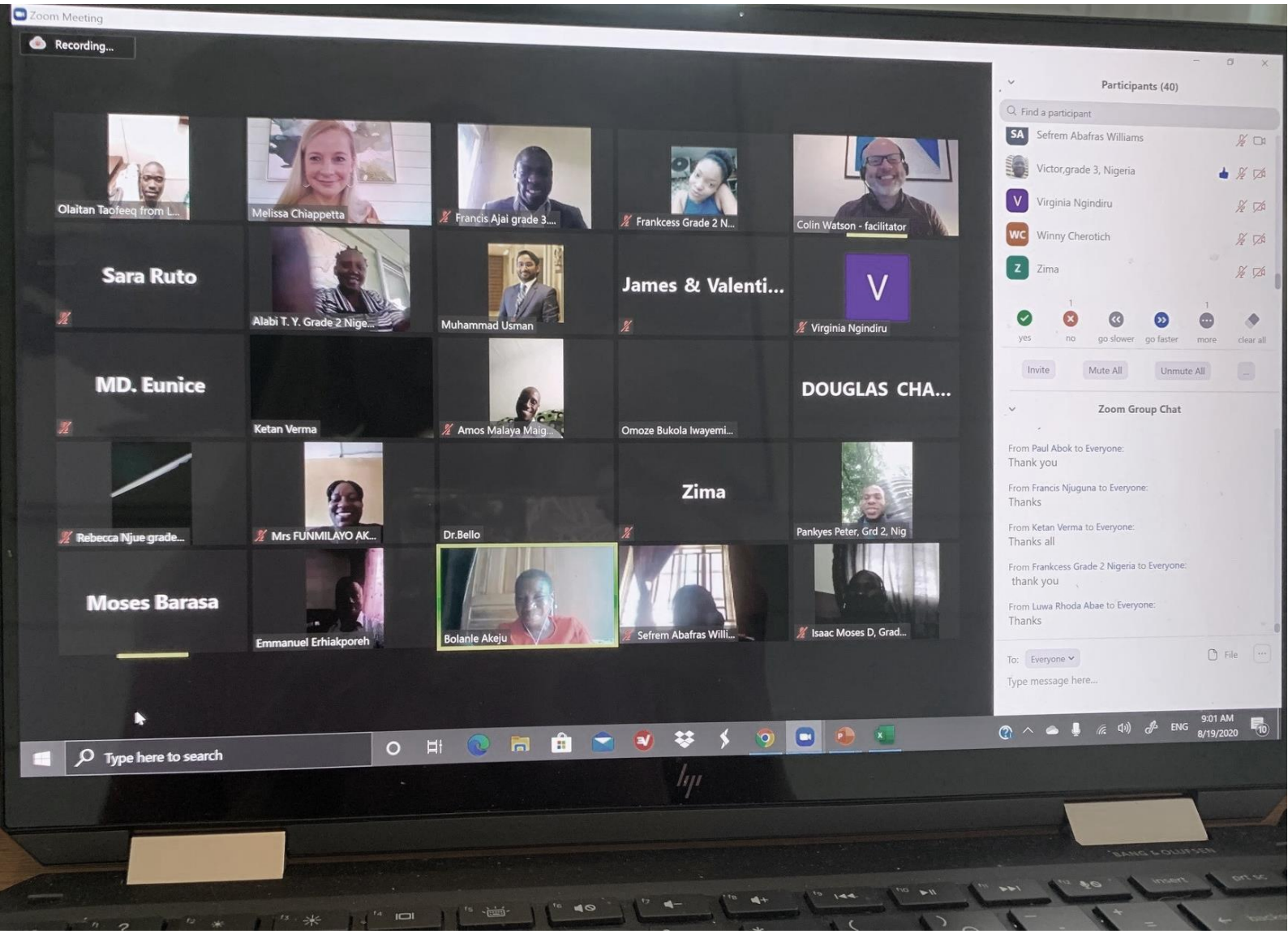


ICAN Policy Linking for Measuring Global Learning Outcomes Workshop Report

December 2020

CONTENTS

LIST OF FIGURES

LIST OF TABLES

BACKGROUND

Policy Linking Background

Policy Linking for Measuring Global Learning Outcomes ("policy linking" for short) is a methodology that allows countries, partners, and assessment organizations to link existing assessments (international, national, and sub-national) to the Global Proficiency Framework (GPF) and Sustainable Development Goal (SDG) 4.1.1:

> *"Proportion of children and young people in Grade 2 or 3 (4.1.1a), at the end of primary education (4.1.1b), and at the end of lower secondary education (4.1.1c) who achieve at least a minimum proficiency level in reading and mathematics."*

The GPF was developed by more than 60 global reading, language, and math content experts based on national content and assessment frameworks across more than 50 countries. The GPF provides performance expectations/standards for students in Grades 1-9 in reading and mathematics. By linking existing assessments to the GPF, countries, partners, and assessment organizations are able to compare learning outcomes across language groups and assessments in countries as well as across countries and over time, assuming all new assessments are subsequently linked to the GPF.

For each assessment, Policy Linking brings together 15-20 panelists, including master teachers and curriculum specialists, to: make judgements on the alignment of the assessment and the GPF, match assessment items to the relevant global proficiency descriptors (GPDs, sometimes called performance standards, which say how much a student needs to be able to demonstrate to prove they have met global minimum proficiency standards), and set benchmarks (also called cut scores) on the assessments that enumerate the score a student must achieve to meet global minimum proficiency standards.

International Common Assessment of Numeracy (ICAN) Background

The Citizen-led Assessment (CLA) model was born in India in 2005 when Pratham, one of India's largest NGOs, designed an innovative approach to assessing the foundational reading and numeracy abilities of all children, regardless of their schooling status. This assessment is the Annual Status of Education Report (ASER) in India. Over the past 15 years, the ASER tools and approach have been borrowed and adapted by many countries across the Global South. The People's Action for Learning (PAL) Network was formally established in 2015 as a South-South partnership of organizations across three continents engaged in CLAs of children's foundational reading and numeracy. These assessments offer a method for assessing learning outcomes that is grounded in the realities of the Global South. Designed and implemented by the PAL Network, ICAN is a simple-to-use tool that measures foundational numeracy. The PAL Network created the ICAN in an effort to provide countries with a cross-national instrument that allows them to measure learning outcomes for all children in their countries. It is conducted at the household level and as a result captures numeracy ability of children regardless of whether they are in school or not. The intention behind the development of the ICAN was to ensure comparability of Grade 2-3 learning outcomes across countries and provide an open-source mechanism for reporting those outcomes to SDG 4.1.1(a), thus filling a gap in global learning measurement not covered by most international assessments. In late 2019 and early 2020, PAL Network members conducted a large-scale household-based assessment using the ICAN tool in 13 countries across Africa, Asia, and the Americas. This first round of the assessment was restricted to one rural district[1] in each participating country to test the feasibility of implementing the assessment in a variety of geographies (PAL Network, 2020). While the ICAN was created to align with SDG 4.1.1 and the Minimum Proficiency Levels (MPLs) established by stakeholders at the UIS September 2018 meeting, as of May 2020, the PAL Network had not set benchmarks on the assessment to enumerate the score needed for children to meet global minimum proficiency.

[1] For ease of communication, 'District' in this report refers to a sub-state/regional/provincial unit, which is known by different names in different countries. For instance, this unit is called a Local Government Area in Nigeria, a sub-county in Kenya, and so on.

While ICAN was being developed and implemented, UNESCO Institute for Statistics (UIS) partnered with the United States Agency for International Development (USAID), with support from other organizations who became members of the Policy Linking Global Working Group (GWG),[2] including: the Australian Council for Educational Research (ACER); the Bill and Melinda Gates Foundation (BMGF); the United Kingdom's Foreign, Commonwealth, and Development Office (FCDO, formerly DFID); and the World Bank to develop the policy linking methodology for linking assessments to SDG 4.1.1. USAID and UIS developed the first version of the Policy Linking Toolkit in September 2019 and began piloting it thereafter, with pilots in Bangladesh, India, and Nigeria between October 2019 and March 2020. However, they had yet to pilot it with a CLA or with a cross-national assessment. CLAs are critical to ensuring countries are able to report reading and math levels to SDG 4.1.1, especially for out-of-school children in their countries.

As such, in talks in May 2020, at the suggestion of FCDO, the PAL Network agreed to pilot the policy linking methodology with the ICAN to check the intended alignment with the GPF (which further elaborates the MPLs) and to set benchmarks for the score children must obtain on the ICAN to meet global minimum proficiency. The BMGF and UIS provided one of the lead facilitators for the workshop: Melissa Chiappetta, an independent consultant working with them to advance the policy linking methodology. And, FCDO provided the other lead facilitator, Colin Watson from the UK Department of Education, who has also been working with the GWG.

Since ICAN was implemented in countries with such diverse contexts and languages of assessment, the lead facilitators suggested that it would be very difficult to host a policy linking workshop that brought together panelists from each of the 13 countries involved in the first round of ICAN implementation. As such, they worked with the PAL Network to identify two countries where both the assessment was administered in the same language and the panelists spoke the same language—Kenya and Nigeria.[3] The PAL Network then brought on its local member organizations in those two countries — Zizi Afrique Foundation and The Education Partnership Centre (TEP Centre), respectively, to help plan the workshops, liaise with local government officials, and invite the panelists from each country.

As shown in Table 1 below, ICAN includes 26 items. Each item on the assessment is worth one point. The final items (items 20-26) are only presented to children who have answered specific questions correctly from the 19 items presented to all children. For example, children are only asked to recognize numbers greater than 10 if they have correctly identified 4 out of 5 numbers less than 10.

Table 1. Number of Items and Score Points for ICAN

Items		Maximum points
Minimum number of items presented	Maximum number of items presented	
19	26	26

[2] UIS and USAID established the GWG in 2019 to ensure close coordination of partners working to advance similar efforts to measure outcomes for SDG 4.1.1.

[3] Note that while the ICAN was administered in English in both Kenya and Nigeria, it was also administered to some children in Kamba in Kenya.

PURPOSE

There were three main purposes of the ICAN Policy Linking Workshop:

1) To test whether the policy linking methodology would work with a cross-national CLA and out-of-school children;
2) To link the ICAN to the GPF;
3) To set two benchmarks on the ICAN – one for the minimum score 2nd-grade students and out-of-school children who are approximately 6-8 years old in Kenya and Nigeria should achieve to prove they have met global minimum proficiency in math and one for the score 3rd-grade students and out-of-school children who are approximately 7-9 years old should achieve.

The benchmarks should allow the governments of Kenya and Nigeria to report outcomes for students who have taken the ICAN to UIS for reporting against SDG 4.1.1, should they wish to do so, noting that currently, they would only be able to do so for one district each (in Mwala, Kenya and Ikorodu, Nigeria).[4] Governments and partners should also be able to use the benchmarks and outcomes to determine where or amongst which populations the gap toward achieving SDG 4.1.1 is the greatest so that they can focus resources in places where they will have the greatest impact.

