage of learners within the boundaries of the calculated ability scale benchmarks can also be established. Having the raters work with the complete set of items is therefore not necessary and could even lead to less valid results, because standard setting with a large set of items can be strenuous.

However, in order for this procedure to lead to valid outcomes, there are several conditions that have to be met. Some of these were already touched upon in the previous paragraph on plausible explanations of the unexpectedly positive results of the workshop.

First of all, the IRT-parameters obtained in the national assessment have to be valid and need to have a small enough standard error to warrant their use in the procedure described. And the items selected have to cover the relevant part of the ability distribution. They have to provide enough information in a statistical sense to prevent the standard errors of the calculated benchmarks becoming too large and to prevent ceiling or floor effects with the Below Partially Meets and Exceeds benchmarks. Note that in this instance, clear differences were found between Round 1 and Round 2 on the one hand and the additional Round 3 on the other hand as far as the percentages are concerned. Khmer all the percentages in Round 3 differ significantly from the results in Round 1 and 2. For Mathematics only the percentages in the Below Partially Meets category differ significantly, but the percentages in the other three categories also differ markedly. This does not imply, however, that the results from the IRT calibration cannot be trusted. Although we should expect comparable benchmarks (i.e. not differing significantly) with a different item set, the changes found can be caused by a different aspect. Raters knew that the outcomes of the workshop were unexpectedly positive and this may have led to a negative bias in the standard setting in Round 3.

Secondly, as we mentioned already, sampling weights and plausible values were not available to us. This could mean that the benchmarks calculated only hold for the sample of learners that took the NLA and cannot be generalized to the Cambodian population as a whole. Furthermore, the ability estimates could be biased, because no use was being made of plausible values. To rule out sampling effects and biases in estimates, these data are necessary.

Thirdly, the confidence intervals we calculated for the benchmarks and the percentages were large. This makes clear that more attention should be given to the selection of items for the standard setting procedure. Visualizing the ability distribution and the position of all items on the ability scale could be a big help in selecting the best suited (i.e. giving the most information in a statistical sense) items.

In addition to this, it is important to mention that there are several IRT models that can be employed and in most of them the sum score is not a sufficient statistic. In other words, for these models it matters *which* items learners answer correctly rather than how many. This has consequences for the benchmarking procedure and the analyses. When an IRT is used for which the sum score is not a sufficient statistic, it is necessary to know exactly which items, according to a panelist, two out of three learners from a JPM, JPM or JE are able to answer correctly.

All in all, there is reason to extend the PLT in several ways when a survey design with IRT modelling is used:

- Consider adding the assessment design, sampling weights, item parameters and ability estimates (or plausible values) to the list of materials that need to be obtained.
- Consider developing a separate description of the analyses that have to be performed in this situation.
- Consider employing different standard setting procedures more suited, like the Bookmark (Mitzel et al., 2001) or 3DC (Keuning, Straat & Feskens, 2017) method.
- Consider expanding the task of the 4.1.1. Review Panel with checking beforehand the quality of the IRT calibration to find out if the assessment proposed is suited for policy linking.

- Consider providing concrete guidance on item selection with a focus on the statistical information which the selected items could provide. This could be done by visualizing the ability distribution and the position of the benchmarks and items on the underlying ability scale.
- Consider including the calculation of confidence intervals on the benchmarks and percentages in the default statistical procedures.
- The analyses should be expanded to include the analysis of outliers and calculation of the confidence intervals of the benchmarks.

.

References

Birnbaum (1968). Some latent trait models and their use in inferring an examinee's ability. In: F.M. Lord & M.R. Novick (Eds.) *Statistical theories of mental test scores*. (pp 397-424). Reading: Addison-Wesley

EQAD (2017). *Results of Grade Six Student Achievement from the National Assessment in 2016.* Ministry of Education, Youth and Sports Cambodia.

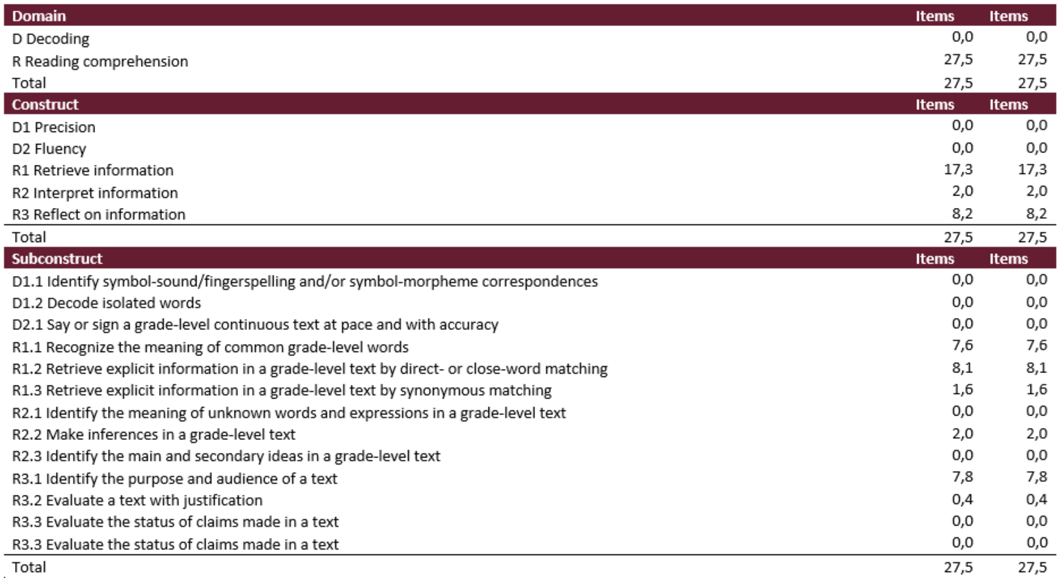
Keuning, J., Straat, J.H. Straat & Feskens, R. (2017). The Data-Driven Direct Consensus (3DC) Procedure: A New Approach to Standard Setting. In: Blömeke, S.K & Gustafsson J.E. *Standard Setting in Education: The Nordic Countries in an International Perspective*. pp. 263–278. Springer International Publishing. doi:10.1007/978-3-319-50856-6.

