unesco
Institute for Statistics

Reporting learning outcomes in basic education: Country's options for indicator 4.1.1

Silvia Montoya
Director
UNESCO Institute for Statistics

Reporting learning outcomes in basic education: Country's options for indicator 4.1.1

Silvia Montoya
Director
UNESCO Institute for Statistics

Abstract[1]

This paper sheds light on the importance of producing good-quality learning data that are comparable at the global level, across countries, and over time. It describes the associated challenges faced and the different options available to countries, including the possibility of participating in a cross-national assessment - international or regional - and the possibility of implementing a national assessment provided the appropriate linking statistical or non-statistical methodology is applied. The paper presents the proven usefulness of some of these methodologies in attaining the desired goal and highlighted the potential of the other methodologies that are still in the piloting phase.

Overall, the paper represents a comprehensive guide to countries on measuring learning outcomes as it compiles all the information related to the production of good-quality learning data covering the different aspects of this highly important and complex topic.

[1] This paper was initially written by Dr. Silvia Montoya, Director of the UNESCO Institute for Statistics, in August 2022 for the 38th Annual Conference for Educational Assessment in Africa (Livingstone, Zambia). It was later updated in May 2023.

Table of Contents

Acronyms

ACER	Australian Council for Educational Research
AMPL	Assessments for Minimum Proficiency Levels
CNA	Cross-National Assessment
EAPRO	East Asia and Pacific Regional Office
EGRA	Early Grade Reading Assessment
EGMA	Early Grade Mathematics Assessment
ERCE	Regional Comparative and Explanatory Study
GEM Centre	Global Education Monitoring Centre
GEMR	Global Education Monitoring Report
GPF	Global Proficiency Framework
ICCS	International Civic and Citizenship Education Study
IEA	International Association for the Evaluation of Educational Achievement
ILSA	International Large-Scale Assessments
LAC	Latin America and the Caribbean
LaNA	Literacy and Numeracy Assessment
LDC	Learning Data Compact
LLECE	Latin American Laboratory for the Assessment of the Quality of Education
MICS	Multiple Indicator Cluster Surveys
MILO	Monitoring Impacts on Learning Outcomes
MPL	Minimum proficiency level
MSI	Management Systems International
NLA	National Learning Assessment
OECD	Organisation for Economic Co-operation and Development
OOS	Out of school children
PASEC	Programme d'Analyse des Systèmes Éducatifs de la CONFEMEN
PIAAC	Programme for the International Assessment of Adult Competencies
PILNA	Pacific Islands Literacy and Numeracy Assessment
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
PL	Proficiency Level
PLD	Proficiency Level Descriptor
SACMEQ	Southern and Eastern Africa Consortium for Monitoring Educational Quality
SDG	Sustainable Development Goal
SEA-PLM	Southeast Asia Primary Learning Metrics
SSA	Sub-Saharan Africa
TIMSS	Trends in International Mathematics and Science Study
UIS	UNESCO Institute for Statistics

1. Objective

Learning data about all children are essential if we want to improve learning for every child and if we want to guide education reform. The data tell us who is not learning, help us to understand why, and can help to channel scarce resources to where they are most needed. A lack of learning data is an impediment to educational progress, and it is in the differences in the learning outcome levels between different groups of students that educational inequality shows up most dramatically. For example, two thirds as many children in low-income countries complete primary schooling as in high- income countries. But, even in some middle-income countries, around 60% of children are at or below minimum learning competency levels, whereas in high-income countries there are essentially no children at these levels: a difference of about 0% to 60%. Moreover, we do not even have the data for many of the low-income countries; we can only estimate differences between high-income countries and low-income countries as a whole. And this is the region where competencies are lower and where up to 80% of children learning at or below minimum competency level that global vulnerability shows up most clearly.

The urgency for establishing concrete steps to obtain high quality, globally comparable learning data that can be used to improve national education systems is now palpable. One of the most important challenges has been to produce global comparable data or, in other words, to harmonise assessments programmes and ensure robust cross-countries comparability, expand the number of comparison points and references for countries, and provide all citizens with an understating of how the schooling for children is going. Note that the call for comparable data is not specific to education nor a mere desideratum for statisticians but is mandated by the SDG process as specified in the report of the Inter-Agency and Expert Group on Sustainable Development Goal Indicators at the 47th session of the Statistical Commission: "Global monitoring should be based, to the greatest possible extent [with few exceptions], on comparable and standardized national data, obtained through well-established reporting mechanisms from countries to the international statistical system" (

Approaches that have been put forward differ most obviously in terms of their technical complexity, financial cost, and implied comparability of national statistics. They differ as well in their sustainability over time, their impact on the politics, planning and operations of national education authorities, their ability to contribute to capacity building within countries, and their impact in the media and policy debates. They also differ, importantly, in how easily they can link (conceptually, not statistically) to measurements at teacher level that can be used not just to track the SDGs but to improve them.

There are several options that could be taken forward in terms of reporting. The most practical options will depend on the fact that the ideal is not immediately possible for a variety of reasons. Instead, a hybrid approach was considered more acceptable. Over time, migration to more robust systems is possible and necessary. The practical solution is to rely, to the degree it is compatible with rigor, on existing measurement systems that counties are already familiar with and use. The current system in use in particular countries will drive the next steps that each country could take. The prioritisation of new data collection programmes should be staggered according to the three levels of the schooling system covered by Indicator 4.1.1. Initially, it may be best to prioritise the

measurement of learning outcomes at the primary level, given the numbers of Out of School (OOS) children, and if there is no measurement system in place yet. Measurement at later levels, where there are highly variable proportions of OOS, will be inherently less reliable or will need expensive in-home data collection. This document aims to describe different strategies to report for indicator 4.1.1 that countries could choose. Following this introduction, section 2 addresses the challenges of achieving consistency in global reporting; section 3 looks at methodological issues and choices; section 4 summarizes the status of current reporting; section 5 discusses current coverage across countries; and finally section 6 summarizes the menu of options and assessments available to countries to report on learning outcomes.

2. Achieving consistency in global reporting

Sustainable Development Goal (SDG) 4 aims to ensure that, by 2030, "*all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes*."

Indicator 4.1.1 refers to the proficiency indicator referring to three levels of schooling: lower primary, upper primary, and lower secondary and two subjects (reading and mathematics). The indicator reads as follows:

> *"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 [MPL] in (i) reading and (ii) mathematics, by sex."*

The reporting format of the indicator aims to communicate two pieces of information:

I. the percentage of students meeting at least minimum proficiency standards for the relevant domains (mathematics and reading) for each point of measurement (grades 2/3; end of primary and end of lower secondary) and

II. whether a program can be considered comparable, and the conditions under which the percentage of children at or above MPL can be considered comparable to the percentage reported from another country.

The indicator needs the following inputs:

- **Domain**: reading and mathematics. Reading and mathematics are measured at the national level in numerous ways;
- **Minimum proficiency level (MPL):** is the benchmark of basic knowledge in a domain (mathematics, reading, etc.) at a given age/grade;
- **Linking to the MPL:** methodologies to harmonize various data sources to a common definition of the MPL;
- **Sample:** the sample needs to be representative of the relevant population.

There are a few critical issues regarding reporting of indicator 4.1.1 that are discussed in detail in Gustafsson (2019) that deserve to be mentioned:

2.1 The comparability of grades and education levels

The fact that primary schooling has a different duration in different countries means a term such as 'the end of primary' can mean different things in different places and the gaps between proficiency benchmarks and reality tends to be systematically correlated to

grade level within countries and regions complicate comparisons across countries and assessment programmes, where the grade is not identical. However, the enormous majority (89%) of countries end their primary cycle in Grades 5, 6 or 7, so the issue should not be unduly exaggerated: some adjustments may need to be made, and are being made, but the matter need not throw into question the basic idea of measuring at the end of the primary cycle.

2.2 Comparability of assessment results across space and time

While the comparability of statistics across countries influences comparability over time, the latter does not imply the former. Cross-country comparison through cross-national assessments helps comparability across countries, and across assessment programmes, at one point in time, through equating or linking methodology. If each assessment programme in addition produces statistics which are comparable over time, then the statistics will be comparable across time and countries.

National assessment programs are not in general comparable to each other, but they can still provide relatively reliable trend data if the measurement is of good enough quality and, when this is not the case, progress towards better quality data happens as part of the SDG agenda. Thus, if all countries, or virtually all countries, are displaying improvements in learning over time, and if assessments are built to be comparable over time, it is almost certain that the world as a whole is displaying improvements.

2.3 Financial costs of assessments for countries

Assessments required to report SDG 4 indicators are relatively costly compared to other data collection systems required for these indicators. However, even for developing countries, the cost of assessing outcomes systematically is extremely low relative to the overall cost of providing schooling and relative to the cost of *not* measuring. Assessment systems, if well-designed, can have positive impacts that go beyond simply producing statistics.

2.4 The timeliness and policy impact of the statistics

Assessments produce national, and often sub-national, statistics which can influence policymaking and policy implementation in positive ways. For these positive impacts to be felt, statistics must not only be accurate, but they must also be widely seen to be credible, and the turnaround time between the assessment and the reporting of results should be as short as possible, without compromising on quality. The need for timely data has been more acute since the SARS-CoV-2 pandemic.