The facilitators, donors, PAL Network, and partners hope to eventually use the benchmarks set in the workshop to establish the threshold that children who take the ICAN need to meet to demonstrate global minimum proficiency in line with SDG 4.1.1. However, given that the policy linking workshop was only conducted with a limited number of panelists from two of the thirteen countries where the ICAN had been implemented as of October 2020, it is not clear whether panelists from other countries with different languages of assessment would set the same benchmarks. As such, the facilitators recommend that before the benchmarks become the standard for the ICAN, the PAL Network conduct one or more additional workshops with other countries to determine whether the benchmarks set remain consistent. Should the benchmarks vary in the second workshop, additional workshops are recommended. Should they remain consistent, the PAL Network might consider one final workshop to ensure the rigor and defensibility of the findings.

[4] Note that the PAL Network plans to scale the ICAN from one to three districts in each of the participating country in 2021-22.

OVERVIEW

The policy linking process includes seven stages, as follows:

1) Initial engagement
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 (if approval received from UIS following Stage 3 to proceed)
5) Implementation of policy linking workshop and documentation of outcomes
6) Review of workshop outcomes by 4.1.1 Review Panel
7) Reporting results for SDG 4.1.1

In summary, the initial engagement stage was productive and led to the workshop and to the decision to focus efforts on linking the ICAN for Kenyan and Nigerian outcomes for Grades 2 and 3 and students who fall in the typical age range for those grades. In Stage 2, the PAL Network submitted documents and background on the ICAN to the lead facilitators since UIS is still working to form the 4.1.1 Review Panel. Note, the lead facilitators were also the authors of the Criteria for Policy Linking Validity, which is the set of criteria that the 4.1.1 Review Panel will use to determine whether an assessment is suitable for policy linking for global learning outcomes. In Stage 3, although the 4.1.1 Review Panel is not yet in place, the lead facilitators reviewed the documents and data to determine if the suitability of the ICAN assessment for policy linking. Ultimately, we found that the assessment met the criteria to be used for policy linking for measuring global learning outcomes. In Stage 4, the lead facilitators worked with the PAL Network, TEP Centre, and Zizi Afrique Foundation to identify four local content facilitators, two from each country, to help lead the workshop and 30 panelists, 15 for each grade with 7-8 of those 15 being from each country, to participate, as shown in Table 2 below. The key stakeholders also made the determination that the workshop would be held remotely as a result of the COVID-19 pandemic, and they made arrangements to ensure panelists had the internet access and materials they would need to participate. Finally, the facilitators also led a two-day content facilitator training. In Stage 5, the facilitators led eight, 2-hour workshop sessions over the period of 23 days. Panelists aligned the ICAN to the GPF and set benchmarks of 17 and 21 for the score children need to achieve to "meet global minimum proficiency" in Grades 2 and 3, respectively. Outcomes that measure the consistency and validity of results met requirements for the Criteria for Policy Linking Validity SDG 4.1.1 for Grade 3 and came close to meeting requirements for Grade 2, with the one exception being the inter-rater consistency, which fell just short of the .80 requirement at .76. More details about this follow.

As mentioned, Stage 6 is not yet possible, as UIS is still working to form the 4.1.1 Review Panel. Stage 7 is dependent on the results of this report, which will inform the PAL Network's next steps and possible conversations with the governments of Kenya and Nigeria.

More detailed results from each of the first five stages are presented in the Results Section below.

Table 2. Content Facilitators and Panelists by Country

Country	Grade 2 Content Facilitators	Grade 2 Panelists	Grade 3 Content Facilitators	Grade 3 Panelists
Kenya	1	8	1	7
Nigeria	1	7	1	8

RESULTS

Stage 1: Initial Engagement

As mentioned above, the idea for the ICAN Policy Linking Workshop first came forward as a way to begin to link assessments that measure outcomes for out-of-school children to SDG 4.1.1 and to pilot test whether policy linking would work for assessments with out-of-school children. Currently, there is limited to no reporting on reading and math outcomes for out-of-school children for SDG 4.1.1. Though the ICAN did not assess many out-of-school children in Mwala or Ikorodu (since most children in these areas are in school) during the 2019 administration of the ICAN, we still hoped to be able to determine if policy linking would be possible for a CLA that also assesses out-of-school children in many countries.

During initial engagement, the PAL Network, UIS, and FCDO agreed that it would be difficult to host a workshop with panelists from all 13 countries where the ICAN has been implemented to date (See **Table 3** for a list of countries, districts, and languages in which the ICAN was conducted in 2019-2020). As such, the group decided to focus in on two countries where the language of the assessment and the lead facilitators matched up, which is how the PAL Network narrowed in on Kenya and Nigeria with English as the language of assessment. It is important to note that while some children in Kenya took the ICAN in English, others completed the assessment in Kamba. The group also made the decision that one policy linking workshop would not be sufficient to link results from all countries where the ICAN was implemented, but that at least one-two additional workshops would be needed to identify if results differed at all by context/language of assessment. As such, the benchmarks presented in this report should only be used to interpret ICAN outcomes for Kenya and Nigeria.[5] The group also made the decision that rather than limiting panelists to the two rural districts where the ICAN was implemented in 2019 in Kenya and Nigeria, they would seek to engage panelists from across both countries with the hopes that the benchmarks set by panelists would be useful nationwide for both countries should the PAL Network and/or the governments of Kenya and Nigeria decide to scale up the assessments.

Table 3. ICAN Districts and Languages of Assessment 2019-2020

Region	Sampled district (Country)	ICAN assessment tool language
Eastern and Southern Africa	Arusha Rural (Tanzania)	Kiswahili
	Larde (Mozambique)	Portuguese
	Mubende (Uganda)	English
	Mwala (Kenya)	Kamba, English
Western Africa	Ikorodu (Nigeria)	English
	Segou (Mali)	French
	Tivaouane (Senegal)	Wolof, French
America	Matagalpa (Nicaragua)	Spanish
	Xalapa Rural (Mexico)	Spanish
South Asia	Betul (India)	Hindi
	Jhenaidah (Bangladesh)	Bangla
	Makwanpur (Nepal)	Nepali
	Toba Tek Singh (Pakistan)	Urdu

Stages 2 and 3: Collation of Evidence of Curriculum and Assessment Validity and Alignment

In Stage 2, the PAL Network, TEP Centre, and Zizi Afrique Foundation gathered evidence and submitted the following to the lead facilitators (given that the 4.1.1 Review Panel has not yet been established):

- The assessment and assessment framework

[5] Note, currently, outcomes only exist for one rural district in each country. However, these benchmarks can also be used for future administrations of the ICAN that may be rolled out more broadly in Kenya and Nigeria. It is important to note, however, that the workshop panel was not representative of teachers in Nigeria or Kenya; thus, it may be prudent for the countries to run a second workshop that is more representative of teachers in the country before using the benchmarks broadly, especially beyond the areas from which panelists participated (3 out of 47 counties in Kenya and 3 out of 6 geopolitical zones in Nigeria).

- The data from the 2019-2020 implementation of the ICAN in Kenya and Nigeria
- The 2019-2020 ICAN Sampling Design

The lead facilitators used the documents and data to determine the acceptability of the ICAN for policy linking and for reporting results to SDG 4.1.1, according to the Criteria for Policy Linking Validity document. While, ultimately, the 4.1.1. Review Panel will grade assessments on a three-point scale of excellent, good, or sufficient for policy linking, since the panel is not yet established, the lead facilitators only considered the minimum criteria, as shown in **Table 4** below. We found that the ICAN met the criteria to be classified as "sufficient" for policy linking, meaning it met each of the below criteria. Even though inter-rater consistency was not checked, raters were quizzed during the training and monitored in the field to make sure they were scoring correctly, and there was only one correct response per question. So, there was not much room for variability in scoring. The lead facilitators found this sufficient to proceed with the policy linking process.

Table 4. Pre-Workshop Criteria of Assessment Acceptability for Reporting on SDG 4.1.1

Criteria	Details	Results
1d) Is the assessment aligned with the *GPF*?	The process for conducting the *alignment study* between an assessment and the *GPF* is set out in the policy linking toolkit. It involves experts reviewing each assessment item and determining whether it aligns (or partially aligns) with any of the knowledge and skills listed in the GPF and then summarizing these results to the subconstruct level for the relevant grade. Once all items have been considered, a decision is made on whether sufficient subconstructs have been covered to agree there is alignment (See **Table 5** for more information on minimum alignment criteria).	Yes, the lead facilitators found that the assessment was Additionally Aligned (as described in **Table 5** below. The ICAN has 26 items, and the lead facilitators determined that all were aligned to the GPF with 14 items linked to the Number domain of the GPF, 10 to the Geometry and Measurement domains combined, and more than 50 percent of the subconstructs covered for those domains in Grades 2 and 3. While the ICAN was developed to include assessment items from Grades 1-4, in the alignment process, the facilitators determined that there were not a sufficient number of items that aligned with the Grade 1 or Grade 4 GPDs from the GPF to set benchmarks for those grades. Further, while the GPF includes four performance levels – below partially meets minimum proficiency, partially meets minimum proficiency, meets minimum proficiency, and exceeds minimum proficiency – the facilitators also found that there were not enough items aligned with each performance level to set so many benchmarks.
2a) Is there evidence that the items in the assessment have been reviewed qualitatively and/or quantitatively to determine their validity?	The assessment should be assessing what it was intended to assess. For instance, a reading comprehension question should not be measuring memory or student understanding of science concepts, such as names of various types of birds. Where data is available, it should be analyzed using either classical test theory or item response theory to investigate how well items performed (e.g. facility or difficulty--percent correct--and discrimination--correlation between item score and total score).	Yes, there is evidence of this, as shown in **Table 6**. The lead facilitators found that item difficulty ranged between 27 percent for the last item (Item 26) and 91 percent for Item 1, meaning 27 percent of children assessed got Item 26 correct, and 91 percent got Item 1 correct. Using the University of Washington's scale of 50 percent or below getting an item correct as being a difficult item, 51-84 percent being moderate items, and 85 percent and above being easy items, this means the ICAN has 6 easy items, 13 moderately difficult items, and 7 difficult items.[6] In terms of item discrimination, lead facilitators calculated two statistics—item discrimination, which

[6] University of Washington. (2020). *Understanding Item Analyses.* University of Washington. https://www.washington.edu/assessment/scanning-scoring/scoring/reports/item-analysis/#:~:text=For%20items%20with%20one%20correct,value%2C%20the%20easier%20the%20question.

