



United Nations
Educational, Scientific and
Cultural Organization



UNESCO
INSTITUTE
FOR
STATISTICS

SDG Data Reporting

Proposal of a Protocol for reporting Indicator 4.1.1

August 2017



Table of Contents

1. Objectives	3
2. Key considerations in interim reporting.....	5
Conceptual framework.....	6
Assessment framework	6
Target population	6
Methodological framework	7
Test design	7
Sampling frame.....	8
Operational design and data generation	8
Data analysis	9
Contextual information	10
Reporting framework	10
Defining scales	10
Benchmarking.....	10
Longitudinal equating.....	11
3. Choosing an assessment program for 4.1.1 reporting	11
4. Format for reporting	15
5. UIS validation process with countries.....	17
Educational cycle: what grades are the defined point of measurement of Indicator 4.1.1?	17
2018 Proposed Metadata - Indicator 4.1.1.	20



1. Objectives

This document aims to inform UNESCO Institute of Statistics (UIS) reporting decisions, and guide discussion at the meeting with regional and international assessments in Hamburg, on September 20-22 2017. The purpose of the discussion is to agree on **interim reporting procedures** for Sustainable Development Goal (SDG) Indicator 4.1.1, in the context of GAML's broader work program for Indicator 4.1.1 reporting:

Indicator 4.1.1: *Proportion of children and young people: (a) in grades 2/3; (b) at the end of primary; and (c) at the end of lower secondary achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex.*

The UIS and its technical partners are developing a process to support international consistency in reporting against Indicator 4.1.1. A key component of this approach is the development of the UIS Reporting Scales (UIS RS), to enable alignment between assessment programs that measure the same domains. The UIS RS will enable countries to pursue different options, depending on the assessment program they choose for 4.1.1 reporting:

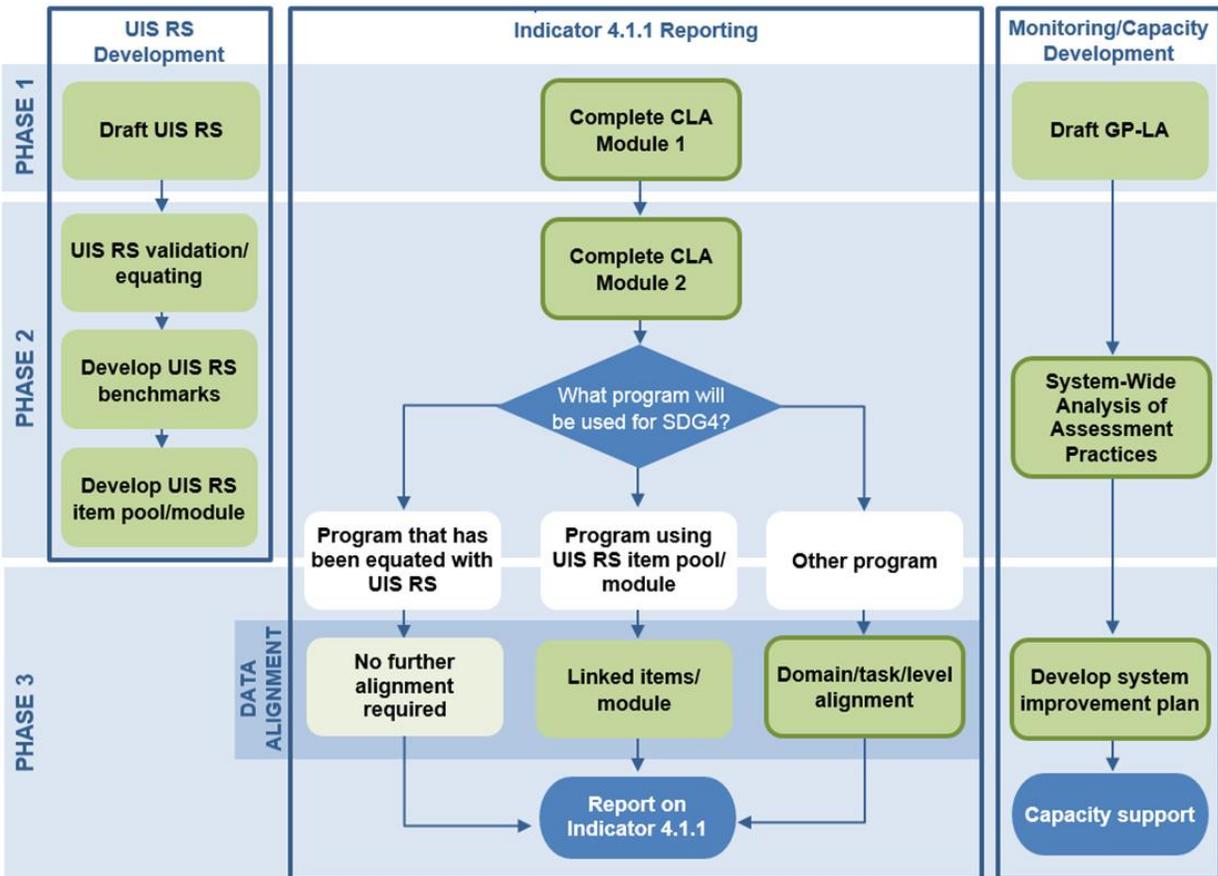
- An equating study will align selected existing assessment programs with the UIS RS, so education systems using these programs can report in consistent ways.
- It is proposed that a pool of items, or modules, is created from the UIS RS, which can then be incorporated into assessment programs to facilitate alignment.
- A Data Alignment process will enable education systems using other assessment programs to examine and report on their degree of alignment with UIS RS.

