BILL & MELINDA GATES foundation

UNESCO INSTITUTE FOR STATISTICS

GLOBAL ALLIANCE TO MONITOR LEARNING

SUSTAINABLE DEVELOPMENT GOALS

DRAFT CRITERIA FOR POLICY LINKING VALIDITY

For Measuring Global Learning Outcomes
For Reading And Mathematics
Grades 2 To 6
January 2020

USAID FROM THE AMERICAN PEOPLE

UKaid from the British people

ACER

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

TABLE OF CONTENTS

Key Terms 2

Section 1: Overview 4

Background 4

Purpose and Audience of this Document 4

4.1.1 Review Panel 5

Document Organization 5

Section 2: 4.1.1 Review Panel Process for Policy Linking 6

Stage 1: Initial engagement 6

Stage 2: Collation of evidence of curriculum and assessment alignment and validity 6

Criterion 1 – Alignment between the assessment, the assessment framework, and the curriculum 7

Criterion 2 – Assessment validity and appropriateness for the population 9

Criterion 3 – Reliability of the assessment 10

Stage 3: Review of evidence by the 4.1.1 Review Panel 11

Stage 4: Preparation for the policy linking workshop 12

Stage 5: Implementation of the policy linking workshop and documentation of evidence on outcomes 12

Stage 6: Review of workshop outcomes by the 4.1.1 QAP 13

Stage 7: Reporting of results against SDG 4.1.1 13

Section 3: Technical Criteria 15

Criterion 1 15

Criterion 2 17

Criterion 3 19

Criterion 4 21

ANNEXES 23

Annex A – Overview of 4.1.1 Review Panel Process 23

Annex B – Decision tree for implementing policy linking 24

Annex C – Assessment quality evidence 26

KEY TERMS

Where a *key term* is used within this document, it is indicated using italics and underlining.

Alignment Study - a study designed to determine how well national assessment items (questions) cover the domains and constructs (skills) that have been determined at a global level to be essential reading and mathematics skills (as depicted in the *Global Proficiency Framework*. Good alignment will allow reporting against SDG 4.1.1. However, poor alignment indicates that a country has a different interpretation of the key elements of reading and mathematics than those established in global standards such that, were the country to report results, they would not be comparable to those reported by other countries.

Assessment Framework – a document that defines the purpose of the assessment, details what should be assessed in terms of domains, constructs, sub-constructs, and skills (countries may have different terms for these classifications), and identifies what percentage of an assessment should be dedicated to assessing which skills. This framework provides a blueprint for developing multiple assessments across years to ensure they are all designed using the same criteria.

Benchmark - a specific assessment score that designates a performance standard has been met on a given assessment; the desired competency or skills learners should developmentally be able to demonstrate when provided appropriate resources and support. Benchmarks should be set based on objective evidence of required performance rather than the current performance of learners.

Classification Accuracy - how precisely children have been classified by the assessment.

Classification consistency - the extent to which children are classified the same way in repeated applications of an assessment.

Coefficient Alpha (or *Cronbach's alpha*) **-** a psychometric test of reliability, or internal consistency, between items on an assessment. It measures whether the items on the assessment seek to measure the same latent variable, which in the case of assessments used for policy linking would be reading or math ability.

Constructed-Response Item - open-ended items where students must generate rather than selecting an option.

Content Domain - the body of knowledge, skills, or abilities being taught in curriculum or measured or examined by a test, experiment, or research study. Aural language comprehension, decoding, reading comprehension, number knowledge, measurement, statistics and probability, geometry, and algebra are all content domains included in the GPF as well as many countries' curriculum and assessment frameworks.

Content Standards – describe what learners should be taught. Countries typically set their own content standards. USAID and other stakeholders used content standards from more than 50 countries to inform the GPF.

Global Proficiency Framework (GPF) – a framework developed by donors and partners based on current country content and assessment frameworks across more than 100 countries, which provides *performance standards* for learners in Grades 2-6 in reading and mathematics.

Global Proficiency Levels (GPLs) – the four levels of proficiency defined by the *GPF* for classifying learner outcomes: does not meet minimum proficiency, partially meets minimum proficiency, meets minimum proficiency, and exceeds minimum proficiency.

Inter-Rater Reliability (IRR) – degree of agreement amongst panelists on scoring of assessment items.

Intra-Rater Reliability - degree of agreement on scoring of items by a single panelist.

Performance Standards – describe how learners should perform on assessments demonstrate they have learned what is presented in the content standard; the GPF includes the internationally-agreed-upon performance standards for grades 2-6 in reading and math. Countries can also set their own performance standards and set *benchmarks* for those standards; but those cannot be used to report against SDG 4.1.1.

Specification / blueprint – a description of the rules that are used to construct an assessment, for example number of items, coverage of domains, question and response formats, scoring arrangements, reporting arrangements and desired psychometric properties.

Standard Error of Measurement - a measure of how much panelists scores are spread around a “true” score**.**

Selected-Response Items - closed item where student chooses from a predetermined list of options e.g., multiple choice, true/false, etc.

Target – a goal for the number or percentage of children that will reach the *benchmark* for a given grade in a given period of time; targets should be altered based on the current performance of learners in schools and should provide a realistic timeline for when learners should be able to achieve minimum proficiency standards. They often vary across populations.

SECTION 1: OVERVIEW

Background

Policy linking is a process[1] to set *benchmarks* (also known as “cut scores” or “thresholds”) on learning assessments that allow those assessments to be aligned across countries and contexts. While the methodology on which the process is based is well established, its use has now been extended to help countries set comparable *benchmarks* using the *Global Proficiency Framework* (GPF). The *GPF* is a framework developed by multilateral donors and partners based on national content and assessment frameworks from across more than 50 countries that provides *performance expectations/standards* for learners in Grades 2-6 in reading and mathematics.

Policy linking allows countries to measure learning outcomes using comparable metrics and also assess relative alignment between the country education standards and the education standards put forth in the *GPF*. By linking their national assessments to the *GPF*, countries and donors are able to compare learning outcomes across language groups and contexts in countries as well as across countries and over time, assuming all new assessments are subsequently linked to the *GPF*. Policy linking also allows countries to use their existing national assessments or other early grade reading and math assessments to report against Sustainable Development Goal (SDG) 4.1.1 as well as some donor-required indicators.

This policy presents a set of guidelines (seven steps that must be taken to implement policy linking as well as criteria for each of those steps) that countries that implement policy linking must follow for their reading and math assessment results to be accepted by the United Nations Educational, Scientific, and Cultural Organization (UNESCO) for reporting against SDG 4.1.1.

This document should be used alongside the Policy Linking Toolkit. The Toolkit provides guidance to countries, donors, and their partners for running policy linking workshops to set global *benchmarks*. While it addresses all seven policy linking steps, its main focus is on the implementation of the workshop. This policy document sets out the steps that must be taken ahead of the policy linking workshop, documentation that should occur during the workshop, and reports that must be submitted following the workshop.

Purpose and Audience of this Document

This document is intended to be used by governments to both:

1) Understand the process that they need to follow to enable reporting against SDG 4.1.1 and
2) Explain the criteria that UNESCO will use to determine whether they will accept country results for SDG 4.1.1 that were obtained through use of policy linking.

This 4.1.1 Review Panel process is designed to ensure that UNESCO can have confidence that results generated through policy linking are robust and comparable to

[1] The policy linking process is based on the Angoff standard-setting methodology, which has a long established use in many countries.

results generated by other countries through policy linking and other methods, which also have validity criteria for reporting to SDG 4.1.1.

