United Nations
Educational, Scientific and
Cultural Organization

UNESCO INSTITUTE FOR STATISTICS

SDG Indicator 4.1.1:

Inputs to the Measurement and Reporting Strategy

Proposal by GAML Task Force 4.1

Global Alliance for Monitoring Learning
Fourth meeting
28-29 November 2017
Madrid, Spain

GAML4/11

Introduction

Sustainable Development Goal (SDG) Target 4.1 focuses on free, equitable, and quality primary and lower-secondary education. The Global Indicator (4.1.1) for Target 4.1 is the "*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*".

Task Force 4.1 of the Global Alliance to Monitor Learning (GAML) was convened to support UNESCO Institute for Statistics (UIS) in thinking through the measurement issues involved in reporting against this indicator and to help them come up with practical solutions. The Task Force's deliberations have run parallel to other work on these issues by UIS and its technical partners and other stakeholders. **The objective of this Task Force 4.1 document is to serve as an input to this ongoing work program[1] by offering Task Force member insights and recommendations on some of the key measurement and reporting challenges.**

Key Challenges and Work Program for Indicator 4.1.1

There are several key challenges involved in measuring and reporting on reading and mathematics outcomes at the global level. These include mapping the content coverage of different assessments onto a common framework (in the absence of a common assessment instrument); developing a relevant learning scale; ensuring a certain level of data quality across assessments; establishing a coherent reporting metric; agreeing on the level of achievement that qualifies as "minimum proficiency" in different national contexts; and building country capacity to produce the needed data and manage financial and human resource allocation.

Task Force 4.1 addressed these challenges in relation to three key phases in an assessment work program:

1. Conceptual framework: Who and what to assess?
2. Methodological framework: How to assess?
3. Reporting framework: How to report?

A summary of Task Force discussions and conclusions for each phase is described in the rest of this report. In the course of these discussions, **Task Force members also arrived at some overall recommendations for next steps in the Indicator 4.1.1 work program. These included:**

- The GAML Secretariat/UIS should convene a group of reading and mathematics content experts, developmental psychologists, assessment experts, and others who can bring the latest research, evidence, and data to bear on the drafting of the longer-term measurement strategy for Indicator 4.1.1, particularly Indicator 4.1.1a. This group of experts should be diverse in terms of regions, languages, and scripts.
- Countries need to be more actively brought into the discussions on Indicator 4.1.1 to ensure that the proposed measurement and reporting approaches are sufficiently adaptive and responsive to their contexts. It's unclear, however, whether GAML is the context in which these country consultations should take place.

[1] For example, see the August 2017 document, *"SDG Data Reporting: Proposal of a Protocol for Reporting Indicator 4.1.1",* as well as the summary report for the technical expert meeting held in Hamburg, Germany from September 20-22, 2017. These documents outline interim and longer-term approaches to measurement and reporting for Indicator 4.1.1.

1. Conceptual Framework: Who and What to Assess?

The general view of Task Force 4.1 members was that existing national and cross-national assessments should provide the basis for determining who and what to assess at the three key measurement points. Table 1 provides an overview of some of these national and cross-national assessment options.

Table 1. Overview of some existing national and cross-national options for who and what to assess

National Assessments	Cross-National Assessments			Related GAML/ UIS Activities	Expected Outcomes/ Product	Timeline
	Name (# countries)	*Reading*	*Mathematics*			
Mapping of national assessment frameworks for mathematics and reading (in progress)	LANA	Grade 4-6	Grade 4-6	Map national assessment frameworks	Content Reference Frameworks (Math, Reading)	2017
	LLECE (15)	Grade 3/6	Grade 3/6	Map assessment characteristics and use of assessment data	Catalogue of Learning Assessment	2017
	PASEC 2014 (10)	Grade 2/6	Grade 2/6			
	PILNA (13)	Grade 4/6	Grade 4/6			
	PISA 2015 (72)	Age 15	Age 15			
	PISA-D 2018 (8)	Age 15	Age 15			
	PIRLS 2016 (61)	Grade 4				
	SEA-PLM 2018 (11)	Grade 5	Grade 5			
	SACMEQ IV (14)	Grade 6	Grade 7			
	TIMSS 2015 (57)		Grade 4/8			

Product was formally reviewed by Task Force 4.1 members.

Most of the national and cross-national assessments shown in Table 1 are designed to provide grade-based data on reading and mathematics performance that is relevant to measurement points 4.1.1b and 4.1.1c. A Task Force subgroup was convened to discuss options for 4.1.1a in more depth. The subgroup determined that assessment at this level needed to focus on the precursor and early reading and mathematics skills required for future academic learning and that key facets of performance to be considered should include accuracy, comprehension, and automaticity/speed. It was recognized, however, that very few cross-national assessment programs currently measure these precursor and early skills (e.g., LLECE/TERCE in grade 3; PASEC in grade 2).