Each of these options offers some level of confidence in the consistency of education systems' reporting against Indicator 4.1.1, while retaining sufficient flexibility for education systems to pursue assessment programs appropriate to their context and needs. The greatest confidence may be provided by reporting using an equated assessment program; while the greatest flexibility is provided in the Data Alignment process, which may be applied to any assessment program in the relevant domains. This flexible approach to reporting will be supported by GAML's capacity development activities, to help education systems identify ways in which their assessment programs can improve over time. This will ensure that Indicator 4.1.1 reporting drives knowledge-sharing, and growth in global capacity to use assessment programs as levers for system improvement.

The options above constitute a substantial program of work by UIS and its partners, to create the UIS RS and design and implement the various processes through which it may be applied. It is therefore necessary to develop interim procedures for reporting against Indicator 4.1.1 while this program of work is in progress. Interim reporting is not only a stop-gap, but a challenge in itself, as it is affected



by many of the issues that currently prevent harmonised global reporting. Creating an intermediate solution to these issues is therefore an important opportunity to lay the groundwork for a more robust longer-term harmonisation approach.



Dark green border = process led by individual education systems



2. Key considerations in interim reporting

The challenges of achieving consistency in global reporting go far beyond the definition of the indicators themselves. In many cases, there is no “one-stop shop” or single source of information for a specific indicator that is consistent across international contexts. Even when there is agreement on the metric to be used in reporting, a harmonising process may still be necessary to ensure that coverage of the data is consistent. Education expenditure is an example of an agreed metric for which international harmonisation of data coverage has dramatically changed how education systems view their results.¹

The measurement of learning required for Indicator 4.1.1 poses particular challenges. Learning is typically assessed through complex processes that require definition of what learning is, and how to measure it. Learning itself is a complex construct, involving cognitive and non-cognitive processes over a sustained period of time. The measurement of learning may be seen as the last stage in this long process, meaning that indicators of learning must be guided by deep understanding of what processes underpin the data.

Broadly, the development and implementation of any learning assessment follows four key phases, illustrated in **Table 1**. These may be applied at an international, regional or national level, depending on the scope of the assessment program.

Table 1: Key phases in an assessment program

Phase	What it addresses	Main components
Conceptual Framework	What and who to assess	<ul style="list-style-type: none"> Assessment framework (cognitive, non-cognitive, and contextual) Target population
Methodological Framework	How to assess	<ul style="list-style-type: none"> Test design Sampling frame Operational design Data generation Data analysis (e.g. classical or Item Response Theory) Contextual information
Reporting Framework	How to report	<ul style="list-style-type: none"> Defining scales Benchmarking Defining progress (longitudinal equating)

¹ See *Who Pays for What in Education?* http://uis.unesco.org/sites/default/files/documents/who-pays-for-what-in-education-national-revealed-through-accounts-2016-en_0.pdf



Each of these phases must be considered in the interim reporting process, to ensure that Indicator 4.1.1 reporting produces the most meaningful possible results. UIS's approach to each of the main components in each of these phases is outlined below, including:

- how they are currently reflected in the international assessment landscape
- how they will be addressed through interim 4.1.1 reporting
- how UIS intends to report on each component in the first round of 4.1.1 reporting (shaded grey)
- how they will be addressed through subsequent 4.1.1 reporting procedures over time.

The issues discussed in this section do not only affect international consistency, but also pertain to the overall quality of the assessment program, and its utility to individual education systems themselves.

Conceptual framework

Assessment framework

Assessment programs differ in the conceptual frameworks that are used to develop their overall assessment framework. For example, depending on the curriculum in a country, national assessments usually have different content coverage for a given grade. Furthermore, even domains can be defined differently. In some cases, programs assess different skills, sometimes they use different content to assess the same domain, and sometimes they do both differently, even for the same grade.

For interim reporting, programs must produce a direct measure of learning outcomes in the learning domains being assessed for Indicator 4.1.1 at the relevant stages of education. The UIS Catalogue of Learning Assessments will enable such programs to be identified.

UIS will report the definition of reading and mathematics as proposed by each assessment.

Over time, once the UIS RS and UIS Global Content Reference Framework are completed, these tools will enable more consistent reporting of conceptual alignment.