4.1.1 Review Panel

To validate the outcomes reported to SDG 4.1.1, UNESCO will appoint an independent panel of 15-20 experts, equally split between reading and math experts and psychometricians (experts in test development, administration, and analysis). UNESCO will have an open call for applications for each of the expert slots and will select the ultimate panel to ensure representation from all regions of the world as well as from those with experience with reading in alphabetic and non-alphabetic languages. The 4.1.1 Review Panel will be tasked with reviewing results submitted for SDG 4.1.1 regardless of the method used to report (Note that countries wishing to report to SDG 4.1.1 can choose from one of four options for reporting: they can engage in a regional or international assessment, statistically link their national assessment to a regional or international assessment, develop an assessment using UNESCO's Global Item Bank, or engage in policy linking.

For policy linking, five members of the QAP will undertake two reviews of country-submitted evidence during the policy linking process:

1) First, they will review the evidence of assessment reliability and validity and alignment with the *GPF* **before** the policy linking workshop takes place to confirm that policy linking using the GPF will be possible before countries spend valuable resources implementing the methodology; and
2) Second, they will review the evidence from the policy linking workshop to confirm that it was conducted in line with the criteria set within and that the results are sufficiently robust for reporting to SDG 4.1.1. Following this review, the 4.1.1 QAP will make a recommendation to UNESCO on whether to accept the results from the workshop for reporting against SDG 4.1.1. UNESCO will make the final decision on whether the results will be accepted.

Document Organization

Given the nature of the process, this document is necessarily technical in parts, and governments may need to either 1) engage their own reading and math content experts and psychometricians or statisticians with experience in test development and assessment or 2) appoint a partner with expertise in assessment and psychometrics to support them in understanding and reporting on the requirements. Section 2, immediately following this section, is intended to provide details of this policy for policy makers. It contains an overview of the seven stages of the policy linking process followed by a detailed, less-technical overview of the requirements countries must meet for UNESCO to accept policy linking results for reporting on SDG 4.1.1. It also suggests what technical expertise might be needed to compile results for UNESCO. Section 3, on the other hand, is intended for technical experts who may be collating evidence for submission to the 4.1.1 QAP on behalf of the government.

Section 2: 4.1.1 Review Panel Process for Policy Linking

There are seven stages for a country interested in engaging in policy linking:

- Stage 1: Initial engagement – leading to decision of whether to move forward with the policy linking process
- Stage 2: Collation of evidence of curriculum and assessment alignment and validity – leading to submission of evidence
- Stage 3: Review of evidence by the 4.1.1 QAP – leading to agreement on whether country conditions support policy linking for reporting against SDG 4.1.1
- Stage 4: Preparation for the policy linking workshop
- Stage 5: Implementation of the policy linking workshop and documentation of evidence on outcomes – leading to submission of evidence to the 4.1.1 QAP
- Stage 6: Review of workshop outcomes by the 4.1.1 QAP – leading to a recommendation to UNESCO of whether results should be accepted for reporting against SDG 4.1.1
- Stage 7: Reporting of results to countries and against SDG 4.1.1

Annex A contains a high-level visual of the process, and more details follows.

Stage 1: Initial engagement

The policy linking process was developed and scaled up through the collaboration of a consortium of global donors, including UNESCO, the World Bank, the United States Agency for International Development (USAID), the Department for International Development (DFID), the Australian Council for Education Research (ACER), and the Bill and Melinda Gates Foundation (BMGF). This consortium is developing a website and communication materials to support and guide governments considering implementing policy linking to enable their reporting against SDG 4.1.1. That website will be shared widely once finalized.

Governments are responsible for making the decision of whether to move forward with policy linking, either at a national or regional/state level. The decision tree in Annex B gives a high-level overview of the likely decision-making process that a government may go through to determine whether policy linking is appropriate or necessary for reporting national assessment results with respect to global standards. However, governments should discuss the options with UNESCO-UIS and other donor organizations that are supporting countries with reporting against SDG 4.1.1.

Stage 2: Collation of evidence of curriculum and assessment alignment and validity

In this stage, governments, with support from donors and/or partners (as relevant), must collate appropriate evidence to confirm the reliability and validity of the assessment they intend to use for policy linking, including carrying out an *alignment study* between the assessment and the *Global Proficiency Framework* (GPF). In addition to conducting an alignment study, governments should also include information on how the assessment aligns with the country's own curriculum, how students were sampled (if a census was not used) to take the assessment, and how the country ensured the assessment was reliable. This information is frequently documented as part of a national assessment framework or a national assessment technical report.

Assessments proposed for policy linking should ideally be developed in line with internationally recognized standards for test development, such as the Standards for Educational and Psychological Testing (2014)[2] to ensure the assessment is sufficiently valid to provide confidence in the outcomes of policy linking. As part of the development and implementation, a wide range of evidence on reliability and validity is usually collected and may have been published in a technical report. The government will need to submit a subset of that evidence to the 4.1.1 QAP, which will review it to determine the suitability of the assessment for policy linking.[3]

The 4.1.1 QAP will be evaluate the evidence using the three criteria described below to ensure a fair and consistent process across countries. While summarized below, the criteria are described in detail in the section 3, which also includes specific questions for governments to answer in relation to each criterion. Some of these questions are starred (*) to indicate that they are essential, whereas questions without a star are desirable. Those that are starred are critical for ensuring a country can report robust and comparable results for SDG 4.1.1, while those that are not starred are likely to assist governments in improving both the percent of learners meeting minimum proficiency standards and a country's overall assessment practice. The criteria include some technical requirements, and governments may decide to appoint a partner with expertise in psychometrics or test development to support them in collating the appropriate evidence.

Governments should use the form in **Annex C** to provide the evidence related to the reliability and validity of the assessment. The following sections provide a high-level overview of the technical criteria. Full details can be found in section 3: Technical Criteria.

Criterion 1 – Alignment between the assessment, the assessment framework, and the curriculum

For any high-quality assessment, it is essential that there is a clear link between what is taught and what is assessed as well as the criteria upon which *benchmarks* are set (in this case, the *GPF*). As discussed above, ideally, assessment systems are developed to meet the Standards for Educational and Psychological Testing (2014).[4] To summarize, as shown in Figure 1, this means that systems should ideally be designed beginning with government experts and partners working together to set the *content standards* before developing the curriculum and *assessment framework*, which should include *performance standards* based on what is being tested. Experts should then work to design an assessment based on the assessment framework. They should then have it

