UNESCO
unesco
Institute for Statistics

4 QUALITY EDUCATION

Credit: Sumit Saraswat/Shutterstock.com

Report of the National Achievement Survey
Policy Linking for Measuring Global Learning Outcomes Workshop
(March 2021)

Setting Global Benchmarks for Grade 8 Hindi Language and Mathematics in India

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.

Acknowledgements

Acknowledgements

Table of Contents

Acronyms and Abbreviations

GPD   Global Proficiency Descriptor

GPF   Global Proficiency Framework

GPL   Global Proficiency Level

JE   Just Exceeds Minimum Proficiency

JM   Just Meets Minimum Proficiency

JP   Just Partially Meets Minimum Proficiency

PLT   Policy Linking Toolkit

SDG   Sustainable Development Goal

SEM   Standard Error of Measurement

USAID U.S. Agency for International Development

NCERT National Council of Educational Research and Training

SCERT State Council of Educational Research and Training

NAS   National Achievement Survey

UNESCO Institute for Statistics (UIS),

Glossary of Terms from the Policy Linking Toolkit

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1. Executive Summary

This document contains the report on the online policy linking workshop that took place from 14 March 2021 until 19 March 2021. The National Council for Educational Research and Training (NCERT) in India and UNESCO Institute for Statistics (UIS) organized this workshop as a pilot. The objective of the workshop was to set global benchmarks on the 2017 National Achievement Survey (NAS) at grade 8 in Hindi language and mathematics using a remote policy linking workshop.

After India had participated in an in-person policy linking workshop in 2019, India this year participated in a blended policy linking workshop. Cito hosted the workshop using a videoconferencing platform (Teams). The participants met in small groups in 10 different locations. The participants performed their tasks with dedication and engaged in lively discussions during the tasks. Every step of the process produced important outcomes. The participants gave very positive feedback, both in person and in their evaluation forms.

The participants' work showed that the NAS for Hindi Language is strongly aligned to the Global Proficiency Framework for grade 8. Mathematics is in depth additionally aligned to the Global Proficiency Framework for grade 8. Furthermore, the panelists managed to reach almost complete consensus on the matching. The final benchmarks of the panelists show a good consistency, which makes the benchmarks useable for comparing, aggregating, and tracking learning outcomes for the NAS in the nine states in which Hindi is the main language. The piloting of the policy linking workshop in a blended mode can be considered a success.

2. Background

Policy Linking Overview

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

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

Global Proficiency Framework

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

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

The policy linking methodology

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

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

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

Policy linking workshop

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

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

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

Overview to the National Achievement Survey (NAS)

The National Council of Educational Research and Training (NCERT) has been conducting the National Achievement Survey (NAS) since 2001. Until 2017, the NAS has been administered over four cycles for grades 3, 5, and 8, along with two cycles for class 10. The latest NAS was conducted in 2017 nationwide for "Classes III, V and VIII on a single day" (NCERT, 2020, p. xi).

The major objective of the NAS is a system level reflection on effectiveness of school education in India. The NAS 2017 was designed for the following objectives (NCERT, 2020, p. 5):

- To report performance of students in different subjects and classes on specific learning outcomes
- To compare the average performance of the different groups of children:
- To identify key learning gaps in achievement of learning outcomes

---

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

- To identify institutional and contextual factors that affect learning achievement of students

Content and design of the NAS in grade 8

The NAS is a system level assessment that summarizes students' achievement at National, State/ UT and District levels. In the first cycles the test questions were based on the common core content across States/UTs. In 2017, NCERT defined the subject and class wise Learning Outcomes (LOs). The survey was designed to measure the student attainment of Learning Outcomes at the end of grades III, V and VIII by assessing students through variety of items measuring skills and competencies (NCERT, 2020, p. 3).

Grade VIII students were tested in Language, Mathematics, Science, and Social Sciences. The paper-and-pencil assessment lasted for 120 minutes for grade VIII (and 90 minutes for the other classes). All the items were multiple choice questions with four response alternatives from which the students were required to select the correct option. The test was translated into 20 languages.

Two test forms were developed for each grade. In grade 8 each test form consisted of 60 items, 15 items per subject. In NAS 2017, tests were linked by placing five common items across the two forms of each class, see Figure 1. Both test forms were equated to adjust for differences in difficulty levels (NCERT, 2020, p. 20). For the Hindi language two forms were used each containing 15 items. The two forms had 5 items in common (see Table 1), so in total 25 unique items were used. Also, for Mathematics two forms were used of 15 items each, with 5 items overlap.

Table 1. Number of items per form in Grade 8

Subject	Form	Unique	Common	Total
Hindi language	1	10	5	15
	2	10	5	15
	Total	20	5	25
Mathematics	1	10	5	15
	2	10	5	15
	Total	20	5	25

Figure 1. Test design of the NAS in grade VIII
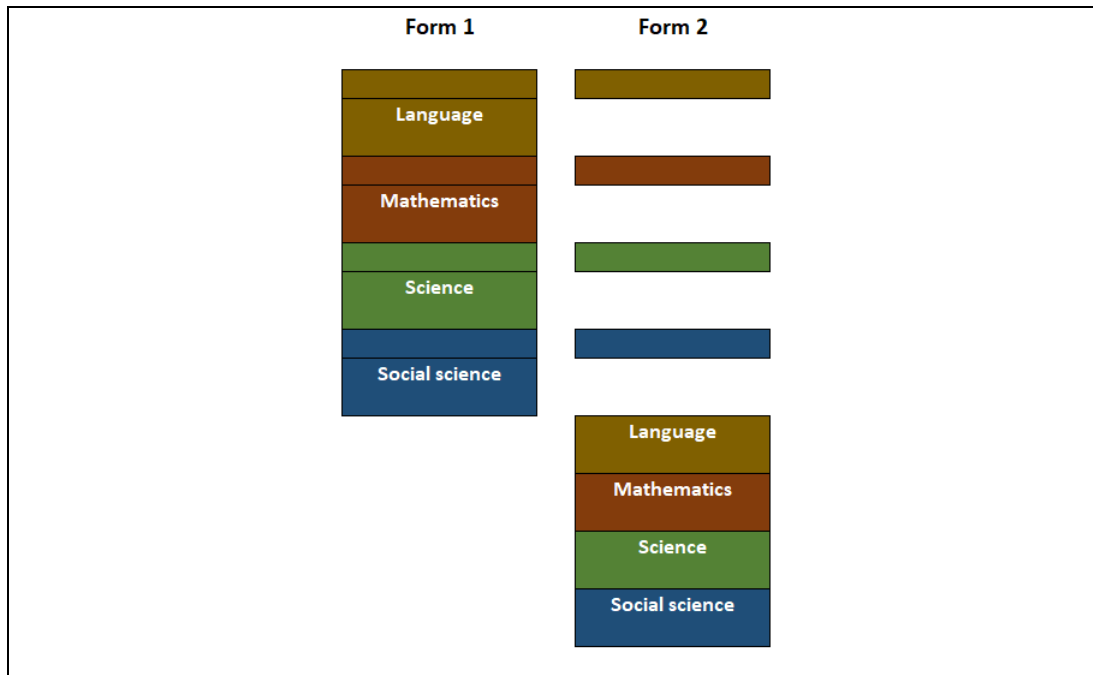
Sample and data analysis

The sampling design used for NAS 2017 is a two-stage stratified cluster sample design. For the first stage of sampling, public schools were stratified, explicitly and/or implicitly. In each district school samples were drawn using a Probability Proportional to Size (PPS) sampling procedure. From each participating District 51 schools were selected for grade VIII. Selection of students was done through a random sampling procedure (NCERT, 2020, chapter 3). Generally, only one section per school was sampled. The desired student sample size per grade was 30. If there were more than 30 students in the sampled section in a class, then only 30 students were randomly selected. To reflect the sampling design and adjust for the different probability of a student being selected from schools of different size school and student weights were determined.

Data were collected "from approximately 110.000 schools, 270.000 teachers and 2.200.000 students through tests and questionnaires from 701 districts of 36 States/UTs of the country" (NCERT, 2020). In grade 8 765.631 students were tested (see Table 2). The reporting scale for the National Achievement Survey (NAS) 2017 data was based on the Item Response Theory (IRT). The IRT model chosen for item calibration was a two-parameter logistic (2-PL) model. Items were calibrated using the strategy that centered the mean of item difficulties to zero and evaluated the distribution of ability estimates in relation to the mean of item difficulties. The item responses of the students were used to estimate their latent ability. The latent ability of students was estimated using the Weighted Maximum Likelihood (WML) method (NCERT, 2020, p. 62). Next, the ability estimates were converted into a reporting scale with a mean of 300 and standard deviation of 50.

3. Pilot Workshop Preparation

Objective of the workshop

The objective of the workshop was setting global benchmarks on the 2017 National Achievement Survey (NAS) at grade 8 in Hindi language and mathematics using a remote policy linking workshop. The workshop had a piloting function and should increase the capabilities of NCERT to conduct similar workshops in the future. From the 36 States/UTs that participated in the NAS, the nine states in which Hindi is the main language[2] participated in the policy linking workshop (see Table 2). NCERT requested to set three benchmarks even though the NAS contained relatively few items (25 unique items for each subject).