Criteria	Details	Results
		indicates the extent to which success on an item corresponds to success on the whole test and the Point-biserial Correlation, which is the Pearson correlation between responses to a particular item and scores on the total test. The results for item discrimination ranged from .23 to .96 for the ICAN. High item discrimination statistics mean that the item is doing a good job of discriminating high-performing children from low-performing children. Many of the interpretation tables suggest that anything above .30 is good, and anything from .1 to .3 is fair, and anything below .1 is poor. The lead facilitators found that 2 items could be classified as "fair" and 24 items as "good," according to this criterion.[7] Like item discrimination statistics, values for point-biserial range from -1.00 to 1.00. Values of 0.15 or higher mean that the item is performing well (Varma, 2006). Facilitators found that the point-biserial statistics for the ICAN were all at .61 or above. In the development phase, ICAN assessments items were field trialed in all the countries of implementation. Administration instructions and item stimulus was modified based on the feedback from field trials. The assessment items were also reviewed by math content experts across the PAL Network to ensure construct, content, and face validity.
2c) Is the cohort that took the assessment representative of the population against which results will be reported?	The assessment should either be census or sample based. If it is sample-based, information should be provided on how the sample was developed. For example, if it is a stratified random sample, countries or assessment organizations should provide details of the strata (which should at least include district or other large administrative units) and any checks they have made on the representativeness of the sample. Where a sample-based approach is used, the margin of error should be 5 percent or less at the 95 percent confidence level.[8]	Yes, the sample is representative of the sampled districts at the 95 percent level. When reporting their results, the PAL Network is clear that the results only relate to the districts in which the assessment was implemented and are not intended to be representative of the countries as a whole.
3a) Is the value of *coefficient alpha*[9] for the grade-level, subject assessment	The *coefficient alpha* for the subject-specific assessment should be greater than or equal to 0.7. Countries may have also calculated values of *coefficient alpha* for individual components of the	The coefficient alpha is a psychometric test of reliability, or internal consistency, between items on an assessment. Coefficient alpha measures whether the items on the assessment seek to measure the same latent variable, which in the case of the ICAN would be

[7] Ibid.

[8] It is accepted that for some countries, defining what 'nationally representative' means may be difficult given a lack of accurate sampling frame. In such cases, governments should make clear how they have attempted to achieve an appropriate sample and identify any known limitations with their approach.

[9] Also known as *Cronbach's alpha*

Criteria	Details	Results
greater than or equal to 0.7?	assessment. These may also be provided, but this criterion will be judged on the value for the entire assessment.	math ability, for the ICAN is .94 overall (and .91 for Mwala and .93 for Ikorodu).
3b(ii) For paper-and-pencil assessments that contain *constructed-response items* and/or oral assessments with *selected-response* and/or *constructed-response items*, how have those responsible for scoring been quality assured to ensure consistency of scoring (*inter-rater consistency*)?	For paper-and-pencil assessments with *constructed-response items* and/or oral assessments with any type of performance-based items, enumerators or those who will score the assessment must achieve an *inter-rater consistency* (IRC) score of .80 or higher using Cohen's Kappa or equivalent statistic. For a country to achieve an excellent or good rating, they should examine IRR for a sample of students assessed in the field for oral assessments or a sample of scored items following a paper-and-pencil assessment. A country may achieve a sufficient rating if they have examined IRC but only during enumerator/rater training.	The IRC was not assessed for enumerators since there was only one right answer per question, and enumerators were quizzed during the training to make sure they understood the data collection process. ICAN's assessment processes are aligned to PAL Network's Data Quality Standards Framework. Specifically, for the enumerators, PAL member organizations in participating countries work with local district level institutions to recruit enumerators. All the enumerators attended a 3-day district-level training. The training had a classroom as well as field component. The PAL Network worked with around 750 enumerators across all the 13 countries. All enumerators were given a quiz to assess their understanding of the data collection process and were evaluated based on their field performance. In addition, there was field monitoring and a field-based back check (recheck) process where District Coordinators and PAL member organization staff assured the quality of assessment. Overall, around 80 percent of surveyed rural communities were either field monitored or field rechecked (after the assessment was completed) or both.

Table 5. Criteria for Level of Assessment Alignment to the GPF for Mathematics Policy Linking Workshops

Level of Alignment	Category	Criteria
Minimally aligned	Domain (depth):	Number (min 5 items)
	Subconstructs (breadth):	Items covering at least 50% of the Number subconstructs
Additionally aligned	Domain (depth):	Number (min 5 items) and Measurement and Geometry (min 5 items)
	Subconstructs (breadth):	Items covering at least 50% of the Number, Measurement, and Geometry subconstructs
Strongly aligned	Domain (depth):	Number (min 5 items) and Measurement and Geometry (min 5 items) and Statistics & Probability and Algebra (min 5 items)
	Subconstructs (breadth):	Items covering at least 50% of all subconstructs

Table 6. Item Discrimination and Difficulty

Item Number	Item Discrimination	Point Biserial Correlation	Item Difficulty Overall	Mwala (Kenya) Grade 2 Item Difficulty	Ikorodu (Nigeria) Grade 2 Item Difficulty	Mwala (Kenya) Grade 3 Item Difficulty	Ikorodu (Nigeria) Grade 3 Item Difficulty
1	0.23	0.63	91%	96%	88%	97%	91%
2	0.59	0.66	70%	77%	44%	93%	52%
3	0.31	0.72	89%	97%	76%	95%	91%
4	0.37	0.61	82%	90%	68%	93%	74%
5	0.36	0.61	83%	83%	83%	79%	76%
6	0.73	0.69	42%	26%	30%	33%	28%
7	0.66	0.64	56%	47%	49%	44%	61%
8	0.77	0.74	39%	9%	24%	14%	37%
9	0.88	0.86	55%	34%	25%	44%	43%
10	0.83	0.80	51%	16%	26%	29%	48%
11	0.36	0.74	87%	93%	77%	94%	82%
12	0.32	0.62	86%	91%	82%	91%	88%
13	0.32	0.73	89%	96%	81%	95%	88%
14	0.60	0.79	76%	80%	58%	88%	68%
15	0.26	0.69	90%	94%	90%	94%	96%
16	0.71	0.94	78%	88%	68%	91%	79%
17	0.82	0.91	71%	71%	51%	85%	72%
18	0.79	0.93	73%	53%	64%	60%	74%
19	0.90	0.93	62%	44%	43%	59%	54%
20	0.50	0.80	82%	89%	78%	93%	88%
21	0.92	0.92	60%	56%	36%	70%	48%
22	0.92	0.88	47%	16%	20%	44%	35%
23	0.96	0.91	44%	10%	17%	17%	23%
24	0.75	0.80	28%	9%	6%	15%	13%
25	0.92	0.88	43%	15%	13%	42%	25%
26	0.76	0.80	27%	11%	7%	12%	9%

Stage 4: Preparation for the Policy Linking Workshop

In preparing for the policy linking workshop, the key stakeholders considered the number of panelists to select per country. The Policy Linking Toolkit suggests 15-20 panelists per assessment/grade. Given that the workshop goal was to set benchmarks for grades 2 and 3, we needed at least 30 panelists. However, as mentioned above, the eventual hope is that the ICAN can be used to link results from many countries to SDG 4.1.1. To make this a possibility, we knew we needed to make sure that benchmarks were representative of all countries using the ICAN. The first step toward doing so would be to see if two countries with the same language of assessment set the same or similar benchmarks. As such, we considered doubling the number of panelists so that we could compare outcomes between the two countries. Unfortunately, though, COVID-19 presented a significant barrier to this plan, as it became difficult to identify a sufficient number of panelists with adequate internet access. The remote nature of the workshop also means that the panelists were necessarily less representative than we would hope.

Another consideration in selecting panelists was deciding how we would ensure panelists were also able to represent the interests of out-of-school children in addition to the in-school population. The Policy Linking Toolkit recommends engaging master teachers and curriculum specialists as panelists, but it does not address what profile of panelists might best represent out-of-school children. However, given the importance of panelists understanding the performance standards represented in the GPF and the knowledge and/or skills

required to answer each of the assessment items on the ICAN, we felt that teachers were likely still best suited and also that they would be able to conceptualize out-of-school children from their communities as easily as other community members. Also, given the high rates of enrollment in the sampled districts in Kenya and Nigeria, the 2019 administration of the ICAN did not include many out-of-school students. When possible, future workshops geared at linking assessments of out-of-school students to SDG 4.1.1 might consider volunteer teachers and informal education teachers as panelists in addition to traditional teachers. More information on panelists' background is included in **Finally,** we prepared materials and sent them to panelists so they could print them ahead of the workshop and created a WhatsApp group to maintain close communication with panelists throughout the workshop.

Table 7 below. Further information on the panelists' background is provided in response to the requirements of the Criteria for Policy Linking Validity provided in **Tables 17 and 18.**

With the support of TEP Centre and Zizi Afrique Foundation, we also engaged four content facilitators, all members of the governments of Kenya and Nigeria, from the Kenyan Institute of Curriculum Development, Kano State Education Resource Department, and Lagos State Curriculum Service Department. Lead facilitators led two four-hour remote sessions with the content facilitators ahead of the workshop to train them on the policy linking methodology and its three main tasks as well as on the GPF and ICAN. Finally, we prepared materials and sent them to panelists so they could print them ahead of the workshop and created a WhatsApp group to maintain close communication with panelists throughout the workshop.

Table 7. Panelist Demographic Information

Characteristic	Grade 2 (n=15)	Grade 3 (n=15)
Gender		
Female	11	4
Male	4	11
Level of education:[10]		
Some college	7 (47%)	6 (40%)
Completed 4-year college	4 (27%)	3 (20%)
Some Master's education	3 (20%)	5 (33%)
Completed Master's education	0 (0%)	1 (7%)
No information	1 (7%)	0 (0%)
Years of experience:		
Average number of years teaching	11	11
Average number of years teaching at the relevant grade level	3	5
Experience teaching the following:		
Private school	3	2
Students with disabilities	5	2
Students affected by crisis or conflict	4	7
Geographic region/county:		
Bungoma, Kenya	2	2
Tana River, Kenya	4	3
Turkana, Kenya	2	2
North Central, Nigeria	2	4
Northeast, Nigeria	1	1
Southwest, Nigeria (region in which Ikorodu local government area is located, where ICAN was administered)[11]	4	3

[10] Percentages do not sum to 100 due to rounding.

[11] Note that no teachers from Machakos County, where ICAN was administered in 2019, were engaged in the policy linking workshop, given accessibility concerns (the workshop had to be held remotely due to the COVID-19 pandemic).

Panelist Representativeness

Given the constraints of the remote workshop, as mentioned above, the panelists were not as representative as we would have liked. Details about the representativeness of panelists follows:

- Kenya: The panelists were from three counties from the Western, Coastal, and Northern regions of Kenya. There are a total of 47 counties in the country. Nationally, the gender ratio for primary school teachers is 51 percent female and 49 percent male, and Kenyan panelists were 53 percent female and 47 percent male. Private school teachers make up 24 percent of the teaching workforce in Kenya, and they made up 13 percent of the panel.
- Nigeria: The panelists were from three geo-political zones - there are six such zones in Nigeria. On average there are 55 percent female teachers and 45 percent male teachers in the country, and Nigerian panelists were 47 percent female and 53 percent male. Private school teachers make up 28 percent of the teaching workforce in Nigeria, and they made up 20 percent of the panel.