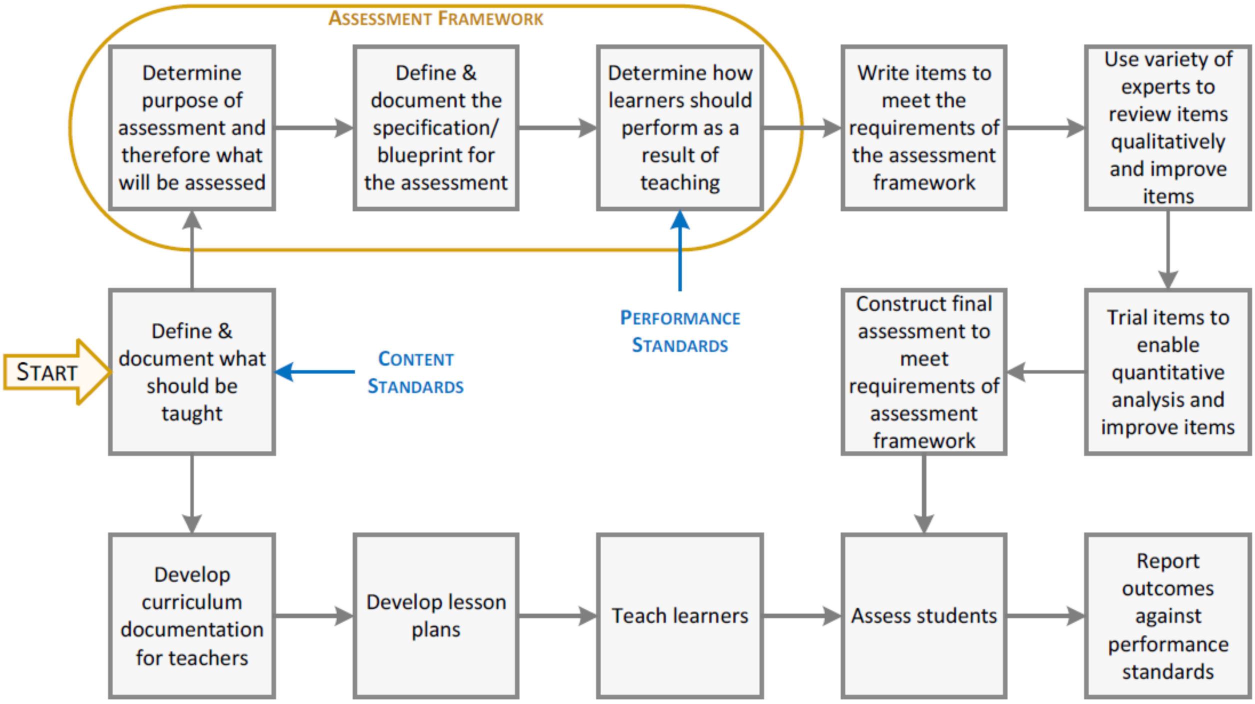
[2] American Educational Research Association, American Psychological Association, National Council on Measurement in Education & Joint Committee on Standards for Educational and Psychological Testing. (2014). Standards for educational and psychological testing. Washington, DC: AERA.

[3] Where possible, documents should be submitted in English. However, documents may be submitted in their original language if necessary and UNESCO-UIS will arrange for translations to be made to enable review by the 4.1.1 QAP.

[4] American Educational Research Association, American Psychological Association, National Council on Measurement in Education & Joint Committee on Standards for Educational and Psychological Testing. (2014). Standards for educational and psychological testing. Washington, DC: AERA.

qualitatively reviewed by a range of experts (including students with special educational needs and disabilities [SEND] experts, language experts, content and assessment experts, etc.) before the assessment is piloted and analyzed quantitatively. Finally, they should design the final assessment. Any subsequent assessments should also be designed using the same process, beginning with using the assessment framework to write the assessment items. They should then be statistically equated with the first assessment if the government wishes to compare results between the different assessments over time. If this is not possible, alternatively, governments can use policy linking with each different assessment to link them at the benchmarks.

Figure 1: Assessment System Design Process


ASSESSMENT FRAMEWORK
Determine purpose of assessment and therefore what will be assessed
Define & document the specification/ blueprint for the assessment
Determine how learners should perform as a result of teaching
Write items to meet the requirements of the assessment framework
Use variety of experts to review items qualitatively and improve items
START
Define & document what should be taught
CONTENT STANDARDS
PERFORMANCE STANDARDS
Construct final assessment to meet requirements of assessment framework
Trial items to enable quantitative analysis and improve items
Develop curriculum documentation for teachers
Develop lesson plans
Teach learners
Assess students
Report outcomes against performance standards

While the 4.1.1 QAP recognizes that not all systems will have been designed in this way, they do still expect to find alignment between the standards, the assessment, and the curriculum. And, for policy linking to work and to allow governments to use their assessments to report against SDG 4.1.1, that system must also align with the *GPF* standards. If there is a misalignment between the curriculum and the GPF, governments will still be able to use policy linking to report to SDG 4.1.1 as long as the assessment aligns with the GPF. However, this misalignment between the curriculum and the assessment means that learners may not perform as well on the assessment overall, especially if they are being assessed on knowledge and skills that they are not reasonably expected to have attained in the classroom. The first criterion is focused on examining this alignment.

The first criterion, therefore, requires evidence of alignment between the curriculum and/or national standards and the *content domain* of the assessment (usually depicted in the *assessment framework*, if one exists) and between the assessment items and the *GPF*. While the QAP will review the alignment between the curriculum and the assessment, countries may also want to go through an alignment study process on their own, using the same methodology as the GPF alignment study (see below). Doing so may highlight gaps in their systems that once addressed will likely improve the quality of reading and math outcomes.

Countries are expected, on the other hand, to complete an alignment study to align their assessment(s) and the *GPF*. Details about how to conduct an alignment study and a template for completing this study are included in the Policy Linking Toolkit. Governments will need to either engage their own reading and math content and grade-level experts or hire a partner with this expertise in the country context to complete this alignment study. If this is not possible, governments may, instead, submit the assessment(s) they wish to use for policy linking along with their national content standards and/or curriculum framework.

MATERIALS TO SUBMIT FOR CRITERION 1:

Preferred:

- Curriculum framework or national content standards
- Assessment framework, including the content domain
- Completed assessment(s) and the GPF alignment study (using X form)

Accepted if preferred option is not possible:

- Curriculum framework or national content standards
- Assessment instrument(s)

Criterion 2 – Assessment validity and appropriateness for the population

For UNESCO to have confidence in assessment outcomes, it is essential that the assessment has been determined to be a valid measure of reading and math skills and to be appropriate for the population taking the test, including students of different genders, those with special educational needs and disabilities (SEND), those of different ethnic or cultural backgrounds, those affected by crisis and conflict, those from rural and urban areas, and those living in poverty. Given that the outcomes of policy linking will be used to report against SDG 4.1.1, the cohort taking the test should be representative of the population against which the results will be reported – this is particularly important where a sampling approach is used. For instance, if a country wishes to report national results, they should use a national census or a sampling approach that is nationally representative.[5] They

MATERIALS TO SUBMIT FOR CRITERION 2:

Preferred:

- Item difficulty statistics (percent of students who get an individual item correct in the most recent assessment(s))
- Correlation between item scores and total scores on the assessment(s)
- Information on sampling methodology and sample frame (population being assessed)

Accepted if preferred option is not possible:

- Data from the most recent assessment(s)
- Technical (outcome) report from most recent assessment(s)

[5] 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.

should also report when some groups are not represented in the reported results, e.g., learners with SEND and/or children who are out-of-school. SDG 4.1.1 seeks to report reading and math outcomes for **all** children from the age groups expected to be in grades 2/3, end of primary, and early secondary school. As such, this should be the goal of governments wishing to report to SDG 4.1.1.

To report on this criterion, governments will need to engage either their own assessment experts or hire a partner with this expertise. The person(s) preparing the evidence for this criterion should, preferably, have been involved in the development of the assessment and in selecting the sample of learners/children assessed. They should also have experience with statistics or psychometrics. Some of the evidence for this criterion (especially the sampling methods) may be available in the country's most recent assessment outcomes report, but data on the validity and appropriateness of assessment items are more likely to be found in assessment design reports.

Criterion 3 – Reliability of the assessment

In order to have confidence in the stability of the assessment outcomes, it is essential that the assessment has appropriate levels of reliability, meaning that if the test were given again to another sample of students with a different set of enumerator or test proctors, results would be similar. Assessment results that fluctuate significantly based on who is administering the assessment or which specific students in a sample take the assessment are not reliable assessments.

There are many ways to estimate the reliability of an assessment, and these will vary with the nature of the assessment, in particular how it is administered and scored. To support countries in providing evidence, this criterion focuses on aspects of reliability that are relatively easy to determine.