First three policy linking stages

Wednesday, 20-01-2021, a kick-off meeting took place between UNESCO, NCERT 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 India, the 4.1.1. Review Panel was not yet in place. "This stage of the process involves the country government sharing standard-, curriculum-, and assessment-related documents (including the most recent round of data) with the project team and examination of those documents by the project team and the 4.1.1 Review Panel to determine whether the assessment(s) meets reliability and validity standards required for a country to proceed with policy linking for reporting global outcomes." (Policy Linking Toolkit, p. 170). The 4.1.1. Review Panel uses three criteria: Alignment between the assessment and the curriculum, Appropriateness of the assessment for the population, Reliability of the assessment.

As the 4.1.1 Review Panel was not in place, Cito made an initial assessment of whether the assessment(s) meets reliability and validity standards required to proceed with policy linking. The Technical Report of the 2017 NAS, shows that the reliability of the assessment (the NAS) is sufficient. Cronbach's alpha for the 15 item forms in language and in mathematics is above .70 (NCERT, 2020, pp. 129-130).

The evidence presented in the Technical Report of the 2017 NAS shows that the NAS also seems appropriate for the population. There is evidence that the items have been reviewed to determine their validity. NCERT piloted the items and tested them also in a field trial (NCERT, 2020, p. 13) and presents the item parameters (both from Classical Test Theory and Item Response Theory) in the Technical Report. The implemented sampling procedure (NCERT, 2020, Chapter 3) ensures that the learners who carried out the assessment are representative of the population against which results are reported.

Cito could not evaluate the alignment of the NAS with the curriculum and the Global Proficiency Framework, because the NAS items had to be kept confidential. But, as India already participated in a policy linking workshop for the NAS Grades 3 and 5 English Language and Mathematics in 2019, the expectation was that the NAS would meet the standards to proceed with policy linking for reporting global outcomes.

---

[2] Hindi is also the main language in two union territories (Chandigarh and the National Capitol Territory of Delhi).

Table 2. Number of grade 8 students in every state and in the nine participating states

State name/Union Territory	Number of grade 8 students	Number of students in states participating in the policy linking workshop
Andaman & Nicobar Islands	2180	
Andhra Pradesh	16007	
Arunachal Pradesh	11241	
Assam	30654	
Bihar	44320	44320
Chandigarh	2888	
Chhattisgarh	32689	32689
Dadra & Nagar Haveli	2758	
Daman & Diu	873	
Delhi	9890	
Goa	3825	
Gujarat	41393	
Haryana	23095	23095
Himachal Pradesh	11003	11003
Jammu And Kashmir	14134	
Jharkhand	27925	27925
Karnataka	47021	
Kerala	17869	
Lakshadweep	870	
Madhya Pradesh	59652	59652
Maharashtra	48096	
Manipur	4252	
Meghalaya	9216	
Mizoram	5202	
Nagaland	4352	
Odisha	35825	
Puducherry	2142	
Punjab	24920	
Rajasthan	38818	38818
Sikkim	4983	
Tamil Nadu	32563	
Telangana	37585	
Tripura	7798	
Uttar Pradesh	70824	70824
Uttarakhand	13998	13998
West Bengal	23835	
Unknown	935	
Total	765631	322324

General preparation of the workshop

UNESCO and Cito planned to facilitate the workshop remotely, due to the COVID-19 pandemic. Given the positive experience with a previous piloting of the policy linking toolkit, NCERT preferred an in-person workshop. For this reason, NCERT proposed a blended workshop in which the panelists were attending in-person in their state (SCERT offices). To limit the days on which panelist had to travel to the State offices, Cito developed an agenda for a 6-day blended workshop. Before finalizing the agenda, this agenda was shared with the stakeholders (NCERT, UNESCO) for suggestions and improvements.

After approval from NCERT in India on Friday the 12th of March, the workshop took place in a blended format from Sunday 14-03-2021 until Friday 19-03-2021. Cito hosted the workshop using the platform Teams.

NCERT sought from each of the nine participating states four teachers: two teachers Hindi and two teachers Mathematics. During the first day of the workshop, NCERT shared the list of panelists. In total 36 panelists participated (see Table 3). From each state also a coordinator was present and at a national level nine experts and two local content facilitators participated. Furthermore, international observers were present during some of the sessions.

NCERT expected most teachers to master English sufficiently, therefore only one interpreter would be present (for Hindi language). The material would not be translated except for the Global Proficiency Levels for Language and some sample items for Language.

Table 3. Panelist' background information

State	Hindi Language	Mathematics
Bihar	2	2
Chhattisgarh	2	2
Harayana	2	2
Himachal Pradesh	2	2
Jharkhand	2	2
Madhya Pradesh	2	2
Rajasthan	2	2
Uttar Pradesh	2	2
Uttarakhand	2	2

Materials for the workshop and pre-workshop analyses

Collecting materials and pre-workshop analyses

Before the workshop, NCERT shared in total eight sample items, two of which were grade 8 items (one for language and one for mathematics). Because of confidentiality, the NAS itself could not be shared with the international facilitators before the workshop and panelists could not administer the NAS to nine learners. Also, during the workshop, the NAS could not be shared with the facilitators from Cito.

The raw data were shared before the workshop. Keys, weights and ability estimates were not included in the data file. For this reason, in preparation for the workshop, the raw data were scored using the key deduced from Table 5.20 (p. 129) and Table 5.21 in the Technical Report. Based on these item scores, the ability of each students was estimated using the item parameters from the 2 PL model published in the Technical report. The ability estimate

calculated was a so-called Warm-estimate (Warm,1989) and reflected the procedure described in the Technical Report.

> "In NAS 2017, student scores were determined by means of the IRT 'pattern-scoring' approach, where a pattern of student responses to items is used to estimate the latent ability (i.e., knowledge and competencies) underlying students' test performance. The techniques used for ability estimation was based on the Weighted Maximum Likelihood (WML) method, which is widely supported in research literature (p. 20)"

In the final step, based on the ability of each student and the item parameters, Cito estimated the expected sum score of the student on the 25 items of the NAS (see Figure 2).

Figure 2. Steps in estimating the sum scores on all 25 items of the NAS for Hindi Language and Mathematics
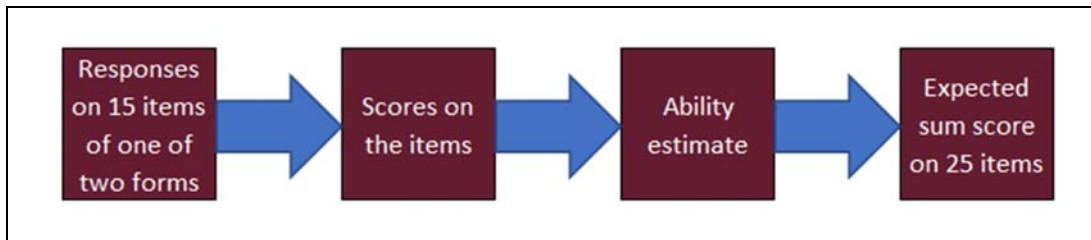
Creating workshop materials

Because panelists were attending in their state and might have to travel from their homes, a revised agenda was requested. To limit the number of travelling days, instead of a 3-week workshop, a six-day workshop was developed (see Appendix A).

Because the policy linking toolkit does not contain digital forms for remote workshops yet, for each of the three tasks Cito developed a digital form (see Appendix B). The digital forms were designed to ease the task of the panelists, to prevent inconsistent ratings and to speed-up the data analyses during the workshop. The digital forms contained macros and were sent to UIS in India for a test prior to the workshop. In the Netherlands, Cito tested the digital forms in old Excel-versions. For the evaluation of the workshop seven short questionnaires were designed in Microsoft Forms (see Appendix E). As the list of panelists was unknown at the start of the workshop, a Technical Test of the platform Teams and the digital forms could not be performed for each panelist in each state.

Cito prepared a package for panelists containing all workshop materials, to be printed on location. The package contained the Global Proficiency Framework for Grades 7 to 9, Glossary and acronym list, a handout of the slides of all presentations, an example of text for Hindi language. Furthermore, the package contained the Alignment rating form, Matching form, Item rating form, and evaluation forms in a printable format.

During the workshop, Cito shared the digital alignment form and item rating form through email on the day the form was needed. The evaluation forms were shared through an email link.

Table 4. Agenda for a 6-day blended workshop

Day 1—14 March 2021	Day 4—17 March 2021
Welcome and introductions	Individual work: Complete Task 2 Matching
Overview Presentation: Policy linking	Task 2 Presentation: Matching results
Overview Presentation: GPF	Task 3 Presentation: Global benchmarking & Angoff
Overview Presentation: NAS	Task 3 Activity: Practice and start Angoff ratings
Day 2—15 March 2021	Day 5—18 March 2021
Individual work: Do NAS & Review GPF	Individual work: Complete Round 1 ratings
Task 1 Presentation: GPF and alignment	Task 3 Presentation: Round 1 results
Task 1 Activity: Align NAS and the GPF	Task 3 Presentation: Discuss round 1 ratings
Day 3—16 March 2021	Day 6—19 March 2021
Individual work: Complete Task 1 Alignment	Individual work: Conduct Angoff Round 2
Task 1 Presentation: Alignment results	Task 3 Presentation: Round 2 results
Task 2 Presentation: Matching NAS and GPLs)	Task 3 Activity: Evaluate workshop
Task 2 Activity: Match NAS and GPDs/GPLs	Closing and logistics

Training the local content facilitators

At the end of February, the local content facilitators were introduced to the Cito content facilitators. As it proved difficult to plan a training for all content facilitators together, the content facilitators of Mathematics and the content facilitators of Hindi, planned three meetings before the start of the workshop. During the meetings the goals and the workflow of the workshop were introduced. The content facilitators discussed the different roles and responsibilities in relation to the planning of the meetings. The counterparts received the workshop materials in the week prior to the workshop. The counterparts also prepared the workshop sessions together in India.