3. Methodological challenges met

3.1 The minimum proficiency level (MPL)

The minimum proficiency level (MPL) is the benchmark of basic knowledge in a domain (mathematics, reading, etc.) at a given age/grade measured through learning assessments. To ensure comparability across learning assessments, a first step was to agree on text defining each MPL, and to agree on the identification of the proficiency level aligned with that text, in terms of typical items and cut scores or proficiency levels in each of the international and regional programs. That has been a vital step toward consensus, now achieved.

It was agreed to report according to the textual definition of the MPL for each domain and levels in the Cross-National Assessments (CNAs). This was established by conducting an analysis of the performance-level descriptors (PLDs) of cross-national, regional, and community-led assessments in reading and mathematics. Based on those definitions, one very important step was to map the contents and curriculum as well as all proficiency levels descriptors in cross-national initiatives to identify the proficiency levels aligned with those definitions. That is, the first step of textual agreement needed some further steps in validation in each assessment to obtain the combination or set of items that better aligned to the MPL through a standard setting exercise.

Table 1 below presents the global MPL definitions for the domains of mathematics and reading.

Table 1. Minimum proficiency levels for reading and mathematics - Indicator 4.1.1

Education level	Mathematics	Reading
Grades 2/3	Students demonstrate skills in number sense and computation, reading simple data displays, shape recognition and spatial orientation.	Students read aloud and comprehend many single written words, particularly familiar ones, and extract explicit information from sentences. They make simple inferences when longer texts are read aloud to them.
End of Primary	Students demonstrate skills in number sense, computation, real world problems, basic measurement, 2D shape recognition, and reading and interpreting simple data displays.	Students independently and fluently read simple, short narrative and expository texts. They locate explicitly-stated information, interpret and give some explanations about the key ideas in these texts. They provide simple, personal opinions or judgements about the information, events and characters in a text.
End of lower secondary	Students demonstrate skills in computation, solving problems in measurement and geometry, interpreting, and constructing a variety of data displays, and making use of algebraic representations.	Students locate and connect multiple pieces of related information across sections of texts to understand key ideas. They make straightforward inferences when there is some competing information. They reflect and draw conclusions based on evidence, in a variety of text types.

Source: *UNESCO Institute for Statistics (2019). Minimum Proficiency Levels: described, unpacked and illustrated. GAML6/REF/2.*

3.2 Linking to the global definition of MPL

A final step needed was to run psychometric linking exercises in order to precisely anchor the verbal definitions to a score or potential score in key assessments.

Linking is the general term used to relate assessment scores on one assessment/form to another/test/form or, in other words, moderating differences between assessments that were designed. The linking of either a national, a regional, or an international assessment to the global proficiency level definition represented by the MPL requires a methodology to identify the same concepts/definition in the national assessment and across assessments, or completely different purposes to express them in the same scale in a way that allows some degree of comparability that, in turn, allows fair inferences about the subjects (countries) compared. The process of making comparable those different assessments, called “moderation” could be based on statistical or non-statistical

calibration.

Statistical methods

Recalibration of existing data

This proposal, by Nadir Altinok, involved applying statistics to score data emerging from cross-national programmes.[2,3] The adjustments take advantage of the fact that some countries, referred to as doubloon countries, participate in more than one cross-national programme. Using several such overlaps has allowed for the identification of roughly comparable cut scores representing global proficiency benchmarks across different programmes, as well as the calculation of confidence intervals around the resulting proficiency attainment statistics.[4] Note that this does not involve working with individual learner data, or the entire datasets, and is thus much less expensive.

Nonetheless, Altinok noted that recalibration of existing data is a second-best approach, and the ideal is comparison of micro or individual learner data, ideally using standard data collection instruments. He noted that while one could use his proposed numbers for approximate group comparisons, the margin of error was too big for higher-stakes individual comparisons (which readers tend to make) and to track over time. Trevino and Ordenes (2017)[5] proposed the utility of this statistical recalibration approach in its ability to provide a reality check against which to compare statistics based on national assessments.

Recalibration by running parallel assessments (Rosetta Stone)

This programme, led by the International Association for the Evaluation of Educational Achievement (IEA) and the TIMSS & PIRLS International Study Center, Lynch School of Education at Boston College, is named after the famous archaeological discovery and linguistic analysis that enabled the reading of Egyptian hieroglyphics by using a translation key: the Rosetta Stone. The Rosetta Stone Study is designed to measure global progress towards SDG 4.1.1 by relating different national and regional assessment programmes to the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS). These are long-standing metrics and benchmarks of achievement6. The goal is to provide countries that participated in regional or national assessments, but not in TIMSS and PIRLS, with information about the proportions of primary school students who have achieved a minimal level of competency in literacy and numeracy (SDG 4.1.1) that allows international comparisons.

The approach involves having the same students take more than one assessment or sub-

---

[2] Altinok, N. (2017). *Mind the Gap: Proposal for a Standardised Measure for SDG 4–Education 2030 Agenda*. Montreal: UNESCO Institute for Statistics.

[3] Altinok, N., N. Angrist and H.A. Patrinos (2018). *Global Dataset on Education Quality (1965-2015).* Washington: World Bank.

[4] idem, p. 77: includes the average scores per country, for the primary and secondary levels, and for a combination of the two, obtained through the adjustments process.

[5] Trevino E. and M. Ordenes (2017). *Exploring Commonalities and Differences in Regional and International Assessments*. Information paper No. 48. Montreal: UNESCO Institute for Statistics.

[6] International Association for the Evaluation of Educational Achievement (2017). *IEA's Rosetta Stone: Measuring Global Progress toward the UN Sustainable Development Goal for Quality Education by Linking Regional Assessment Results to TIMSS and PIRLS International Benchmarks of Achievement*. Chestnut Hill.

assessment belonging to a different assessment (e.g., TIMSS and PASEC) and then producing concordance or translation tables between them.

The Rosetta Stone solution appears as very valuable as one component of the future Indicator 4.1.1 reporting system, though it would be inadequate as the core of the system, largely due to coverage and costs. The Rosetta Stone solution is more assessment-focused and aims to equate scores across different programmes and in the medium term is expected to enhance the comparability of scores and proficiency statistics across the cross-national programmes focusing on the primary level.

Module to measure the MPL (AMPL-a, -b, -c)

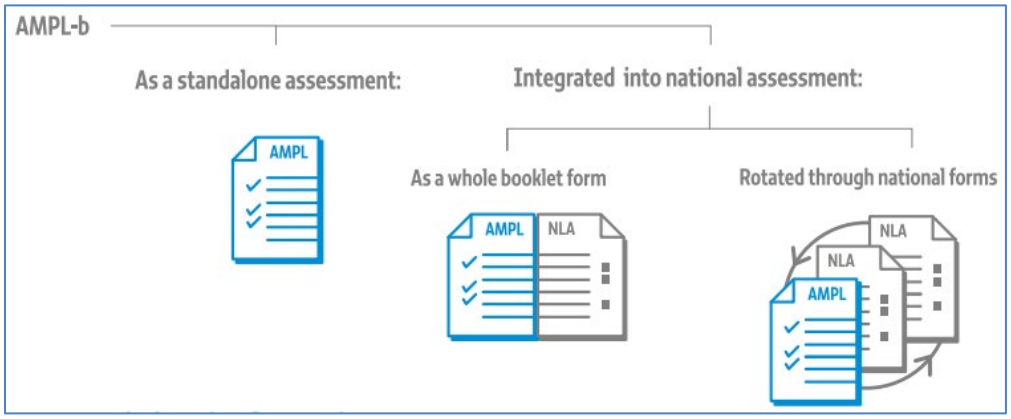
The Assessments for Minimum Proficiency Levels (AMPL) are ground-breaking and robust tools developed to measure learning outcomes against SDG 4.1.1b (i.e. at the end of primary). AMPL assessments were administered in 2021 alongside national or regional assessments and aligned to the Global Proficiency Framework, which defines the mathematics and reading learning domains.

The AMPL material was selected from the UIS Global Item Bank. The Item Bank is comprised of items from a range of sources, including 300 mathematics and 300 reading items provided by the Global Education Monitoring (GEM) Centre. The items to measure the attainment of the MPLs were selected from the Item Bank to match the benchmark definitions published by the GEM Centre in its paper: 'Minimum Proficiency Levels: Described, unpacked and illustrated' which were written to provide a more concrete definition of the MPLs, along with detailed explanatory information and exemplars.

The AMPL-b (see below for AMPL-a and AMPL-c) is independently scaled - call it the AMPL-b scale for example. On the AMPL-b scale, the AMPL-b cut point has been located, using the standard setting exercise conducted in Monitoring Impacts on Learning Outcomes (MILO) (and validated by the International Standard Setting Exercise conducted by the ACER GEM Centre recently)[7].

Figure 1 summarizes the potential alternatives to implement AMPLs: a) as a standalone assessment; and integrated into the national assessment, either as a b) whole booklet form or c) rotated through national forms.

Figure 1 - AMPL-b as a resource



[7] The Australian Council for Educational Research (2022)

Depending on the goals of the program, a national assessment could be designed to incorporate the AMPL in different ways and there are different scaling possibilities for the national assessment. For example, the national assessment could be equated to the AMPL using a 'common students' method as the basis for the equating. This would mean the establishment of a national scale upon which the MPL was located. This national scale could then be used again in the future, with or without AMPL booklet. As mentioned above, the AMPL booklet could also be incorporated as a rotating booklet.