Target population

Assessment programs may be either age or grade based; and may also vary in the point within a grade that is assessed. For example, some programs assess at the middle of an education level, some at the end of an education level, while others assess at both mid- and end points of an education level. Furthermore, the number of years of schooling (or duration of schooling) represented by a particular education level may vary across education systems. For example, some systems have six years of primary education, so testing at the mid- and end point of an



education level may represent Grades 3 and 6 respectively. Others have four years of primary school, so the mid- and end point of the same education level may be Grades 2 and 4 respectively.

A second limitation is that when assessments are school-system based - usually referred as school-based learning assessments - the indicators cover only those in school. The proportion of in-school target populations varies from country to country due to differences in out-of-school children and populations of young people in the country. Assessing competencies of children and young people who are out-of-school would require household-based surveys. Assessing children and young people in households is under consideration, but may be very costly and difficult to administer.

For interim reporting, programs must be large-scale learning assessments based on a representative sample or universal administration at the country level. This may include in-school assessments, or programs including out-of-school children and young people.

In- and out-of-school children: In general assessments are available only for in-school children. While in-school assessment offers greatest consistency across countries, some countries also have high proportions of children out-of-school. UIS would report on other sources of information that cover the out-of-school population, such as citizen-led assessment (CLA), provided that the country has officially authorized the publication at the global level. GAML and the Technical Coordination Group may discuss options for using the CLA for this purpose.

UIS will report on in-school students, with the exclusions taken by each assessment, as well as the target grade (with -1/+1 grade, if the target grade is not cleanly defined). UIS will identify any assessment used in 4.1.1 reporting that include children or young people outside-of-school (OOSCi).

Over time, efforts to improve school participation will increase the number of students included in in-school assessments. UIS will continue to work with relevant countries to improve out-of-school assessment.

Methodological framework

Test design

Assessments can be built in different formats, from multiple choice questions only to a combination of multiple choice and constructed response items. Over the years, technology has enabled more dynamic assessment design. With improved psychometric modelling, with which reasonable estimations can be done using a smaller number of items and target populations, different implementation platforms and operational procedures can be used. These possibilities have led to more complex test design, which must be carefully examined to ensure that it provides appropriate coverage of the learning domains under assessment.



The levels of learning progress represented in the test design is another important issue. For example, items in a cognitive test with different levels of difficulty have implications in terms of international consistency. It is therefore important that some effort is made to compare the difficulty level of different assessment programs used in reporting against Indicator 4.1.1.

For interim reporting, it is recognised that assessment programs may vary in the level of learning progress that they represent, even when administered to the same age or grade.

UIS will preface Indicator 4.1.1 reporting with a clear explanation that assessment programs may measure varying levels of learning progress, in this round of reporting for 2018 and probably 2019.

Over time, the Data Alignment process² will support more consistent reporting, using the draft UIS RS as a point of reference. It will include options for empirical alignment of existing tests with the UIS RS, as well as a method for countries to examine other assessment programs' coverage of UIS RS domains and levels.

Sampling frame

The nature of the sample is critical to the robustness of the assessment program as a measure of student learning progress, independent of any considerations of international consistency. Indicator 4.1.1 clearly implies a national representative sample, so a valid random sampling method needs to be adopted. Likewise, sample survey data must be reported along with standard errors, so inference is allowed and proper confidence intervals can be inferred. The only exception is where an assessment program includes all students at the relevant age or grade.

For interim reporting, education systems will be asked to provide clear information about the sampling frame, method and standard error as part of their reporting on Indicator 4.1.1.

UIS will report the standard error, and any limitations in sampling method, with 4.1.1 data.

Over time, sampling information and standard errors will continue to be an essential component of Indicator 4.1.1 reporting, as a key component of Data Alignment.

Operational design and data generation

Robust, consistent operations and procedures are an essential part of any large-scale assessment, to maximise data quality and minimise the impact of procedural variation on results. Examples of procedural standards may be found in all large-scale international assessments, and for many large-scale assessments at regional level, where the goal is to establish procedural consistency across international contexts. Many national assessments also set out clear procedural guidelines, to support consistency in their operationalisation. It is nevertheless recognised that some global

² See Data Alignment Concept Note



variation in operations and procedures is necessary, to reflect the resources and circumstances of each education system.

For interim reporting, UIS will generally assume that education systems are operationalising assessment in ways that are appropriate to their resources and context, notwithstanding the specific requirement for reporting sampling methods above. Education systems are encouraged to report any major operational issues that may affect the validity of the data.

UIS will report any major operational issues, in consultation with education systems.

Over time, GAML's global capacity building program will help education systems to improve the quality of their assessment operations and data generation processes.

Data analysis

In terms of data analysis, some countries may use more sophisticated modelling and reporting methods - like item response theory - to scale and report scores. Others may use simple classical theory descriptive statistics, like the proportion of correct scores. Depending on the model used, reporting scores will differ in their scales and metrics.

Data analysis typically includes disaggregation by student demographic characteristics such as age or age-group of students, sex, location, socio-economic status, migrant status and ethnicity. This supports confirmation of the representativeness of the sample for cohorts for whom learning outcomes may differ, and also assists education systems to understand which student cohorts are best served by their schools. Disability status is not currently available in most national and cross-national learning assessments.