4. Implementing the blended workshop

Familiarization

Following feedback from other policy linking workshops, the workshop started with a preparation session. After the formal welcome, the first day focused on familiarizing panelists with policy linking, the Global Proficiency Framework and the National Achievement Survey. Because of the confidentiality of the National Achievement Survey (NAS), the panelists had not seen the NAS before the workshop. The workshop materials (like the Global Proficiency Framework) had been shared with the panelists one day before the workshop, because a formal approval for the workshop arrived only two days before the workshop.

During the sessions, the panelists were provided with background information on policy linking, including a chronology of the development of the method in response to the global indicators. The facilitators then provided the panelists with training on the Global Proficiency Framework and its role in policy linking. The example of the benchmarks and the proficiency levels is shown in Figure 3. In the breakout rooms, the content facilitators introduced each of the domains, constructs, subconstructs, statements of knowledge and/or skill(s), and GPLs and GPDs. An example from part of the mathematics GPF is shown in Table 5.

Figure 3. Example of three benchmarks and the global proficiency levels
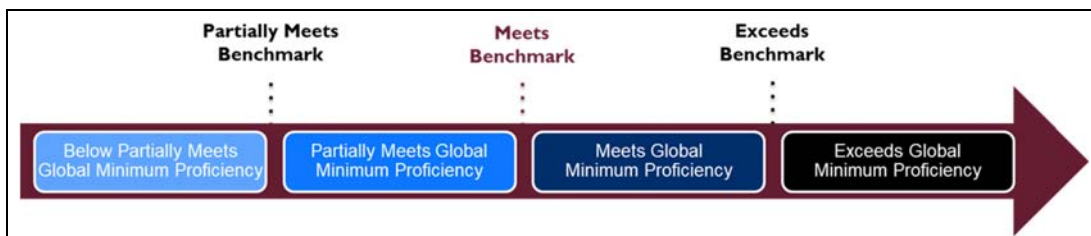
Table 5. Part of the Global Proficiency Framework of Mathematics describing the domain, constructs and subconstructs

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

The day closed with an introduction to the National Achievement Survey and discussing the first five items of the NAS in the subject-specific break-out rooms. In the morning of the second day the panelists were asked to study the Global Proficiency Framework and fill-out the NAS themselves. While answering the items of the NAS the panelists were asked to note stumble blocks and aspects of the items that would make the item easy or difficult for Grade 8 students.

Observations

Based on their previous experience with policy linking, NCERT did not expect a familiarization phase in which, for example, the lead facilitator would give a presentation about the NAS. Furthermore, NCERT expected that the panelists would start immediately with the first task (Alignment) in the morning. The conflicting expectations and instructions caused confusion among the local content facilitators and panelists about the individual work in the morning. Consequently, the panelists started later with studying the Global Proficiency Framework and filling-out the NAS. The panelists also started later, because NCERT decided to start every day one hour later than originally planned.

For most panelists, this was the first time they participated in an online workshop using a videoconferencing platform. The participants consequently experienced several technical difficulties. Because several panelists and experts were together in the same room without headsets, a disturbing echo occurred. Also, panelists had trouble with virtually changing rooms.

In Teams (and Zoom) the options are limited when participating without a license. Most participants were participating without license, which limited their options (such as choosing a background), but more importantly most participants were therefore only indicated as "Guest" in the main list. Therefore, it proved to be impossible to redirect the panelists automatically to the break-out rooms as intended.

Task 1: Alignment

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

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

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

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

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

- Complete Fit (C) signifies that all content required to answer the item correctly is contained in the statement of knowledge and/or skill(s), i.e., if the learner answers the

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

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

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

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

All results were summarized at the subconstruct level. Only the subconstructs were considered with knowledge and/or skill(s) expected at the grade level for which alignment was being conducted (grade 8). The data analyst took the average of the number of items that the panelists aligned to each grade 8 subconstruct, construct and domain. Each item was counted only once (even if it was a partial fit), non-fitting items were not counted towards alignment.

Alignment NAS Hindi Language

Averaging the panelists' ratings, we see that on average 24 out of 25 items aligned to Reading comprehension. At least 5 items were aligned to each of the three constructs of reading comprehension and 80% of all subconstructs were covered (Table 19). The NAS Language assessment was therefore strongly aligned in depth and breadth (see the criteria in Table 6).

Table 6. Reading Alignment Criteria for Grades 1–9

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

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

Alignment NAS 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 17 of the 25 items, aligned to grade 8 subconstructs. In the Global Proficiency Framework 18 subconstructs are mentioned for grade 8 and the NAS covered 14 of those subconstructs (see Appendix C). In breadth the NAS is strongly aligned to the Global Proficiency Framework for Grade 8 as the items covered more than 50% of all grade 8 subconstructs.

The NAS Mathematics items covered all five domains and ten out of 12 constructs for grade 8. According to the new criteria in the Policy Linking Toolkit, at least 5 items should align to the domain Number and Operations (see Table 7). On average only 1.6 items covered the domain of Number and Operations, for this reason the NAS did not align in depth to the Global Proficiency Framework. If we consider not only the grade 8 subconstructs, but also the grade 7 and grade 9 constructs, we find that the NAS is also strongly aligned in depth to the Global Proficiency Framework.

Table 7. Mathematics Alignment Criteria for Grades 1–9

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

Observations

During their individual work, panelists still needed a lot of guidance from the (local) content facilitators. At the beginning, the panelists had difficulty in understanding the terms "construct" and "subconstruct" and the many acronyms. A lot of the discussions of the panelists and experts present focused initially on the comparison between the Global Proficiency Framework and their own curriculum. Panelists had the tendency to compare the grade 8 constructs to their own grade 8 curriculum. Facilitation of the discussions was difficult, because no interpreter was present (Mathematics) or not all the time (Language).

Some panelists used the alignment rating forms without drop-down menus as the forms with macros did not work for them. Even after helping panelists to enable macros, on some computers the forms did not work. These panelists received a form without macros. As a consequence, the returned forms contained typos, often lacked a reference to the domain, construct or subconstruct (see the circles in

) or the fit, thus preventing automatic processing and fast analyses. In one case, two panelists handed in a form together, showing that they did not work independently.

Table 8. Reference to domain, construct and subconstruct to use in the alignment rating form.

Domain		Construct		Subconstruct		Knowledge or Skill
N	Number and	N1	Whole numbers	N1.1	Identify and count in whole numbers, and identify their relative magnitude	Count, read, and write whole numbers
						Compare and order whole numbers
						Skip count forwards or backwards
				N1.2	Represent whole numbers in equivalent ways	Determine or identify the equivalency between whole numbers represented as objects, pictures, and
						Use place-value concepts
						Round whole numbers
				N1.3	Solve operations using whole numbers	Add and subtract whole numbers
						Find the double or half of a set of objects
						Multiply and divide whole numbers
						Demonstrate fluency with basic addition and
						Demonstrate fluency with basic multiplication and
						Identify factors and multiples of whole numbers
						Perform calculations involving two or more
				N1.4	Solve real-world problems involving whole numbers	Solve real-world problems involving the addition and subtraction of whole numbers, including with
						Solve real-world problems involving the multiplication and division of whole numbers,

Task 2: Matching

On the third day, after the panelists completed task 1, they received training for the next task: Matching the NAS 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 4. Global Proficiency Levels (GPLs) and Global Proficiency Descriptors (GPDs) in the Global Proficiency Framework
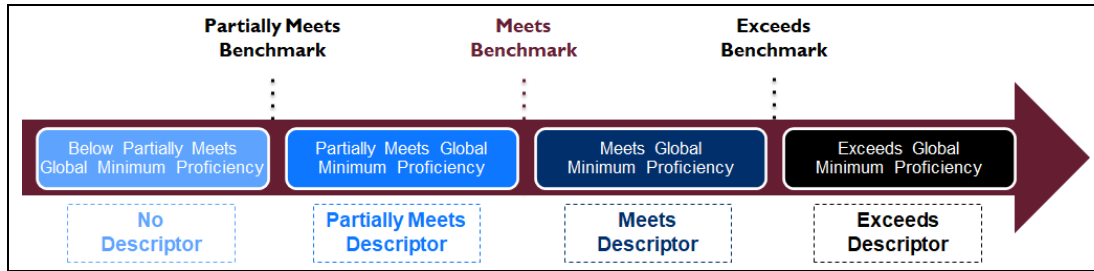
A Global Proficiency Descriptor is a detailed definition crafted by subject matter experts that clarifies how much of the content described under the statements of knowledge and/or skill(s) in the Global Proficiency Framework a learner should be able to demonstrate within a subject at a grade level. The Global Proficiency Descriptors (GPD) describes the minimum proficiency for the Global Proficiency Levels (GPLs), i.e., the minimum knowledge or skill(s) necessary for classification into each GPL (by grade and subject), see Figure 4.