Non-statistical methods

What is discussed here as a single proposal is actually two separates yet overlapping proposals that have been put forward and experimented with. They have, in common, the work by a team of experts to arrive at comparable cut scores in the various cross-national assessment programmes, at least in part through pedagogically informed evaluation of documents and items from the existing programmes. The first proposal is **policy linking** that requires a group of experts (mainly teachers) in a country to evaluate the difficulty of assessment items and set cut scores. The second proposal is the **pairwise comparison** that consists of a group of experts in pedagogy and psychometrics doing the same evaluation but in an independent way.

The approaches have been originally proposed for increasing comparability across countries. But the reporting (or proficiency) scale could in fact facilitate comparability over time within countries. If national teams of experts remain consistent over time, and/or the processes and criteria are carefully documented, it is likely that the reporting scale would measure consistently across years.

Policy linking is implemented though a toolkit that still is evolving and, as happens with any innovation, will have iterative cycles of development, piloting, refinement, implementation and then re-development. Pairwise comparison has not been tested as well, so we will not go into it at this stage.

The main contribution of this type of methodology is that it could expand considerably the coverage in terms of the student population relative to a scenario where only cross-national assessments were counted, or where only Rosetta-stone or AMPL booklet approaches were used. Coverage at the primary level would double, in terms of the population-weighted, if national assessments were included.

3.3 Linking Strategies: Cost-Benefit

Table 2 below summarizes the costs and the status of execution of the different linking strategies, the main milestones executed and pending, and their time frame. Two different set of costs are estimated. The first one is the set of fix costs that need to take each one of the alternatives to a full working status. For instance, in the case of common students, it is needed one regional assessment, 3 participating countries and an assessment tool adequate to measure to that the student population plus two days of administration; in the case of common item linking, the investment in a technological solution and the compilation/elaboration of items could be critical in investment size to kick-off with very low marginal costs, where the marginal unit is a country. Non-statistical methods, especially policy linking, require the development of a toolkit, with a set of clear guidelines to

standardize administration and the piloting in a few countries and, although the development and piloting have already been made, it would demand various iterations to fine-tune the methodology.

unesco
Institute for Statistics

Table 2. Comparing linking options

	Statistical				Non-Statistical	
	Ex-post calibration (Altinok)	**Common Students**	**Common items**	**AMPL module**	**Policy Linking**	**Pairwise comparison**
Data collection	Ex-post	Ex-ante	Ex-ante	Ex-ante	Ex-post	Ex-post
Students	Different	Same	Different	Different	Different	Different
What	Set different assessments on a common scale.	Concordance table of one scale into other.	Common items are inserted in the assessment.	A module calibrated to the MPL is inserted either as an additional booklet or by running parallel assessments.	Matches up definitions of the MPL descriptor using subjective judgement and, under certain conditions, allow those assessments to be aligned across countries.	
Items/Test	Different assessments	Different assessments	Common items in different assessments	Same module across different assessment program	Different assessments	
Calibration	Puts all information in the same scale.	Calibration needs a various step and builds a concordance table.	Joint calibration of assessment forms.	Accurate to report on the MPL.	Depends on assessment program.	
Alignment with Global MPL	No	Yes, but needs standards setting to define accurate alignment	Depends on alignment and sufficiency	Yes	Depends on alignment and sufficiency of items	
Sufficient # of items	n/a	Yes	Depends on choice	Yes	Depends on each assessment tool	
Measurement skills continuums	No	Yes	Depends on the assessment programs	Not now but possible with current and future developments	Depends on each assessment tool	
Track progress over time	Unclear	Yes	Yes	Yes	Not clear depends on quality of tools and the longitudinal equating	
Frequency	n/a	Cycle depending on each assessment	On demand	On demand	n/a	n/a
Output	Common scale using a modelling strategy	Concordance table	Allows to report on selected cut off points for both scales (e.g. MPL)	Calibrated to the MPL	Identifies the MPL cut-off points	Identifies the MPL cut-off points
How	-	Relies on the participation of countries in two assessments. Students take the two	Construction of a single reporting scale for each domain with items	Insert the booklets either as a standalone running parallel	Experts judge each item and set initial cut scores based on their understanding	Group of experts provide judgement about difficulty of each item on the

		assessments to help link between the results of both assessments.	from assessment programs.	assessment or as rotating booklet.	of the levels and the population.	assessment relative to items that have already been calibrated to MPL.
Country ownership	None	Very low	Medium to low	high	high	Medium
Needs	Tests measure the same latent construct.	Tests have enough quantity of items that could identify linking.	A common subset of calibrated items to be piloted to proof utility.	A tool built with items that are aligned and sufficient to measure the MPL.	Good quality cognitive tools and procedures. Strong alignment of assessment tools to GPF.	
Pros	Inexpensive	Technically rigorous	Technically rigorous	Technically rigorous	Cost-effectiveness	
Cons	Unless there are equivalent tools not accurate for higher stakes uses, may be suitable for group and approximate Uses.	Costly. Efficient if done between a regional and a global assessment.	Costlier financially and operationally.	Does not allow deep investigation of the construct.	Relatively subjective (less for pairwise). Depends on the quality of the assessment tool and implementation of the linking process.	
Achieved so far	Many attempts explored but most notably all the work of Altinok (2017).	Rosetta Stone: ERCE (LAC) and PASEC (SSA) participated with idea in the Rosetta Stone Exercise.	--	AMPL-b administered. AMPL-c ready to be administered (PISA). AMPL-a under Preparation.	First phase of Pilots around 10 countries run.	Standard setting exercise for MILO (ACER, 2022).
Next/remaining steps	--	Potentially expansion to other regions and national assessments	--	Scale-up depends on country's interest and development partners support	Revision of toolkit	Methodology guidance and analysis
National Cost	None	Between US$ 250,000 and 400,000	--	Printing cost of a booklet. Extra administration costs depends on modality.	Between US$30,000 to 50,000 for national workshop	none
International Cost	100,000 to 250,000	International US$ 1 million per region. Regional – US$ 500,000	--	US$ 100,000 on average for technical assistance	Between US$ 50,000 and 75,000 per country	US$ 40,000

Source: *UNESCO Institute for Statistics.*

4. Current reporting on indicator 4.1.1

In this section, the focus Is on the current reporting of indicator 4.1.1. Given the coverage so far, and the challenges created by the fact that there is no unique source of information to report on learning assessments, the global community faces a consistency challenge.

Current reporting is based on large-scale assessments. Large-scale assessments are designed to describe the achievement of students in a curriculum area in an aggregated form to provide an estimate of the achievement level in the education system at a particular age or grade level. Their design is organized based on a curriculum area, although in some cases they are designed based on a set of cognitive skills (math or reading) that a person should have at a specific age. Normally, these assessments involve the administration of achievement assessments to a sample of students

To provide statistically valid results in sample-based assessments, a representative sample of schools (usually 150 to 200 schools) is drawn from each country, and a sample of students is randomly drawn from within each of the sampled schools, either by sampling entire classrooms or by sampling students across classrooms. Although the best-known cross-national assessments feature a few similarities, there are also some substantial differences that need to be considered when comparing the results for different education systems.

A hierarchy of assessment types has been developed from more to less reliable, where reliability can in part be thought of in terms of comparability across space (for instance countries) or time. It is useful to think of three types of learning assessments, each offering specific opportunities and challenges. The first two categories are the cross-national assessments that allow participating countries comparability among them at a different scale: the three large international programmes (PISA, TIMSS and PIRLS) and the five regional assessments. The third group comprises the national assessments programs.

4.1 Cross national assessments

The two main organizations implementing large international assessments are the International Association for the Evaluation of Educational Achievement (IEA), which organizes studies like TIMSS, the Progress in International Reading Literacy Study (PIRLS) and International Civic and Citizenship Education Study (ICCS); and the OECD, which conducts studies like PISA and the Programme for the International Assessment of Adult Competencies (PIAAC).

There are, however, other organizations conducting or supporting regional assessments, such as UNESCO's Regional Comparative and Explanatory Study (ERCE) in Latin America, the Southeast Asian Ministers of Education Organization and UNICEF's Southeast Asia Primary Learning Metrics (SEA-PLM) in South-East Asia, the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) in southern and eastern Africa, the Pacific Islands Literacy and Numeracy Assessment (PILNA) by the Educational Quality and Assessment Programme of the Pacific Community, and the Programme d'Analyse des Systèmes Educatifs de la CONFEMEN (PASEC).

Characteristics and costs

Table 3 elaborates on the characteristics of large-scale learning assessment domain or area of assessment and the intended population by age and grade. Assessments differ not only in costs

but also in intended population assessed, the capacity development activities included and the inclusion or not of the national report. In some cases, the contribution varies according to the economic capacity of the country

For instance, capacity development is usually done through meetings with all participating countries where they are taken through all the steps of the studies. They learn, for example, item development and review, scoring of items, data management, as well as getting lectures about sampling, scaling, and so on. A few example details on several assessments' capacity building efforts follow.