There was general agreement among Task Force members that ongoing **GAML/UIS activities to map assessment frameworks and capture the characteristics and uses of assessment data should continue, but with more focus on ensuring that these efforts include attention to grades 2/3 assessments in reading and math (4.1.1a).** Given the relative dearth of national and cross-national assessment options for the early grades, the Task Force 4.1 subgroup also suggested that countries consider drawing on early-grades reading assessment (EGRA), early-grades mathematics assessment (EGMA), household-based (e.g., Multiple Indicator Cluster Survey [MICS]), and citizen-led (e.g., Annual Status of Education Report [ASER] and UWEZO) tools that measure key aspects of the constructs of early reading and math. **For the longer-term, the subgroup recommended developing a set of purpose-built tools that countries could draw on/adapt.**

One of the key outputs produced by UIS and its technical partners to support countries' efforts in reporting against Indicator 4.1.1 is a set of Content Reference Frameworks for mapping mathematics and reading assessment frameworks. Task Force 4.1 members were invited to submit feedback on the draft Content Reference Framework for mathematics. **Task Force 4.1 members signaled general agreement with the approach taken to developing the mathematics content reference framework, but also:**

- Concern about the possible influence of a restricted number of language groups and cultures on the current version of the framework in terms of the relevance and adequacy of its domain coverage.
- Requests for a more explicitly research-based approach to delineating key subdomains/levels and for the inclusion of more concrete examples for each sublevel.
- Recommendations to apply the framework to a greater variety of national assessment frameworks in order to further refine and validate it. This might include determining how well the reference framework applies to a national assessment in a top-ranked PISA country and whether certain aspects of national assessment frameworks are deemed unsuitable for the Content Reference Framework (and why).
- Requests to provide more information on how the framework might be adapted over time.

The draft Content Reference Framework for reading ("*Method for Developing an International Curriculum and Assessment Framework for Reading and Writing*") was circulated at a much later date. Task Force 4.1 members were not formally required to submit feedback on this framework, but were invited to do so if they had time. **Submissions by Task Force members on the draft reading framework methodology paper can be summarized as follows:**

- Appreciation for the presentation of reading interlinked with writing as part of the broader construct of literacy.
- Recommendation that given Indicator 4.1.1's focus on reading, that aspect of the literacy construct should be emphasized moving forward.
- Recommendation to extend/test the framework against other languages, apart from alphabetic and European; at the very least, the framework should be tested against the remaining United Nations' languages of Arabic, Chinese, and Russian.
- Concern that the framework is based on the perspective of one discipline (psychology) and one school of thought within that discipline (cognitive psychology) and that other perspectives and evidence bases (e.g., linguistic and sociolinguistic) should be incorporated.
- Request for a more explicitly research-based approach to constructing the Content Reference Framework and for more specialist input.

2. Methodological Framework: How to Assess?

Task Force 4.1 members did not focus as much on the "how to assess" aspect of indicator 4.1.1, which seemed to end up falling more under the purview of the Assessment Implementation Task Force. Task Force 4.1 was not formally requested to review any of the technical outputs in this area. Table 2 provides an overview of some national and cross-national assessment approaches. Most emphasize sample-based and group-administered approaches, and also primarily focus on children and youth in school.

Table 2. Overview of some existing national and cross-national options for how to assess

National Assessments	**Cross-National Assessments***		
	Name	*OOSC*	*Individual/group administration*
UIS Catalog of Learning Assessments to provide information on methods used by national assessments, including sampling, administration, and quality checks	LANA	No	Group
	LLECE	No	Group
	PASEC 2014	No	Both
	PILNA	No	Group
	PISA 2015	No	Group
	PISA-D 2018	Yes	Both
	PIRLS 2016	No	Group
	SEA-PLM 2018	No	Group
	SACMEQ IV	No	Group
	TIMSS 2015	No	Group

* All sample-based

GAML/UIS Activities	**Expected Outcomes/Products**	**Timeline**
Develop and compile good practices in learning assessment	Good Practices for Learning Assessment Manual	2017
Evaluate alignment in assessment content	Content Alignment Framework	2017-2018
Evaluate data collection process	Data Quality Framework	2017-2018

A Task Force 4.1 subgroup was convened to discuss 4.1.1a measurement approaches in more depth. Most early-years assessments are designed for one-on-one administration. EGRA, EGMA, and all household-based and citizen-led assessments use one-on-one approaches for this age/grade level, although school-based assessments employ a mix of one-on-one and group-administered approaches. In the short-term, it was felt that all of these should be viewed as options for countries to consider. **In the longer-term however, it was noted that there might be value in moving towards more school-based and group-administered approaches given the attendant savings in cost, time, and efficiency.**

Three key "how to" issues were addressed by Task Force 4.1 members during their virtual group discussions. These included:

How to include out-of-school children in measurement and reporting?