The Global Proficiency Descriptors are organized by domain, construct and subconstruct, with descriptors for each subconstruct. In Table 9 an example is displayed of Global Proficiency Descriptors for the three GPLs (partially meets, meets and exceed global minimum proficiency). For language the panelists reached consensus on all but one item. For mathematics, the panelists managed to reach complete consensus regarding the proficiency levels matching with each item. However, with one item the panelists agreed to disagree about the primary (sub-)constructs.

Table 9. Example of the Global Proficiency Descriptors for three Proficiency Levels.

| Partially Meets Global Minimum Proficiency	Meets Global Minimum Proficiency	Exceeds Global Minimum Proficiency
G1: PROPERTIES OF SHAPES AND FIGURES		
G1.1: Differentiate shapes and figures by their attributes		
G1.1.1_P Use the defining attributes (i.e., type of angle, parallel and perpendicular lines) of complex two-dimensional shapes to classify them.	G1.1.1_M Recognize and name parts of the circle (i.e., radius, diameter, circumference) and identify the relationship between the radius and diameter.	G1.1.1_E N/A
G1.1.2_P 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°)*.	G1.1.2_M Use the angle sum of a triangle to solve problems *(e.g., determine the missing angle of a triangle where two angles are given)*.	G1.1.2_E Use the angle relationships associated with intersecting lines, and with parallel lines intersected by a transverse line to solve problems *(e.g., calculate missing angles on a diagram with parallel and intersecting lines)*.
G1.1.3_P Recognize single-step, two-dimensional shape transformations expressed quantitatively (i.e., rotation by a given fraction of a turn, reflection along a given mirror line, or enlargement by a given scale factor).	G1.1.3_M Describe and implement two-dimensional shape transformations (i.e., reflection, rotation, translation, enlargement/reduction).	G1.1.3_E Describe and implement sequential two-dimensional shape transformations (i.e., reflection, rotation, translation, enlargement/reduction).

Observations

Especially for language, it initially proved to be a difficult task for the panelists to match the items to Global Proficiency Descriptors. To aid the Matching, the international content facilitators joined the break-out rooms in the morning. In both groups, the panelists did not manage to reach consensus on the fourth day. For this reason, the panelists continued longer and used some time on the morning of the fifth day to reach consensus on almost all items.

The digital matching form developed for this purpose was not used, because of the macros contained therein. Instead the local content facilitator kept track of the consensus on paper.

Task 3: Benchmarking

On the fourth day the panelists received training in setting global benchmarks using the Angoff method, even though the Matching task was not finalized yet. The facilitator first presented a hypothetical example of how the benchmarking method would link a national assessment to the GPF, thus allowing for the calculation of the percentages of students attaining minimum proficiency (see Figure 5). This example was extended to three national assessments of different difficulties, and how this would lead to a different benchmark for each assessment. The facilitators discussed how the benchmarking results – when applied to the assessment data sets – could be used for comparing and aggregating assessment results, as well as tracking those results over time.

Figure 5. Example of an assessment and a benchmark
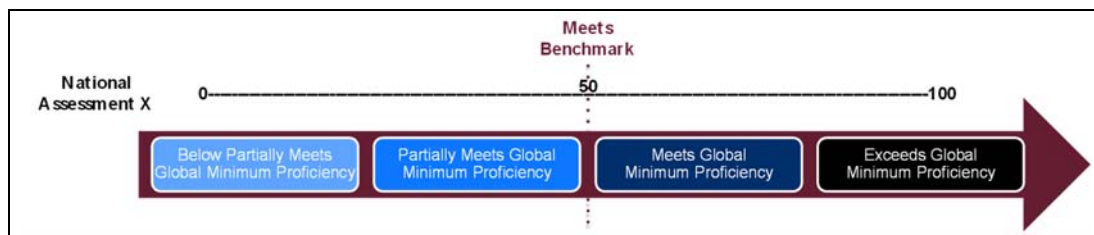
The panelists then received an introduction to their third task: setting benchmarks with the Angoff benchmarking method. The lead facilitator emphasized that the ratings for task 3 should be individual and independent and that, in contrast to task 2, consensus on the rating is not needed, even though consistency is desired.

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

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

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

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

Step 4: Based on an understanding of Steps 1–3, follow this procedure (displayed in Figure 6): Ask whether minimally proficient JP learners would be able to answer the item correctly, i.e., are

17

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 the second Task (Matching) was completed on the fifth day, 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 be able to answer at the respective GPL, rather than what a learner should be able to answer.

Figure 6. Steps for Rating Items
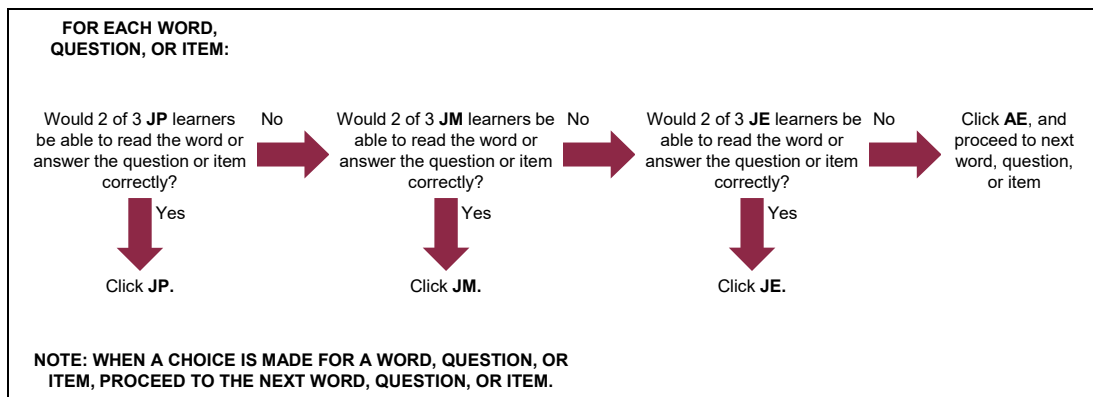
After the panelists conducted their first ratings in the morning of the fifth day, they sent the Item Rating form to Cito. The data analyst kept track of the forms sent, removed doubles (keeping the last one), renamed forms using the panelist ID and sent reminders regarding missing forms.

Once the forms were complete, Cito checked the forms to establish whether:

- The panelist rated all items
- The panelist had filled in the ID at the top (rather than the name, or missing)
- The panelist had used the correct form (the Language form for Languages and the Mathematics form for Mathematics)
- The panelist had used the correct symbol for entering the rating (an "X")
- The panelist had not accidentally deleted the calculations inside the form

After data cleaning, Cito merged the forms and the data analysis could start. The analyses were preliminary, as a data set with weights was not available prior to the workshop and not all forms were handed in in time for the analyses. We also worked with expected raw scores on all 25 items

of the NAS, as the pattern scoring (weighted scores) underlying the 2 PL model is very hard to understand for panelists.

The data-analyst and lead facilitator performed all 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 item difficulty parameters from the Technical Report (NCERT, 2020)
- 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 content facilitators presented the preliminary results of Round 1 and facilitated an item-wise discussion. The content facilitators focused during the discussion on those items where panelists strongly disagreed. The facilitators invited the panelists to share their views during the discussion.

Round 2

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

After receiving the final forms containing both ratings after the workshop, all checks and analyses were redone for the final results. During the workshop, NCERT sent a data file containing the sampling weights and the ability estimates of the students. Cito used these data to calculate the impact of the benchmarks reported in the next section.

Observations

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

Working remotely (and in part without interpreter), it proved to be difficult to ascertain that the ratings were individual and independent. Working in pairs on location and the presence of experts might have had an influence on the alignment ratings and item ratings.

Collecting the forms took quite some checking and time, because panelists worked at very different rates, sometimes did not submit the form, submitted the same forms several times and through different email-addresses, submitted revised forms, submitted forms using each other's email-addresses and submitted forms without ID's.

As we had to use forms without macro's, more data cleaning had to be done. For example, we received forms containing strange characters and forms in which the panelist had deleted accidentally necessary calculations.

Workshop evaluation

Near the end of the sixth day, after returning the Round 2 ratings, all panelists were asked to share their opinion about the workshop. Their evaluations are completely anonymous. They were informed that their opinion was important to improve the workshop and to evaluate the validity and reliability of the standard setting process. Panelists each received a link, through which they could fill in a questionnaire about the workshop. In the original agenda a daily evaluation was planned, but as NCERT had decided to start each morning one hour later, the evaluation was shifted to the end of the workshop. 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 Achievement Survey
c) The training on the alignment methodology
d) The training on the matching methodology
e) The training on the benchmark-setting (Angoff) methodology
f) Benchmark Round 2 evaluation
g) Overall evaluation

The questions included are presented in the policy linking toolkit (see also Appendix E). The questionnaire, made in Microsoft Forms, consists of Likert-type scales and open-ended questions on the panelists' satisfaction with the orientation, training, and process.

5. Results of the benchmarking

Round 1

The data analyst and lead facilitator produced summary tables and graphs from the first round, which showed the initial benchmarks, score ranges, and impact data for each Minimum Proficiency Level (see Table 10 and Table 11). During the workshop the tables and graphs shown to the panelist were based on the estimated score of leaners on the 25 items of the NAS (as described in chapter 2 on the pre-workshop analyses).