To report on this criterion, countries will need to submit evidence on how their scoring is quality assured (this information is generally included in assessment outcome reports) and will also need to conduct some statistical tests on the actual data. To accomplish the latter, governments should either engage their own psychometricians or statisticians or appoint a partner with expertise in statistics and/or psychometrics to support them. Where this is not possible, governments may submit the raw data from their most recent assessment, and the data will be analyzed by the 4.1.1 QAP.

MATERIALS TO SUBMIT FOR CRITERION 3:

Preferred:

- Inter-rater reliability statistics (percent agreement between enumerators when scoring the same learner) OR details on quality assurance of scoring for close-ended items
- Results of a statistical test of assessment coherence (whether all items seek to measure the same thing – reading or math ability)

Accepted if preferred option is not possible:

- Data from the most recent assessment(s)
- Details on how enumerator or scorer reliability is determined

Stage 3: Review of evidence by the 4.1.1 Review Panel

The 4.1.1. Review Panel is the body appointed by UNESCO-UIS to review the rigor and quality of results reported for SDG 4.1.1 to ensure the validity and comparability of outcomes. The panel will make recommendations to UNESCO-UIS on whether a country is well placed to engage in policy linking (based on a review of the above criteria), and UNESCO will make the final decision about whether to recommend countries proceed with policy linking.

The panel is involved twice in the process: stage 3 – to review the reliability and validity of the assessments and their alignment to the *GPF* before the workshop (i.e. the evidence from stage 2); and stage 6 – to review the outcomes of the policy linking workshop (i.e. the evidence from stage 5).

During Stage 3, the 4.1.1 QAP will review the evidence and make one of three recommendations:

1) Assessment is suitable for policy linking; if this decision is made, the 4.1.1 QAP will also assign the assessment a grade of excellent, good, or sufficient:
   - **Excellent** - all criteria met with suitable responses for all questions
   - **Good** - all criteria reported on with suitable responses for all starred (*) questions, including adequate *inter-rater reliability* scores (criterion 3bii) during the administration/ scoring of the assessment
   - **Sufficient** - all starred criteria suitably met but adequate *inter-rater reliability* scores (criterion 3bii) only calculated during enumerator or rater training
2) More evidence required to confirm if assessment is suitable for policy linking
3) Assessment is not suitable for policy linking

For the final two categories, the assessment will be graded as ‘insufficient’ since insufficient criteria have been met to continue with policy linking for reporting to SDG 4.1.1**.** Governments will be allowed to submit further evidence to address any omissions identified by the 4.1.1 QAP if a grade of ‘insufficient’ is awarded.

Governments will submit the evidence collated in stage 2 to UNESCO-UIS, including completed forms from Annex C and any supporting information. Governments should be submit documents by March 31, June 30, September 31, or December 31 to receive a decision from UNESCO before the end of the next quarter on whether the country is well placed to move forward with policy linking. If countries wish to submit results from policy linking to report against SDG 4.1.1 in the current year, they will need to submit documents for Stage 3 by December 31 of the previous year, as shown in Table 1 below, as the deadline to submit results for SDG 4.1.1 is September 31 every year.

Table 1: Timeline for Submitting Results and Receiving Responses

Submission of Documents for Stage 3	Decision from the QAP and UNESCO (Stage 4)	Policy Linking Workshop (Stage 5)	Submission of Documents for Stage 6	Decision from QAP and UNESCO (Stage 7)
January	March 31	April – June	By June 30	September 31
February	March 31	April – June	By June 30	September 31
March	March 31	April – June	By June 30	September 31
April	June 30	July – Sept.	By Sept. 31	December 31

May	June 30	July – Sept.	By Sept. 31	December 31
June	June 30	July – Sept.	By Sept. 31	December 31
July	September 31	Oct. – Dec.	By Dec. 31	March 31
August	September 31	Oct. – Dec.	By Dec. 31	March 31
September	September 31	Oct. – Dec.	By Dec. 31	March 31
October	December 31	Jan. - March	By March 31	June 30
November	December 31	Jan. - March	By March 31	June 30
December	December 31	Jan. - March	By March 31	June 30

The 4.1.1. QAP will produce a report to explain the rationale for their decision, including stipulating any additional documents that a government must submit before moving forward with policy linking. UNESCO-UIS will review the report and recommendation before making a final decision. UNESCO-UIS will share the outcomes with the government concerned and confirm next steps.

Stage 4: Preparation for the policy linking workshop

Once an assessment has been determined to be suitable by the 4.1.1 QAP, the government can implement policy linking using the Policy Linking Toolkit. There are a number of activities that need to take place before the policy linking workshop, and sufficient time should be planned to undertake them. These activities include:

- Coordination between governments, donor organizations, and partner organizations
- Sourcing suitable venues and agreeing on logistical arrangements
- Identifying/recruiting both process and content facilitators
- Recruiting/inviting panelists
- Finalizing materials for use in the workshop.

Full details of the preparation activities required, and step-by-step instructions, can be found in the Policy Linking Toolkit.

MATERIALS TO SUBMIT FOR CRITERION 4:

Preferred:

- Details on panellist demographics/ qualifications
- Data on the ratings of each panellist across both rating rounds & statistics on intra-rater reliability, inter-rater reliability, and standard error of measurement (described in detail in Section 3)
- Data from the evaluation forms required of panellists

Accepted if preferred option is not possible:

- Just the above data, without the actual statistics

Stage 5: Implementation of the policy linking workshop and documentation of evidence on outcomes

The Policy Linking Toolkit provides step-by-step instructions for administering the policy linking workshop. Once the workshop has been completed, the government must submit evidence to support the validity of the policy linking workshop outputs.

Specifically, the government must submit evidence regarding the profile of each of the panelists and details to demonstrate that they meet the qualification criteria listed in the toolkit and are representative of the target population of schools being assessed. Governments also need to submit statistics to show individual panelist consistency of ratings, cross-panelist consistency of rating, accuracy of panelist ratings, and post-workshop panelist evaluation data on their understanding of the process and confidence in their ratings.

The criterion for this stage (criterion 4) is described in detail in the section 3, including questions for governments to answer. The process facilitators who lead the workshop should have the skills necessary to produce the statistics required for this stage. More details are available in the toolkit.

Stage 6: Review of workshop outcomes by the 4.1.1 QAP

Governments will submit the evidence collated in stage 5 to UNESCO-UIS. Please inquire how by emailing UIS.lo@unesco.org. Governments should be submit documents by March 31, June 30, September 31, or December 31 to receive a decision from UNESCO before the end of the next quarter on whether the results from the country's policy linking workshop can be used to report against SDG 4.1.1. If countries wish to report results to UNESCO for the current year, they will need to complete their policy linking workshop and submit results by June 30 of that year.

The 4.4.1 QAP will review the evidence and make one of three decisions:

1) Policy linking carried out appropriately and reported outcomes are validated; as with in Stage 2, the 4.1.1 QAP will also provide a grade for the adequacy of the policy linking workshop. Grades follow:
    a. **Excellent –** All six criteria are met.
    b. **Good –** Four of the six criteria are met, two of which must be criteria b and c.
2) More evidence required to confirm whether policy linking was carried out appropriately before outcomes can be validated
3) Policy linking not carried out appropriately and/or outcomes cannot be validated

The 4.1.1 QAP will produce a report to explain the rationale for their recommendation, including stipulating any additional documentation that must be submitted before they can recommend acceptance of the results by UNESCO. UNESCO-UIS will review the report and recommendation before making a final decision. UNESCO-UIS will share the outcomes with the Government concerned and confirm next/final steps.