Stage 5: Implementation of Policy Linking Workshop and Documentation of Outcomes

The ICAN Policy Linking Workshop took place over the course of about three weeks, as shown in the agenda in Annex A: ICAN Policy Linking Workshop Agenda, and was presented via Zoom teleconference as a result of the COVID-19 pandemic. Following feedback from other policy linking workshops, facilitators hosted two preparation sessions focused on familiarizing panelists with the GPF and the ICAN tool ahead of the workshop as well as giving panelists time to use the ICAN tool to assess children in their communities and/or students from their classes ahead of the workshop. In these sessions, we worked to ensure all panelists understood the terms and intent of each of the GPDs (also called performance standards) from the GPF since these form the basis for the link between assessments and SDG 4.1.1. We also led the panelists through a training on how to implement the ICAN so that they could select three to five children from their communities/students from their classes who they knew just barely met the requirements of "meeting global minimum proficiency" for their grade/age group, according to the GPF. We then gave the panelists nearly two weeks to identify any questions they might have about the GPF as well as to assess students (if it was safe to do so and followed social distancing requirements in their communities) and record the results.

During the six regular workshop sessions, facilitators followed the first draft of the Policy Linking Toolkit, engaging panelists in three tasks –

1) Task 1 – Panelists made independent and individual judgements on the alignment of each ICAN item to the knowledge and/or skill(s) needed to correctly answer the item
2) Task 2 – Panelists made independent and individual judgements on the match between each ICAN item to the lowest GPD needed to correctly answer the item[12]
3) Task 3 – Set one benchmark for meeting global minimum proficiency by rating (through two rounds) whether a child who just barely meets the requirements of the "meets global minimum proficiency level" as described by the GPDs in the GPF would correctly answer each ICAN item

The results of each of these tasks follows in Table 8 through Table 11.

Task 1 – Alignment Results

The first task in the policy linking workshop asks panelists to align each item from the GPF to one or more of the statements of knowledge and/or skill(s) in the GPF. Given that the GPF was newly revised ahead of the ICAN Policy Linking Workshop, and the first version of the Policy Linking Toolkit did not mention the need to align to knowledge and/or skills but rather to subconstructs, in this task, ICAN Policy Linking Workshop panelists worked to align ICAN items to one or more GPF subconstructs (rather than statements of knowledge

[12] Note that the newer version of the Policy Linking Toolkit, which was released after the ICAN workshop, makes it clear that Task 2-matching should be a group activity where panelists work toward consensus. Facilitators believe this step would have helped to improve some of the results from the ICAN Policy Linking Workshop.

and/or skill(s), which is the level of alignment required in the toolkit version to be released in December 2020). **Table 8** below shows the results of the panelist alignment activity. The percentages provided are based on the modal response of panelists during the alignment exercise undertaken during the workshop. It should be noted that, at the time of the workshop, the statements of knowledge and/or skills that are included in the current version of the GPF were not available to panelists. This could account for less precise alignment among panelists, given that the subconstructs do not contain as much information as the statements of knowledge and/or skill(s). The lead facilitators felt that the ICAN had greater coverage of subconstructs than is indicated in the table. In either case, the coverage of domains, constructs, and subconstructs was sufficient for the assessment to be classified as "additionally aligned" (See **Table 5** for details) since it contained five number items; five measurement and/or geometry items; and at least 50 percent of the number, measurement, and geometry subconstructs for Grade 2 or 3, respectively.

Table 8. Panelist Alignment of Items with Domains, Constructs, and Subconstructs

Alignment (percentages)			
Items to GPF	GPF Domains Covered	GPF Constructs Covered	GPF Subconstructs Covered
100%	80%	60% (Grade 2) 56% (Grade 3)	57% (Grade 2) 53% (Grade 3)

Task 2 – Matching Results

The second policy linking workshop task is matching, where panelists take Task 1 one step further to align to the GPDs and GPLs that describe how much children should demonstrate to prove they have met expectations for minimum proficiency. Given that a decision was made early on that panelists would only set one benchmark per grade, as opposed to optional three advocated in the Policy Linking Toolkit, this task had to be adapted from the version in the Toolkit to ask panelists to determine whether each item aligned to the grade 2 or grade 3 GPD, and not to determine which GPL each item aligned to within a grade. Facilitators believe that more thought is needed to determine the best approach to this task when only a single benchmark is being set. Also, as mentioned above, Task 2 was completed by panelists individually and independently as opposed to by group consensus. Note that the newer version of the Policy Linking Toolkit, which was released after the ICAN workshop, makes it clear that Task 2- matching should be a group activity where panelists work toward consensus. Facilitators believe this adjustment to methods would have helped to improve some of the ICAN Policy Linking Workshop results, most notably consistency of panelists ratings, described below.

The results of this task were not collected centrally by the facilitators, though the task was reviewed as a grade group. It is therefore not possible to determine the level of agreement between panelists, though there did appear to be broad agreement amongst panelists in the group sessions.

Task 3 – Benchmarking Results

The final task in the policy linking workshop includes two rounds where panelists make individual and independent judgements on whether minimally proficient students would answer items correctly. The results of this activity are individual panelist benchmarks, which are essentially the number of yeses the panelist marked by item, and panel-level benchmarks, which are the average of the individual benchmarks for the grade level. The results from Round 1 benchmarks are included in **Table 9** below.

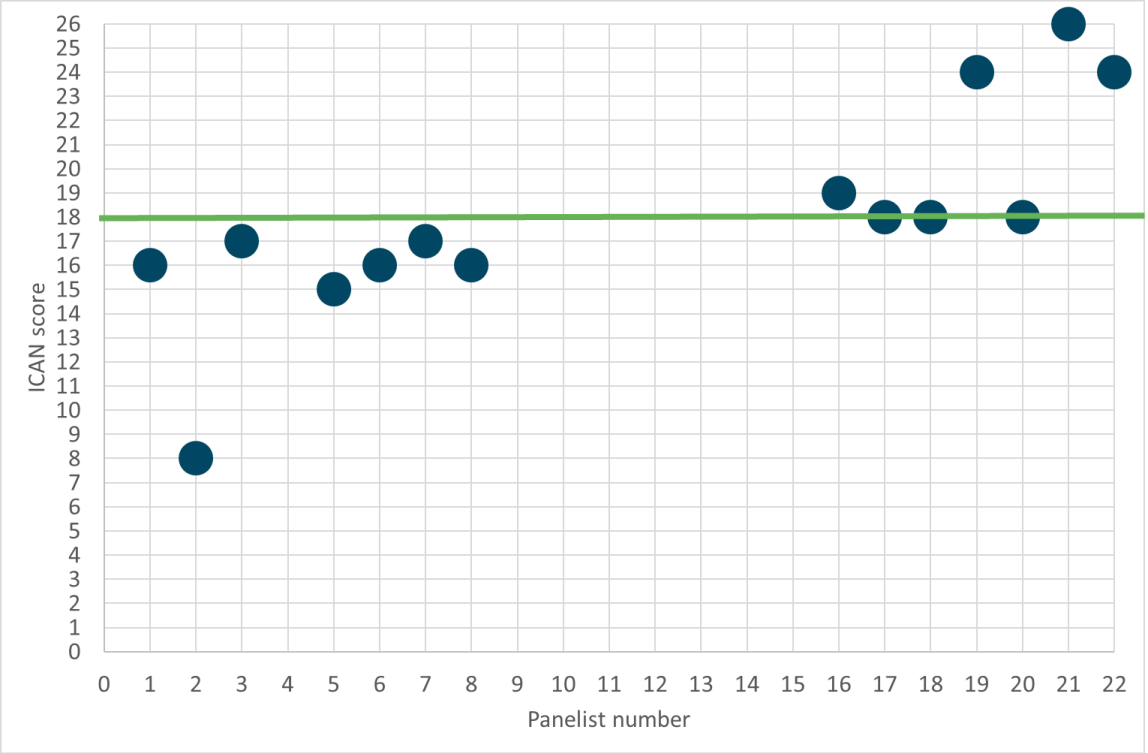
Table 9. Round 1 Benchmarks, Ranges, and Impact Data, by Grade

Grade	Benchmark (in points)	Benchmark Score Ranges (in points)	Impact Data[13] (in percentages)	
			Mwala Sub-County in Machakos, Kenya	Ikorodu Local Government Area in Lagos, Nigeria
2	18	8-26	21%	20%
3	21	15-26	16%	15%

As required in the Policy Linking Toolkit, facilitators rounded the average benchmark for all panelists down to determine the final benchmark for Round 1. For Grade 2, the average was 18.0; so, no rounding was required. However, for Grade 3, the average was 21.9 and was rounded down to 21. It should also be noted that one of the panelists for Grade 2 completed the exercise for Round 1 as if she was considering Grade 3. Her results have been removed from the Round 1 calculations. **Table 9** also includes impact data, which shows the percentage of children assessed in the 2019 ICAN in Mwala (Kenya) and Ikorodu (Nigeria) who would have met requirements for "meeting global minimum proficiency" given the benchmarks set by panelists in Round 1. The impact data provided for the two districts included from Nigeria and Kenya in the ICAN pilot are not intended to be representative of the countries as a whole, just those districts.

- Following Round 1 ratings, facilitators presented panelists with impact data, information on the populations represented in that impact data from Mwala and Ikorodu, item difficulty statistics (as shown in Table 6 above), and anonymous normative information on panelist ratings in Round 1 (shown in **Figure 1** and **Figure 2** below). They also had an opportunity to discuss items in which there was considerable disagreement on ratings.