In the break-out room the panelists were presented with anonymous normative information on the panelists ratings (see Figure 7 and Figure 8). We saw that the ratings of panelists varied considerably, especially for the lowest benchmark (Partially meets).

Figure 7. Anonymous information on the panelists' ratings of Hindi Language Round 1
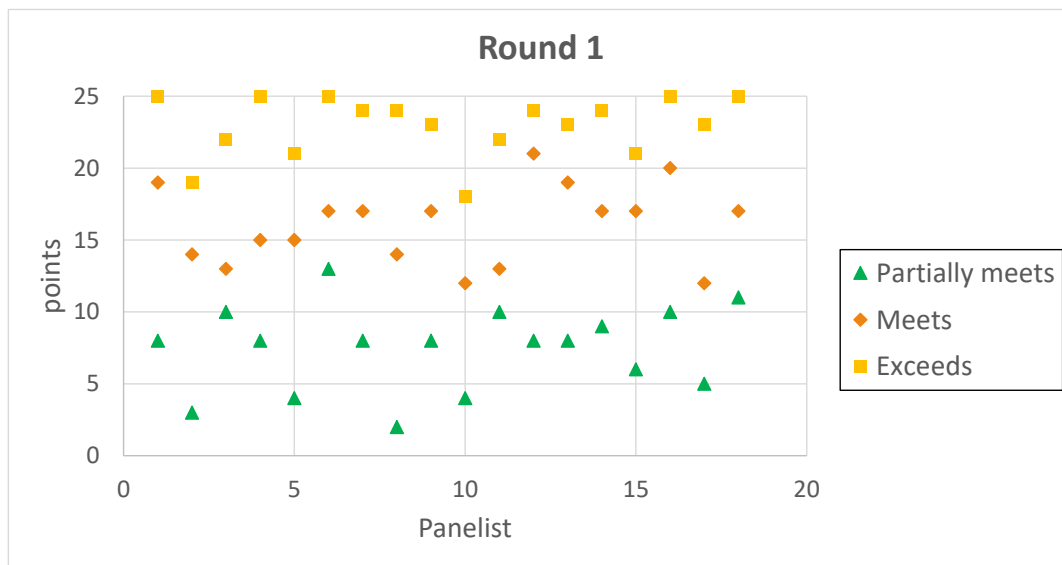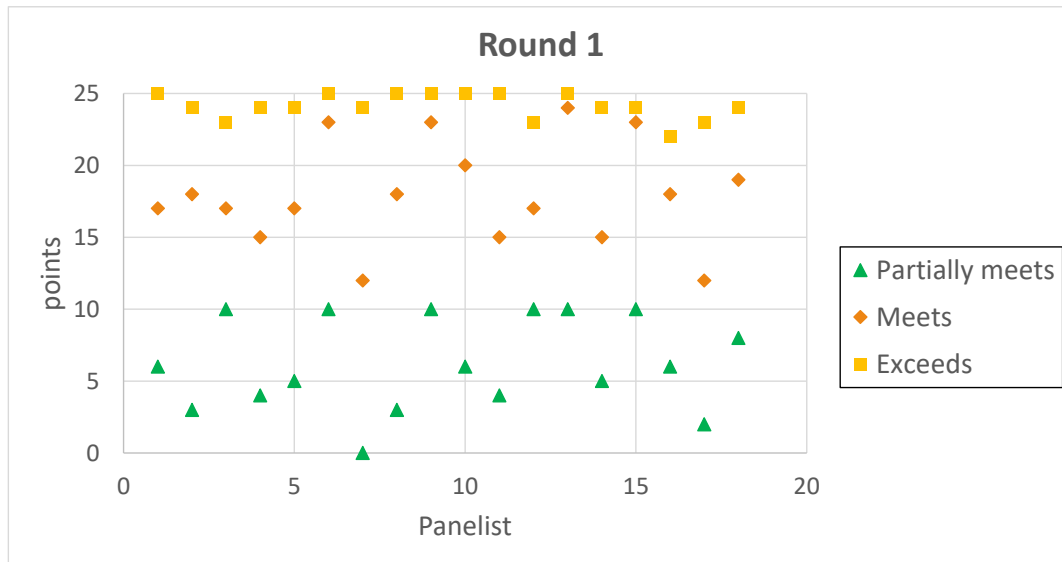
Figure 8. Anonymous information on the panelists' ratings of Mathematics Round 1

After round 1 the benchmark was calculated as the average of the panelists' benchmarks. The average benchmark was rounded down, as stipulated in the policy linking toolkit. For Hindi Language, the impact information shows 42.1% of the learners in the nine States would fall in the Partially Meets Proficiency level and almost the same percentage (42.2%) in the Meets level (see Table 10).

For Mathematics, the impact information shows that over sixty percent (61.6%) in the nine States would fall in the Partially Meets Proficiency level and 21.9% would fall in the Below Partially Meets level (see Table 11) using round 1 benchmarks.

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

Minimum Proficiency Levels	Round 1 Benchmark	Score Range	Percentage of Learners (using sampling weights[3])		
			Male	Female	Total
Below Partially Meets	N/A	0 - 6	9.2%	9.7%	9.3%
Partially Meets	7.56	7 - 15	42.2%	41.7%	42.1%
Meets	16.11	16 - 22	42.2%	42.1%	42.2%
Exceeds	23	23 - 25	6.5%	6.6%	6.4%

---

[3] The percentage of learners was estimated using the data that could be matched to the received sampling weights

Table 11. Round 1 benchmarks, score range and impact for Mathematics

Minimum Proficiency Levels	Round 1 Benchmark	Score Range			Percentage of Learners (using sampling weights)
			Male	Female	Total
Below Partially Meets	N/A	0 - 5	15.5%	15.6%	15.7%
Partially Meets	6.22	6 - 16	62.3%	61.1%	61.6%
Meets	17.94	17 - 23	21.4%	22.5%	21.9%
Exceeds	24.11	24 - 25	0.8%	0.9%	0.8%

Round 2

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

We see that in Round 2 the ratings of panelists varied less than in Round 1, especially for Mathematics (Figure 9 and Figure 10).

Figure 9. Anonymous information on the panelists' ratings of Hindi Language Round 2
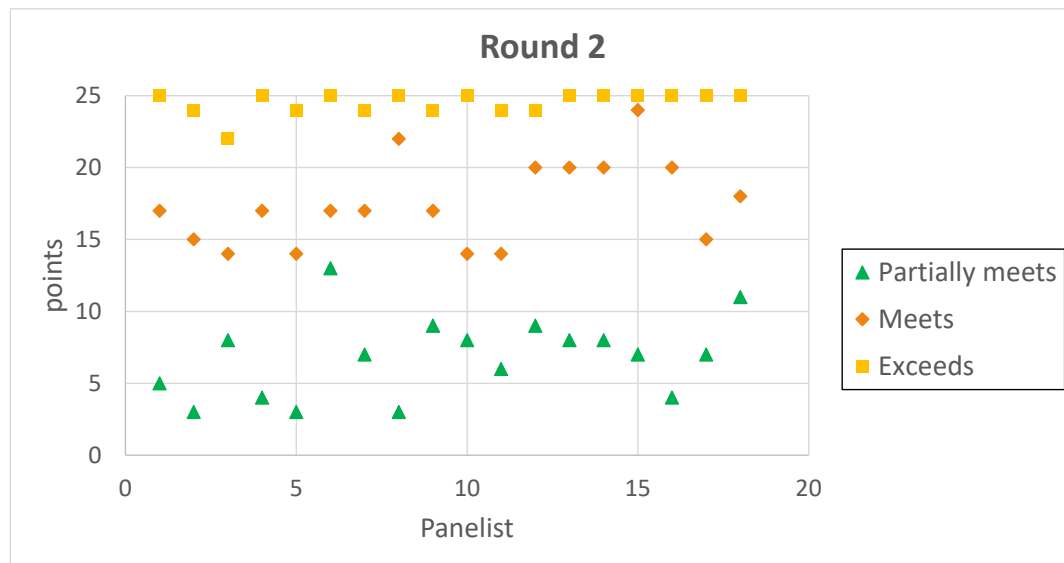
Figure 10. Anonymous information on the panelist's ratings of Mathematics Round 2
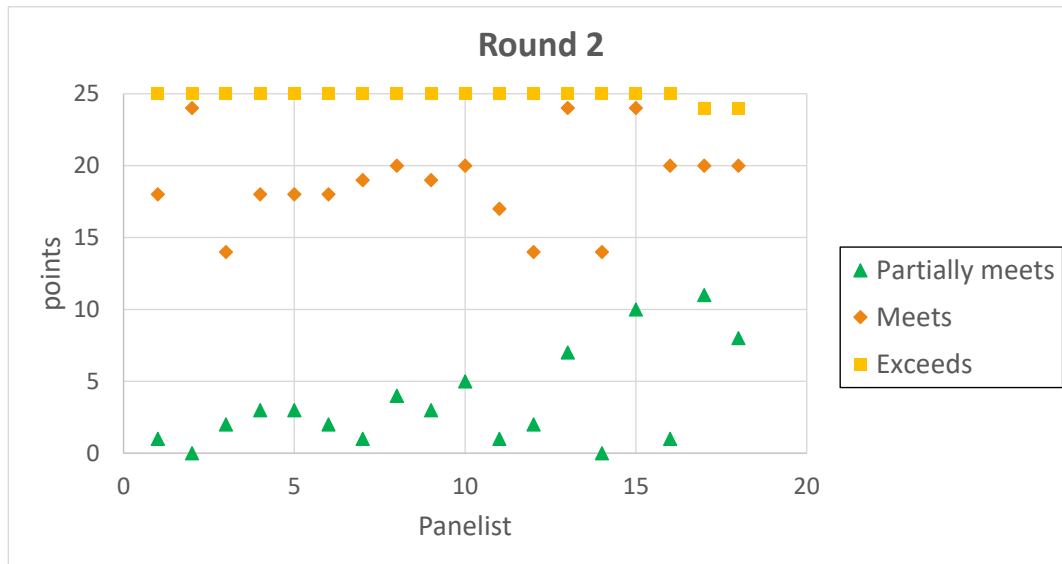
For Hindi language, the results show that in Round 2 less than 7% of the learners fall in the Below Partially Meets level (see Table 12). Half of the students (50.8%) fall in the Partially Meets level, 39.4% in the Meets level and only 3.0% in the Exceeds level. The Below Partially Meets benchmark was set lower in round 2 than in round 1 and the Meets and Exceeds benchmarks higher (see Table 13). As a consequence, after round 2 a higher percentage of learners falls in the Partially meets proficiency level. A lower percentage of a learners fall in the other levels than was the case in Round 1.