Stage 7: Reporting of results against SDG 4.1.1

Once the outcomes of policy linking have been validated by the 4.1.1 QAP and accepted by UNESCO-UIS, the government can submit the data for reporting against SDG 4.1.1. Data will be reported with associated grades, assigned as follows:

- **Excellent –** Country received an "excellent" rating on both the suitability of the assessment used for policy linking and the adequacy of the policy linking workshop.

- **Good –** Country either received “good” ratings for both the suitability of the assessment and the adequacy of the policy linking workshop or a “good” rating for one and an “excellent” rating for the other.
- **Sufficient –** Country received a “sufficient” rating for the suitability of the assessment and a “good” or “excellent” rating for the adequacy of the policy linking workshop.

SECTION 3: TECHNICAL CRITERIA

This section includes the technical details of the criteria explained in Section 2 above. As described in that section, governments, in coordination with their partners, must provide responses to the first three criteria during stage 2 for review in stage 3 by the 4.1.1 QAP. Governments and their partners, if relevant, should then provide responses to the fourth criterion following stage 5 for review by the 4.1.1 QAP during stage 6.

Technical details and requirements for all four criteria are included below. Each table includes the questions for the relevant criterion by which the 4.1.1 QAP will judge responses and details on what materials countries must provide to fulfill the reporting requirement. As mentioned in Section 2, the stars indicate required materials.

Criterion 1

This criterion is related to stages 2 and 3. To demonstrate the necessary alignment between the curriculum, assessment and *GPF*, countries must provide responses to the questions listed in Table 1, with supporting evidence where appropriate.

Table 1: Criterion 1 Requirements

Question	Criteria	Materials
1a) Are the expectations for the grade/subject clearly defined in the curriculum?	Countries should have a curriculum framework that includes details on domains, constructs, subconstructs[6], and skills that are expected to be taught in classrooms by grade. Descriptors should be detailed enough to make it clear what should be taught.	Countries should provide a curriculum framework (or equivalent document)[7] that includes *content standards* such as, "Students in Grade 3 should fluently add and subtract numbers within 100."
1b) Is the *content domain* for the assessment clearly defined?	The assessment framework should make it clear what skills a test seeks to measure.	Countries should provide the assessment framework (or equivalent document), including details on the *content domain* such as, which domains (e.g. reading comprehension), constructs (e.g. retrieve information) and which sub-constructs (e.g. retrieve explicitly stated information) are assessed.

[6] It is not expected that all countries will make use of these terms within their curriculum frameworks, but rather that there is an attempt to detail the topics they expect to be taught.

[7] Where a country has a highly decentralized system in relation to curriculum arrangements, they should submit a small number of examples of local curricula that are considered broadly representative.

1c) Do the items in the assessment appropriately sample from the assessment *content domain* such that the assessment can be considered a comprehensive assessment of reading or mathematics as defined in the assessment framework?	The assessment items must match the requirements in the assessment specification, for example in terms of number of items required on each domain. This also means there must be a sufficient number of items present to measure the *content domain* effectively.	Countries should provide the assessment[8] and assessment specification and show how the two align. If there are deviations, they should be explained.
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 matches (or partially matches) any of the knowledge and/or skills for the relevant grade at each proficiency level. Once all items have been considered, a decision is made on whether sufficient subconstructs have been covered to agree there is alignment.	Countries should provide both the assessment and evidence of the alignment ratings for each item as well as the country's overall decision on alignment using the form provided in the toolkit, also included in Annex X.

[8] The 4.1.1 QAP is able to provide signed non-disclosure agreements, if requested by governments, to ensure the security of the assessment. The QAP also guarantees that the assessment and other provided materials will only be used for the purposes described in this policy.

Criterion 2

This criterion is related to stages 2 and 3. To demonstrate the appropriateness of the assessment, countries must provide responses to the questions listed in Table 2, with supporting evidence where appropriate.

Table 2: Criterion 2 Requirements

Question	Criteria	Materials
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).	Countries should provide details of how they qualitatively and quantitatively reviewed assessment items, including any outputs (e.g., facility or difficulty, and discrimination) from the data analysis.
2b) Have the items been reviewed to ensure fairness to all relevant subgroups of the population, including students with SEND?	The assessment should appear free from bias. Items should not ask questions about foreign concepts or concepts familiar to only some cultural, ethnic, ability, socioeconomic, or geographic groups. For instance, reading comprehension passages that discuss holidays that may be celebrated only by some groups or that discuss snow in countries that do not get snow would be inappropriate. Countries should also demonstrate what, if any, test adaptations they have made for students with SEND.	In addition to providing the assessment, countries must provide a general description of the population being assessed as well as a description of what steps they took to ensure fairness of the assessment to relevant subgroups. Where available, countries should provide outputs of any statistical analyses to compare subgroups (for example differential item functioning (DIF) analysis).

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 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 (i.e. in terms of sex or students with SEND). Where a sample-based approach is used, the margin of error should be 5 percent or less at the 95 percent confidence level (see footnote 5).	Countries must include a description of what population the assessment is meant to represent. In many cases this will be national, but it might also be nationally representative of the formal school population only, or it might be representative only of a specific language group in the country, etc. Countries should provide data from their Education Management Information System (EMIS) or other systems showing the total students in the target population and the breakdown of their demographics by sex, location, etc. Countries should also explain whether the assessment was meant to be a census or a sample-based assessment. If it is the latter, they should provide details on how they identified the sample size and selected their sample as well as how they weighted the data (including what weights they applied and why), if relevant, to ensure the data is representative of the population for which results will be reported (see footnote 5).
2d) If a sample is used, is the sample appropriately powered to detect reasonable differences over time?	Samples should be sufficiently powered to allow countries to capture changes in outcomes over time. The minimum detectable effect size (MDES) should have been calculated and thought through ahead of finalizing sample size calculations.	Countries must submit evidence of their MDES, including both the most recent assessment dataset as well as their power calculations. They must also provide evidence to show why they believe their MDES is low enough to effectively capture expected changes over time (e.g., past effect sizes over time).

Criterion 3

This criterion is related to stages 2 and 3. To demonstrate the reliability of the assessment, countries must provide responses to the questions listed in Table 3, with supporting evidence where appropriate. To support countries in providing evidence, this criterion focuses on aspects of reliability that are relatively easy to determine even though there may be more psychometrically appropriate conceptions of reliability for a particular assessment. As such, the QAP will also accept the other conceptions, as described in 3c if governments wish to submit additional details.

Table 3: Criterion 3 Requirements

Question	**Criteria**	**Materials**
3a)* Is the value of *coefficient alpha*[9] (see definition above) for the grade-level, subject assessment greater than or equal to 0.7?	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 assessment. These may also be provided, but this criterion will be judged on the value for the entire assessment.	Countries should provide the dataset from the most recent facilitation of the assessment, and, if possible, their calculation of *coefficient alpha*.
3b(i)* For paper-and-pencil assessments that contain selected-response items, how has the scoring been quality assured to ensure appropriate scores for each student? (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 reliability*)?	For paper-and-pencil assessments with *selected-response items*, there must be details on how the scoring has been quality assured either through backchecks, statistical validation methods, etc. 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 reliability* (IRR) 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	Countries must provide either details on how they have quality assured their scoring or evidence of *inter-rater reliability*, including data of multiple raters scoring the same assessments, or both (if relevant).

[9] Also known as *Cronbach’s alpha*

	they have examined IRR but only during enumerator/rater training.	
3c) Is there any additional evidence relating to the reliability of the assessment?	If alternative measures of reliability have been developed for the assessment, these should be provided. This may include estimates of *classification consistency* or *classification accuracy*.	Countries should provide copies of technical reports where reliability statistics are published.