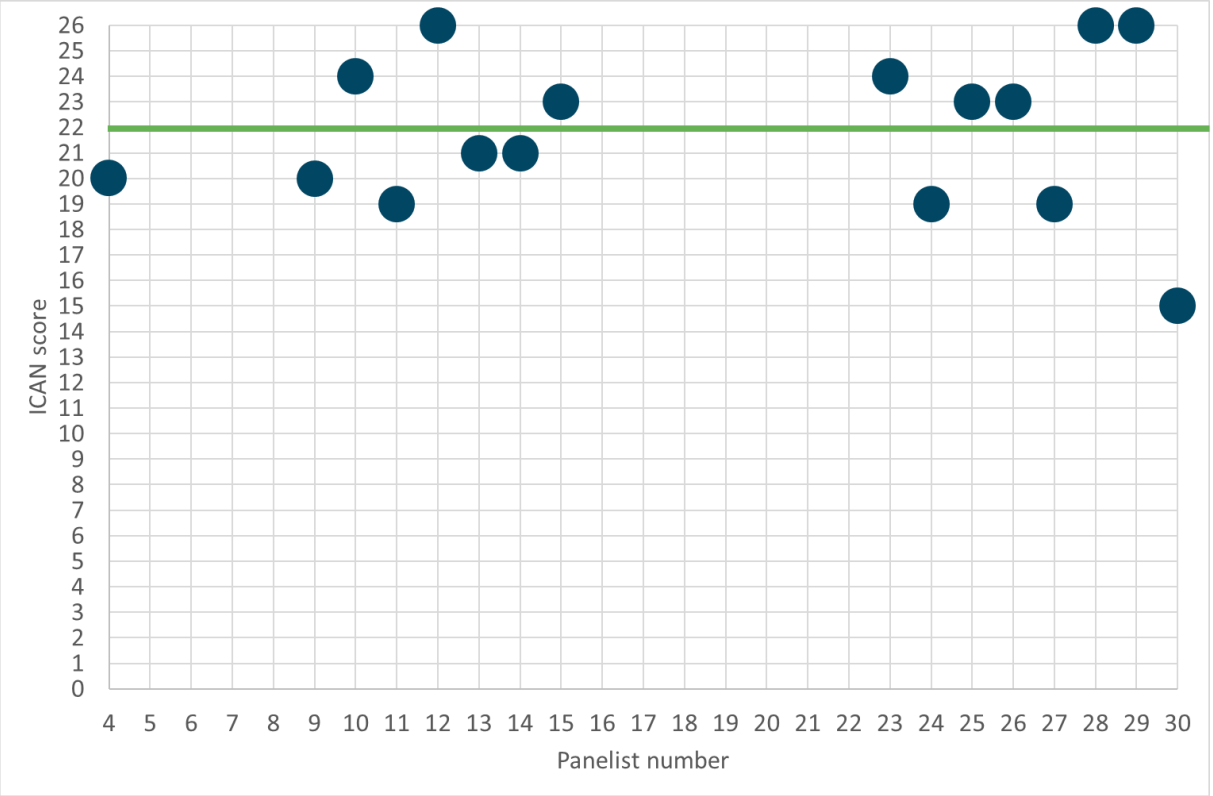
Figure 1. Grade 2 Normative Panelist Information (Benchmarks) by Panelist Number – Round 1[14]



[13] Impact data is the proportion of children assessed to have met the requirements for "meeting global minimum proficiency" given the benchmarks set by panelists.

[14] Panelist 4's score is missing, as that was the panelist who rated the wrong grade in Round 1.

Figure 2. Grade 3 Normative Panelist Information (Benchmarks) by Panelist Number – Round 1[15]



Following a review and discussion about that information, panelists had the opportunity to conduct a second round of individual and independent ratings. The results of that second round of ratings are presented in **Table 10** below, with equivalent charts to show the panelists ratings in **Figure 3** and **Figure 4**.

Table 10. Round 2 Benchmarks, Ranges, and Impact Data, by Grade

Grade	Benchmark (in points)	Benchmark Score Ranges (in points)	Impact Data (in percentages)	
			Mwala Sub-County in Machakos, Kenya	Ikorodu Local Government Area in Lagos, Nigeria
2	17	12-24	27%	22%
3	21	18-26	16%	15%

[15] Panelist 4's Round 1 benchmark is included in the Grade 3 graph since that panelist scored Grade 3 in Round 1 rather than Grade 2. However, the panelist's Round 1 benchmark is not included in the average benchmark score for the panel, shown in Table 9.

Figure 3. Grade 2 Normative Panelist Information (Benchmarks) by Panelist Number – Round 2

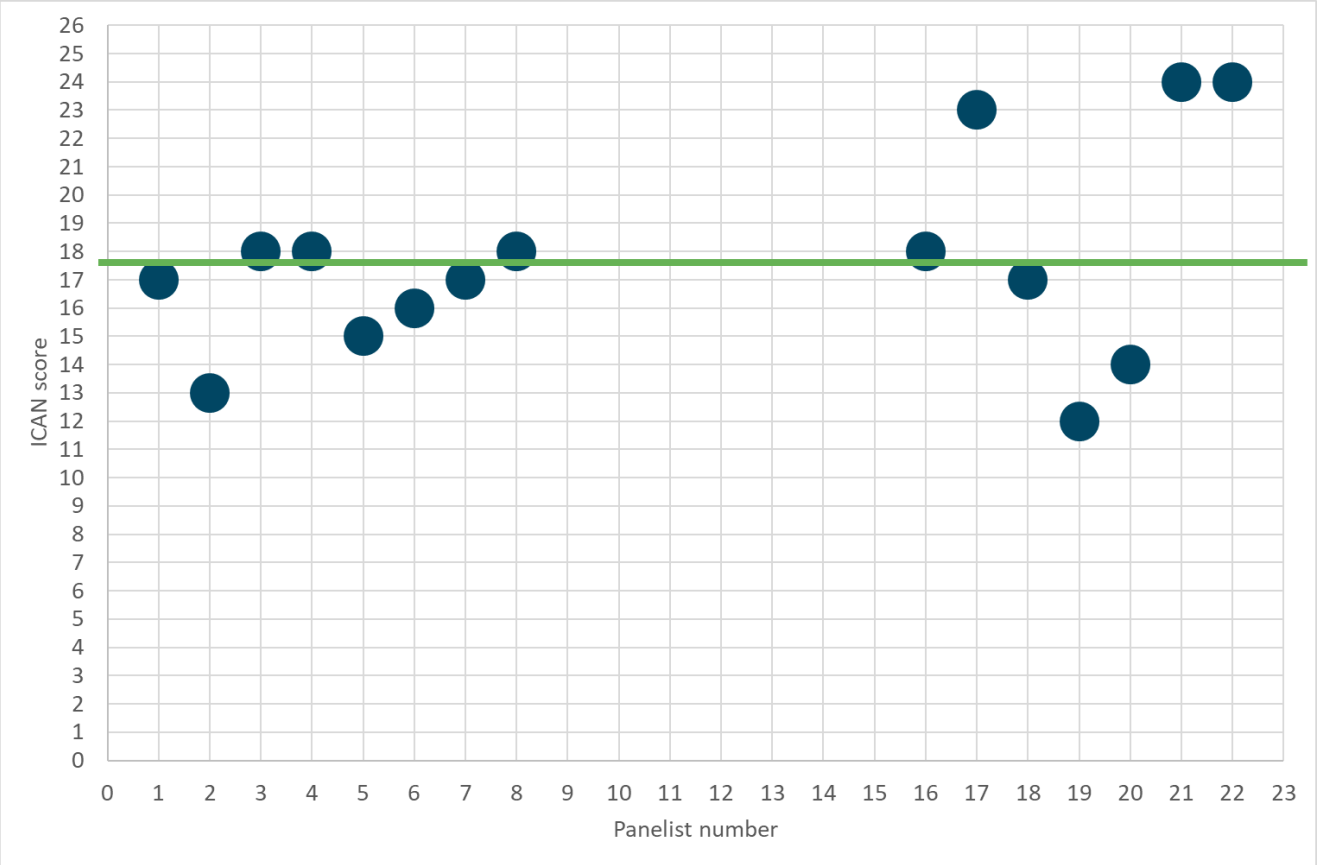
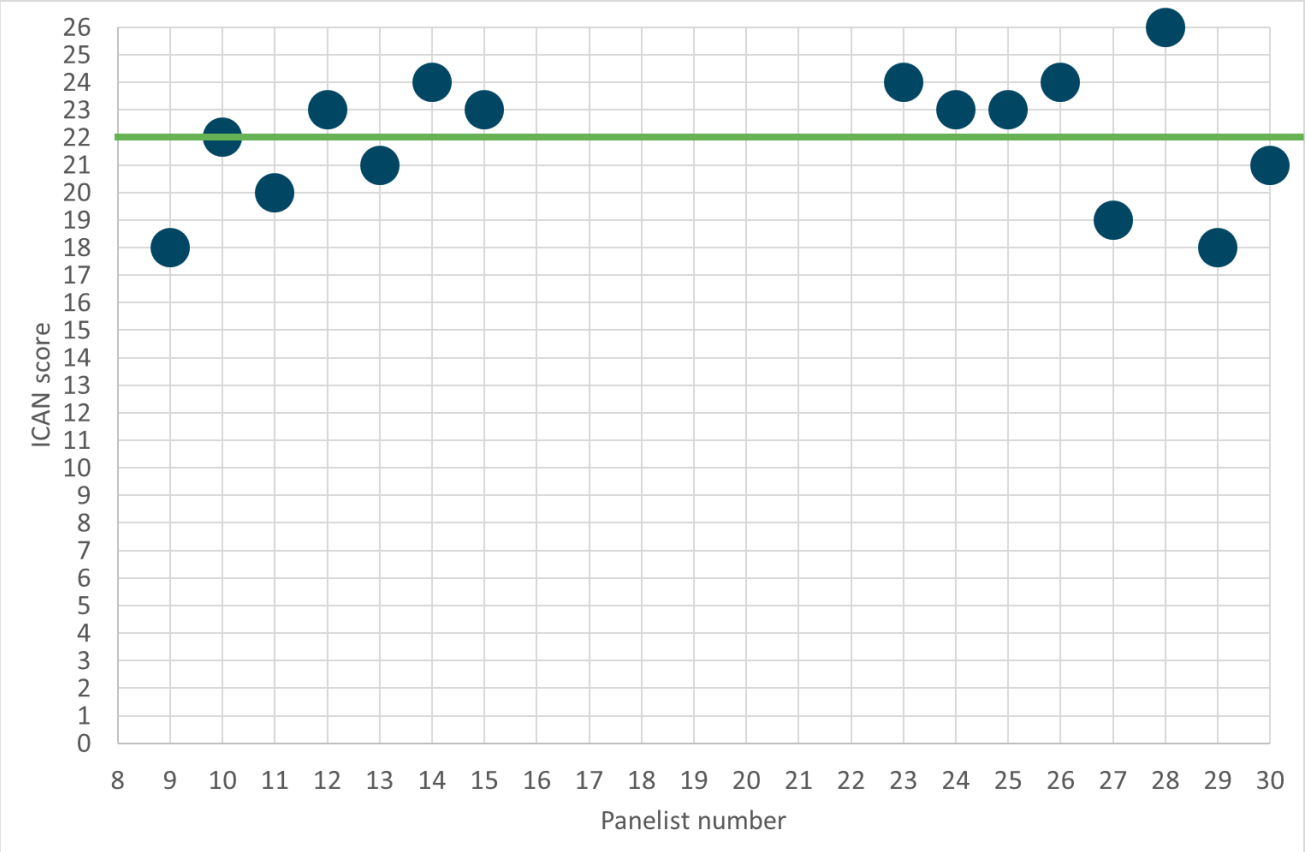


Figure 4. Grade 3 Normative Panelist Information (Benchmarks) by Panelist Number – Round 2



Facilitators rounded the benchmark for Grade 2 down from 17.6 to 17, per the toolkit and rounded the Grade 3 benchmark down from 21.9 to 21. The wide range in panelist ratings for Grade 2 is mostly related to three panelists who determined the benchmark to be 23 or 24 (out of 26). The remaining 12 panelists were in the range 12-18, as shown in **Table 11**.