In the case of IEA, meetings are bi-annual and all exhibits of the international report are available to the countries in editable format to help writing their national reports. There are usually two workshops on how to analyze the data to facilitate their analysis for their national reports. One also includes a module on developing themes and questions for their national reports. Altogether, we can understand PIRLS and TIMSS not only as assessments but also as capacity development projects for developing and conducting large-scale assessments. National reports are not part of the fee, but IEA gives the participating countries all tools needed for this. National reports could be found at: https://www.iea.nl/publications/study-reports/national-reports-iea-studies.

A special mention goes to a new IEA's initiative, the Literacy and Numeracy assessment (LaNA) that is a shorter, less demanding assessment in comparison to TIMSS and PIRLS, meant to be administered at the end of primary school. LaNA, rooted in the comprehensive assessment frameworks of TIMSS and PIRLS (IEA, 2022), is designed for low- and middle-income countries with the aim of producing national data to monitor foundational literacy and numeracy goals. LaNA could benefit countries by producing reliable data on student performance in literacy and numeracy based on a representative sample; developing experience in implementing assessment procedures; and capacity building in planning and administering assessments. With a lower international fee, it could be administered on demand, or, in other words, does not have to follow the specific cycle, giving thus more flexibility to countries[8].

In the case of ERCE, the fee of US$ 61,000 USD funds the regional report and 4 national reports on: learning achievements and associated factors; socio-emotional skills module; writing module; and the national report of the curricular analysis.

As part of the implementation of its regional assessment, PASEC is building country capacity. The aim is to ensure sustainable capacity building in countries by providing them with access to high quality capacity building activities, although PASEC is aware of the high turnover of teams within the ministries of education and the fact that many staff are not suited to the profiles required.

In general, capacity building of national teams is based on specific training related to the implementation process of standardized assessments. This includes training in item design, booklet assembly, sampling, psychometric analysis of assessments and questionnaires, construction of indicators, data analysis, data processing, use of data software such as Stata, etc.

[8] LaNA is comprised of only multiple-choice items and has lower costs in terms of international fees relative to TIMSS and PIRLS but has less capacity development activities associated (IEA, LaNA brochure 2022). LaNA's literacy assessment consists of a reading comprehension test, in which students read simple passages and answer related questions. The passages encompass the PIRLS dual purposes for reading: reading for literary experience (stories) and reading to acquire and use information. LaNA's numeracy assessment includes adjusted items from TIMSS, covering topics such as recognizing and comparing simple fractions, whole number computation, and reading graphs. IEA (2022), *LaNA brochure*.

The second component involves capacity building and the autonomy of national evaluation systems according to the needs expressed by the countries themselves and by the partners[9].

An issue not reflected in the table is the degree of institutionalization within national budgets in the participating countries of the fees and cost of the national assessments and also related to the national cost of administration. In general terms, IEA's and OECD's, as well as ERCE's, fees are, in general, paid with domestic resources which also cover all the related activities such as translation, printing of the assessment tools, travelling and data entry and cleaning. PILNA, PASEC and SEA-PLM have different funding sources for countries and for their Secretariat. For instance, in the case of SEA-PLM, UNICEF regional funding paid for almost all the participating countries. It could be the case, that international fees are related to the income level of the country. For instance, for the next round of PASEC, CONFEMEN member countries bear 70% of the global cost of the evaluation (1.8 million USD). This covers the collection of data in grades 2, 6, and 9/10 (and on all the teachers and school directors at the sampled schools) and CONFEMEN itself bears 30%. For non-member countries, it is 100% of the cost of the evaluation.

Table 3. Main characteristics of Cross-National Learning Assessments

Assessment	Domain, Area	Grade/Age	Cycle every ... Years	Estimated fees per round (in thousand USD)	Capacity development		Number of countries-
Global					**Test Related**	**Country report**	
Progress in International Reading Literacy Study (PIRLS)	Reading	Fourth	4	227	included	not included	60
Trends in International Mathematics and Science Study (TIMSS)	Mathematics and Science	Fourth and eighth	4	222	included	not included	60
Literacy and Numeracy Assessment (LaNA)	Reading and Math	End of primary	on demand	100 to 150	not included	not included	
Programme for International Student Assessment (PISA)	Reading and Math	15-year-olds	3	199	with extra costs	with extra costs	79
Regional							
ERCE	Language (reading and writing) and Mathematics.	Third and sixth	6	300	included	included	16
Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ)	Literacy and numeracy	Sixth grade	6	150	included	not included	14
Programme d'Analyse des Systèmes Educatifs de la CONFEMEN (PASEC)	French and mathematics	two and sixth	5	630	included	not included but supported	15
The Southeast Asia Primary Learning Metrics (SEA-PLM)	Reading, Mathematics, Writing, Global Citizenship	Fifth	4	119	included	not included	6
Pacific Islands Literacy and Numeracy Assessment (PILNA)	Literacy and Numeracy	4th and 6th	3	97	included	not included	15
Calibrated module							
AMPL	Reading and Math	Upper Primary	on demand	80*	included	analysis and short report only AMPL	8

Note: * on average; PILNA: Secretariat Costs paid by Australia and New Zealand; Department of Foreign Trade and Affairs (DFAT) Australia pay the technical partners costs; Country costs are estimative.
SEA-PLM: UNICEF- EAPRO and UNICEF Country offices paid for the SEA PLM Expenses of participating countries and co-shared in regional expenses (regional workshops and field trial and main survey expenses) and staff support.

Source: *UIS based on assessment program information.*

[9] PASEC's objective is for the teams to master all the steps of an evaluation, from its preparation to the sampling of schools and students, the training of administrators, the collection and processing of data, and the analysis of data to produce a report.

4.2 National assessments

National learning assessments (NLA) are a diagnostic tool that can establish whether students achieve the learning standards expected in the curriculum by a particular age or grade, and how this achievement changes over time for subgroups of the population. The label 'assessment' in a programme is no guarantee that standard linking procedures across years, using common items, are used.

Implementing national learning assessments has the advantage of helping strengthen national assessment capacity and being better aligned with national curricula. However, national assessments need to be comparable over time to be able to monitor progress and aligned to global definitions in order to allow comparability and facilitate peer learning processes. In order to develop processes to align reporting with SDG benchmarks, the Global Alliance to Monitor Learning (GAML) and the Technical Cooperation Group (TCG) for Education 2030 found agreement on the definition of MPLs for SDG 4.1.1 and also have developed, in collaboration with partners, the Global Proficiency Framework (GPF) to guide progress towards and report results against SDG 4.1.1. These are both the bases for enabling national assessments to report on SDG 4.1.1.

Some mechanisms are in place to be able to understand national assessments and proficiency levels according to international benchmarks as described in section 2.2. The tools could be separated into two groups depending on whether one is trying to align past data collection or future data collection.

To date, Policy Linking (as described above) has been implemented in some national assessments with the objective of exploring potential use for reporting learning data while encouraging reflection on existing assessments and tools and building system capacity. Bangladesh, Ghana, India, Lesotho, Nigeria, Cambodia, Nepal, and Zambia are among the countries that have engaged in policy linking. Unfortunately, in most of the cases, the tools need further development as well as the procedures to be suitable for reporting.

Examinations (usually formally called "Public Examinations") would deserve a separate chapter. They are often high-stakes assessments taken by all students at the end of the primary or secondary cycle and serve a certification purpose for the labour market and for university entrance. One advantage of the use of examinations for gauging trends is that they already feature prominently in the policy debates of many countries. Examinations undoubtedly provide some guidance to policymakers and the public in relation to the extent to which children are or are not acquiring basic skills. They are almost certainly, in most cases, better than having nothing. And there is some evidence that countries have, in the past, used them to good effect to improve the quality of their education systems.

Methodologically, there is no reason why examinations cannot follow an alignment process to explore their suitability for tracking over time and for reporting. The fact that stakeholders are familiar with them, and understand them, is a plus. However, at this time, little is known about the reliability of most public examinations to establish trends over time. In fact, there are good reasons to suspect that they should not be used in this manner. Since they play a gate-keeping function in allocating scarce study opportunities at higher levels of the system, and since these opportunities do not change fast, the more that students take the examination, the lower the pass rate must be, in a sense. Thus, pass rates by themselves are often not a suitable indicator of

quality, yet they get used in this manner. More can be done to analyse trends in the total points received by the students, or other scoring methods, but even that is relatively meaningless if the difficulty of the assessments is not equated psychometrically, and the extent to which this happens, systematically, is unknown.

4.3 Weighing options

Table 4 provides an evaluation of the 3 mains sources of learning assessment data: international assessments, regional assessments, and national assessments according to a few parameters, compromises, and trade-offs.