Table 12. Round 2 benchmarks, score range and impact for Hindi Language

Minimum Proficiency Levels	Round 2 Benchmark	Score Range			Percentage of Learners (using sampling weights)
			Male	Female	Total
Below Partially Meets	N/A	0 - 5	6.5%	6.9%	6.7%
Partially Meets	6.83	6 - 16	51.0%	50.8%	50.8%
Meets	17.56	17 - 23	39.5%	39.2%	39.4%
Exceeds	24.56	24 - 25	3.0%	3.1%	3.0%

Table 13. Comparison of Round 1 benchmarks and Round 2 benchmarks for Hindi Language

Minimum Proficiency Levels	Round 1 Benchmark	Round 1 Percentage of Learners (using sampling weights)	Round 2 Benchmark	Round 2 Percentage of Learners (using sampling weights)
Below Partially Meets	N/A	9.3%	N/A	6.7%
Partially Meets	7	42.1%	6	50.8%
Meets	16	42.2%	17	39.4%
Exceeds	23	6.4%	24	3.0%

For Mathematics, we see that in Round 2 the Partially Meets benchmark was set at a lower score and the Meets benchmark slightly higher (see Table 15). As a consequence, after round 2 a higher percentage of learners falls in the Partially Meets proficiency level. A lower percentage of a learners than was the case in Round 1 fall in the other levels. Only 3.5% of the learners fall in the Below Partially Meets level (Table 14). More than three quarters of the students (78.3%) fall in the Partially Meets level, 17.4% in the Meets level and less than 1 % in the Exceeds level.

Table 14. Round 2 benchmarks, score range and impact for Mathematics

Minimum Proficiency Levels	Round 2 Benchmark	Score Range	Round 2 Percentage of Learners (using sampling weights)		
			Male	Female	Total
Below Partially Meets	N/A	0 - 2	3.4%	3.6%	3.5%
Partially Meets	3.56	3 - 17	79.0%	77.6%	78.3%
Meets	18.94	18 - 23	16.9%	17.9%	17.4%
Exceeds	24.89	24 - 25	0.7%	0.9%	0.8%

Table 15. Comparison of Round 1 benchmarks and Round 2 benchmarks for Mathematics

Minimum Proficiency Levels	Round 1 Benchmark	Percentage of Learners	Round 2 Benchmark	Percentage of Learners
Below Partially Meets	N/A	15.7%	N/A	3.5%
Partially Meets	6	61.6%	3	78.3%
Meets	17	21.9%	18	17.4%
Exceeds	24	0.8%	24	0.8%

6. Evaluation of the Standard Setting Process

Internal Evaluation SEM, Panelist Consistency and Panelists' Agreement

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

As shown in Table 16, the Standard Error of Measurement (SEM), which measures how much panelists' benchmarks are spread around a "true" benchmark, was under 1.0 for both Language and Mathematics in both Rounds. For an assessment of 25 items like the NAS, the SEMs can be considered appropriate. The results show that the Standard Error of Measurement is smaller for the Exceeds benchmarks. This is a consequence of a ceiling effect for this benchmark. For both the Hindi language and Mathematics the Exceeds benchmark is located at score 24, while the maximum score is 25 (see previous section).

Table 16. Standard Error of Measurement by Round

	SEM by Benchmark					
	Round 1			Round 2		
Subjects	Partially Meets	Meets	Exceeds	Partially Meets	Meets	Exceeds
Language	0.70	0.63	0.52	0.75	0.78	0.13
Mathematics	0.79	0.87	0.22	0.81	0.75	0.08

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

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

Table 17. Inter-rater consistency and intra-rater consistency by subject and round

	Round 1		Round 2	
Subjects	Inter-Rater Consistency	Intra-Rater Consistency	Inter-Rater Consistency	Intra-Rater Consistency
Language	0.778	0.752	0.844	0.767
Mathematics	0.759	0.848	0.847	0.887

Procedural Evaluation

All panelists shared their opinion about the workshop through a questionnaire (see Appendix E). The panelists indicated on a five-point scale (Strongly Disagree-Disagree-Neutral-Agree-Strongly Agree) how strongly they agreed with several statements about six aspects of the workshop. On average, we see that the respondents were positive about the workshop. All six

aspects received an average score above 4 (on a scale of 1 to 5). The overall evaluation shows that the panelists are overall very positive: 3.46 on a scale of 1 to 4.

Table 18. Workshop evaluation results

Part of the workshop	Scale	Number of statements	Average	Standard deviation	N
The training on the Global Proficiency Framework	1-5	8	4.42	0.29	36
The training on the National Achievement Survey	1-5	5	4.23	0.34	36
The training on the alignment methodology	1-5	5	4.28	0.35	36
The training on the matching methodology	1-5	5	4.19	0.30	36
The training on the benchmark-setting (Angoff) methodology	1-5	11	4.32	0.37	36
Benchmark Round 2 evaluation	1-5	8	4.12	0.33	36
Overall evaluation	1-4	4	3.46	0.39	36

7. Conclusions and Recommendations

After India had participated in an in-person policy linking workshop in 2019, this year India participated in a blended policy linking workshop. Due to the travel restrictions of COVID-19, the international facilitators hosted the workshop using a videoconferencing platform (Teams). The participants met in small groups in 10 different locations. It was the first time such a blended policy linking workshop was held in which the panelists were attending in-person in their state and the international facilitators joined remotely. For many of the participants, this was the first time they participated in an international workshop and the first time using a videoconferencing platform.

After getting used to this mode the first day, the participants engaged in lively discussion regarding the alignment of the NAS 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 a blended mode can be considered a success.

The participants' work showed that the NAS for Hindi Language is strongly aligned to the Global Proficiency Framework for grade 8. Mathematics is in depth additionally aligned to the Global Proficiency Framework for grade 8. Furthermore, the panelists managed to reach almost complete consensus on the matching. The final benchmarks of the panelists show a good consistency, which makes the benchmarks useable for comparing, aggregating, and tracking learning outcomes for the NAS in the nine states in which Hindi is the main language.

Recommendations

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

Workshop Preparation

Collecting workshop materials and pre-workshop analyses

In the policy linking toolkit, the materials to be collected are clearly described. We would like to emphasize that the following materials should be obtained prior to the workshop:

- For facilitating the policy linking workshop, it is crucial that all the facilitators have been able to study the assessment prior to the workshop.
- Whenever a complex survey design has been used, the dataset including weights, ability estimates, item parameters should be provided prior to the workshop. To calculate the correct impact information (on population level) this information is indispensable. Without item parameters and sample weights, only impact at the level of the sample which answered a particular booklet/form can be presented.

Creating workshop materials

- A technical test should include all locations and preferably also include all participants. It allows us to establish whether all panelists have a head-set and a computer with the necessary software and settings (e.g. to enable the macros). A very short training can be given during the Technical test in turning on-and-off the audio and switching between break-out rooms. Such a short training will help to prevent technical issues during the workshop. A technical test with all locations and participants will also make clear in advance if back-up forms (e.g. without macros) are necessary and to troubleshoot any technology issues.

- A list of participants with their contact details should be available at least a week ahead of the workshop. The contact details and demographic information can be checked, and a panelist ID can be provided individually. This also would allow inviting panelists to a technical test and providing them with the Global Proficiency Framework prior to the workshop.
- It would be much easier for the panelists to familiarize themselves with the Global Proficiency Framework and to execute the tasks, if they received key documentation in the form of a hand-out translated in their own language, especially the Global Proficiency Framework.
- In Teams (and Zoom) the options are limited when participating without a license. Another platform might be considered (one for which the country has a license). Also, panelists need some instruction when participating for the first time. This instruction can be given during a Technical test or, for example, during the registration on the first day of the workshop.

Training the local content facilitators

- The local content facilitators might benefit from a more intensive training or general rehearsal. Conducting a rehearsal with the local content facilitators might help in raising the awareness of the goals of the workshop and of the tasks panelists must perform. The local content facilitators should also receive the (translated) Global Proficiency Framework well ahead of the workshop.
- In case countries have participated before in a policy linking workshop, during a rehearsal with the local content facilitators also clear instructions can be given on the changes in the policy linking toolkit to prevent conflicting instructions and expectations.