For interim reporting, the indicator is calculated as the percentage of children and/or young people at the relevant stage of education achieving or exceeding a pre-defined proficiency level in a given subject, disaggregated by sex for each year level. This may be written as:

Performance above the minimum level, $PL_{tn,s,above\ minimum} = p$, where p is the percentage of students in a learning assessment at stage of education n , in subject s in any year $(t-i)$ where $0 \leq i \leq 5$, who has achieved the level of proficiency that is greater than a pre-defined minimum standard, S_{min} .³

UIS will report the results of this analysis, which will be collected through CLA Module 2.

³ UIS (2017) Meta Data, <http://uis.unesco.org/sites/default/files/documents/sdg4-metadata-global-thematic-indicators.pdf>, page 7.



Over time, this analysis will remain the mainstay of Indicator 4.1.1 reporting. Education systems may also choose to enhance data analysis through GAML's capacity development activities, to increase the impact of their assessment programs on evidence-based policy and practice.

Contextual information

Contextual information is usually collected during national or cross-national assessments through in-school assessments or household assessment surveys. As the only contextual information relevant to Indicator 4.1.1 is the sex and year level of the student (see data analysis above), this component of the methodological framework is not addressed in interim reporting. It may nevertheless form part of education systems' broader objectives for their assessment programs, to maximise their relevance for policy and research.

Reporting framework

Defining scales

Assessment programs typically report using different scales. Analysis of results therefore remains contained to their particular test, linked to one methodology and one scale. While methodologies tend to converge between international and regional assessments, it is still difficult to situate an individual student's learning progress on an indicative pathway, as there is no reference curve for the acquisition of learning outcomes.

For interim reporting, UIS will generally assume that education systems are constructing scales in a way that is appropriate to the chosen analytic method (for example, Item Response Theory or classical test theory). Technical information about scaling should be publicly available, to support transparency in reporting.

UIS will work with education systems to ensure that technical documentation about scaling is available in the public domain, for any assessment programs used in 4.1.1 reporting. The Catalogue of Learning Assessments is the natural vehicle for collection.

Over time, the UIS RS and Data Alignment process will enable reporting from a variety of assessment programs on a consistent scale, for the 4.1.1 learning domains.

Benchmarking

Currently, there are no common standards validated by the international community as a global benchmark. While data from many national learning assessments is available now, every country sets its own standards so the performance levels defined in these assessments may not always be consistent. This is also true with cross-national learning assessments, including international and regional learning assessments. For education systems who participated in the same cross-national learning assessments, results are comparable, but not across different cross-national learning assessments.



For interim reporting, countries will report the percentage of students who have achieved at least the minimum proficiency level as defined by each assessment. Until there is agreement on a global scale and its associated benchmarks, there will be no expectation for consistency.

UIS will preface Indicator 4.1.1 reporting with a clear explanation that assessment programs may define minimum standards of proficiency in different ways.

Over time, benchmarks will be identified on the UIS RS at the three points identified in 4.1.1, and the UIS RS equating study will help to map these to existing assessment programs. The three measurement points will have their own established minimum standard, which will divide students into below minimum or at or above minimum proficiency levels.

Longitudinal equating

How periodical an assessment should be does not necessarily affect consistency, and should depend on each education system's needs. Based on current practice, three to five years is seen as an appropriate period between large-scale assessments. This balances the need for current data against the costs of administering a high-quality assessment program.

Assessments that are repeated periodically may have potential for longitudinal equating. This may involve comparison of results from a single country at different periods of time (that is, for country X at periods A and B), or between countries at multiple periods (for example, comparing country X to country Y at period A, and again at period B, even if A and B are not comparable over time for a single country). Longitudinal linkability is a worthwhile property of assessment programs to consider in any international monitoring.

For interim reporting, UIS will consider which assessment programs may have potential for longitudinal equating, to help prepare for monitoring progress towards 4.1.1 over time.

UIS will report on the periodicity of each assessment, and if it is longitudinally equated.

Over time, longitudinal equating possibilities for Indicator 4.1.1 will be investigated further.

3. Choosing an assessment program for 4.1.1 reporting

The current learning assessment data landscape comprises two broad categories of large-scale assessment that may be suitable for reporting on Indicator 4.1.1: Cross-National Large-Scale Assessments (CN-LSAs) and National Large-Scale Assessments (N-LSA):

- **Cross-National Large Scale Assessments** allow comparison between participating countries and (if designed appropriately) will allow for trend comparison; that is, comparison over time within countries and across countries. Reporting trends will allow countries to monitor and track the



progress of their education system and children learning. They also allow countries to compare among themselves overtimes, if they choose to do so. In order to report on trend and allow comparison across time, the need of an established baseline and a well-conceived psychometric design, even if learning developed and change there is still basis to link across time.