Table 4. Pros and cons of every type of assessment

	International Assessments	**Regional assessments**	**National Assessments**
Comparability between countries	High within each programme, relatively easy to equate across programmes, at least for groups of countries and/or in approximate fashion, insofar as technical documentation is comprehensive and there are many doubloon countries. Certainly, more could be done here than is being done.	Almost as high within each programme, less easy to equate across programmes. Differences across programmes in the selected grade complicates comparisons.	Low due to a large variety of sampled populations, different methodologies, possible interference by some governments, lack of documentation of sampling and psychometric properties, often lack of equating over time even within any given national assessment.
Comparability over time	Mostly high.	As for previous column although only comparable for the last 2 cycles.	Could be high for those countries following rigorous methods but this is the case only in a minority of the countries.
Timeliness of the statistics	One year lag with respect to cycle.	Varies – one to four years.	Would vary by country, but likely to be the timelier than any cross-national program
Scope for public buy-in and policy impact	The fact that the assessments are seen as fair and independent and the fact that they allow for international comparisons, make the results highly influential.	Largely as for the previous column, though concerns around the accuracy of the statistics, and the transparency of methods used, are more prevalent.	If rigorous, improves the chances results will influence curriculum design and teacher training, at least indirectly by being part of a proper assessment system. If not, national results may not be taken seriously or risk using weak information to inform policy.
Scope for national capacity building	Limited, in general restricted to learning by doing, but could be paired with more explicit training in the different stages of the assessment cycle.	The regional nature of the programme increases countries' direct involvement in technical aspects.	If country experts have access to good materials and training programmes, national assessments can play a large role in building capacity at the national level.
Alignment to the MPL	There is agreement about the levels that align to the MPL although it should be completed with a standard-setting exercise that identifies precisely the MPL in each case.		Could use AMPL to align to the global MPL

Source: UNESCO Institute for Statistics. Based on: Gustafsson, M. (2019). *Costs and Benefits of Different Approaches to Measuring the Learning Proficiency of Students (SDG Indicator 4.1.1). Information Paper No. 53, January 2019.*

4.4 Global reporting

For global reporting, the UIS currently accepts the assessments listed in **Table 5** for the grades described.

Table 5. Assessments currently used for reporting by level of education

Assessments	Grade 2/3	End of primary	End of lower secondary
ERCE/UNESCO	X	X	
PASEC	X	X	
PILNA	X	X	
PIRLS	X	X	
PISA/ PISA-D			X
SACMEQ IV		X	
SEA-PLM		X	
TIMSS 4th grade - Math	X	X	
TIMSS 8th grade- Math			X
National assessments	Subject to statistical linking		
Modules that measure only one Proficiency Level			
AMPL-b (MILO testlet)		X	

Note: As mentioned in section 3.2.1, there are various developments such as LaNA, AMPL-a and PISA module.

Source: *UNESCO Institute for Statistics.*

International programmes that collect learning outcomes data from children through household surveys can serve as a vital reality check when data derived from schools-based programmes are evaluated. Household-based data is generally not prioritised as a primary source for reporting Indicator 4.1.1. One disadvantage with permitting the use of household-based data for Indicator 4.1.1 would thus be an undesirable shift away from the core focus of establishing effective assessment programmes within a schooling system. Yet household data, where available, ought to be used when schools-based data are validated. A major development has been the inclusion of reading and mathematics tests in Version 6 of UNICEF's Multiple Indicator Cluster Surveys (MICS).

There are other assessments that are widely used by the global education community and that have become influential with countries, as they are often usable for, and are in fact used, for informing classroom practice and for generating public discourse and dialogue. These include the Early Grade Reading Assessment (EGRA) and Early Grade Mathematics Assessment (EGMA) family of assessments, the PAL or "citizen-led" family of assessments, and UNICEF's MICS Foundational Skills Module. However, none of these were designed to enable cross-country comparisons and in fact some of them discourage such comparisons. However, with more work, they might be able to be used for this purpose. Some of them can track over time, at least with further refinement or if adjusted. If they could be used, this would increase the level of reporting significantly at relatively low cost. And it would be a form of reporting with clear conceptual links to how learning data can be used to not just report but for local improvement or at least dialogue. UIS has worked with the relevant stakeholders on this issue, but more could be done.

Some of the assessments discussed in the paragraph above could become part of or at least be used by the Learning Data Compact developed by UNESCO, UNICEF, and the World Bank. The Learning Data Compact is a commitment to ensure that all countries, especially low- and middle-income ones, have at least two quality measures of learning by the year 2025[10].

Assessment program proficiency levels used to report indicator 4.1.1

To report to the global level, in each assessment program, the proficiency level (PL) whose descriptor is better aligned to the global definition of MPL was identified. This first step, that has to be completed by a standard-setting exercise, is summarized in **Table 6.** One possible outcome of the alignment process is that the proficiency level descriptor (PLD) identified in the assessment program as aligned to the global MPL is different from the one used to report in the assessment program; this implies that a different definition of MPL is used in the assessment program and could create some confusion in the reporting, if not clarified properly.

In general, the assessment programmes assign a number that describes the PL and to that PL is attached a PLD that defines the skills and contents the students that are in that level master. In general, the higher the level, the more proficient the students in each domain. The only exception is IEA where the different PL have a name associated: Low, Intermediate, High, and Advanced benchmarks.

Table 6 below summarizes this information by describing the PL that is used to report Indicator 4.1.1 in each of the assessment programmes (column 3). Column 4 reflects the PL used in each assessment program in their own reporting. For instance, for grades 4-6, SACMEQ's PL aligned with the global MPL (used to report on the global indicator 4.1.1) is level 5; however, the PL used by SACMEQ as the MPL for its regional report is level 3. For the same grades, PASEC's PLs aligned with the global MPL is level 3 for mathematics and level 4 for reading, whereas the assessment's PLs used for reporting in its regional report are, respectively, levels 2 and level 3.

What is the implication for a country of using different proficiency levels? When the global MPL is higher than the assessment's own MPL, the percentage of proficient students reported as proficient is lower than the percentage of students reported by the assessment program based on a less stringent threshold. The size of the impact in the percentage of students reported as above the MPL would depend on the underlying distribution of students by levels of proficiency. In other words, ceteris paribus, a higher cut-off point would imply a lower percentage of students proficient, it is not possible to foresee the impact on the proficiency levels of the population from moving from a less demanding to a more demanding proficiency level.

[10] UNESCO, UNICEF, and The World Bank. April 2022. *The Learning Data Compact (LDC)*. Brochure.

Table 6: Identification of MPLs in different assessments by subject and grade/age

Grade/Age *(1)*	**Assessment name** *(2)*	**Assessment Proficiency Level Descriptor (PLD) aligned to SDG MPL descriptor** *(3)*	**MPLs in assessment program** *(4)*
Mathematics			
Grades 2-3	ERCE 2013-2019	2	2
	PASEC 2014-2019	2	2
Grades 4-6	ERCE 2013-2019	3	3
	PASEC 2014-2019	3	2
	PILNA 2012-2018	6	5
	SACMEQ 2006-2013	5	3
	SEA-PLM 2019	6	6
	TIMSS 1995-2019	Intermediate Benchmark	Intermediate Benchmark
Grades 7-10 Age 15	PISA 2000-2018	2	2
	TIMSS 1995-2019	Intermediate Benchmark	Intermediate Benchmark
Reading			
Grade 2/ Grade 3	PASEC 2014-2019	3	3
	ERCE 2013-2019	2	2
Grades 4-6	ERCE 2013-2019	3	3
	PASEC 2014-2019	4	3
	PIRLS 2001-2016	Low Benchmark	Low Benchmark
	PILNA 2012-2018	5	4 (grade 4) and 5 (grade 6)
	SACMEQ 2006-2013	5	3
	SEA- PLM 2019	6	6
Grades 7-10 Age 15	PISA 2000-2019	2	2

Source: *UNESCO Institute for Statistics.*

Selection of reporting source when various sources are available

For each of the indicators listed above for global reporting, the sources of data selected should be prioritized according to the following order of assessments, provided that a mapping of grades to SDG 4.1.1 a, b, or c, has guided a first selection of sources:

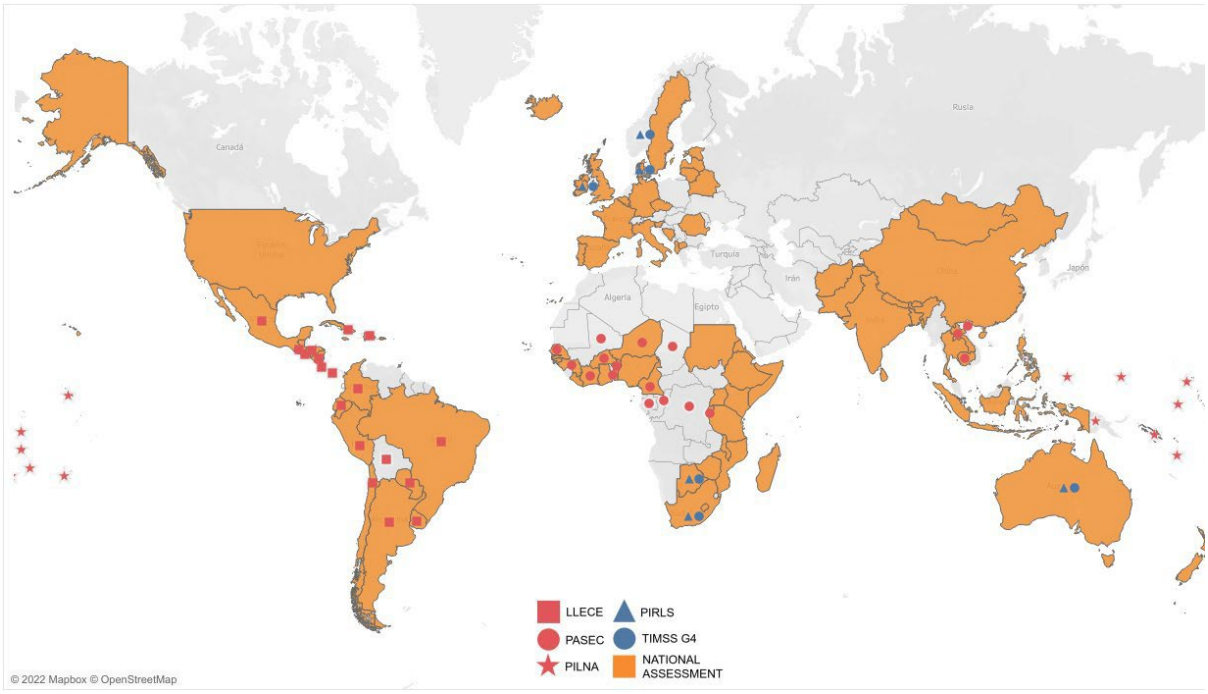
i. International assessments
ii. Regional assessments
iii. National assessments if they comply with the alignment process.