Implementing the blended workshop

To facilitate the sessions and discussions, it is essential that everything is translated (from English to the local language and vice versa). The presence of an interpreter (not the local content facilitator) should be planned for all sessions.
- A three-week workshop as is described in the policy linking toolkit is the preferred option. The schedule in the six-day remote 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 with the digital forms.
- In a remote workshop, more time is needed collecting, checking, merging, analyzing and reporting the results of the alignment and two Rounds of Item rating. To ease the process, Cito suggests that the collecting and checking of the forms is done locally.

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.
- In general, the panelists need more time to get acquainted with the Global Proficiency Framework and to get a good understanding of the framework, specifically GPD's and GPL's. The presence of the international content facilitators during this phase (also during the individual work) might help.
- The presentations could be more targeted to the panelists, e.g. the familiarization with the Global Proficiency Framework could be more focused on practicing than on presenting. The panelists seem to have difficulty with the many acronyms and technical words. A didactical approach can help in making the slides clearer and less word-based

aiming at more language independent information. A translation of the slides would help as well.

Tasks

- It is important that the Review Panel 4.1.1 is in place and has established that the assessment meets the standards required for policy linking. Especially the criteria for alignment need to be met before panelists must do the alignment in the workshop.
- It takes a lot of time to reach consensus during the Matching Task. We foresee serious problems with longer assessments. Within a six-day workshop, we estimate that about 30 items can be aligned, matched and rated.

8. References

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

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

NCERT (2020). *NAS 2017 Class III, V and VIII: National Report to inform Policy, Practices and Teaching Learning*. National Council of Educational Research and Training.

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

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

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

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

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

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

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

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

9. Annexes

Appendix A: Agenda for the blended 6-day workshop

India Policy Linking Workshop
For the Grade 8 Reading and Math National Achievement Survey
March 14 – 19, 2021

Overview

Day	Time	Activity
Sunday, March 14	9:30 – 17:00	Welcome, Introductions, Overview of the Workshop Purpose, Global Proficiency Framework (GPF), and the National Achievement Survey (NAS)
Monday, March 15	9:00 – 17:00	Review of the GPF, NAS Practice, Introduction to Task 1: Alignment, Practice Alignment
Tuesday, March 16	9:00 – 17:00	Complete Task 1, Review Task 1 Results and Discuss, Introduction to Task 2: Matching, Practice Matching
Wednesday, March 17	9:00 – 18:30	Complete Task 2, Review Task 2 Results and Discuss, Introduction to Task 3: Benchmarking, Practice Benchmarking, Facilitator-Panelist Consultations
Thursday, March 18	9:00 – 18:30	Complete Round 1 Benchmark Ratings, Review and Discuss Round 1 Benchmark Ratings, Review and Discuss Item Difficulty and Impact Data, Presentation on Task for Round 2, Facilitator-Panelist Consultations
Friday, March 19	9:00 – 17:00	Complete Round 2 Ratings, Present and Discuss Round 2 Ratings and Workshop Outcomes, Presentation of Certificates, Closing

India Policy Linking Workshop
For the Grade 8 Reading and Math National Achievement Survey
March 14 – 19, 2021

Sunday, March 14, 2021

Start	End	Activity	Facilitation
9:30 – 10:00		Registration	Project team
10:00 – 11:00		Welcome and introductions	NCERT, UNESCO
		Opening Remarks about the program: Prof Indrani Bhaduri, Educational Survey Division(ESD), NCERT	
		Background, objectives of the workshop: Silvia Montoya, Director, UNESCO Institute for Statistics (UIS)	
		Keynote speaker : Eric Falt, Director, UNESCO New Delhi Key messages for the workshop Dr. Sridhar Srivastava Director, NCERT	
11:00 – 11:20		Morning tea break	
11:20 – 12:00		Presentation: Overview of policy linking	Lead facilitator
12:00 – 12:30		Lunch break	
12:30 – 13:30		Presentation: Overview of the GPF	All facilitators
13:30 – 15:00		GPF Review	Content facilitators
15:00 – 15:20		Afternoon tea break	
15:20 – 16:00		Presentation: Overview of the assessment	All facilitators
16:00 – 16:40		Review each item on the NAS	Content facilitators
16:40 – 17:00		Explanation of individual work next day & closing	Lead facilitator

India Policy Linking Workshop
For the Grade 8 Reading and Math National Achievement Survey
March 14 – 19, 2021

Monday, March 15, 2021

Start		End	Activity	Facilitation
9:00	-	10:00	Individual work: Do the NAS	Local content facilitators
10:00	-	10:30	Individual work: Review GPF and identify any elements that are still unclear	Local content facilitators
10:30	-	10:50	Morning tea break	
10:50	-	12:00	Individual work: Review GPF and identify any elements that are still unclear (Continued)	Local content facilitators
12:30	-	13:00	Lunch break	
13:00	-	13:10	Welcome and purpose of session 2	Lead facilitator
13:10	-	13:40	Discussion of NAS activity	All facilitators
13:40	-	14:15	Review GPF activity and provide clarification	Content facilitators
14:15	-	15:00	Task 1 Presentation: GPF and alignment	Lead facilitator
15:00	-	15:20	Afternoon tea break	
15:20	-	16:00	Small group discussions on first 5 items	Content facilitators
16:00	-	16:40	Plenary discussion on questions that came up in the groups	Content facilitators
16:40	-	17:00	Explanation of individual work next day & closing	Lead facilitator

India Policy Linking Workshop
For the Grade 8 Reading and Math National Achievement Survey
March 14 – 19, 2021

Tuesday, March 16, 2021

Start	End	Activity	Facilitation
9:00 -	09:20	Individual work: Complete evaluation 1	Local content facilitators
9:20 -	10:30	Individual work: Task 1 Alignment of NAS and the GPF	Local content facilitators
10:30 -	10:50	Morning tea break	
10:50 -	12:30	Individual work: Task 1 Alignment of NAS and the GPF (cont.)	Local content facilitators
12:30 -	13:00	Lunch break	
13:00 -	13:10	Welcome and purpose of session 3	Lead facilitator
13:10 -	15:00	Task 2 Presentation: Matching NAS and GPDs/GPLs	Content facilitators
15:00 -	15:20	Afternoon tea break	
15:20 -	16:00	Task 1 Presentation: Alignment results	Lead facilitator
16:00 -	16:40	Task 2 Activity: Matching assessment items and GPDs/GPLs	Content facilitators
16:40 -	17:00	Explanation of individual work next day & closing	Lead facilitator

India Policy Linking Workshop
For the Grade 8 Reading and Math National Achievement Survey
March 14 – 19, 2021

Wednesday, March 17, 2021

Start		End	Activity	Facilitation
9:00	-	09:30	Individual work: Complete evaluation 2	Local content facilitators
9:30	-	10:30	Small groups complete Task 2 together	Local content facilitators
10:30	-	10:50	Morning tea break	
10:50	-	12:30	Welcome and plenary discussion: Matching NAS items and GPDs/GPLs and results of matching	Content facilitators
12:30	-	13:00	Lunch break	
13:00	-	13:10	Purpose of session 4	Lead facilitator
13:10	-	13:40	Task 3 Presentation: Global benchmarking	Lead facilitator
13:40	-	14:30	Task 3 Presentation: Angoff method	Lead facilitator
14:30	-	15:00	Task 3 Activity: Angoff practice	Content facilitators
15:00	-	15:20	Afternoon tea break	
15:20	-	16:00	Plenary discussion of questions that arose in small groups	Content facilitators
16:00	-	16:40	Task 3 Activity (individual): Angoff Round 1	Content facilitators
16:40	-	17:00	Explanation of individual work next day & closing	Lead facilitator
17:00	-	18:30	Consultation hour in which panelists of each state can consult the content facilitator	Content facilitators

India Policy Linking Workshop
For the Grade 8 Reading and Math National Achievement Survey
March 14 – 19, 2021

Thursday, March 18, 2021

Start	End	Activity	Facilitation
9:00 -	09:20	Individual work: Complete evaluation 3	Local content facilitators
9:20 -	10:30	Individual work: Complete Round 1 ratings on all remaining items	Local content facilitators
10:30 -	10:50	Morning tea break	
10:50 -	12:30	Individual work: Complete Round 1 ratings on all remaining items (continued)	Local content facilitators
12:30 -	13:00	Lunch break	
13:00 -	13:40	Welcome and purpose of session 5	Lead facilitator
13:40 -	15:00	Review and discuss Round 1 ratings in plenary	All facilitators
15:00 -	15:20	Afternoon tea break	
15:20 -	16:10	Review Round 1 ratings in small groups, going through each item where there was disagreement	Content facilitators
16:10 -	16:40	Share and discuss item difficulty and impact data	All facilitators
16:40 -	17:00	Explanation of individual work next day & closing	Lead facilitator
17:00 -	18:30	Consultation hour in which panelists of each state can consult the content facilitator	Content facilitators

India Policy Linking Workshop
For the Grade 8 Reading and Math National Achievement Survey
March 14 – 19, 2021