- For **National Large-Scale Assessments**, the potential for comparability (across countries or across time) is, in general, defined at the design level. N-LSAs do not allow comparison with other either CN-LSAs or other N-LSAs unless they are linked by design, or by practice due to the inclusion of a pool of common items that allow cross-country comparability. Comparability over time is possible if they are designed with an adequate linking process between waves of administration. This is not necessarily the practice, and programs should be examined carefully for consistency over time before any inference is made. **Table 2** lists current CN-LSAs, and their coverage of the Indicator 4.1.1 learning domains.



Table 2: Current Cross-National Large-Scale Assessments

Grade	Assessment	Number of countries	Reading	Numeracy and mathematics
2	PASEC	10	X	X
3	TERCE	15	X	X
4	PIRLS	48	X	
	PILNA	13	X	X
	TIMSS	63		X
5	SEA-PLM	7 (Field Trial)	X	X
6	TERCE	15	X	X
	PASEC	10	X	X
	PILNA	13	X	X
	SACMEQ	15	X	X
4th to 6th	LaNA	Pilot Phase	X	X
7	N/A	N/A	N/A	N/A
8	TIMSS	63		X
9 or 10	PISA (15-year olds)	65	X	X

PASEC – Programme d'analyse des systèmes éducatifs de la CONFEMEN

TERCE – Tercer Estudio Regional Comparativo y Explicativo

PIRLS – Progress in International Reading Literacy Study

PILNA – Pacific Islands Literacy and Numeracy Assessment

TIMSS – Trends in International Mathematics and Science Study

SEA-PLM – Southeast Asia Primary Learning Metrics

SACMEQ – Southern and Eastern Africa Consortium for Monitoring Educational Quality

LaNA - Literacy and Numeracy Assessment

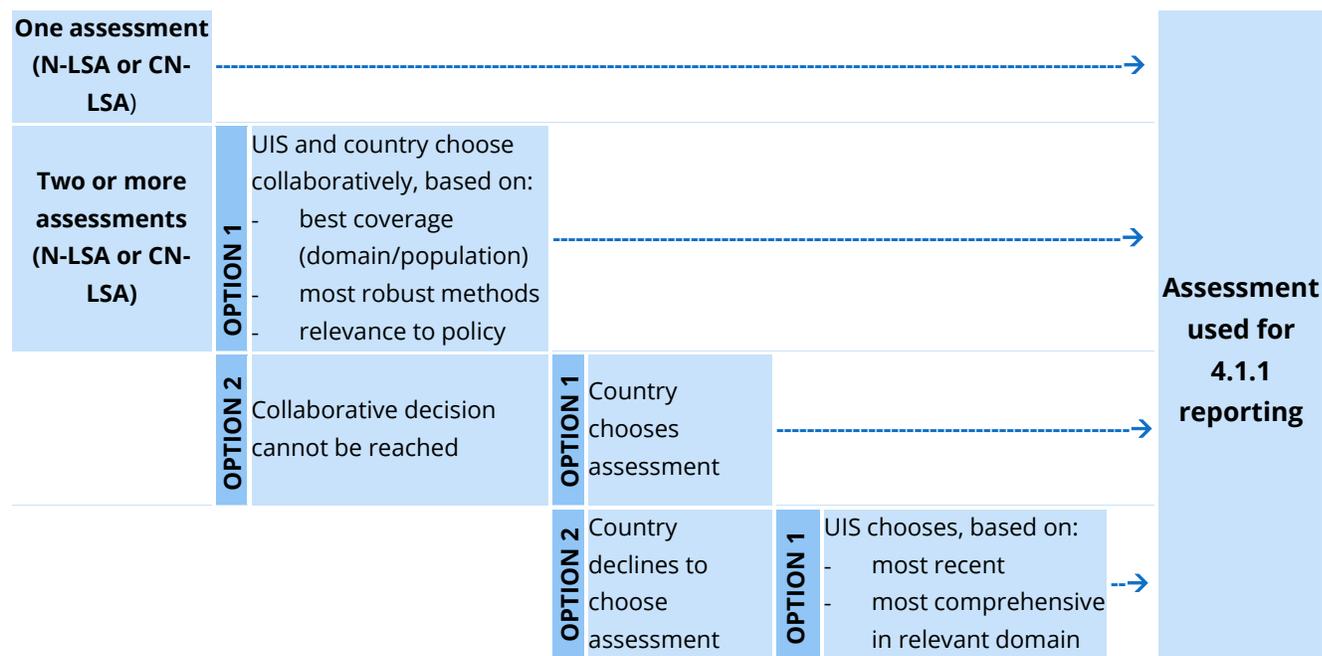
PISA – Programme for International Student Assessment

It is possible for a country to have multiple assessment programs that may be suitable for reporting on Indicator 4.1.1 in the same domain, at the same level. For example, a country may have a N-LSA for students in a particular grade, and also participate in a CN-LSA that targets a sample of the same students. Some countries may even participate in more than one CN-LSA covering the same grade.