Global comparability would lead to choose the international assessment that best maps to the required level of reporting, then followed by the regional assessment in order to find the highest possible degree of comparability. National learning assessment are the reporting option only if alignment to the Global MPL had been run. More on this topic follows in the next sections.

5. Coverage of student population

Figures 2 and 3 show the scope of coverage of international assessment program by level. Both figures confirm the pattern of the presence of international and regional assessments in primary education. It is important to note that an important disparity exists in Africa: Francophone countries covered by PASEC have statistics at the lower primary level, but this is not the case for the (mostly) anglophone SACMEQ countries. Hence, for the SACMEQ countries, it becomes necessary to rely rather heavily on national assessments at this level. Both figures indicate as well that the international assessments provide the best coverage at the lower secondary level if only international assessments are considered. The second best-compared level is end of primary, where more than half of the world's countries are covered. Participation in either international or one of the five regional programmes expands coverage for the end of primary.

Figure 2 - Coverage by region and type of cross-national assessment - Primary level (Grades 2/3 and End of Primary)



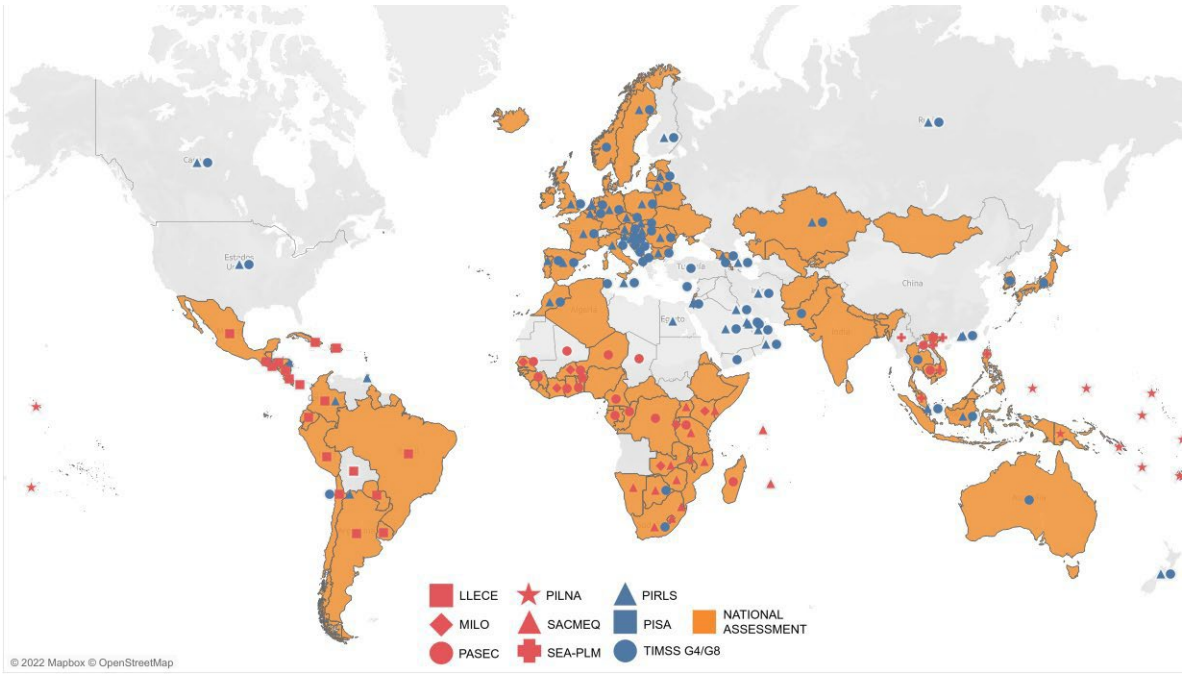
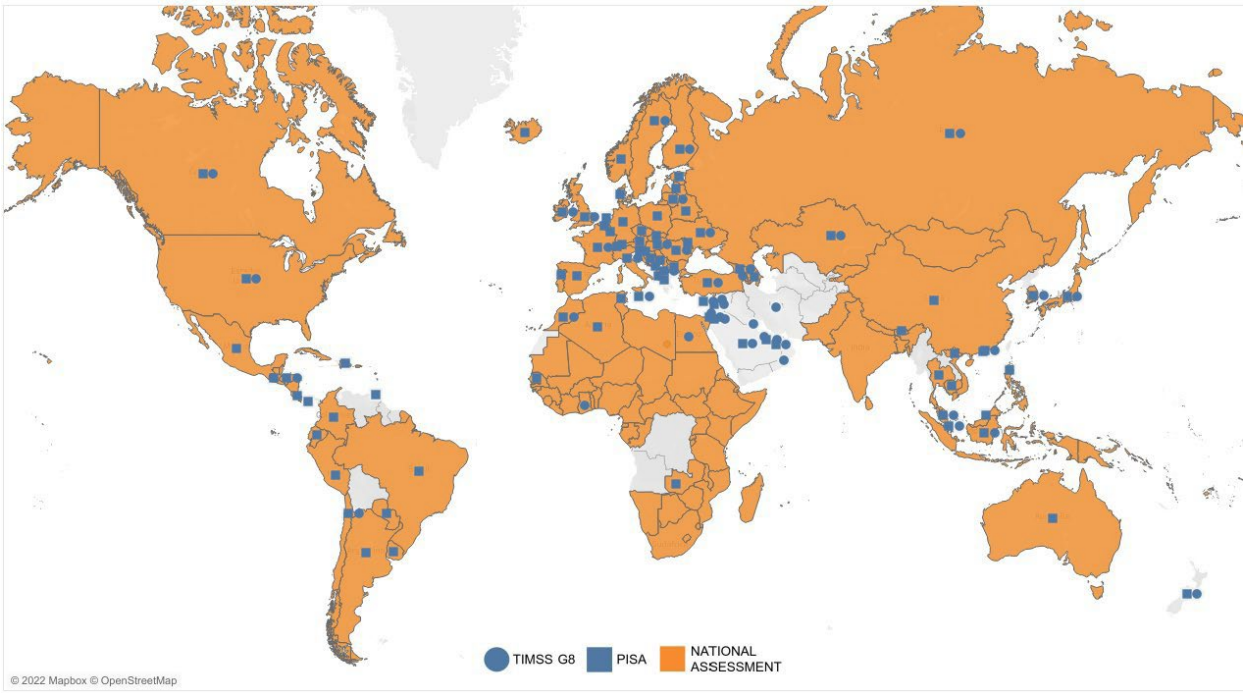
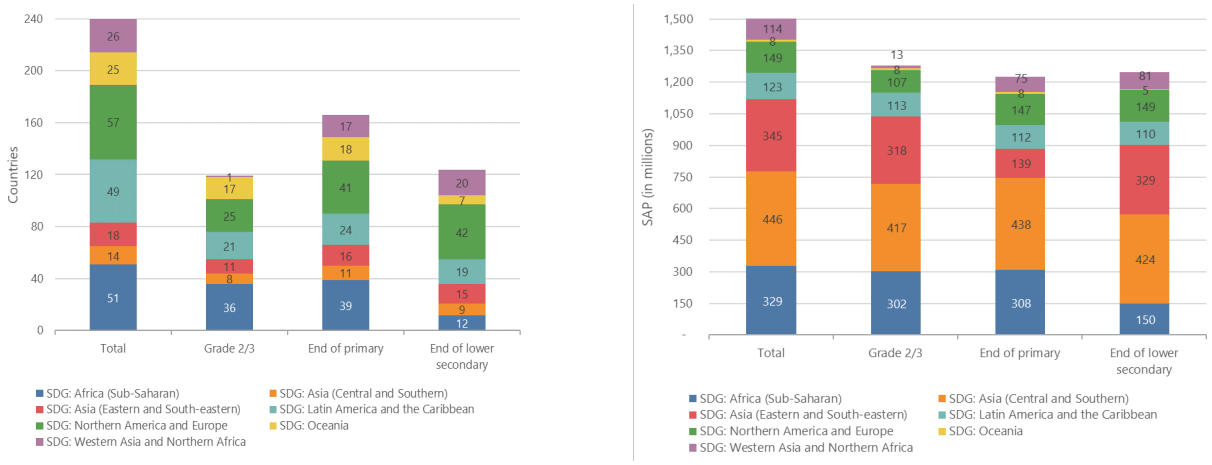