Task Force 4.1 members discussed whether and how to adjust school-based assessment results for countries with sizeable out-of-school populations as a way to reduce the bias produced by a non-representative sample. This included discussion on whether citizen-led assessments could be used to complement school-based assessment in such contexts given their coverage of both in- and out-of-school populations. Task Force 4.1 members felt that countries with this issue could be encouraged, but not required, to report data from their citizen-led assessments (if available) as an additional source of information on learning levels. In addition, countries could be encouraged to report the percentage of their student populations that are actually in school, but this statistic should not be used to adjust

assessment results for (or otherwise 'punish') countries, at least not in the first phase of reporting under 4.1.1.

How to determine "minimum acceptable requirements" for assessment data?

Suggested requirements ranged from very specific technical and psychometric criteria (e.g., reliability and validity coefficients, sample size requirements) to more content-related requirements regarding the breadth and appropriateness of the content being assessed. Task Force members noted, however, that it would not be fair to prescribe very precise technical criteria that countries are unlikely to have been aware of ahead of time. Instead, "minimum acceptable requirements", at least initially, would be more along the lines of ensuring that the submitted data are nationally representative and consistent with the national curriculum/standards. Evidence that the data are comparable over time also would be critical. More detailed technical and psychometric criteria could be used as a basis for country capacity building and system strengthening over time. It also was suggested that UIS request countries to submit their data sets, in addition to their assessment instruments, as part of the validation process. Reporting of results would then be accompanied by a "report card" of sorts on the quality of the underlying data. This would signal to the global community the extent to which the data could be "trusted" while at the same time providing a basis for countries and donor partners to determine capacity building needs.

How to decide which assessment data should be used for reporting?

This issue is likely to come up for countries that have participated in international and/or regional assessments in addition to their own national learning assessment. Task Force members considered whether countries should be given the freedom to choose which assessment data to report, or whether the decision should be made more centrally.

The sense among Task Force 4.1 members was that it would be important to be flexible on these and other decisions early on and focus more on encouraging countries to get in the habit of submitting data on learning. At the same time, efforts should be made to create incentives for countries to participate more systematically in international and regional assessments. From the UIS perspective, it would make sense to have a standardized protocol for making decisions about which of the data sources available for certain countries should be used for reporting against indicator 4.1.1. If all of the assessments meet basic technical quality requirements, then perhaps UIS could let countries choose which to use?

3. Reporting Framework: How to Report?

Table 3 provides an overview of some national and cross-national assessment reporting options. Most cross-national assessments convert raw scores to scaled scores using IRT approaches. In general, results are reported both in the form of scaled scores and/or as the percentage of students reaching specific proficiency levels or benchmarks on the scale. Each proficiency level tends to be accompanied by a description of what students at these levels are likely to know and be able to do. At the national level, the situation is more varied. Many national assessments, particularly in developing contexts, still report results as a mean raw score or percentage. Many do not have proficiency level descriptors or any benchmark for what constitutes "minimum proficiency".

Table 3. Overview of some existing national and cross-national options for how to report

National Assessments	**Cross-National Assessments**	
	Name	Proficiency Levels (#)
UIS Catalog of Learning Assessments to provide information on reporting methods used by national assessments, including use of scales, proficiency levels, and other benchmarks	LANA	To be determined
	LLECE	5
	PASEC 2014	4 (numeracy); 5 (literacy)
	PILNA	9
	PISA 2015	6
	PISA-D 2018	6
	PIRLS 2016	4
	SEA-PLM 2018	To be determined
	SACMEQ IV	8
	TIMSS 2015	4

GAML/UIS Activities	**Expected Outcomes/Products**	**Timeline**
Define indicators and metadata	Glossary of Common Language and Terminology	2017-2018
Develop reporting protocol	Interim reporting	2017
Develop UIS reporting scale	Learning Progression Explorer and Reporting Scale	2017-2019
Benchmark and define minimum proficiency level	Proficiency Level Definition	2018

Product was formally reviewed by Task Force 4.1 members.

Task Force members noted that a key challenge in reporting, particularly in relation to Indicator 4.1.1a, was comparability across systems and languages. Early-years assessments tend to focus on precursor or early reading and math skills. If these instruments have to be translated into different languages, it can affect their relative difficulty and hence the comparability of results. Because of this, some early-years assessments (e.g., EGRA) avoid comparing results (e.g., precursors, fluency measured in words correct per minute) across languages and others (e.g., MICS) focus on skills that are less affected by differences across languages (e.g., accuracy, comprehension). Task Force member suggestions included: (i) possibly using a hybrid approach of translation and adaptation to balance the relative difficulty of instruments across languages and enhance comparability, and (ii) prioritizing comparisons within languages, at least to start with.