If a country has two potentially suitable assessments in the same domain, a discussion between the country and the UIS would determine the choice of the assessment that will be used to report. It might be the case that one assessment has more robust methods than another, better coverage in terms of the national population or the learning domain, or better relevance for informing policy. Countries would be invited to engage with discussion with UIS about which assessment will be used, with the preference being for UIS and the country to choose the program for reporting collaboratively. As a second option, the country will be given the opportunity to choose which is reported, either on the



both domains or only one. If they decline, UIS will select the assessment based on which is most recent, and most comprehensive in the relevant domain.



Where a choice has been made between two assessments, UIS will record in a footnote alongside the 4.1.1 data if this was a UIS choice (UIS), or country choice (CC). No footnote will indicate that the assessment program was either the only one available, or a collaborative choice. This will help to assure transparency in decisions about the assessment programs that are used for 4.1.1 reporting.



4. Format for reporting

The Indicator 4.1.1 reporting format aims to communicate two pieces of information:

1. the percentage of students meeting minimum proficiency standards for the relevant domain and measurement point; and
2. the conditions under which the percentage can be considered comparable to the percentage reported from another country, including any caveats that may affect comparability.

In the first round of reporting, the number of caveats on comparability (limitations) is likely to outweigh the number of conditions under which cross-country comparability can be considered (possibilities). This does not detract from the value of interim reporting, recalling that the primary goal of 4.1.1 reporting is not to compare results across countries, but to inform system improvement within individual countries or country groups. Over time, possibilities for international comparability may increase, but this primary purpose will remain.

Assuming that only assessment programs with nationally representative samples will be reported (see key considerations above), the following footnotes will be used alongside the 4.1.1 data:

Grade is not exact: +1/-1 from exact ISCED grades level is reported (letter +1=a; -1=b)

Choice of Assessment: CC: country has chosen assessment; UIS: UIS has chosen assessment.

Longitudinal equating: assessment program is longitudinally equated (letter e)

OOSCi included letter f

Data Source: letters from g to s to describe the source.

- g. National Large Scale Assessment
- h. Programme d'analyse des systèmes éducatifs de la CONFEMEN (PASEC)
- i. Progress in International Reading Literacy Study (PIRLS)
- j. Programme for International Student Assessment (PISA)
- k. Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ)
- l. Tercer Estudio Regional Comparativo y Explicativo (TERCE)
- m. Trends in International Mathematics and Science Study (TIMSS)
- n. Literacy and Numeracy Assessment (LANA)
- o. Pacific Islands Literacy and Numeracy Assessment (PILNA)
- p. Southeast Asia Primary Learning Metrics (SEA-PLM)
- q. Citizen Led Assessments (CLA)
- r. UNICEF Multiple Indicator Cluster Survey (MICS)
- s. Early Grade Reading Assessment (EGRA) / Early Grade Mathematics Assessment (EGMA)

Table 3 shows what a UIS reporting table of 4.1.1 for each single point of measurement will look like.

**Table 3: UIS reporting table – End of Primary**

	Time	1999	2000	2001	2002	2003	2004
Country							
Afghanistan	
Åland Islands	
Albania		A					
Algeria		B ^{b,CC} I					
American Samoa		C ^{a,c,d} , UIS ^f					
Andorra		..					
Angola		..					
Anguilla		..					
Antigua and Barbuda		..					
Argentina							
Armenia							

Note **CC**: country has chosen assessment; **UIS**: UIS has chosen assessment. (letter CC or UIS), no letter implies ; a: +1 of the exact ISCED grade for the country is reported; b: -1 from the exact ISCED grade is reported; e: the test is longitudinally equated; f: OOSCI included in sample or census; Letters that follow describe the Large Scale Assessment used as Data Source; g- National Large Scale Assessment; h: Programme d'analyse des systèmes éducatifs de la CONFEMEN (PASEC); i: Progress in International Reading Literacy Study (PIRLS); j: Programme for International Student Assessment (PISA); k: Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ); l: Tercer Estudio Regional Comparativo y Explicativo (TERCE); m: Trends in International Mathematics and Science Study (TIMSS); n: Literacy and Numeracy Assessment (LANA); o: Pacific Islands Literacy and Numeracy Assessment (PILNA); p: Southeast Asia Primary Learning Metrics (SEA-PLM); q: Citizen Led Assessments (CLA); r: UNICEF Multiple Indicator Cluster Survey (MICS); s: Early Grade Reading Assessment (EGRA) / Early Grade Mathematics Assessment (EGMA)



5. UIS validation process with countries

This section presents UIS's initial work on mapping existing CN-LSAs for the purposes of 4.1.1 reporting, in their definitions of the relevant educational cycles respecting countries' educational structures and minimum levels of proficiency. This analysis focuses on CN-LSAs, because they have the greatest potential for consistent reporting against Indicator 4.1.1 across education systems, which has been the main reason to have indicator 4.1.1. (b and c) on Tier II. N-LSAs are still relevant to 4.1.1 reporting, however, especially as the creation of the UIS RS and implementation of Data Alignment will allow them to be aligned to CN-LSAs using a common scale.