Excluding these three, whose benchmarks would have placed them at the high end of the grade 3 panelists (where only 1 out of 15 panelists selected a benchmark that was higher than 24), brings the average down to 16 (rounded from 16.1). Their results may reflect the fact that some panelists appeared to be making judgements based on their 'average' student rather than those described by the GPDs for meets Global Minimum Proficiency.

Table 11. Round 2 Ratings for Grade 2

Benchmark	Number of Panelists
12	1
13	1
14	1
15	1
16	1
17	3
18	4
19	-
20	-
21	-
22	-
23	1
24	2

Overall Workshop Outcomes

In addition to calculating benchmarks and impact data, the Policy Linking Toolkit and Criteria for Policy Linking Validity also require calculating measures of workshop validity and consistency. There are seven criteria that need to be evaluated:

1) **Intra-rater reliability** – how consistent panelists are in their decision making
2) **Inter-rater consistency** – how consistent panelists are with each other
3) **Standard error** – how much spread there was amongst panelists' benchmarks compared to the 'true' benchmark score
4) **Representativeness of panelists** – how representative panelists were of the teaching population (details presented above in **Table 7** and the Sub-section on
5) Panelist Representativeness)
6) **Panelist experience** – how much relevant experience panelists had (details presented in **Table 7**)
7) **Panelist evaluation** – how confident were panelists in the process and outcomes
8) **Agreement and consistency coefficients** – what proportion of students (on a scale from 0 to 1) would be likely to achieve the same outcome (either meeting global minimum proficiency or not) on a repeated administration of the assessment

At an overall level, the Grade 3 results meet the requirements for "good" workshop validity, but the Grade 2 results do not quite rise to this level due to an issue of inter-rater consistency (consistency of rater judgements). Details follow, including the outcomes for each criterion, reported in **Table 12** through **Table 15** and summarized according to the Criteria for Policy Linking Validity in **Table 17** and **Table 18** below.

Table 12 provides details on the inter-rater consistency (IRC) and intra-rater reliability (IRR) from Rounds 1 and 2 of the workshop by grade. The Policy Linking Toolkit only includes a formula for calculating the IRR when three benchmarks are being set ('partially meets', 'meets' and 'exceeds'). In the ICAN workshop, only one benchmark was set, and, therefore, facilitators had to adapt the formula. The formula requires that we determine the absolute value of when subtracting the empirical item difficulty level (p-value) from the conditional item difficulty levels for students with 0-25% (partially meets), 26-50% (meets), 51-75% (exceeds), and 76-100% (above exceeds) scores on each item given the panelists' judgement. When making yes-no judgements to set a single benchmark, we cannot know whether the panelist believed the items they rated as 'no' should be classified as 'exceeds' or 'above exceeds' and we, therefore, cannot determine which category should be used in the formula. In **Table 12**, we have assumed category 3 (exceeds), though further work is needed to determine whether this is the most appropriate solution. The IRC fell within the required level for Grade 3 after Round 2 but not for Grade 2. As discussed above, we believe this issue may have been corrected

by the changes that have just recently been incorporated into the toolkit—namely aligning to the statements of knowledge and/or skill(s) rather than just subconstructs in Task 1 and working to reach consensus in Task 2. In terms of IRR, the acceptable threshold is still to be set by the 4.1.1 Review Panel when it is convened. However, the values for both grades 2 and 3 appear acceptable.

Table 12. Inter-Rater Consistency and Intra-Rater Reliability by Grade and Round

Grade	Inter-Rater Consistency		Intra-Rater Reliability	
	Round 1	Round 2	Round 1	Round 2
2	0.71	0.76	0.91	0.91
3	0.77	0.83	0.82	0.86

As shown in Table 13, the Standard Error (SE), which measures how much panelists scores are spread around a "true" score, was under 1.0 for both Grades 2 and 3 in Round 2, which is considered appropriate for an assessment of this length.

Table 13. Standard Error by Grade and Round

Grade	Standard Error	
	Round 1	Round 2
2	1.24	0.98
3	0.85	0.63

Details of the representativeness of the panelists was provided in the section on Stage 4: Preparation for the Policy Linking Workshop, on pages 9 and 10 above).
Table 14 below includes the results of the daily evaluations completed by panelists. All ratings are above 4, which indicates high levels of satisfaction amongst panelists with each session. Several evaluation questions after each session related to potential IT issues that panelists might have faced (for example, poor connections). When these questions related to IT issues are removed from the average, satisfaction levels increase. This demonstrates the importance with remote workshops of ensuring strong IT support for panelists, as it can have a major effect on panelist understanding and satisfaction.

Table 14. Workshop Evaluation Results by Session

Category	Average ratings (maximum of 5.0)		Average ratings excluding IT issues (maximum of 5.0)	
	Grade 2	Grade 3	Grade 2	Grade 3
Session 1	4.3	4.1	4.4	4.2
Session 2	4.3	4.3	4.3	4.3
Session 3	4.2	4.5	4.3	4.5
Session 4	4.1	4.2	4.2	4.3
Session 5	4.3	4.5	4.4	4.6
Session 6	4.3	4.6	4.4	4.7

Table 15 shows the results from the final evaluation, which includes questions about the overall process and panelists' confidence in the overall results. These are critical numbers for ensuring the validity of the workshop. Overall, these key questions in the evaluation indicate high levels of satisfaction with the workshop and confidence in the outcomes. The Criteria for Policy Linking Validity requires average results of 4 or above on a 5-point scale for panelist confidence. The workshop met this criterion for both Grades 2 and 3.

Table 15. Workshop Evaluation Results Regarding the Overall Process

Question	Average ratings (maximum of 5.0)	
	Grade 2	Grade 3
The GPF was easy to use	4.4	4.4
I agree with the decisions the group made in aligning the ICAN to the GPF	4.1	4.4
It was easy to reach the final recommended outcomes	3.9	4.0
Recommended outcomes are a good representation of the discussion	4.1	4.3
I feel confident about the outcomes	4.1	4.4

In addition to the requirements in the Criteria for Policy Linking Validity, the Policy Linking Toolkit requires facilitators estimate agreement and consistency coefficients. These are measures of what proportion of students (on a scale from 0 to 1) would be likely to achieve the same outcome (either meeting global minimum proficiency or not) on a repeated administration of the assessment. The difference between the two statistics is that the consistency coefficient takes account of the likelihood of the same outcome being achieved by chance, whereas the agreement coefficient does not. To calculate the coefficients, we used data from the whole samples in Mwala (Kenya) and Ikorodu (Nigeria) separately. These data were used to calculate separate values of coefficient alpha (r) for each sample and values of Z, where:

$$Z = \frac{\text{(Benchmark for the test} - 0.5 - \text{Mean observed test score)}}{\text{Standard deviation of observed test score}}$$

These values are used to estimate the agreement coefficient and consistency coefficient using the method proposed by Subkoviak.[16] Results are included in Table 16 below.

Table 16. Estimated Student Consistency in Meeting Benchmarks

Statistic	Mwala Sub-County in Machakos, Kenya		Ikorodu Local Government Area in Lagos, Nigeria	
	Grade 2	Grade 3	Grade 2	Grade 3
r	0.91	0.91	0.93	0.93
Z	-0.29	0.39	0.06	0.65
Agreement coefficient	0.86	0.87	0.86	0.89
Consistency coefficient	0.71	0.71	0.71	0.70

Results indicate a high proportion of consistency would be expected through repeat administrations.

Table 17 and Table 18 summarize the outputs required to determine whether the workshop met criteria for policy linking validity. According to the Criteria for Policy Linking Validity, to achieve a grade of "excellent" for the workshop, the workshop must meet all six of the criteria laid out in the tables. To achieve a rating of "good" for the workshop, the workshop need only achieve four of the six criteria. But, those four criteria must include items "b" and "c" below. Overall, the Grade 2 results met the requirements for workshop validity in all areas except for with regards to panelist representativeness (as a result of COVID-19) and inter-rater consistency, which was just under the requirement of .80 at .76. As such, though the Grade 2 results meet the requirements for four of the criteria, they do not meet the requirements for item "b" below. As such, the Grade 2 results do not quite rise to the level of "good" for reporting. The Grade 3 results, on the other hand, meet the requirements for "good" workshop validity. The only criteria missed in Grade 3 was panelist representativeness, as described in more detail in the Limitations Section of this report below.

[16] Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability for mastery tests. Journal of Educational Measurement, 25, 47-55.

Table 17. Summary of Results for Criteria for Policy Linking Validity – Grade 2

Question	Criteria	Response
4a) What was the *intra-rater reliability* for the second round of ratings?	The criterion for *intra-rater reliability* is still to be determined.	0.91
4b) What was the *inter-rater consistency* for the second round of ratings?	The *inter-rater consistency* should be at least .80.	0.76[17]
4c) What was the *Standard Error* (SE) at each *global proficiency level*?	*SE* should be appropriate for each *global proficiency level* reported. There is no maximum *SE* provided in this document, since it will depend on the number of items in the assessment.	0.98

4d) To what extent were the panelists representative of the target population of schools being reported on?	Panelists should be selected to ensure:
	• Gender representation – The panelists must be selected to ensure gender balance, both for the teachers and non-teachers.
	• Geographical representation – The teachers (and non-teachers, if possible) must be selected to ensure representation from regions, provinces, and/or states.
	• Ethnic and/or linguistic representation (where applicable) – The panel must have diversity that reflects the population; there must be native speakers of assessment languages, as well as classroom teachers who understand learning in second or third languages.
	• Representation of crisis-and-conflict-affected areas.
		• Due to logistical issues related to the remote workshop, panelists were not fully representative of Nigeria/Kenya; more details follow
		  • Gender representation was 50% female, 50% male, which is close to the ratios in Kenya and Nigeria (more details are included in the text below **Table 7** above)
		  • Panelists were not representative of the geographical regions, as describe below **Table 7**,
		  • Nigerian panelists were representative of private/public school breakdowns, but Kenyan panelists were not.
		  • All panelists taught in English
		  • 27% had experience with crisis-and-conflict-affected children

4e) To what extent did the panelists meet the other selection criteria described in the Policy Linking Toolkit?	Panelists should all have: • Several years of teaching experience in the grade level for which they are providing ratings (classroom teachers) • Skills in the subject area (all panelists)	• Average teaching experience = 10.9 years (range 1-30 years) n = 14 (1 non-response) • 100% teach subject at appropriate grade • 100% teach in English (some combined with another language) n = 14 (1 non-response)

[17] Although the inter-rater consistency statistic is lower than 0.8, this appears to be related to the three outlier panelists (see table 8b). Excluding these judges increases the inter-rater consistency to 0.81.