Criterion 4

This criterion is related to stages 5 and 6. To demonstrate the robustness of the policy linking workshop and the reliability of the outcomes, countries must provide responses to the questions listed in Table 3, with supporting evidence where appropriate. Formulas for each of these criterion are included in the Policy Linking Toolkit.

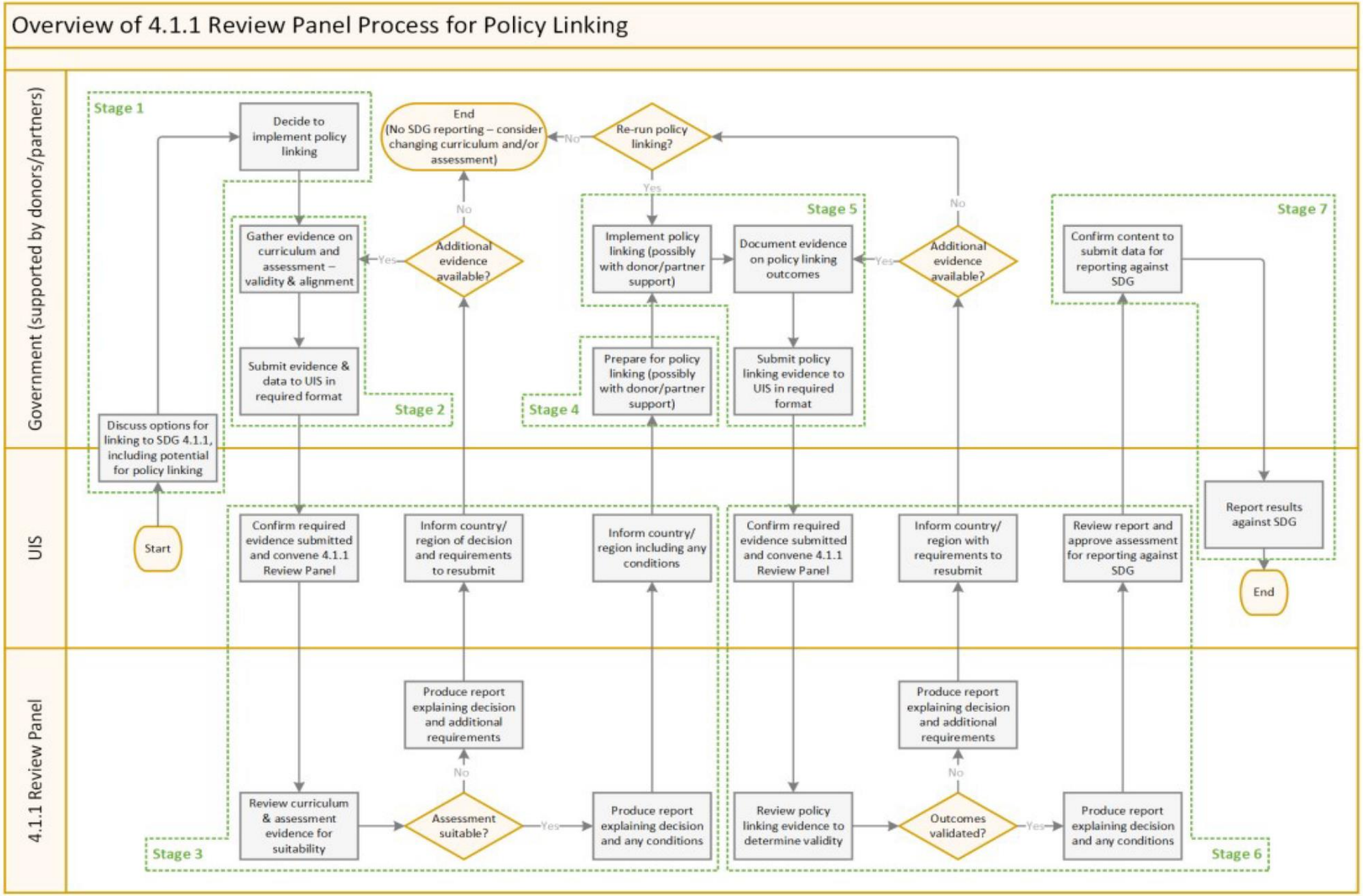
Table 4: Policy Linking Workshop Reliability and Validity Requirements

Question	Criteria	Materials
4a)* What was the *intra-rater reliability* for the second round of ratings?	The *intra-rater consistency will depend on the length of the assessment. Acceptable levels will be determined by the 4.1.1 Review Panel*.	Countries should provide statistics on intra-rater reliability as well as data that include the scores of each of the raters for both rounds of ratings. Each rater should be assigned a rater number so that his/her scores can be identified across rounds.
4b)* What was the *inter-rater reliability* for the second round of ratings?	The *inter-rater reliability* should be at least .80.	Countries should provide statistics on inter-rater reliability and the scores of each of the raters for both rounds of ratings.
4c)* What was the *Standard Error of Measurement* (SEM) at each *global proficiency level*?	*SEM* should be appropriate for each *global proficiency level* reported. There is no maximum *SEM* provided in this document, since it will depend on the number of items in the assessment.	Countries should provide the *SEM* and details of how the *SEM* was calculated (either using classical test theory or item response theory) and an explanation of why they believe this to be appropriate given the test features.
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	Countries should provide an explanation of what criteria they used to select panelists as well as demographic details about each of the panelists and how

	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.	they meet the requirements listed for this criterion.
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 learners of different proficiency levels, including at least some who would meet the requirements of the meets minimum proficiency level and some who would meet the requirements of the exceeds minimum proficiency level (all panelists) ● Knowledge of the instructional environment (all panelists) ● Experience administering the assessment(s) being used for the policy linking workshop.	Countries should provide demographic details about each of the panelists and how they meet the requirements listed under this criterion. Panelists should fill out workshop evaluation forms that include questions about their exposure to the assessment ahead of the workshop and during the workshop, assess their knowledge of the instructional environment, etc.
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.	Countries should share all panelist evaluation forms as well as a database of their Likert scale responses and average scores for each of the categories listed in this question.