Mitzel, H.C., Lewis, D.M., Patz, R.J., & Green, D.R. (2001). The Bookmark procedure: Psychological perspectives. In: Cizek, G.J. (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives*. Lawrence Earlbaum, Mahwah, NJ, pp 249-281.

Annex

Alignment of the complementary NLA items with the domains, constructs and subconstructs

Table 43. Khmer: Number of complementary NLA items (total of 28) aligned to each grade 6 domain, construct and subconstruct



Domain	Items	Items
D Decoding	0,0	0,0
R Reading comprehension	27,5	27,5
Total	27,5	27,5
Construct	**Items**	**Items**
D1 Precision	0,0	0,0
D2 Fluency	0,0	0,0
R1 Retrieve information	17,3	17,3
R2 Interpret information	2,0	2,0
R3 Reflect on information	8,2	8,2
Total	27,5	27,5
Subconstruct	**Items**	**Items**
D1.1 Identify symbol-sound/fingerspelling and/or symbol-morpheme correspondences	0,0	0,0
D1.2 Decode isolated words	0,0	0,0
D2.1 Say or sign a grade-level continuous text at pace and with accuracy	0,0	0,0
R1.1 Recognize the meaning of common grade-level words	7,6	7,6
R1.2 Retrieve explicit information in a grade-level text by direct- or close-word matching	8,1	8,1
R1.3 Retrieve explicit information in a grade-level text by synonymous matching	1,6	1,6
R2.1 Identify the meaning of unknown words and expressions in a grade-level text	0,0	0,0
R2.2 Make inferences in a grade-level text	2,0	2,0
R2.3 Identify the main and secondary ideas in a grade-level text	0,0	0,0
R3.1 Identify the purpose and audience of a text	7,8	7,8
R3.2 Evaluate a text with justification	0,4	0,4
R3.3 Evaluate the status of claims made in a text	0,0	0,0
R3.3 Evaluate the status of claims made in a text	0,0	0,0
Total	27,5	27,5

Table 44. Mathematics: Number of complementary items (total of 39) aligned to each grade 6 domain, construct and subconstruct

Domain	Items	Items
N Number and operations	21,1	21,1
M Measurement	5,2	5,2
G Geometry	3,9	3,9
S Statistics and probability	0,5	0,5
A Algebra	4,5	4,5
Total	35,3	35,3
Construct	**Items**	**Items**
N1 Whole numbers	4,4	4,4
N2 Fractions	9,6	9,6
N3 Decimals	7,1	7,1
M1 Length, weight, capacity, volume, area, and perimeter	3,4	3,4
M2 Time	1,9	1,9
G1 Properties of shapes and figures	2,9	2,9
G2 Spatial visualizations	0,9	0,9
G3 Position and direction	0,1	0,1
S1 Data management	0,5	0,5
S2 Chance and probability	0,0	0,0
A1 Patterns	0,0	0,0
A3 Relations and functions	4,5	4,5
Total	35,3	35,3
Subconstruct	**Items**	**Items**
N1.1 Identify and count in whole numbers, and identify their relative magnitude	0,0	0,0
N1.2 Represent whole numbers in equivalent ways	0,5	0,5
N1.3 Solve operations using whole numbers	2,9	2,9
N1.4 Solve real-world problems involving whole numbers	1,0	1,0
N2.1 Identify and represent fractions using objects, pictures, and symbols, and identify relative magnitude	6,0	6,0
N2.2 Solve operations using fractions	3,4	3,4
N2.3 Solve real-world problems involving fractions	0,2	0,2
N3.1 Identify and represent decimals using objects, pictures, and symbols, and identify relative magnitude	0,0	0,0
N3.2 Represent decimals in equivalent ways (including fractions and percentages)	4,5	4,5
N3.3 Solve operations using decimals	1,7	1,7
N3.4 Solve real-world problems involving decimals	0,9	0,9
M1.1 Use non-standard and standard units to measure, compare, and order	0,1	0,1
M1.2 Solve problems involving measurement	3,2	3,2
M2.2 Solve problems involving time	1,9	1,9
G1.1 Recognize and describe shapes and figures	2,9	2,9
G2.1 Compose and decompose shapes and figures	0,9	0,9
G3.1 Describe the position and direction of objects in space	0,1	0,1
S1.1 Retrieve and interpret data presented in displays	0,5	0,5
S2.1 Describe the likelihood of events in different ways	0,0	0,0
A1.1 Recognize, describe, extend, and generate patterns	0,0	0,0
A3.1 Solve problems involving variation (ratio, proportion, and percentage)	3,8	3,8
A3.2 Demonstrate an understanding of equivalency	0,7	0,7
Total	35,3	35,3