Question	Criteria	Response
	• Skills in the different languages of instruction and assessment (all panelists) • Knowledge of students of different proficiency levels, including at least some who would meet the requirements of the meets minimum proficiency level and some who would meet the requirements of the exceeds minimum proficiency level (all panelists) • Knowledge of the instructional environment (all panelists) • Experience administering the assessment(s) being used for the policy linking workshop.	• 100% teach children at appropriate grade • 100% teachers so have knowledge of instructional environment n = 14 (1 non-response) • 100% administered the assessment as part of their pre-workshop activity, though none were part of the original administration of the ICAN
4f) To what extent did panelists report understanding the *GPF*, assessment, and policy linking methodology? And, to what extent did they feel comfortable with their Round 2 evaluations and final *benchmarks*?	On a five-point Likert scale, with 1 being strongly disagree, very uncomfortable, etc. and 5 being strongly agree, very comfortable, etc., the average rating for each of these criteria should be 4 on average or above.	GPF • I understand the GPF (session 1) – 3.9 • I have a good understanding of the GPF and what it means for expectation about students in the grade I work with (session 5) – 4.2 • The GPF was easy to use (session 6) – 4.4 ICAN assessment • I understand the ICAN assessment (session 1) – 4.2 Policy linking methodology • I understood how to complete the tasks discussed today (session 4) – 4.1 • I feel well-prepared to complete the 'homework' (inter-session) tasks (session 4) – 4.3 • I found the inter-session task (given in Session 4 session) easy to do (session 5) – 4.4 • I understand the Angoff ratings discussed today (session 5) – 4.4 Comfortable with Round 2 evaluations and final benchmarks • It was easy to reach the final recommended outcomes (session 6) – 3.9 • The recommended outcomes are a good representation of the discussion in the virtual room (session 6) – 4.1 • I feel confident about the outcomes (session 6) – 4.1

Table 18. Summary of Results for Criteria for Policy Linking Validity – Grade 3

Question	Criteria	Response
4a) What was the *intra-rater reliability* for the second round of ratings?	The criterion for *intra-rater reliability* is still to be determined.	0.86
4b) What was the *inter-rater consistency* for the second round of ratings?	The *inter-rater consistency* should be at least .80.	0.83
4c) What was the *Standard Error* (SE) at each *global proficiency level*?	*SE* should be appropriate for each *global proficiency level* reported. There is no maximum *SE* provided in this document, since it will depend on the number of items in the assessment.	0.63
4d) To what extent were the panelists representative of the target population of schools being reported on?	Panelists should be selected to ensure: • Gender representation – The panelists must be selected to ensure gender balance. • Geographical representation – The teachers (and non-teachers, if possible) must be selected to ensure representation from regions, provinces, and/or states. • Ethnic and/or linguistic representation (where applicable) – The panel must have diversity that reflects the population; there must be native speakers of assessment languages, as well as classroom teachers who understand learning in second or third languages. • Representation of crisis-and-conflict-affected areas.	• 27% female; 73% male • Due to logistical issues related to the remote workshop, panelists were not fully representative of Nigeria/Kenya • All panelists had experience teaching in English • 47% had experience with crisis-and-conflict-affected children
4e) To what extent did the panelists meet the other selection criteria described in the Policy Linking Toolkit?	Panelists should all have: • Several years of teaching experience in the grade level for which they are providing ratings (classroom teachers) • Skills in the subject area (all panelists) • Skills in the different languages of instruction and assessment (all panelists) • Knowledge of students of different proficiency levels, including at least some who would meet the requirements of the meets minimum proficiency level and some who	• Average teaching experience = 10.9 years (range 3-25 years) n = 15 • 100% teach subject at appropriate grade • 100% teach in English (some combined with another language) n = 15 • 100% teach children at appropriate grade • 93% teachers; 7% content experts so have knowledge of instructional environment n = 15 • 100% administered the assessment as part of their pre-workshop activity, though none were part of the original administration of the ICAN

Question	Criteria	Response
	would meet the requirements of the exceeds minimum proficiency level (all panelists) • Knowledge of the instructional environment (all panelists) • Experience administering the assessment(s) being used for the policy linking workshop.	
4f) To what extent did panelists report understanding the *GPF*, assessment, and policy linking methodology? And, to what extent did they feel comfortable with their Round 2 evaluations and final *benchmarks*?	On a five-point Likert scale, with 1 being strongly disagree, very uncomfortable, etc. and 5 being strongly agree, very comfortable, etc., the average rating for each of these criteria should be 4 or above.	GPF • I understand the GPF (session 1) – 4.1 • I have a good understanding of the GPF and what it means for expectation about students in the grade I work with (session 5) – 4.5 • The GPF was easy to use (session 6) – 4.4 ICAN assessment • I understand the ICAN assessment (session 1) – 4.4 Policy linking methodology • I understood how to complete the tasks discussed today (session 4) – 4.4 • I feel well-prepared to complete the 'homework' (inter-session) tasks (session 4) – 4.2 • I found the inter-session task (given in Session 4 session) easy to do (session 5) – 4.6 • I understand the Angoff ratings discussed today (session 5) – 4.7 Comfortable with Round 2 evaluations and final benchmarks • It was easy to reach the final recommended outcomes (session 6) – 4.0 • The recommended outcomes are a good representation of the discussion in the virtual room (session 6) – 4.3 • I feel confident about the outcomes (session 6) – 4.4

LIMITATIONS

As described above, there were several limitations to the ICAN Policy Linking Workshop. These follow in more detail:

- Alignment was not done using the statements of knowledge and/or skills but rather subconstructs – As mentioned above, the Policy Linking Toolkit did not include details about the need to align ICAN assessment items to the statements of knowledge and/or skills. As such, panelists aligned to subconstructs. Since the new version of the toolkit describes aligning to knowledge and/or skills but summarizing results of the alignment exercise at the subconstruct level (since knowledge or skill is a subcategory of subconstructs), this shouldn't affect the general validity of the results. However, had we had knowledge or skills included in the GPF and aligned to those, we may have been able to achieve more consistent panelist results, which may have improved the IRR.

- Matching was not completed by consensus – As described above, panelists completed the matching process individually and independently, which may have affected the consistency of ratings in Task 3 as well since at that point, different panelists may have been judging whether a minimally proficient student would answer an item correctly using different GPDs. Had they used the same GPDs, this might also have improved consistency and the IRR for Grade 2 to an acceptable level.

- COVID-19-related challenges with assessing students ahead of the workshop – While all panelists were able to assess children ahead of the workshop using the ICAN, they weren't all able to assess students from their respective classes whom they knew just barely met the requirements of the "meets global minimum proficiency" level, as a result of COVID-19, which prevented panelists from travelling outside of their communities. This meant that many panelists were only able to assess children from their communities, whom they did not know in advance of assessing them whether they met the requirements for just meeting global minimum proficiency. As such, the act of assessing children did not prove as valuable for panelists since they largely could not use it to determine how children who just barely meet global minimum proficiency requirements from the GPF would perform on the assessment, which is very helpful when making ratings.

- Remote access issues – The need to host the workshop remotely as a result of COVID-19 was a challenge for several reasons:
    - Representativeness of panelists – Since not all teachers and curriculum experts in Kenya and Nigeria have equivalent access to the internet, the sample of panelists was skewed. We were unable to identify panelists in some of the more remote regions of both countries, which may have had a slight effect on the overall outcomes if there is variation between teachers in different regions of a country. Since panelists are asked to envision students who meet the proficiency levels, theoretically, this shouldn't be cause for major differences in ratings between teachers from different parts of the country. But, in practice, that assumption does not always hold. And, the three counties represented by Kenyan panelists are three that tend to have lower math scores than many of the other counties in the country.[18]
    - Hardware issues – Some panelists joined the sessions from Smart Phones and had a harder time seeing some of the slides. We knew in advance that this might be an issue; so, we increased the font size as much as possible and also sent copies of the slides to the panelists in advance.
    - Technical issues – Several of the panelists reported challenges, and the facilitators observed the same challenges, with panelist connections. Several panelists lost connection on and off throughout the workshop sessions, could not speak during the sessions, or were breaking up

[18] Uwezo. (2016). *Are Our Children Learning? Uwezo Kenya Sixth Learning Assessment Report.* http://www.uwezo.net/wp-content/uploads/2016/12/UwezoKenya2015ALAReport-FINAL-EN-web.pdf

during their participation. This meant that panelists may have missed key points or not been able to get all of their questions answered during sessions. To try to mitigate this issue, the facilitators recorded each session, uploaded them to YouTube (creating a private link that would allow panelists to watch even if they had low bandwidth) and shared the recordings with the panelists via WhatsApp, to which all panelists had access. The hope was that the panelists could watch sections of sessions that they missed and then follow up with the facilitators via email or WhatsApp with questions. Many panelists reported doing this, and the facilitators fielded a series of questions between sessions and even engaged panelists individually to catch them up on missed sessions when requested.

  - Difficulty observing panelists – Because we met remotely, the facilitators could not walk around and observe group work or look over panelists' shoulders as they completed their tasks. This means that facilitators were unable to catch some of the common errors that panelists were making in the moment and, instead, had to attempt to address them between tasks. Even still, some questions/issues may have gone overlooked entirely since they are often only highlighted by overhearing panelists talking with one another in the workshop room. We address possible solutions for this challenge in the Lessons Learned Section below.

- Panel size issues – As described above, it would have been helpful to have Kenyan Grade 2 panelists set one benchmark and Nigerian Grade 2 panelists set the same benchmark simultaneously (and the same for Grade 3) to see if/how much country context issues affect benchmarks and policy linking workshop outcomes. However, this was not possible as a result of COVID-19 and our somewhat limited ability to find sufficient panelists with adequate internet access. As such, it is not clear how much the panelists' country backgrounds affected the benchmarks because the sample size (number of panelists) was too small. Descriptive statistics show that, though the results were almost identical between Kenyan and Nigerian panelists at Grade 3 in Round 1, with Kenyan panelists setting a benchmark of 22 and Nigerian panelists setting the benchmark at 21.9, in Round 2, that difference increased slightly, with the benchmarks set at 21.3 and 22.3, respectively. At Grade 2, there was a significant difference between the benchmarks set by the panelists from the two different countries. Kenyan Grade 2 panelists set an average benchmark of 15.6 in Round 1, while Nigerian Grade 2 panelists set an average benchmark of 21 in Round 1. In Round 2, that difference was somewhat reduced with benchmarks of 16.5 and 18.9, respectively. To know whether the country context really makes a significant difference, more research and workshops will be necessary (either in-person or remote but with additional panelists—preferably 15 or more per country, per grade—from each country).

CONCLUSIONS, RECOMMENDATIONS, AND LESSONS LEARNED

Conclusions

Overall, given that this was the first remote Policy Linking for Measuring Global Learning Outcomes Workshop, the facilitators consider it a success. Panelists were engaged throughout the workshop and expressed excitement and appreciation for the process, even suggesting that they may benefit from other trainings and engagements held remotely in the future. The workshop also met the requirements for Policy Linking validity for Grade 3 and only narrowly missed meeting the requirements for Grade 2.