ANNEXES

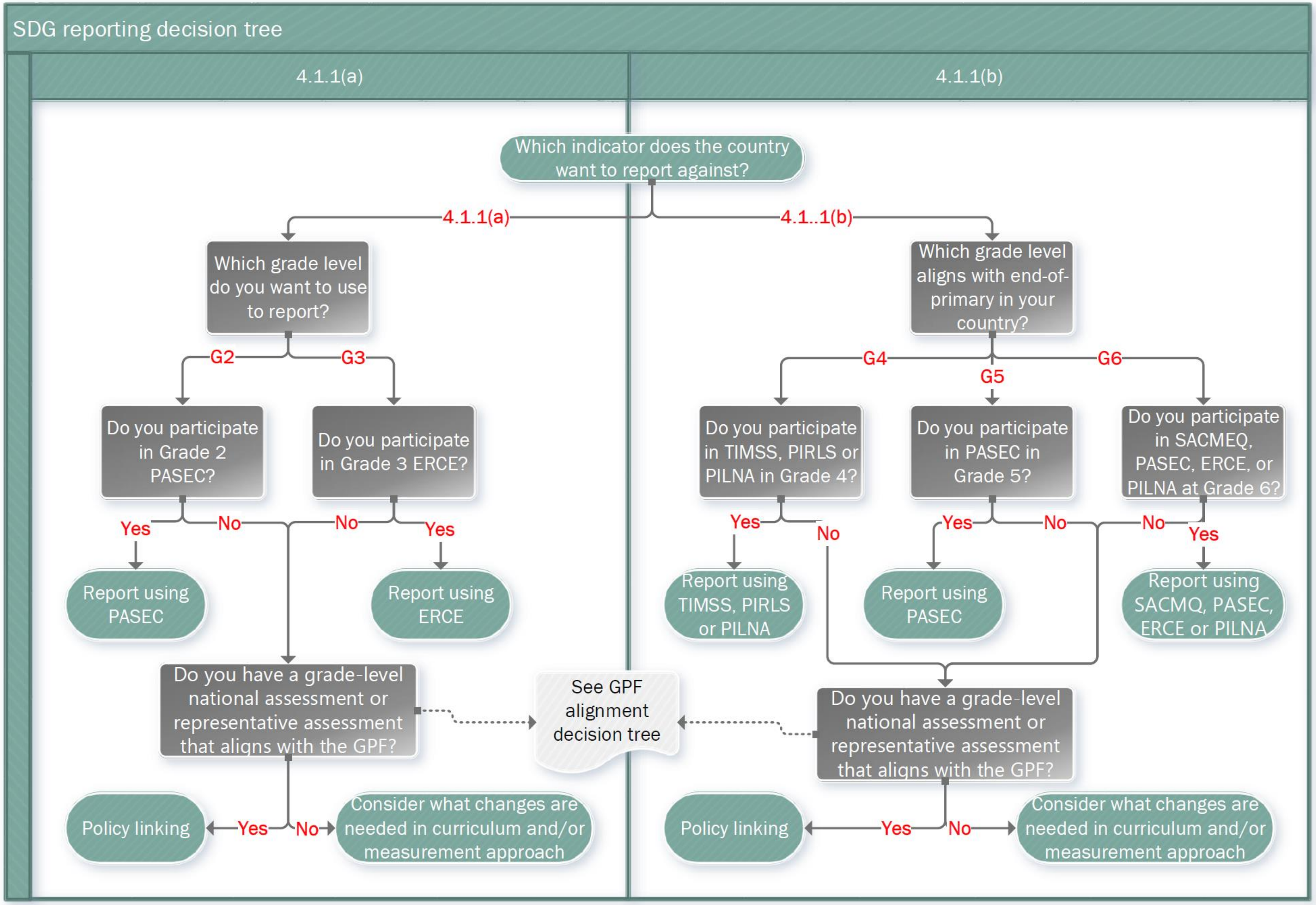
Annex A – Overview of 4.1.1 Review Panel Process

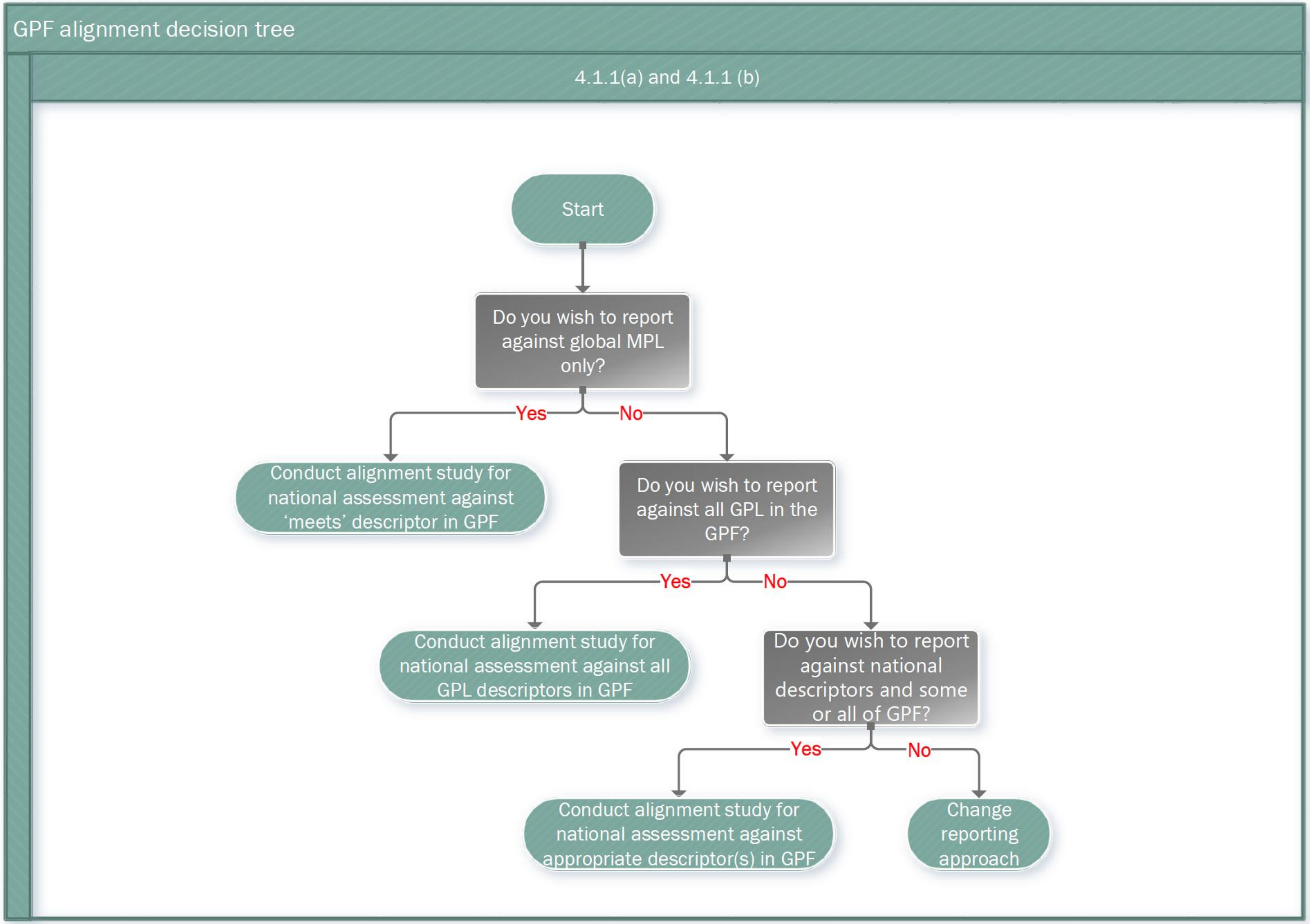


Overview of 4.1.1 Review Panel Process for Policy Linking

Government (supported by donors/partners)
UIS
4.1.1 Review Panel
Stage 1
Stage 2
Stage 3
Stage 4
Stage 5
Stage 6
Stage 7
Start
Discuss options for linking to SDG 4.1.1, including potential for policy linking
Decide to implement policy linking
Gather evidence on curriculum and assessment – validity & alignment
Submit evidence & data to UIS in required format
Confirm required evidence submitted and convene 4.1.1 Review Panel
Review curriculum & assessment evidence for suitability
Assessment suitable?
No
Yes
Produce report explaining decision and additional requirements
Inform country/ region of decision and requirements to resubmit
Additional evidence available?
Yes
No
End (No SDG reporting – consider changing curriculum and/or assessment)
Re-run policy linking?
No
Yes
Produce report explaining decision and any conditions
Inform country/ region including any conditions
Prepare for policy linking (possibly with donor/partner support)
Implement policy linking (possibly with donor/partner support)
Document evidence on policy linking outcomes
Submit policy linking evidence to UIS in required format
Confirm required evidence submitted and convene 4.1.1 Review Panel
Review policy linking evidence to determine validity
Outcomes validated?
No
Yes
Produce report explaining decision and additional requirements
Inform country/ region with requirements to resubmit
Additional evidence available?
Yes
No
Produce report explaining decision and any conditions
Review report and approve assessment for reporting against SDG
Confirm content to submit data for reporting against SDG
Report results against SDG
End