Educational cycle: what grades are the defined point of measurement of Indicator 4.1.1?

The first analysis considers the questions that define the stages of learning in Indicator 4.1.1:

- ✓ What does it mean to be in grades 2/3?
- ✓ What is the end of primary schooling?
- ✓ What is the end of lower secondary schooling?

Table 4 shows whether each assessment is informing on the exact levels (for example, PISA is age-based, so we reflect the grade/course where most of the children or young people are). The table shows the exact grade and plus minus 1 and 2 to get a sense of the coverage that it implies, extending beyond the exact grade.

Table 4: Correspondence between ISCED levels and CN-LSA

Assessment	ISCED grade															
	Grade 2/3				End of primary						End of lower secondary					
	-1	Exact	1	2	-4	-3	-2	-1	Exact	1	-3	-2	-1	Exact	1	2
(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)	(m)	(n)	(o)	(p)	
PASEC 2014		10							10							
PILNA 2015			11	3			1		11	3						
PIRLS 2011					1	5	21	7	10	1						
PISA 2012												1	4	17	32	8
PISA 2015												1	4	19	35	10
SACMEQ III								10	4							
TERCE 2013		15							13	2						
TIMSS 2011					1	3	24	6	13		1	4	28	7		
TIMSS 2015					1	3	26	5	10		2	3	25	7		
SEA-PLM 2015								4	2	1						
Total count		25	11	3	3	11	72	32	73	7	3	9	61	50	67	18
Total count (Most recent)		25	11	3	2	8	48	22	69	7	2	4	29	26	35	10



Table 4 has been built according to the following process for each country:

- a. Map what grade is end of primary and end of lower secondary school
- b. Record in what grade the country has taken the assessment
- c. Identify if the match between a and b is exact; if this were the case, then the country is mapped as having an exact match
- d. If the match in point c is not exact, then it is defined as the distance between the grade in which the test is administered and the grade that corresponds to the point of measurement (either end of primary or end of lower secondary school). The distance could be from - 3 to + 3.
- e. For instance, Argentina has administered TIMSS it in 8th grade and the end of lower secondary school is in 9th grade, then the country is mapped as informing for -1 as the match is not exact. Argentina is mapped in this example in column (I) of the table. The same process was run for each country and point of measurement.
- f. For reporting in 2017, UIS has decided to take only the exact grade and -1/+1, and the proposal is to keep the same criteria towards future.

The second analysis considers the benchmarks, definition of associated proficiency levels and minimum levels used in each assessment.⁴ In general, Cross National Large Scale Assessment, either regional or international, define various performance levels, and report as well the mean and standard deviation. They choose as well one level as the cut-off point that defines what children/youth are below or above level. Table 5 summarizes this information. Unless otherwise indicated, the same cut-off level is used for both reading and mathematics.

⁴ UIS (2017) Meta Data, <http://uis.unesco.org/sites/default/files/documents/sdg4-metadata-global-thematic-indicators.pdf>, page 7.



Table 5: Proficiency levels corresponding to each international and regional student achievement test

Harmonized proficiency levels	Original name of levels by assessment																					
	PISA		TIMSS		PIRLS	PASEC				TERCE				SACMEQ		PILNA						
	8		4	8	4	2	2	6	6	3	3	6	6	6	6	4	4	6	6			
	Reading	Math	Math	Math	Reading	Reading	Math	Reading	Math	Reading	Math	Reading	Math	Reading	Math	Reading	Math	Reading	Math			
Level 8 (L8)	Level 6													Level 8								
Level 7 (L7)	Level 5	Level 6													Level 7							
Level 6 (L6)	Level 4	Level 5													Level 6							
Level 5 (L5)	Level 3	Level 4	Advanced International Benchmark	Advanced International Benchmark	Advanced International Benchmark	Level 4			Level 4					Level 5								
Level 4 (L4)	Level 2	Level 3	High International Benchmark	High International Benchmark	High International Benchmark	Level 3	Level 3	Level 3	Level 3	Level IV	Level IV	Level IV	Level IV	Level 4								
Level 3 (L3)	Level 1a	Level 2	Intermediate International Benchmark	Intermediate International Benchmark	Intermediate International Benchmark	Level 2	Level 2	Level 2	Level 2	Level III	Level III	Level III	Level III	Level 3								
Level 2 (L2)	Level 1b	Level 1	Low International Benchmark	Low International Benchmark	Low International Benchmark	Level 1	Level 1	Level 1	Level 1	Level II	Level II	Level II	Level II	Level 2								
Level 1 (L1)	Below Level 1b	Below Level 1b	Below Low International Benchmark	Below Low International Benchmark	Below Low International Benchmark	Below Level 1	Below Level 1	Below Level 1	Below Level 1	Level I	Level I	Level I	Level I	Level 1								

 : achieve at least a minimum proficiency level

 : not applicable



2018 Proposed Metadata - Indicator 4.1.1.

4.1 *By 2030, ensure that all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes*

4.1.1 *Proportion of children and young people (a) in Grade 2 or 3; (b) at the end of primary education; and (c) at the end of lower secondary education achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex*

Definition:

Percentage of children and young people in Grade 2 or 3 of primary education, at the end of primary education and the end of lower secondary education achieving at least a minimum proficiency level in (a) reading and (b) mathematics. The minimum proficiency level will be measured relative to new common reading and mathematics scales currently in development.