Figure 3 - Coverage by region and type of cross-national assessment - End of Lower Secondary



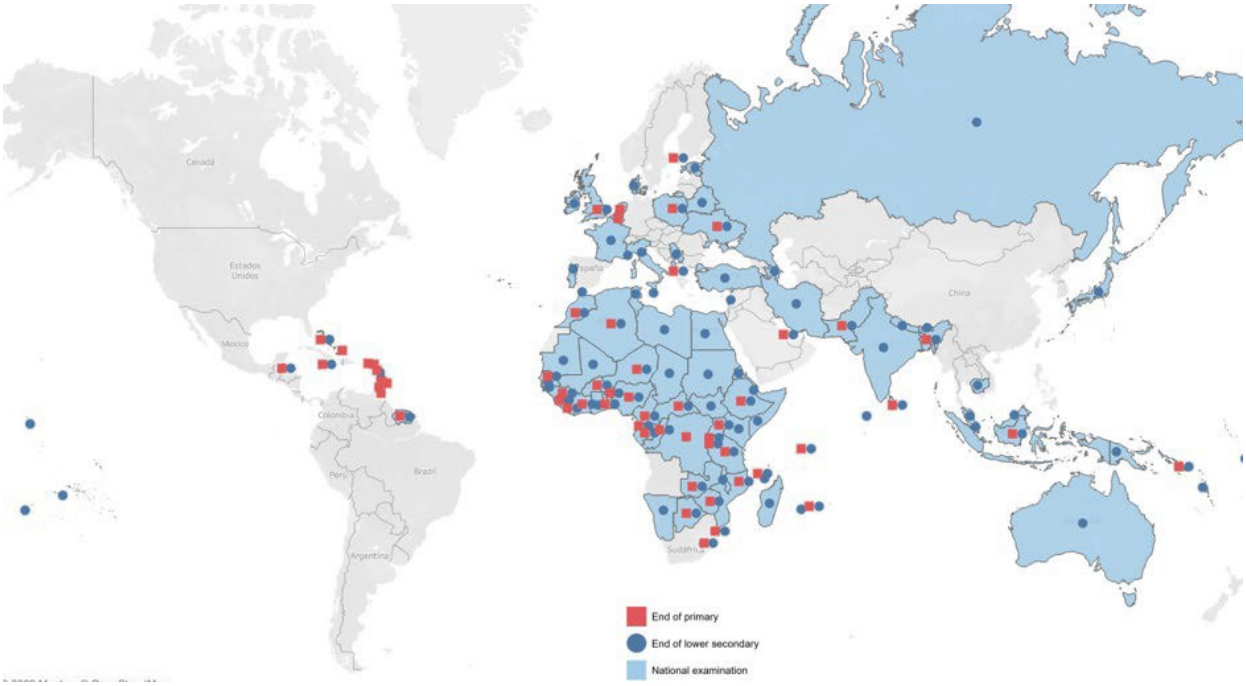
What are the key messages? 96% of the population-weighted world has some type of assessment at some education level. **However, despite this high number, note that for any *one* of the three education levels, there is no current internationally comparable assessment, and therefore over half of the world would have to be monitored using national assessments** (or examinations), at least given current levels of coverage of the cross-national programmes (**Figure 4**).

Figure 4 - Coverage of cross-national assessments by world region



As a complementary information on coverage, **Figure 5** shows the national assessments for end of primary and lower secondary by country.

Figure 5 – National assessments at the end of the school cycle by country



6. How can a country produce comparable data for Indicator 4.1.1?

Data of good quality frequency and comparability over time are key to contribute towards a better quality of schooling around the world and could the possibility to measure change *over time* with respect to learning outcomes and the attainment of proficiency benchmarks. In their quest for improving the availability and quality of data to monitor learning, countries should be guided by a number of principles and are offered a varied menu of options.

6.1 Principles to guide choice

To guide the choice of learning measurement, and to ensure assessment data are consistent with long-term strategic goals of effective decision-making, the UIS, UNESCO, World Bank and UNICEF have developed a set of principles on which this section is based. The following shared principles are important not just for designing assessments or deciding which assessment to buy "off the shelf", but for developing an assessment system for one's own country. The system should be good not just for reporting but for managing improvement at all levels of education, for developing the capacity to guide decision making, and for linking the system-level assessments to formative assessments and classroom practices.

Principle 1. Build on what exists

It is key to **(develop) and build on existing capacity** of data producers, analysts, and users. Most countries can build on some existing capacity. Assessing and reporting with given frequency and regularity can foster habit and expectation.

Principle 2. Allow flexibility to ensure alignment with country needs (not one-size-fits-all)

It is important to know **what to assess and how to measure it**. The learning data ought to measure against a clear standard of what the learner must know, comprehend and be able to do at a specific age/grade, criteria that can be laid out in the national curriculum and/or anchoring on the GPF and MPLs.

Principle 3: Foster country ownership through a demand-driven approach

The approach should be **demand-driven to foster** strong country ownership. Through data reuse, and the use of parallel data coherent with the measurement for reporting, to drive actual improvement in the classroom, it is possible to enhance stakeholders' perceived values of collecting data.

Principle 4. Ensure data is relevant for decision-making

Assessments must be **relevant for monitoring progress in order to inform decision-making**. The assessment results must be comparable, which means that questions have to be designed at the same level of difficulty across time and administered to students at similar grades or ages.

To ensure that assessments can accurately monitor progress for decision making, data also must be **internationally comparable** for benchmarking. Every country ought to have an assessment that, in one way or another, was designed for, or can be used for, international comparability - a commitment in the SDG process (not just in education).

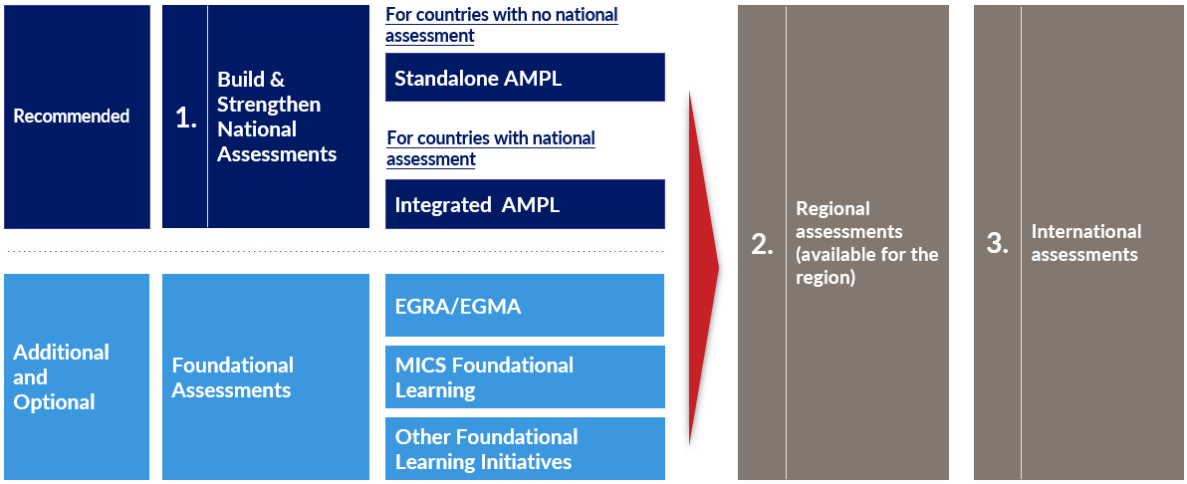
6.2 Options depending on country's initial situation

A diverse 'menu of options' is available to countries to help them determine the most adequate journey and tools to improve both availability and quality of data to monitor learning. The decision should be based on countries' initial conditions, including the current position or starting point with respect to learning outcomes, and the technical capacity at the national level. A country falls into one of the following two categories:

1) **Country with no learning data available**:
   A country with no learning data available is recommended to start the journey building and strengthening its national system and progressing towards participating in regional or international assessments. Foundational assessments presented earlier in section 4.4 are additional and optional resources that a country with no learning data could resort to: these include the EGRA/EGMA assessments, the MICS Foundational Learning module and other Foundational Learning initiatives.

   Figure 6 presents all the resources available for a country belonging to this category.