With regards to the purpose of the workshop, the workshop outcomes suggest strong evidence that the policy linking methodology is viable for use with cross-national CLAs. Alignment results show that ICAN is "Additionally Aligned" with the GPF, as described in Table 5 above, suggesting a fairly robust link between ICAN and the GPF. Finally, we believe the benchmarks set are useable for comparing, aggregating, and tracking learning outcomes for ICAN assessment results both from the 2019 implementation round in two rural districts and then more widely in the future, though it would be beneficial to run a confirmatory workshop in both countries to validate.

Recommendations for PAL Network

We recommend that the PAL Network consider hosting a follow-on workshop in Kenya and/or Nigeria to validate the results of this workshop. The workshop could be held in-person (preferred) or remotely but should include a more representative set of panelists, which may mean it needs to be held either in-person or with at least the panelists gathering in one place with strong wifi.

Further, following on the Kenyan or Nigerian workshop, we recommend at least one additional workshop be conducted with panelists from at least one additional country. Depending on the results of that workshop, benchmarks may be validated or require follow-on workshops.

In terms of using the benchmarks, the benchmarks can be used to interpret current ICAN results with some caution, mostly related to Kenya given that panelists did not represent the area in which the assessment was conducted. Benchmarks are best used to interpret results for Grade 2 and 3 students as well as out-of-school children who are 6-9 years old. However, they might also be used to interpret outcomes for students from higher grades and ages. For instance, a Grade 5 student who meets the recommended Grade 3 benchmark of 21 could be considered to be performing at a Grade 3-level.

Lessons Learned for Policy Linking

Given that this was the first remote policy linking workshop, we learned several lessons that we believe will be useful for future remote policy linking workshops. Some of these lessons are based on things that worked well for the workshop, and others are based on changes we would make now. All follow.

Logistics

- Ensure panelists have the printed documents they will need to complete the workshop.
- Ensure panelists are able to join via a laptop (strongly preferred) or smartphone so that they can see slides and submit tasks. Allow panelists to submit tasks either as soft copies, photos/scans of forms, or (depending on the task) in the body of the text through email or WhatsApp to ensure panelists are able to complete tasks with limited IT challenges.
- Provide data cards to panelists to ensure they have sufficient data to connect to the sessions, and encourage panelists to assess their service far in advance of the workshop in case they need to explore changing providers (if possible), etc.
- Set up a WhatsApp group in advance of the workshop to facilitate announcements, remind panelists of sessions, and ensure ease of communication between workshop sessions when many panelists do not have regular access to email communications.
- Send out calendar invitations for all panelists for the sessions.

- Use a teleconference platform that allows for: 1) presenting slides and sharing one's screen, 2) assigning panelists to break-out groups; 3) recording the sessions (for panelists who miss portions of the workshop due to technological issues to listen to after the sessions; if possible, ensure the platform, host computer, and wifi are strong enough to ensure short processing time for recordings so they can be released to panelists quickly); 4) muting everyone upon entry in the meeting; 5) typed chats; 6) raising one's hand to indicate a question or comment; registration of participants to help track attendance (if the latter is not possible, administrative staff should be on hand to track changing attendance throughout each session - possibly noting who is there at the beginning, middle, and end; this allows facilitators to follow up with panelists who missed significant portions of the workshop due to technological issues).
- Host a series of short pre-workshops calls to check small groups of panelists' abilities to connect and troubleshoot any technology issues.
- Have an administrative assistant (NOT a facilitator) manage the teleconference platform, letting participants in, assigning panelists to small groups, etc., as this task can be quite difficult to manage while leading sessions.

Lead facilitator(s)

- Engage two (or at least one per grade/subject/language of assessment) lead facilitators to help facilitate the small-group break-out sessions, to allow panelists to hear from more than one person, and to allow for one person to be tracking questions that come up in the chat while the other facilitator is presenting.

Content facilitator training and interaction

- Plan for a minimum of an 8-hour remote content facilitator training, split into two sessions. However, if it is possible to increase the length of this training to ensure the content facilitators have time to complete each of the activities themselves, it is recommended.
- Have the lead facilitators lead all plenary sessions unless the content facilitators have previous experience with standard setting.
- In addition to the general content facilitator training, scheduling short preparation sessions with the content facilitators to remind them of key issues just before the sessions where they are leading breakout groups is highly recommended.

Pre-sessions

Remote workshops have an advantage in that they can be extended out over a somewhat longer period of time since project teams need not be concerned with hotel and per diem arrangements (unless panelists are meeting in person with only the lead facilitators attending remotely).

- Plan pre-sessions to allow panelists to become more familiar with the GPF and the assessment before undertaking the student assessment task with students who meet the requirements for each GPL.
- Note, in some cases, it may not be possible for panelists to complete the student assessment task (e.g., due to security concerns related to COVID-19). In those cases, ensure panelists have an opportunity to take the assessment themselves during one of the pre-sessions or to administer the assessment to children in their homes or communities (e.g., outside using masks) between the pre-sessions and the regular session.
- To aid with the later tasks, ask panelists to write down the names of students in their class who are described the by "meets" GPDs as part of their inter-session activity.

Discussions

One major disadvantage of remote workshops is that panelists don't have the opportunity to engage in informal discussions with their neighbors, which often highlight misunderstandings or questions, nor do facilitators have the ability to walk around while panelists complete the tasks and look over panelist shoulders to identify potential misunderstandings. The tips below are focused on trying to address these shortcomings.

- If possible, it would be helpful to identify a way of allowing panelists to have conversations between themselves and then come back together to ask facilitators questions. This might be done by going into breakout groups for 10 minutes after every set of slides to discuss and identify any questions/issues. Sessions may need to be extended to accommodate this possibility.
- If possible, it would also be helpful to identify a way of "looking over panelists' shoulders." This might be done by scheduling individual one-on-one 15-30 minute sessions between a lead facilitator and each panelist after the end of the plenary sessions. During these calls, the facilitators can ask panelists to explain the task and describe how they are aligning/matching/ rating each item. This should help to identify and correct misunderstandings. It should also ensure panelists who missed portions of the workshop due to technology issues have time to ask questions and become clear on the task.
- Finally, lead facilitators might stay on the call for each workshop session that includes a task assignment (Task 1 and 3, for both rounds) for an hour or so after the session to allow people to do the task on their own but re-join the call if they have questions.

ANNEX A: ICAN POLICY LINKING WORKSHOP AGENDA

Preparation session 1 – Wednesday, August 19

Timing	Activity	Facilitator
0-15 mins	Welcome and introductions	Lead facilitator
15-40 mins	Overview of policy linking	Lead facilitator
40-55 mins	Purpose of preparation session	Lead facilitator
55-60 mins	Comfort break	
60-80 mins	Overview of the GPF	Content facilitator
80-100 mins	Grade 2 and 3 mathematics GPF	Content facilitator
100-110 mins	Explanation of inter-session activities	Lead facilitator
110-120 mins	Closing remarks	Lead facilitator

Panelist inter-session activities

- Review Grade 2 and Grade 3 GPF and identify any elements that are unclear (submit 1 week prior to workshop)

Preparation session 2 – Friday, August 21

Timing	Activity	Facilitator
0-15 mins	Welcome and purpose of the preparation session	Lead facilitator
15-30 mins	Overview of the ICAN	Content facilitator
30-55 mins	Review each item on the ICAN	Content facilitator
55-60 mins	Comfort break	
60-100 mins	Continue reviewing items and discuss ICAN administration	Content facilitator
100-110 mins	Explanation of inter-session activities	Lead facilitator
110-120 mins	Closing remarks	Lead facilitator

Panelist inter-session activities

- Administer the ICAN to up to 5 students (of appropriate age and performance level)

Workshop session 1 – Tuesday, September 1

Timing	Activity	Facilitator
0-10 mins	Welcome and purpose of session 1	Lead facilitator
10-55 mins	Review GPF activity and provide clarification	Content facilitator
55-60 mins	Comfort break	
60-105 mins	Discussion of ICAN administration activity	Content facilitator
105-120 mins	Evaluation approach and completion of evaluation 1	Lead facilitator

Workshop session 2 – Wednesday, September 2

Timing	Activity	Facilitator
0-10 mins	Welcome and purpose of session 2	Lead facilitator
10-20 mins	Address any concerns raised in evaluation 1	Content facilitator
20-55 mins	Introduction to alignment review	Lead facilitator
55-60 mins	Comfort break	
60-90 mins	Small group discussions on first 5 items	Content facilitators[19]
90-110 mins	Plenary	Content facilitator
110-120 mins	Explanation of inter-session activities and close	Lead facilitator

Panelist inter-session activities

- Complete alignment review on all remaining items (submit 4 hours after close of session)
- Complete evaluation 2 (submit with alignment review).

[19] Each small group will have a lead facilitator and two content facilitators (one from each country)

Workshop session 3 – Friday, September 4

Timing	Activity	Facilitator
0-10 mins	Welcome and purpose of session 3	Lead facilitator
10-40 mins	Review inter-session activities and provide clarification	Content facilitator
40-55 mins	Matching presentation	Lead facilitator
55-60 mins	Comfort break	
60-110 mins	Matching practice and beginning of the matching activity	Lead facilitator
110-120 mins	Explanation of inter-session activities and close	Lead facilitator

Panelist inter-session activities

- Complete evaluation 3 (submit 1 hour after close of session).

Workshop session 4 – Monday, September 7

Timing	Activity	Facilitator
0-10 mins	Welcome and purpose of session 4	Lead facilitator
10-55 mins	Angoff methodology presentation	Lead facilitator
55-60 mins	Comfort break	
60-90 mins	Small group Angoff ratings using practice items	Content facilitators
90-110 mins	Start Round 1 Angoff ratings, individually and independently	Independent work
110-120 mins	Explanation of inter-session activities and close	Lead facilitator

Panelist inter-session activities

- Complete Round 1 ratings on all remaining items (submit 4 hours after close of session)
- Complete evaluation 4 (submit with Round 1 ratings).

Workshop session 5 – Wednesday, September 9

Timing	Activity	Facilitator
0-10 mins	Welcome and purpose of session 5	Lead facilitator
10-55 mins	Review Round 1 ratings	Content facilitator
55-60 mins	Comfort break	
60-90 mins	Review Round 1 ratings (continued)	Content facilitator
90-110 mins	Share impact data	Lead facilitator
110-120 mins	Explanation of inter-session activities and close	Lead facilitator

Panelist inter-session activities

- Complete Round 2 ratings (submit 4 hours after close of session)

Workshop session 6 – Friday, September 11

Timing	Activity	Facilitator
0-10 mins	Welcome and purpose of session 6	Lead facilitator
10-30 mins	Review Round 2 ratings and share final outcomes	Content facilitator
30-40 mins	Complete evaluation 5	Independent work
40-60 mins	Thanks and close	Lead facilitator
60-120 mins	Focus group	Evaluation team