Minimum proficiency level is the benchmark of basic knowledge in a domain (mathematics, reading, etc.) measured through large-scale learning assessments based on representative samples at the country level. Currently, there are no common standards validated by the international community as a global benchmark as there is not agreement at the national

The indicator shows data published by each of the agencies and organizations specialised in cross-national learning assessments as well as information provided by countries. Comparability is limited to the countries that have participated in that particular assessment in the case of Cross National Large scale assessment. There is no comparability across the different cross-national learning assessments unless they are linked in a way that allows draw inference.

Purpose:

The indicator is a direct measure of the learning outcomes achieved in the subject areas being assessed at the relevant stages of education.

Calculation method:

The indicator is calculated as the percentage of children and/or young people at the relevant stage of education achieving or exceeding a pre-defined proficiency level in a given subject.

Performance above the minimum level, $PL_{tn,s,above\ minimum} = p$

where p is the percentage of students in a learning assessment at stage of education n , in subject s in any year $(t-i)$ where $0 \leq i \leq 5$, who has achieved the level of proficiency that is greater than a pre-



defined minimum standard, S_{\min} . The minimum standard will be defined by the global education community taking into consideration regional differences.

Interpretation:

The three measurement points will have their own established minimum standard. There is only one benchmark that divides students into below minimum or at or above minimum proficiency levels.

(a) Below minimum is the proportion or percentage of students who do not achieve a minimum proficiency level as established by countries according to the globally defined minimum proficiency level.

(b) At or above minimum is the proportion or percentage of students who have achieved at least the minimum proficiency level as defined by each assessment. Until there is agreement on a global scale and its associated benchmark there is no comparability across assessments.

Type of data source:

Large Scale Learning assessments at the country level based either on the universe or in a national representative sample.

Disaggregation:

By age or age-group of students, sex, location, socio-economic status, migrant status and ethnicity. Disability status is not currently available in most national and cross-national learning assessments.

Data required:

Performance level data from national and cross-national assessments with clear indication of the performance level that defines the minimum level of proficiency for the LSA.

Data sources:

Large scale learning assessment based on representative sample or universal administration at the country level.

Cross-national learning assessments including: *Programme d'analyse des systèmes éducatifs de la CONFEMEN* (PASEC), Progress in International Reading Literacy Study (PIRLS), Programme for International Student Assessment (PISA), Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ), *Tercer Estudio Regional Comparativo y Explicativo* (TERCE) and Trends in International Mathematics and Science Study (TIMSS).

Data collection



UIS Catalogue for Learning Assessment. Modules I and II and reporting to UIS from Cross National Organization and agencies.

Limitations and comments:

While data from many national learning assessments are available now, every country sets its own standards so the performance levels may not always be directly comparable. This is also true with cross-national learning assessments which include international and regional learning assessments. Countries who participated in the same cross-national learning assessments the results are comparable but these results are comparable across different cross-national learning assessments.

A second limitation is that assessments are school-system based which is usually referred as the school-based learning assessment, the current indicators cover only those in school. The proportion of in-school target populations varies from country to country due to differences in out-of-school children and young people populations in country. Assessing competencies of children and young people who are out-of-school would require household-based surveys. Assessing children and young people in households is under consideration but may be very costly and difficult to administer.

Footnoting

Grade is not exact: +1/-1 from exact ISCED grades level is reported (letter +1=a; -1=b)

Choice of Assessment: plain not note assessment was joint decisions of country and UIS. **CC:** country has chosen assessment; **UIS:** UIS has chosen assessment. (letter CC or UIS)

Robustness: the test is reporting methods and procedures (**letter c**)

Alignment of contents: The test is not aligned (**letter d**)

Longitudinal equating: test if longitudinally equated (**letter e**)

OOSCi included letter f

Data Source: letters from g to s to describe the source

- g. National Large Scale Assessment
- h. Programme d'analyse des systèmes éducatifs de la CONFEMEN (PASEC)
- i. Progress in International Reading Literacy Study (PIRLS)
- j. Programme for International Student Assessment (PISA)
- k. Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ)
- l. Tercer Estudio Regional Comparativo y Explicativo (TERCE)
- m. Trends in International Mathematics and Science Study (TIMSS)
- n. Literacy and Numeracy Assessment (LANA)
- o. Pacific Islands Literacy and Numeracy Assessment (PILNA)
- p. Southeast Asia Primary Learning Metrics (SEA-PLM)
- q. Citizen Led Assessments (CLA)
- r. UNICEF Multiple Indicator Cluster Survey (MICS)
- s. Early Grade Reading Assessment (EGRA) / Early Grade Mathematics Assessment (EGMA)