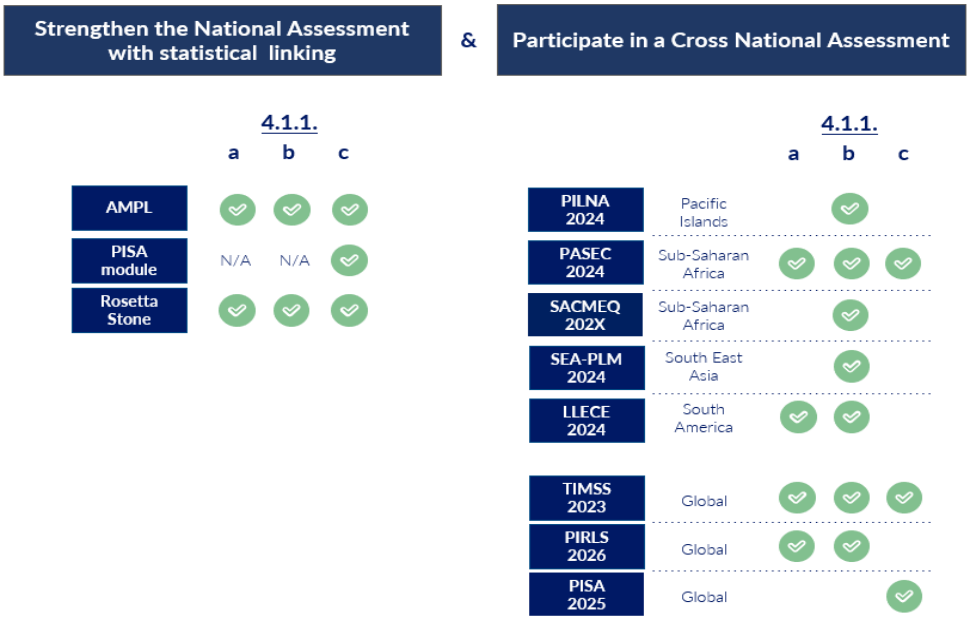
Figure 6 – Resources available to countries with no learning data



2) **Country with 'assessment capacity':**
A country with 'assessment capacity' is a country that has already implemented at least two cycles of national assessments and/or that has participated in at least one cross-national assessment. It can strengthen its national assessment systems through linking national assessments or through participating in international or regional assessments. Linking options were presented in detail earlier in section 3.2.

Figure 7 presents all the resources available for countries belonging to this category.

Figure 7 – Resources available to countries with 'assessment capacity'



In sum, the diverse menu of options available to countries to report on indicator 4.1.1 could be summarized in the following three points:

1) **If a country wishes to join an existing regional or international assessment** to report on 4.1.1 for a selected level of education, it should ideally choose the appropriate assessment that takes into account data points collected from previous participation, if applicable. This will allow the country to estimate trends for the indicator.

2) **If the country wishes to implement a national assessment for the first time**, it should ensure that it is aligned with global reporting:

a. In case the data has not been collected yet, the country could add a booklet aligned to global reporting in the design of the national assessment. For example, it could add AMPL-b for the end of primary.

b. Otherwise, the country should apply an appropriate and rigorous methodology to ensure alignment to the global definition of the minimum proficiency level (MPL).

3) **If the country wishes to implement a national assessment for the second or third time**, it should ideally be in the same grade as the previous time. In order to ensure the longitudinal anchoring of the National Learning assessment, the following steps should be followed:

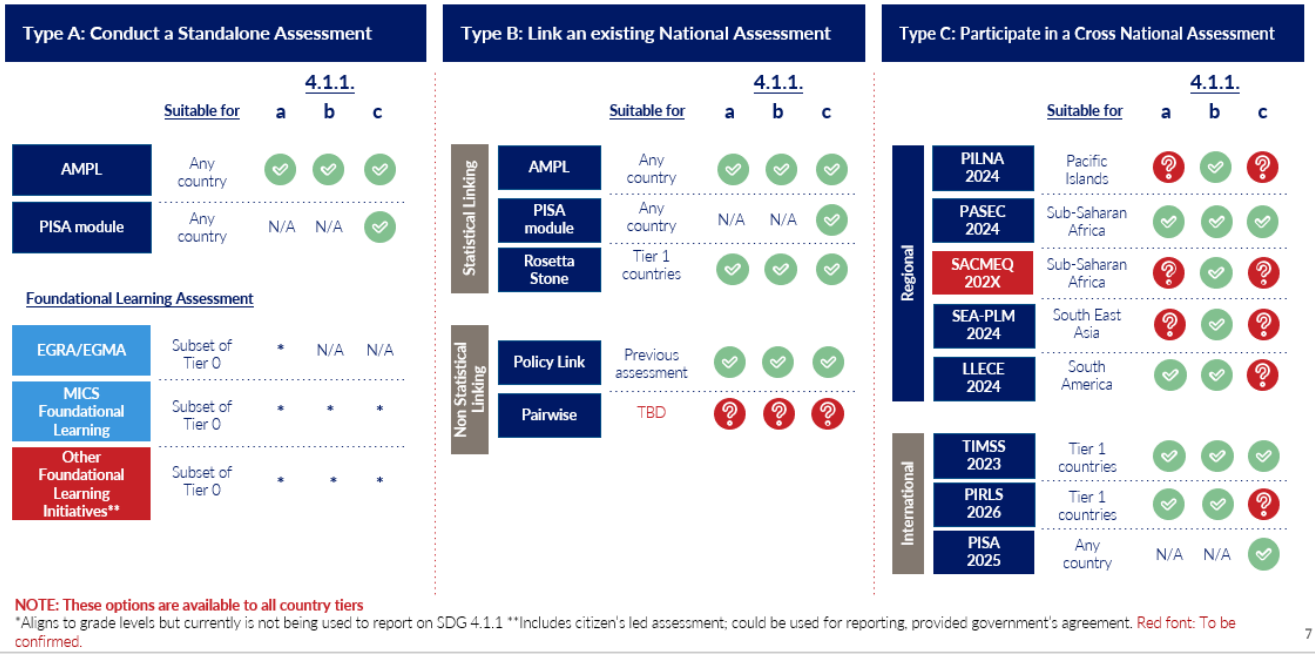
- **a. For previous rounds:** run a pedagogical calibration to identify alignment of curriculum, assessment and PLDs as a minimum. Sampling and data procedures need also to be reviewed to allow alignment with procedures.
- **b. For future data collection**: once the needed adjustment is identified, use one of the two options:

  i. The first and preferred option is to add a booklet or items that are aligned to specific levels of global reporting (GPF and MPLs), such as AMPL-b for the end of primary, to allow the linking of the current and previous rounds using the items as anchors.

  ii. The second option is to run policy linking once data are collected if the country was not able to choose option a.

Figure 8 presents the full landscape of assessments available by type of country and level of education to be reported, specifying in particular the options available if a country decides to conduct a standalone assessment, to link an existing national assessment through statistical or non-statistical methods or to participate in cross-national assessments. **Appendix A** also shows the types of assessments available to countries along with the corresponding links for more information. **Appendix B** offers a list of some of the major resources helpful for reporting on SDG Indicator 4.1.1 along with their respective links.

Figure 8 – Landscape of assessments available to countries



Appendix A: Types of assessments available to countries and the corresponding links for more information.

Type	Assessment	Name	Webpage
Build & Strengthen National Assessment	AMPL	Assessments for Minimum Proficiency Levels	https://milo.uis.unesco.org/wp-content/uploads/sites/17/2022/10/ampl.pdf
Cross National Assessments: Regional assessments	PILNA	Pacific Islands Literacy and Numeracy Assessment	https://eqap.spc.int/PILNA
	PASEC	Programme for the Analysis of Education Systems	https://pasec.confemen.org/
	SAQMEQ	Southern and Eastern Africa Consortium for Monitoring Educational Quality	http://www.sacmeq.org/
	SEA-PLM	Southeast Asia Primary Learning Metrics	https://www.seaplm.org/index.php?lang=en
	LLECE (ERCE)	Latin American Laboratory for the Assessment of the Quality of Education	https://gaml.uis.unesco.org/wp-content/uploads/sites/2/2019/05/GAML6-Session5-LLECE.pdf
Cross National Assessments: International Assessments	TIMSS	Trends in International Mathematics and Science Study	https://timssandpirls.bc.edu/index.html
	PIRLS	Progress in International Reading Literacy Study	https://timssandpirls.bc.edu/index.html
	PISA	Programme for International Student Assessment	https://www.oecd.org/pisa/
Foundational Learning Assessment	EGRA	Early Grade Reading Assessment	https://earlygradereadingbarometer.org/
	EGMA	Early Grade Mathematics Assessment	https://shared.rti.org/content/early-grade-mathematics-assessment-egma-toolkit
	MICS/ Foundational Learning Module	Multiple Indicator Cluster Surveys Foundational Learning Module	https://data.unicef.org/resources/guidelines-adapting-foundational-module-non-mics/

Appendix B: Resources helpful for reporting on SDG Indicator 4.1.1

Resource	Link
Global Alliance to Monitor Learning (GAML) website: Indicator 4.1.1	https://gaml.uis.unesco.org/4-1-1/
Minimum Proficiency Levels	http://gaml.uis.unesco.org/wp-content/uploads/sites/2/2021/03/Minimum-Proficiency-Levels-MPLs.pdf
Metadata	http://tcg.uis.unesco.org/wp-content/uploads/sites/4/2020/09/Metadata-4.1.1.pdf
Monitoring of the Sustainable Development Goals using large-scale international assessments	https://tcg.uis.unesco.org/wp-content/uploads/sites/4/2022/04/Monitoring-of-the-SDGs-Using-Large-Scale-International-Assessments_April-2022.pdf
Aligning and reporting on indicator 4.1.1: UIS annotated workflow	https://gaml.uis.unesco.org/wp-content/uploads/sites/4/2020/03/4.1.1_Aligning-and-reporting_SDG-4.1.1_2023.03.28.pdf