Friday, March 19, 2021

Start	End	Activity	Facilitation
9:00 - 09:20		Individual work: Complete evaluation 4	Local content facilitators
9:20 - 10:30		Individual work: complete Task 3 Activity Angoff Round 2	Local content facilitators
10:30 - 10:50		Morning tea break	
10:50 - 12:30		Individual work: complete Task 3 Activity Angoff Round 2 (continued)	Local content facilitators
12:30 - 13:00		Lunch break	
13:00 - 13:10		Welcome and purpose of session 6	Lead facilitator
13:10 - 15:00		Task 3 Presentation: Round 2 results	All facilitators
15:00 - 15:20		Afternoon tea break	
15:20 - 16:20		Discuss outcomes and final panelist questions	All facilitators
16:20 - 16:40		Complete evaluation 5	All facilitators
16:40 - 17:00		Closing and logistics	NCERT Director and UNESCO

38

Appendix B: Example of the forms

Figure 11. Alignment rating form – print format

Figure 12. Alignment rating form – digital format with drop-down menus

Figure 13. Matching form – print format

Panelist ID	

Question	Agreed alignment	Primary domain, construct and subconstruct	Fit
1	Domain		
	Construct reference		
	Subconstruct reference		
	Knowledge or skill		
	Difficulty		
	Lowest GPL		
2	Domain		
	Construct reference		
	Subconstruct reference		
	Knowledge or skill		
	Difficulty		
	Lowest GPL		
3	Domain		
	Construct reference		
	Subconstruct reference		
	Knowledge or skill		
	Difficulty		
	Lowest GPL		
4	Domain		
	Construct reference		
	Subconstruct reference		
	Knowledge or skill		
	Difficulty		
	Lowest GPL		

Figure 14. Item rating form – print format (with first four items rated)

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								

Appendix C: Alignment of the NAS items with the domains, constructs and subconstructs

Table 19. Language: Number of items aligned to each grade 8 domain, construct and subconstructs

Domain	Average nr. of items
D Decoding	0.0
R Reading comprehension	24.2
Total	24.2
Construct	
D1 Precision	0.0
D2 Fluency	0.0
R1 Retrieve information	10.1
R2 Interpret information	9.1
R3 Reflect on information	5.1
Total	24.2
Subconstruct	
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	1.9
R1.2 Retrieve explicit information in a grade-level text by direct- or close-word matching	7.8
R1.3 Retrieve explicit information in a grade-level text by synonymous matching	0.3
R2.1 Identify the meaning of unknown words and expressions in a grade-level text	2.6
R2.2 Make inferences in a grade-level text	5.1
R2.3 Identify the main and secondary ideas in a grade-level text	1.4
R3.1 Identify the purpose and audience of a text	1.8
R3.2 Evaluate a text with justification	2.2
R3.3 Evaluate the status of claims made in a text	0.9
R3.4 Evaluate the effectiveness of a text	0.2
Total	24.2

Table 20. Mathematics: Number of items aligned to each grade 8 domain, construct and subconstructs

Domain	Average nr. of items
N Number and operations	1.6
M Measurement	6.4
G Geometry	2.1
S Statistics and probability	4.9
A Algebra	2.6
Total	17.6
Construct	
N3 Decimals	1.3
N4 Integers	0.1

N5 Exponents and roots	0.3
M1 Length, weight, capacity, volume, area, and perimeter	5.4
M2 Time	1.0
G1 Properties of shapes and figures	1.8
G2 Spatial visualizations	0.3
G3 Position and direction	NA
S1 Data management	4.9
S2 Chance and probability	NA
A2 Expressions	2.0
A3 Relations and functions	0.6
Total	17.6
Subconstruct	
N3.2 Represent decimals in equivalent ways (including fractions and percentages)	1.0
N3.3 Solve operations using decimals	0.3
N3.4 Solve real-world problems involving decimals	NA
N4.2 Solve operations using integers	NA
N4.3 Solve real-world problems involving integers	0.1
N5.1 Identify and represent quantities using exponents and roots, and identify the relative magnitude	0.3
M1.1 Use non-standard and standard units to measure, compare, and order	0.9
M1.2 Solve problems involving measurement	4.4
M2.2 Solve problems involving time	1.0
G1.1 Recognize and describe shapes and figures	1.8
G2.1 Compose and decompose shapes and figures	0.3
G3.1 Describe the position and direction of objects in space	NA
S1.1 Retrieve and interpret data presented in displays	3.4
S1.2 Calculate and interpret central tendency	1.5
S2.1 Describe the likelihood of events in different ways	NA
A2.1 Evaluate, model, and compute with expressions	2.0
A3.1 Solve problems involving variation (ratio, proportion, and percentage)	0.2
A3.3 Solve equations and inequalities	0.4
Total	17.6

Appendix D. Difficulty Level of the Items

Table 21. Item parameters, p-value and Item-Total correlation of the NAS Hindi Language (NCERT, 2020, p. 129)

Form	Question	p-val	It-correlation	a-par	b-par
Form 1 & 2	Q1	0.76	0.38	0.74	-1.21
	Q2	0.45	0.32	0.51	0.25
	Q3	0.65	0.40	0.70	-0.68
	Q4	0.70	0.41	0.79	-0.86
	Q5	0.40	0.23	0.34	0.77
Form 1	Q1_6	0.54	0.38	0.59	-0.19
	Q1_7	0.49	0.45	0.79	0.01
	Q1_8	0.52	0.43	0.75	-0.10
	Q1_9	0.62	0.51	1.04	-0.42
	Q1_10	0.65	0.51	1.06	-0.52
	Q1_11	0.66	0.46	0.87	-0.63
	Q1_12	0.42	0.35	0.56	0.37
	Q1_13	0.68	0.48	0.95	-0.70
	Q1_14	0.75	0.49	1.11	-0.91
	Q1_15	0.69	0.42	0.81	-0.82
Form 2	Q2_6	0.70	0.46	0.96	-0.75
	Q2_7	0.31	0.15	0.23	2.10
	Q2_8	0.60	0.44	0.85	-0.41
	Q2_9	0.52	0.33	0.51	-0.10
	Q2_10	0.31	0.18	0.26	1.96
	Q2_11	0.40	0.21	0.32	0.78
	Q2_12	0.70	0.49	1.18	-0.71
	Q2_13	0.76	0.48	1.28	-0.91
	Q2_14	0.29	0.20	0.33	1.65
	Q2_15	0.52	0.31	0.51	-0.12

Table 22. Item parameters of the NAS Mathematics (NCERT, 2020, p. 130)

Form	Question	p-val	It-correlation	a-par	b-par
Form 1 & 2	Q1	0.40	0.34	0.56	0.49
	Q2	0.52	0.43	0.83	-0.07
	Q3	0.42	0.38	0.64	0.38
	Q4	0.59	0.33	0.56	-0.47
	Q5	0.43	0.27	0.41	0.46
Form 1	Q1_6	0.42	0.35	0.56	0.43
	Q1_7	0.41	0.40	0.67	0.42
	Q1_8	0.39	0.46	0.87	0.40
	Q1_9	0.35	0.41	0.72	0.67
	Q1_10	0.47	0.46	0.85	0.12
	Q1_11	0.28	0.17	0.26	2.28
	Q1_12	0.45	0.21	0.29	0.41
	Q1_13	0.53	0.40	0.71	-0.14
	Q1_14	0.32	0.25	0.38	1.27
	Q1_15	0.38	0.38	0.63	0.54
Form 2	Q2_6	0.44	0.46	0.88	0.20
	Q2_7	0.34	0.21	0.31	1.29
	Q2_8	0.38	0.19	0.28	1.06
	Q2_9	0.24	0.11	0.17	4.12
	Q2_10	0.40	0.44	0.81	0.38
	Q2_11	0.50	0.38	0.67	-0.01
	Q2_12	0.40	0.20	0.28	0.91
	Q2_13	0.39	0.38	0.66	0.47
	Q2_14	0.40	0.34	0.54	0.50
	Q2_15	0.50	0.44	0.87	-0.03

Appendix E. Questions and instructions in the Evaluation form of the workshop

EVALUATION OF THE WORKSHOP

We kindly ask you to share your opinion about the policy linking workshop. Please complete this short questionnaire inquiring about your experience. Your answers will be used to improve the workshop and the training. Your feedback will not be shared widely or reflect on your participation in the workshop.

TRAINING ON THE GLOBAL PROFICIENCY FRAMEWORK

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.

1. 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 8					

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
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					

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

TRAINING ON THE NAS

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.

5. 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					

6. Please list any questions you have about the assessment(s).
7. 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.

8. 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					

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

TRAINING ON MATCHING METHODOLOGY

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

11. 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					

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

TRAINING ON THE BENCHMARK-SETTING (ANGOFF) METHODOLOGY

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

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

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

16. Please describe the benchmarking methodology in your own terms.
17. Please list any questions or areas of confusion you have about the benchmarking methodology/process.
18. 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.

19. 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					
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					

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

OVERALL EVALUATION

21. How comfortable are you with your final performance predictions?
    a) Very uncomfortable
    b) Somewhat uncomfortable
    c) Fairly comfortable
    d) Very comfortable
22. If you marked either of the uncomfortable options, please explain why.
23. Overall, how would you rate the success of the policy linking workshop?
    a) Totally Successful
    b) Successful
    c) Unsuccessful
    d) Totally Unsuccessful
24. How would you rate the organization of the workshop?
    a) Totally Successful
    b) Successful
    c) Unsuccessful
    d) Totally Unsuccessful

25. 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.

Figure 15. Screen print of the evaluation form