Annex B – Decision tree for implementing policy linking


SDG reporting decision tree
4.1.1(a)
4.1.1(b)
Which indicator does the country want to report against?
4.1.1(a)
4.1..1(b)
Which grade level do you want to use to report?
G2
G3
Do you participate in Grade 2 PASEC?
Do you participate in Grade 3 ERCE?
Yes
No
No
Yes
Report using PASEC
Report using ERCE
Do you have a grade-level national assessment or representative assessment that aligns with the GPF?
See GPF alignment decision tree
Policy linking
Yes
No
Consider what changes are needed in curriculum and/or measurement approach
Which grade level aligns with end-of-primary in your country?
G4
G5
G6
Do you participate in TIMSS, PIRLS or PILNA in Grade 4?
Do you participate in PASEC in Grade 5?
Do you participate in SACMEQ, PASEC, ERCE, or PILNA at Grade 6?
Yes
No
Yes
No
No
Yes
Report using TIMSS, PIRLS or PILNA
Report using PASEC
Report using SACMQ, PASEC, ERCE or PILNA
Do you have a grade-level national assessment or representative assessment that aligns with the GPF?
Policy linking
Yes
No
Consider what changes are needed in curriculum and/or measurement approach


GPF alignment decision tree
4.1.1(a) and 4.1.1 (b)
Start
Do you wish to report against global MPL only?
Yes
No
Conduct alignment study for national assessment against 'meets' descriptor in GPF
Do you wish to report against all GPL in the GPF?
Yes
No
Conduct alignment study for national assessment against all GPL descriptors in GPF
Do you wish to report against national descriptors and some or all of GPF?
Yes
No
Conduct alignment study for national assessment against appropriate descriptor(s) in GPF
Change reporting approach

Annex C – Assessment quality evidence

General information

Country	
Region (if not whole country)	
Language(s) of administration	
Assessment	
Subject	Reading / Mathematics *[delete as appropriate]*
Grade	
Cohort	Sample / Census *[delete as appropriate]*
Number of items/marks in assessment	
Have copies of the most recent assessment instruments been provided?	Yes / No *[delete as appropriate]*
Has the most recent data set from the assessment been provided?	Yes / No *[delete as appropriate]*
Has a technical report on the most recent assessment been provided?	Yes / No *[delete as appropriate]*

Quality evidence

<table>
<tr><th>Criterion 1</th><th>Alignment between the assessment and the curriculum</th></tr>
<tr><td>1a Is there a common curriculum for all pupils taking part in the assessment</td><td>Yes / No *[delete as appropriate]*
[Where different local curricula are in place, please indicate what work, if any, has been undertaken to consider the alignment of the different curricula and how this may affect student performance]</td></tr>
<tr><td>1a Are the expectations for the grade/subject clearly defined in the curriculum?</td><td>Yes / No *[delete as appropriate]*
[Please provide a copy of the relevant curriculum documentation (or a representative selection where a common curriculum is not in place) for the grade and subject]</td></tr>
<tr><td>1b Is the content domain for the assessment clearly defined?</td><td>Yes / No *[delete as appropriate]*
[Please provide a copy of the relevant assessment documentation for the grade and domain]</td></tr>
<tr><td>1c Do the items in the assessment appropriately sample from the assessment content domain such that the assessment can be considered a comprehensive assessment of reading or mathematics as defined in the assessment framework?</td><td>Please complete the following table:
<table>
<tr><th>Content domain area</th><th>Number of items required in specification (if appropriate)</th><th>Number of items in assessment</th></tr>
<tr><td><i>Example – calculation</i></td><td><i>15-20</i></td><td><i>18</i></td></tr>
<tr><td><i>Example – retrieval of simple information</i></td><td><i>10</i></td><td><i>10</i></td></tr>
<tr><td></td><td></td><td></td></tr>
</table>
[Please delete examples and add additional rows to the table as required]</td></tr>
<tr><td>1d* Is the assessment aligned with the Global Proficiency Framework?</td><td>Yes / No *[delete as appropriate]*
[Please provide a copy of the outputs of the alignment study to support this assessment]</td></tr>
</table>

Criterion 2	Appropriateness of the assessment for the population
2a* Is there evidence that the items in the assessment have been reviewed qualitatively and quantitatively before administration to determine their appropriateness for the population?	Yes / No *[delete as appropriate]* *[Please provide details of the process used to review the items qualitatively and/or quantitatively. This could include details of any reviews of the items by teachers prior to their administration or data analysis conducted on trials/pilots of the items. Please also include information on how the outcomes from any qualitative or quantitative reviews were fed into the item development process]*
2b Have the items been reviewed to ensure fairness to all relevant subgroups of the population, including students with SEND?	Yes / No *[delete as appropriate]* *[Please provide details of the process used to review the items qualitatively and/or quantitatively for fairness to all relevant sub-groups of the population. This could include details of any reviews of the items by cultural experts or inclusion experts prior to their administration or data analysis conducted on trials/pilots of the items to consider potential bias. Please also include information on how the outcomes from any qualitative or quantitative reviews were fed into the item development process]*
2c* Is the cohort that took the assessment representative of the population against which results will be reported?	Yes / No *[delete as appropriate]* *[For census assessments, please indicate how the population was determined e.g. in-school population, specific language group.* *For sample-based assessments, please indicate the sampling methodology used, including how the sampling frame was determined and the sampling approach e.g. stratified random sample of all government schools. Please include details of any strata and/or weightings used and provide details of the margin of error and confidence level]*
2d If a sample is used, is the sample appropriately powered to detect reasonable differences over time?	Yes / No *[delete as appropriate]* *[Please include the minimum detectable effect size (MDES) and the power calculations. Please also provide evidence to demonstrate that the MDES is sufficiently small to effectively capture changes over time]*

Criterion 3	Reliability of the assessment
3a* Is the value of coefficient alpha for the grade-level, subject assessment greater than or equal to 0.7?	Yes / No / Not calculated *[delete as appropriate]* Value of alpha: __________________ *[Please include any calculated values of alpha for individual components of the assessment if available]*
3b(i)* For paper-and-pencil assessments that contain selected-response items, how has the scoring been quality assured to ensure appropriate scores for each student?	*[Please provide details of how the scoring has been quality assured e.g. through backchecks/sampling, statistical validation etc.]*
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 reliability)?	*[Please provide details of how the scoring has been quality assured e.g. through backchecks/sampling, statistical validation etc. either in the field or during enumerator/rater training]* Value of kappa: __________________ (or equivalent – please indicate statistic)
3c (optional) Is there any additional evidence relating to the reliability of the assessment?	*[Please provide details of any alternative measures of reliability that have ben developed for the assessment e.g. classification consistency or classification accuracy]*

For completion by 4.1.1 Review Panel

Recommendation

Recommendation	Tick as appropriate
Assessment is suitable for policy linking	
More evidence required to confirm suitability for policy linking	
Assessment is not suitable for policy linking	

Evaluation against criteria

Criterion	Recommendation *[delete as appropriate]*	Comments
1a	Met / Not met	
1b	Met / Not met	
1c	Met / Not met	
1d*	Met / Not met	
2a*	Met / Not met	
2b	Met / Not met	
2c*	Met / Not met	
2d	Met / Not met	
3a	Met / Not met	
3b (i)/(ii)*	Met / Not met	
3c	Met / Not met	

Grade

Recommendation	Tick as appropriate
Excellent (all criteria met)	
Good (all starred criteria met – including 3b(ii) during administration[10])	
Sufficient (all starred criteria met – including 3b(ii) during training[8])	
Insufficient (not all starred criteria met)	

[10] Only for assessments where oral responses or constructed responses are included

Recommendations where a grade of ‘insufficient’ has been awarded