One of the key outputs produced by UIS and its technical partners to support countries in reporting against Indicator 4.1.1 are the UIS Reporting Scales for mathematics and reading. **Task Force 4.1**

members were invited to submit feedback on the draft UIS Reporting Scale Concept Note (July 2017 version). Task Force member feedback can be summarized as follows:

- Recognition of the huge amount of work that had gone into developing the reporting scale, but, at the same time, noting some serious conceptual issues:
    - Whether such a scale is even required – Indicator 4.1.1 does not refer to a metric per se
    - Whether such a scale could ever truly allow for comparisons of student outcomes across countries
    - Whether such a scale might inadvertently dominate 4.1.1 discussions entirely, excluding a focus on the more important, broader learning agenda
- Task Force members also voiced concern about the lack of clarity regarding the relationship between the UIS Reporting Scales and the Content Reference Frameworks and requested further clarification on how these would work in unison.
- The Task Force was divided as to whether work on the scale should proceed or if alternatives should be sought.
    - Those in favor of continuing work on the scale suggested being clearer about the objective and target audience; e.g., is this primarily a "formative tool" for education systems to assist in monitoring and developing educational quality, or is it primarily a tool for international reporting?
    - Those in favor of alternatives to the current scale suggested:
        - Using a more traditional reporting scale that uses descriptors (such as below basic, basic, proficient, and advanced) to describe different achievement levels. This would involve first agreeing on the scale and proficiency levels/descriptors against which student performance should be measured, then identifying the instruments or items that fit the respective levels, and then dealing with the empirical part.
        - A methodology that allows for comparisons across assessments at each of the three points (4.1.1a, 4.1.1b, and 4.1.1c), but not necessarily spanning/connecting the three points.
        - Giving more attention to further development of existing cross-national assessments, in order to use these as a stepping stone for capacity strengthening and development of national assessments in countries.
- Task Force members were generally supportive of the proposed empirical approach to validating the UIS Reporting Scale and offered the following additional suggestions:
    - Provide more detail on the country-level implementation workplan.
    - Ensure that in-country Task Teams include teacher union representatives and academics as well as specialists.
    - Carry out the proposed work in close cooperation with existing cross-national assessment programs and give a prominent role to regional assessment programs.
    - Ensure that key stakeholders (including international and regional assessment organizations) have an opportunity to review the scale once it has been prepared.
    - Consider how to address the potential risks incurred by using larger countries as "representative" of any particular region, perhaps by using a regional rather than a country-level approach in instances such as Oceania.

- Conduct the assessments needed to validate the scale in countries where international and regional assessments have already taken place – this would reveal how the scales perform in different country contexts with the same assessments, and whether the performance levels match up across countries.
- Explore a test-based linking exercise for each of measurement points 4.1.1a, 4.1.1b, and 4.1.1c, instead of an item-based linking exercise.

Another of the key outputs produced by UIS and its technical partners to support countries' efforts in reporting against Indicator 4.1.1 is guidance for the Setting Benchmarks on the UIS Reporting Scale. **Task Force 4.1 members were invited to submit feedback on the draft Concept Note on Setting Benchmarks on the UIS Reporting Scale (June 2017). Task Force member feedback on the draft Concept Note can be summarized as follows:**

- Should there be global or national "minimum proficiency" benchmarks on the scale?
  - There was an even split among Task Force 4.1 members on this issue, with similar numbers in favor of each option.
- Should there be 1 or 3 "minimum proficiency" benchmarks per domain (i.e., mathematics and reading)?
  - The overwhelming majority of Task Force members were in favor of 3 benchmarks per domain; i.e., a "minimum proficiency" benchmark for each of the three measurement points – 4.1.1a, 4.1.1b. and 4.1.1c.
- Should existing "minimum proficiency" benchmarks be adopted or should new benchmarks be set?
  - Task Force members offered arguments in favor of both options. There were slightly more Task Force members in favor of adopting existing "minimum proficiency" benchmarks, although there was also recognition that over time there might be a need to set more customized benchmarks as a result of lessons learned from countries' data and experiences.
- Should there be global or national performance expectations for the percentage of students expected to reach "minimum proficiency"?
  - There was an even split among Task Force 4.1 members on this issue, with similar numbers in favor of each option.
- Should there be status- or progress-based expectations for the percentage of students expected to reach "minimum proficiency"?
  - Task Force members offered arguments in favor of both options. However, more Task Force members were in favor of having status-based expectations for the percentage of students expected to reach "minimum proficiency".

An overriding concern of Task Force Members was how to ensure that the benchmarking and other reporting strategies adopted for Indicator 4.1.1 would optimize the relevance and utility of the results for schools. In other words, how can we ensure that the results will have meaning for schools and that they will be able to take action on them?